Appeared online at CEUR-WS.org/Vol-1341

ArgNLP 2014
Frontiers and Connections between Argumentation Theory and Natural Language Processing

Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing

Forlì-Cesena, Italy, July 21-25, 2014.

Edited by

Elena Cabrio *
Serena Villata *
Adam Wyner **

* INRIA, Sophia Antipolis, France
** The University of Aberdeen, Aberdeen, UK

Table of Contents

Papers

Preface

Elena Cabrio[1], Serena Villata[1], and Adam Wyner[2]
[1]INRIA Sophia Antipolis - Mediterranee
[2]Department of Computing Science, University of Aberdeen

Large amounts of text are added to the Web daily from social media, web-based commerce, scientific papers, eGovernment consultations, and so on. Such texts are used to make decisions in the sense that people read the texts, carry out some informal analysis, and then (in the best case) make a decision: for example, a consumer might read the comments on an Amazon website about a camera, then decide which camera to buy; a voter might read various political platforms, then vote. An analyst or consumer of such corpora of text is confronted by several problems. The information in the corpora is distributed across texts and unstructured, i.e. is not formally represented or machine readable. In addition, the argument structure - justifications for a claim and criticisms - might be implicit or explicit within some document, but harder to discern across documents. As well, the sheer volume of information overwhelms users. Given all these problems, extracting and reasoning with arguments from textual corpora on the web is currently infeasible.

To address these problems, we need to develop tools to aggregate, synthesize, structure, summarize, and reason about arguments in texts. Such tools would enable users to search for particular topics and their justifications, trace through the argument (justifications for justifications and so on), as well as to systematically and formally reason about the graph of arguments. By doing so, a user would have a better, more systematic basis for making a decision. Clearly, deep, manual analysis of texts is time-consuming, knowledge intensive, and thus unscalable. Thus, to acquire, generate, and transmit the arguments, we need scalable machine-based or machine-supported approaches to extract and reason with arguments. The application of tools to mine and process arguments would be very broad and deep given the variety of contexts where arguments appear and the purposes they are put to.

On the one hand, text analysis is a promising approach to identify and extract arguments from text, receiving attention from the natural language processing community. For example, there are approaches on argumentation mining of legal documents, on-line debates, product reviews, newspaper articles, court cases, scientific articles, and other areas. On the other hand, computational models of argumentation have made substantial progress in providing abstract, formal models to represent and reason over complex argumentation graphs. The literature covers alternative models, a range of semantics, complexity, and formal dialogues.

Yet, there needs to be progress not only within each domain, but in bridging between textual and abstract representations of argument so as to enable reasoning from source text. To make progress and realize automated argumentation, a range of interdisciplinary approaches, skills, and collaborations are required, covering natural language processing technology, linguistic theories of syntax, semantics, pragmatics and discourse, domain knowledge such as law and science, computer science techniques in artificial intelligence, argumentation theory, and computational models of argumentation.

To begin to address these issues, we organized the seminar *Frontiers and Connections between Argumentation Theory and Natural Language Processing*, which was held July 21-25, 2014 at the University Residential Center, Bertinoro, Italy. It was attended by 29 participants. The papers in this CEUR Workshop Proceedings are the outcome of the workshop, ranging over a host of topics, empirical approaches, and theoretical frameworks.

Argumentation for Scientific Claims in a Biomedical Research Article

Nancy L. Green
Dept. of Computer Science
University of North Carolina
Greensboro
Greensboro, NC 27402, USA
nlgreen@uncg.edu

Abstract

This paper provides an analysis of some argumentation in a biomedical genetics research article as a step towards developing a corpus of articles annotated to support research on argumentation. We present a specification of several argumentation schemes and inter-argument relationships to be annotated.

1 Introduction

This paper provides an analysis of some argumentation in a biomedical genetics research article (Schrauwen et al., 2012), as a step towards developing a corpus of articles annotated to support research on argumentation (Green, 2014). For each argument for a scientific claim in an article, we would like to annotate its premises, conclusion, and argumentation scheme. In addition we would like to annotate certain relationships between pairs of the arguments, e.g., where one provides support for the premise of another. In order to develop an annotation system that can be consistently applied by different coders or by the same coder at different times, it is necessary to develop a precise specification of each argumentation scheme and inter-argument relationship. In this paper, we present a specification of several argumentation schemes and inter-argument relationships to be annotated.

The main claim of (Schrauwen et al., 2012), summarized in its title, is that a certain variant, c.637+1G>T of the *CABP2* gene, is a cause of moderate-to-severe, autosomal recessive non-syndromic hearing loss (arNSHL). According to our analysis, the argumentation in the article serves at least four types of discourse goals. The first is to persuade peer reviewers that the article is worthy of publication. The second is to persuade the audience that the scientific methodology used by the authors was sound and that the evidence so acquired is reliable. Arguments for the third type support or defend the scientific claims of the article. Arguments for the fourth type support the practice implications, i.e., the authors' suggested application of the scientific contribution to medical practice. The planned corpus will be annotated for arguments of the third and fourth type. In the next section, we briefly discuss the first two types, before focusing on the third and fourth types.

2 Discourse Goals

2.1 Novelty and Significance

The Knowledge Claim Discourse Model (KCDM) (Teufel, 2010) provides a multi-level description of consecutive text segments of a scientific article in terms of the "knowledge claims", or purported scientific contribution of the article. "The top level … formalizes the authors' high-level rhetorical goals, which serve to defend the new knowledge claim of an article against possibly hostile peer review … For instance, authors must argue that their new knowledge claim is novel and significant, and sufficiently different from already existing knowledge claims to warrant publication" (p. 102). According to Teufel, these arguments are not "directly textually expressed", but can be inferred by the reader from lower-level rhetorical moves that "often contain meta-discourse phrases such as 'In contrast to traditional approaches'.

In the Introduction section of (Schrauwen et al., 2012) the significance of the search for causes of arNSHL can be inferred from text such as "Hearing loss is a common sensory disorder that can significantly impact quality of life" (p. 636). An argument for novelty is given in Excerpt 1.

Excerpt 1:
"Most families segregating arNSHL typically have a prelingual, bilateral, severe-to-profound hearing loss. An exception is found with mutations in *TECTA* … and *STRC* …; these mutation cause moderate-to-severe hearing loss … Recently, we identified a locus associated with arNSHL on 11q12.3-11q13.3 (DFNB 93) in an Iranian family that also presents a similar moderate-to-severe hearing loss phenotype … Here, we report that a mutation in *CABP2* … is the cause of DFNB93 moderate-to-severe hearing loss and reveal a role for CaBP2 in the mammalian auditory system." (p. 636)

By design, the KCDM does not address argumentation whose identification requires understanding of scientific content. Thus, the KCDM is not concerned with characterizing the other uses of argumentation that we found in the genetics article.

2.2 Methodological Soundness

The Results section of (Schrauwen et al. 2012) employs a narrative style reporting the sequence of events in the authors' scientific investigation, the reasons for the actions taken during the investigation, and the results of those actions. In so doing, the authors provide implicit arguments for the soundness of their scientific methodology. (The Materials and Methods section of the article provides more details about the methodology.) For example, Excerpt 2 provides a reason for the authors' decision to sequence a certain region of the genome of a certain individual (V:14).

Excerpt 2:
"The DFNB93 region contains more than 300 annotated and hypothetical genes, and several genes are expressed in the mouse and human inner ear. Because there are many strong candidates in the region, we sequenced all genes and noncoding genes in this region by using a custom DNA capture array to identify the disease-causing mutation in one affected individual from the family." (p. 639)

This passage can be analyzed as an instance of a type of Practical Reasoning argument whose discourse goal is to justify the authors' action (sequencing the DFNB93 region by using a custom DNA capture array) in order to achieve the authors' goal (to identify the disease causing mutation in one affected individual). In addition, as will be discussed in the next section, the excerpt contains a causal argument.

2.3 Scientific Claims and Practice Implications

The focus of our planned annotation efforts is on argumentation for scientific claims and practice implications. In this section we present our analysis of several examples of this type of argumentation, given mostly in the Results section of (Schrauwen et al. 2012). In addition to the instance of Practical Reasoning discussed in 2.2, we analyze Excerpt 2 as making the causal argument shown in Argument 1.

Argument 1.
a. Premise: Several genes in the DFNB93 region are expressed in the human inner ear.
b. Premise (implicit generally accepted assumption): A mutation of a gene that is expressed in a human tissue or system may lead to an abnormality in that tissue or system.
c. Premise: A certain individual (identified as V:14) has arNSHL.
d. Conclusion: The mutations occurring in DFNB93 of V:14 are strong candidates for the cause of V:14's arNSHL.

Argument 1 can be represented more abstractly, for purposes of annotation of similar arguments in the corpus, by the following argumentation scheme. In addition to specifying the premises and conclusion, we have added a critical question. Critical questions associated with an argumentation scheme provide a way to challenge arguments instantiating the scheme (Walton et al. 2008). The use of critical questions in our annotation efforts is discussed in section 3.

Effect to Some Cause in Candidate Set
Premise: There is a causal pathway from G-type events to P-type events.
Premise: An individual has experienced P (a P-type event).
Conclusion: Some G-type event experienced by that individual may be the cause of P.
Critical Question: *What if the set of candidates G does not include the actual cause of the event?*

Excerpt 3 contains the argument described in Argument 2.

Excerpt 3:

"After the identified homozygous variants were filtered through the 1000 Genomes Project November 2010 release and dbSNP131, 47 previously unreported variants remained…" (p. 639)

Argument 2.

a. Premise (same as 1d): The mutations occurring in DFNB93 of V:14 are strong candidates for the cause of V:14's arNSHL.

b. Premise (implicit generally accepted assumption): If a variant is a frequent polymorphism then it is not likely to be the cause of a deleterious condition.

c. Premise: All but 47 of the homozygous variants in DFNB93 of V:14 are frequent polymorphisms.

d. Conclusion: One of the remaining 47 homozygous variants may be the genetic cause of V:14's condition.

Excerpt 4 contains the argument described in Argument 3.

Excerpt 4:

"… 47 previously unreported variants remained and included two exonic mutations, one splicing mutation, six nontranslated mutations, 16 intergenic (downstream or upstream) mutations, and 22 intronic mutations. The two exonic variants included one nonsynonymous variant … in *PFIA1* and synonymous variant … in *GAL3ST3* … The splice-site variant, c.637+1G>T … was located at … of *CABP2* … The variants in *PPFIA1* and *CABP2* were subsequently validated by Sanger DNA sequencing, which only confirmed the splicing variant in *CABP2*. (p. 639).

Argument 3.

a. Premise (same as 2d): One of the remaining 47 homozygous variants may be the genetic cause of V:14's condition.

b. Premise (implicit generally accepted assumption): Only exonic or splice-site variants confirmed by Sanger DNA sequencing could be the cause of a genetic condition.

c. Premise: Of the remaining 47 homozygous variants, only the c.637+1G>T splicing variant in *CABP2* was confirmed.

d. Conclusion: The c.637+1G>T variant in *CABP2* may be the genetic cause of V:14's condition.

Arguments 2 and 3 can be described as instances of the following argumentation scheme.

Elimination of Candidates

Premise: There exists a set of candidates C, one of which may be the cause of event E.

Premise: One or more members of C can be eliminated as candidates.

Conclusion: One of the remaining members of C may be the cause of E.

Excerpt 5 contains two arguments, described in Arguments 4 and 5.

Excerpt 5:

"Next, we checked the inheritance of the *CABP2* variant in the entire Sh10 family … and screened an additional 100 random Iranian controls to ensure that the variant is not a frequent polymorphism. The mutation was not detected in any of the controls, and inheritance was consistent with hearing loss in the family." (p. 639).

Argument 4.

a. Premise: The c.637+1G>T variant in *CABP2* segregates with arNSHL in Sh10 (V:14's pedigree).

b. Premise (implicit generally accepted principle): A variant may be the cause of an autosomal recessive condition if it segregates with the condition in a pedigree, i.e., occurrence of the condition and the variant are consistent with an autosomal recessive inheritance pattern.

c. Conclusion (implicit): The c.637+1G>T variant in *CABP2* may be the cause of arNSHL in Sh10.

Although Argument 4 is in some respects similar to Mills' Joint Method of Agreement and Difference (described in Jenicek and Hitchcock, 2004), its premise (4b) provides a causal explanation that is not part of that type of argument. Argument 4 can be described more precisely as an instance of the following argumentation scheme.

Causal Agreement and Difference

Premise: There exists a set of individuals I-present that have a feature F and property P.

Premise: There exists a set of individuals I-absent that do not have feature F and property P.

Premise: There is a plausible causal link from F to P that could account for the presence of P in I-present.

Conclusion: F may be the cause of P in I-present.

Critical Question: *Is there some other feature G in I-present that could account for P, or is there some other factor G in I-absent that could account for the absence of P?*

Argument 5 from Excerpt 4 is as follows.

Argument 5.
a. Premise: The c.637+1G>T variant in *CABP2* is present in the arNSHL affected members of Sh10.
b. Premise: The variant does not occur in a control group.
c. Conclusion (implicit): The c.637+1G>T variant in *CABP2* may be the cause of arNSHL in Sh10.

Argument 5 can be described as an instance of the following argumentation scheme, based upon Mills' joint method of agreement and difference. Note that its first critical question is shared with Causal Agreement and Difference, but its second critical question is not needed for that argumentation scheme, one of whose premises is that there is a causal mechanism that may account for the differences between I-present and I-absent.

Joint Method of Agreement and Difference
Premise: A set of individuals I-present have a feature F and property P.
Premise: A set of individuals I-absent (distinct from I-present) do not have F and P.
Conclusion: F may be the cause of P in I-present.
Critical questions:

- *Is there some other feature G in I-present that could account for P, or is there some other factor G in I-absent that could account for the absence of P?*
- *Is there a plausible causal mechanism that explains how F leads to P?*

Excerpt 6 contains a causal argument for how the c.637+1G>T variant of *CABP2* could lead to hearing loss, as shown in Argument 6.

Excerpt 6:
"… we evaluated the effect of the c.637+1G>T mutation on splicing … Analysis … revealed … indicating that the mutation of c.637+1G>T leads to a complete skipping of exon 6 … Skipping of exon 6 is expected to lead to a shifted reading frame and a premature truncation of the protein" (p. 639-0).

Argument 6.
a. Premise: The c.637+1G>T mutation of *CABP2* may have a deleterious effect on CaBP2 protein during synthesis by *CABP2*.
b. Premise (implicit): CaBP2 protein plays a role in the auditory system.
c. Premise (implicit generally accepted principle): Damage to a protein can lead to disease of the tissue or biological system in which that protein plays a role.
d. Conclusion (implicit): A c.637+1G>T mutation of *CABP2* may result in a disease of the auditory system.

Argument 6 can be described by the following argumentation scheme.

Cause to Effect
Premise: There is a partially known causal pathway from events of type G to events of type P.
Conclusion: The occurrence of a G-type event may result in a P-type event.

Excerpt 7 contains Argument 7, which is similar to Argument 4 and can likewise be described as an instance of Causal Agreement and Difference.

Excerpt 7:
"We identified two families (Sh11 and He) with affected individuals who were homozygous in this region … Affected family members presented with an audiogram similar to the affected individuals in the Sh10 family… Sanger sequencing … revealed the same c.637+1G>T mutation in these families." (p. 640)

Argument 7.
a. Premise: Affected members of two families, Sh11 and He, have audiograms similar to those of affected family members of Sh10 and the c.637+1G>T variant in *CABP2* segregates with hearing loss in those two families.
b. Premise (implicit generally accepted principle): A variant may be the cause of an autosomal recessive condition if it segregates with the condition in a pedigree.
c. Conclusion (implicit): The c.637+1G>T variant in *CABP2* may be the cause of arNSHL in Sh11 and He.

Perhaps because they expect it to be obvious to the intended audience, the authors do not explicitly state Argument 8.

Argument 8.
a. Premise (generalizing 4c, 5c, 7c): The c.637+1G>T variant in *CABP2* may be the cause of arNSHL in several pedigrees.
b. Premise (implicit generally accepted assumption): A homozygous mutation known to have a certain effect in some families will have a similar effect in anyone who inherits it.
c. Conclusion (implicit): Anyone having homozygous c.637+1G>T variants of *CABP2* may be affected by arNSHL.

Such an argument could be described by the following argumentation scheme.

Induction/Generalization
Premise: P is true of some members S of a class C.
Conclusion: P is true for all members of C.
Critical question: *What if the individuals in S are exceptional with respect to P?*

The conclusion of Argument 8 is needed as a premise of Argument 9 for the practice implications of the article given in Excerpt 8 (which, unlike the other excerpts in this paper, comes from the article's Discussion section).

Excerpt 8:
"In conclusion, we identified mutations in *CABP2* in individuals with moderate-to-severe hearing loss. Mutations in *CABP2* cause an audiometric phenotype that is seen in most families segregating arNSHL. Our results suggest the importance of screening for mutations in *CABP2*, as well as in *TECTA*, in families with this milder audiometric phenotype." (p. 644)

Argument 9.
a. Premise (implicit): The reader's goal is to prevent or mitigate the occurrence of arNSHL.
b. Premise (implicit, same as 8c): Someone having homozygous c.637+1G>T variants of *CABP2* may be affected by arNSHL.
c. Premise (implicit): Screening may determine if someone has homozygous c.637+1G>T variants.
d. Premise: (implicit) Knowing if someone has homozygous c.637+1G>T variants is necessary to prevent or mitigate the occurrence of arNSHL.
e. Conclusion: It is desirable to screen for c.637+1G>T variants in *CABP2*.

Argument 9 can be described as a form of Practical Reasoning.

Practical Reasoning
Premise: Agent's goal is to prevent or mitigate the occurrence of D.
Premise: The occurrence of G may result in D.
Premise: Doing Act may result in Agent's knowing if G.
Premise: Knowing if G is necessary to prevent or mitigate D.
Conclusion: It is desirable for Agent to do Act.

3 Inter-Argument Relationships

The previous section illustrates a chained relationship in Arguments 1-3, i.e., the conclusion of Argument i is a premise of Argument i+1. Arguments 4 and 5 share the same conclusion: *The c.637+1G>T variant in CABP2 may be the cause of arNSHL in Sh10*. The conclusions of Arguments 4, 5, and 7 (*The c.637+1G>T variant in CABP2 may be the cause of arNSHL in Sh11 and He*) in combination support the premise of Argument 8, whose conclusion is: *Anyone having homozygous c.637+1G>T variants of CABP2 may be affected by arNSHL*. The conclusion of Argument 8 is further supported by the conclusion of Argument 6: *A c.637+1G>T mutation of CABP2 may result in a disease of the auditory system.*

To provide an explanation for why the authors chose to provide various arguments, rather than merely observing their presence in the text, we must consider how the authors expect their arguments to be challenged or evaluated by the intended audience. Note that the chain of Arguments 1-3 could be challenged by posing the instantiated critical question of Argument 1: *What if the cause of V:14's genetic condition was not in the set of candidates that were tested?* Rather than directly responding to that critical question, the authors continue with Argument 4 whose claim is that the c.637+1G>T variant is the cause of arNSHL in V:14's family, Sh10. In other words, Argument 4 makes a broader claim, a claim that subsumes the claim of Argument 3.

Argument 4 can itself be challenged by posing its critical question: *Is there some other feature G in I-present that could account for P…?* Then one could explain why the authors include Argument 5, in which the Sh10 family is compared to a control group.

Argument 8 can be challenged by posing its critical question: *What if the individuals in S are exceptional with respect to P?* The biochemical argument 6 that a c.637+1G>T mutation of

CABP2 may result in a disease of the auditory system provides a response to that challenge.

Dialogue games have been used to model argumentation between intelligent agents (McBurney and Parsons, 2009) and in human-human dialogue (Budzynska and Reed, 2012). A dialogue game could be used to represent this aspect of discourse structure in scientific articles. (See Figure 1.) We shall refer to this new game as SDG (Science Dialogue Game). As in the ASD game (Walton et al., 2008), SDG incorporates argumentation schemes and critical questions. The locutions of SDG are Argue (Author supports a claim with reasons Ri), Challenge (Reader requests an argument for a reason Ri given in the author's argument), Pose (Reader requests an answer to address an instantiated critical question of the argumentation scheme of the author's argument), and Reject (Author rejects a hypothesis given elsewhere in the text). Reflecting a writer's reliance on discourse context and expected background knowledge and inferential capabilities of the reader, the reasons of an argument may be implicit in SDG.

The Dialogue Rules of SDG reflect weaker ordering constraints in text than in dialogue and the fact that the reader is imaginary: The permissible replies of the reader to Argue are: implicit Challenge, implicit Pose, or silence. The permissible reply of the author to Challenge or Pose is Argue.

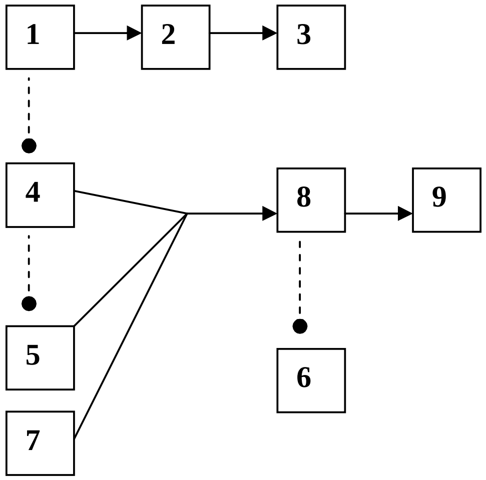


Figure 1. SDG structure. Arrows show conclusion-to-premise support between arguments. Lines ending in circles show responses to critical questions. Conclusions of arguments 4, 5, and 7 are aggregated into premise of argument 8.

4 Discussion

This paper described our analysis of some argumentation schemes and inter-argument relationships in a research article as part of our initial effort to create an annotation scheme. We are continuing to analyze representative articles as preparation for developing and evaluating the annotation scheme. Our longer term goal is to create a freely available corpus of open-access, full-text scientific articles from the biomedical genetics research literature, annotated to support research on argumentation.

Acknowledgments

We thank Dr. Malcolm Schug of the Biology Department of the University of North Carolina Greensboro for verifying our interpretation of the arguments in the genetics article.

References

Budzynska, K. and Reed, C. 2012. The Structure of Adhominem Dialogues. In Verheij, B., Szeider, S., and Woltran, S. (eds.), Computational Models of Argument: Proc. of COMMA 2012. Amsterdam, IOS Press, 454-461.

Green, N. L. 2014. Towards Creation of a Corpus for Argumentation Mining the Biomedical Genetics Research Literature. In *Proc. of the First Workshop on Argumentation Mining*, ACL 2014, Baltimore, MD.

McBurney, P. and Parsons, S. 2009. Dialogue Games for Agent Argumentation. In I. Rahwan, G.R. Si mari (eds.), Argumentation in Artificial Intelli gence, Springer, Dordrecht, pp. 261-280.

Schrauwen et al. 2012. A Mutation in CABP2, Ex pressed in Cochlear Hair Cells, Causes Autosomal-Recessive Hearing Impairment. *The American Journal of Human Genetics* 91, 636-645, October 5, 2012.

Teufel, S. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. Stanford, CA, CSLI Publications.

Walton, D., Reed, C., and Macagno, F. 2008. *Argumentation Schemes.* Cambridge University Press.

Counter-Argumentation and Discourse: A Case Study

Stergos Afantenos
IRIT, Univ. Toulouse
France
stergos.afantenos@irit.fr

Nicholas Asher
IRIT, CNRS,
France
asher@irit.fr

Abstract

Despite the central role that argumentation plays in human communication, the computational linguistics community has paid relatively little attention in proposing a methodology for automatically identifying arguments and their relations in texts. Argumentation is intimately related with discourse structure, since an argument often spans more than one phrase, forming thus an entity with its own coherent internal structure. Moreover, arguments are linked between them either with a support, an attack or a rebuttal relation. Those argumentation relations are often realized via a discourse relation. Unfortunately, most of the discourse representation theories use trees in order to represent discourse, a format which is incapable of representing phenomena such as long distance attachments and crossed dependencies which are crucial for argumentation. A notable exception is Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). In this paper we show how SDRT can help identify arguments and their relations. We use counter-argumentation as our case study following Apothéloz (1989) and Amgoud and Prade (2012) showing how the identification of the discourse structure can greatly benefit the identification of the argumentation structure.

1 Introduction

People use arguments to persuade others to adopt a point of view or action they find beneficial to their interests, or alternatively to prevent others from adopting a position or action that they find contrary to their interests. Of course an agent may find it in her interest to convince an interlocutor to adopt a position she herself does not believe; for instance, a seller may want to persuade a buyer that a product is worth more than she believes it is worth. Because argumentation involves an interaction between an arguer and an addressee, it involves game theoretic aspects: it is the means in language for getting an agent to a position of agreement with the position one is advocating, or in game theoretic terms it is an equilibrium in a persuasion game in which the addressee adopts an optimal action based on the conversational history and in which the arguer adopts her conversational strategy based on the addressee's strategy for adopting an action (Glazer and Rubinstein, 2004). Yet, despite its importance in human communication and behavior and despite the fact that textual realizations of arguments and debates are numerous on the web, it is surprising that this area has received very little attention by the Computational Linguistics community.

One domain of research in Computational Linguistics that is of particular interest for argumentation is that of discourse. In a typical argumentation process, which takes the form of a dialogue, every argument has an internal coherence meaning that it can be represented by a discourse graph. Moreover arguments are linked between themselves either with support, attack or rebuttal relations which are realized once again as discourse relations linking either the whole discourse subgraphs representing the arguments or parts of them. Any attempt to automatically extract the argumentation structure from a given text cannot afford to ignore discourse. Our goal in this paper is to show how argumentation is intimately involved with discourse structure. We achieve this by using counter-argumentation (following (Apothéloz, 1989; Amgoud and Prade, 2012)) as a case study.

The remainder of this paper is structured as follows. In section 2 we present the current work in the so-called argumentation mining, the subfield of computational linguistics that deals with the automatic extraction of the argumentation structure from texts. In section 3 we tell a few words on discourse and in section 4 we show how SDRT (Segmented Discourse Representatuin Theory, (Asher and Lascarides, 2003)) can be applied in a case study focused on counter-argumentation. In section 5 we present the future work and we conclude this paper.

2 Argumentation in Computational Linguistics

Despite its general neglect, argumentation has been the focus of some work in Computational Linguistics. Teufel (1999), Teufel and Moens (2002) aim at identifying what they call the argumentative zones of scientific articles. The zones they have used include the aim of the paper, general scientific background, description of the authors' previous work, comparison with other works, etc. They are using a naive bayes model trying to classify each sentence into one of the predefined categories using mostly surface features (position, length, etc) and whether the sentence contains title words or words scoring high in terms of $tf.idf$.

Palau and Moens (2009) have recently attempted argumentation mining, or the identification of arguments in a text. They assume that an argument consists of a series of premises and a conclusion. Premises and conclusions are represented by propositions in the text. Of course, not all propositions in a given text are part of an argument. In order to tackle the problem of argumentation mining the authors break it into a series of subtasks. Initially they are interested in performing a binary classification of each proposition into either a proposition participating in an argument or not. Propositions that are positively classified are then sent to a second classifier which determines whether it is a premise or a conclusion. For both classification tasks they use a maximum entropy model and the Araucaria corpus[1] as well as a corpus extracted from the european court of human rights. The features they use for the first classifier include surface features ($\{1, 2, 3\}$-grams, punctuation, sentence and word length), POS information (adverbs, verbs and modal auxiliaries) and syntactic parsing. The second classifier uses again surface features, POS tags for the subject and main verb, simple rhetorical and argumentative patterns as well as the results of the first classifier (although no structured prediction is attempted which would probably be more appropriate, given that the two classifiers are not independent). Of course, once one has identified the propositions that are premises and conclusions, one does not yet have the full arguments. In order to get them, the authors create a simple CFG grammar which tries to get the tree structure of arguments. The authors do not attempt to detect the relations (e.g. support, attack, rebuttal) that connect the arguments between each other.

The Araucaria corpus is used by Feng and Hirst (2011) as well but their goal is not performing argumentation mining. Instead they focus on the task of classifying arguments into *argumentation schemes* (Walton et al., 2008). Araucaria arguments contain enthymemes annotated by human subjects which Feng and Hirst (2011) remove. Moreover, each argument is annotated with various argumentation schemes but the authors keep only the ones that are annotated with Walton's schemes. They keep only the 5 more frequent schemes. In total they have 393 arguments which they classify into one of five schemes. Concerning the classification method, they use the C4.5 algorithm implemented in Weka in order to perform either a *one-vs-all* classification or a pairwise classification. The features they use are divided into general ones concerning all schemes (features reflecting textual surface form) or specific ones for each scheme (mostly cue phrases and patterns).

Cabrio and Villata (2012a; 2012b) take a different stance. Their goal is to use Dung's (1995) abstract argumentation framework in order to detect a set of accepted arguments from online debates. They extract arguments from Debatopedia[2] using textual entailment techniques. More precisely, if a sentence T entails another sentence H then they consider that there is a support relation between the two sentences (and thus points of views) otherwise there is an attack relation. They use the open source software package EDITS[3] in order to perform textual entailment. In order then to identify the set of arguments that would be acceptable by a an external observer the authors use Dung's (1995) abstract argumentation framework. In essence an argument belongs to the aforementioned set if all the arguments attacking it are rejected. An argument is rejected if at least one accepted argument attacks it.

3 Discourse

The little prior work on argumentation has ignored discourse structure, and we think this is a mistake. A complete discourse structure of a dialogue will determine how each interlocutor's contribution relates to other contributions, both her own and that of other dialogue participants. This structure already by itself is crucial to determining the structure of an argument—which attacks are directed towards which other contributions. Moreover, an argument is not just a sequence of attacks but a much more complex structure. For one thing, arguments contain support moves as well; a good persuasion strategy is to explain why one's claims are true, but another is to provide background that will enable the addressee to understand one's reasons, and yet another is to provide more details about the claims themselves. All of these ”strategies” involve in fact rhetorical moves that are different and that may be appropriate in different situations. A discourse structure makes plain these different types of moves through the use of different discourse relations.

In effect, discourse structure has the promise to give a much more detailed picture of the nature and structure of argumentation. At the moment, we don't know exactly what that picture is. But by pursuing the analysis of dialogues in terms of argument structure and discourse structure we can find out.

4 Counter-Argumentation: A Case Study

To illustrate our point in the previous section, we illustrate how constructed examples of different sorts of

[1] http://araucaria.computing.dundee.ac.uk/

[2] http://debatepedia.idebate.org

[3] http://edits.fbk.eu

arguments given by Apothéloz (1989) look from a discourse structure point of view. Apothéloz (1989) identified four different modes of arguing against a given argument. In this work an argument is simply a pair $\mathcal{C}(x) : \mathcal{R}(y)$ where $\mathcal{R}$ represents the *function of reason* and x its content and $\mathcal{C}$ the *function of conclusion* and x its content. x and y can be either propositions, conclusions or enthymemes. Given the above, Apothéloz (1989) distinguishes between four different modes of arguing against a given argument $\mathcal{C}(x)$:

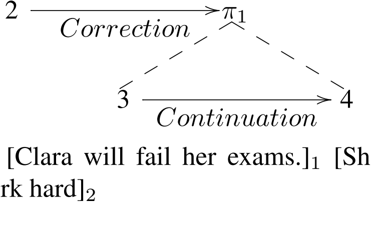
1. disputing the plausibility or the truth of a reason, that is the propositions used in y
2. disputing the completeness of the reason
3. disputing the relevance of the reason with respect to the conclusion, and
4. disputing the argumentative orientation of the reason by showing the reason presented is rather in favor of the conclusion's opposite.

Nonetheless, Apothéloz (1989) completely ignores the internal structure that the arguments have. In the following we analyse the different modes of counter-argumentation that Apothéloz (1989) provides, giving examples found in (Amgoud and Prade, 2012). Our goal is to show how discourse analysis can help the field of computational linguistics not only detect relations between arguments but also analyse the internal structure of an argument. In the following, we are using the Segmented Representation Discourse Theory (SDRT) (Asher and Lascarides, 2003). For the sake of representation, discourse is represented as a hypergraph with discourse relations being the edges of the graph and Elementary Discourse Units (EDUs) being nodes containing only one element, while Complex Discourse Units (CDUs) are nodes containing more than one simple elements (Asher et al., 2011).

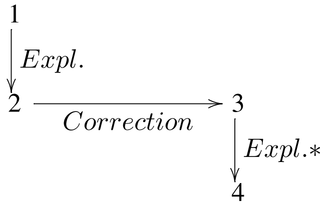
Disputing the plausibility of a reason

When one disputes the plausibility of a reason essentially it amounts to proving that the reason is false. Apothéloz (1989) provides three different ways of showing that; we illustrate them with the following examples.

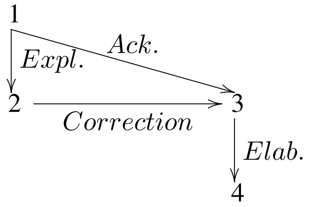
(1) — [Clara will fail her exams.]$_1$ [She did not work hard]$_2$
— [Clara?!]$_3$ [She worked non-stop.]$_4$



(2) — [Clara will fail her exams.]$_1$ [She did not work hard]$_2$
— [No, she worked hard.]$_3$ [Her eyes have bags underneath them.]$_4$



(3) — [Clara works hard]$_1$ [because she is ambitious.]$_2$
— [It is not out of ambition that Clara works hard.]$_3$ [She is not ambitious.]$_4$

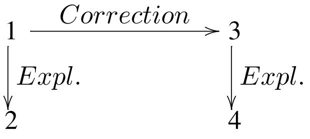


In all three examples, the second speaker does not challenge her interlocutor concerning her conclusion (EDU 1 in all three cases). In fact, in the example (3) the second speaker explicitly acknowledges the content of the conclusion ($Acknowledgment(1, 3)$). Instead the second speaker's disagreement is always with the truth value of the reason behind the conclusion. This takes the form of a $Correction$ relation between the first speaker's EDU representing the reason (EDU 2 in all cases) and the second speaker's counter-argument (EDU 3 for examples (2) and (3) and CDU π_1 for example (1)). For the last two examples the speaker provides additional reason for her beliefs either by means of an $Elaboration$ relation or an $Explanation{*}$ relation. This last relation signals an explanation of why b said that Clara worked hard. It is an explanation of a speech act and provides epistemic grounds for the content of the assertion. Note that in all the above examples the $Correction$ discourse relation amounts to an attack relation.

Disputing the completeness of a reason

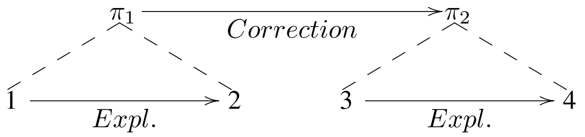
In the second mode of counter-argumentation that Apothéloz (1989) has identified, the second speaker does not attack the truthfulness of the reason but rather its completeness. Here are some examples.

(4) — [Clara will fail her exams.]$_1$ [She did not work hard]$_2$
— [Clara will not fail her exams.]$_3$ [She is very smart.]$_4$



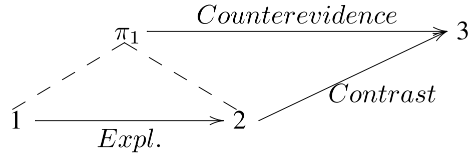
In this example, the second speaker neither affirms neither denies the reason, i.e. the fact that Clara didn't work hard. Instead, she is ignoring it (manifested by the fact that no discourse relation exists between EDUs 2 and 3 or 4). Instead she corrects the conclusion of the first speaker by providing more evidence which lead to the contrary. Again, the $Correction$ discourse relation connects two arguments and serves as an attack argumentative relation.

(5) — [Paul is in his office]$_1$ [because his car is in the carpark.]$_2$
— [But the car is in the carpark]$_3$ [because it has a mechanical problem and is undriveable.]$_4$



In this case both arguments (as before) are thoroughly supported by an $Explanation$ discourse relation. Moreover the second speaker even explicitly agrees with the reason given by the first one ($Acknowledgment(2, 3)$) but she disagrees with the whole argument (note the $Correction$ relation between the two CDUs) since she judges that the reason is not enough and provides more evidence (EDU 4) to back her disagreement up.

(6) — [This object is red]$_1$ [since it looks red.]$_2$
— [But the object is illuminated by a red light.]$_3$

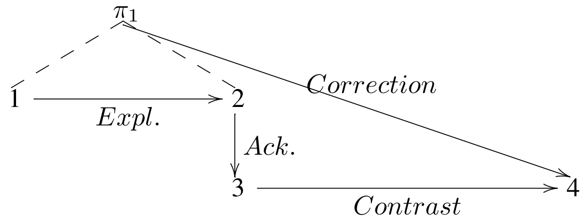


Now, this example is quite more complicated to analyze. There is a contrast between the object's looking red, which generates the expectation that it is red, and the fact that the object is illuminated by a red light, which would tend to put that expectation in doubt. But putting the expectation into doubt also puts into doubt the causal relation supposed by the first speaker between 1 and 2.

Disputing the relevance of a reason

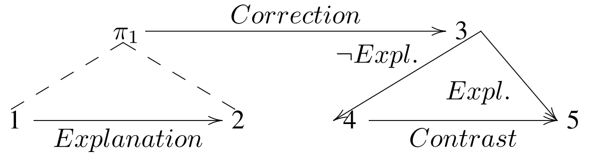
In the third mode of counter-argumentation that Apothéloz (1989) has identified concerns the second speaker does not attack the truthfulness or the completeness of a reason but instead its relevance. Below are some examples of this mode of counter-argumentation.

(7) — [Clara will fail her exams.]$_1$ [She did not work hard]$_2$
— [Indeed, she did not work hard,]$_3$ [but not working hard is not a reason to necessarily fail one's exams.]$_4$



Here the second speaker acknowledges the reason of the first person, as seen by the discourse relation between EDUs 2 and 3, but then shows that there is a contrast between this and her conclusion, disagreeing thus with the whole argument. It is important to note once again that in this example, as the preceding ones, the discourse analysis enables us to clearly pinpoint which elements of the first argument are accepted and which are attacked by the second speaker.

(8) — [Clara will fail her exams.]$_1$ [She did not work hard]$_2$
— [She will not fail her exams]$_3$ [because she did not work hard,]$_4$ [but rather because of the stress.]$_5$

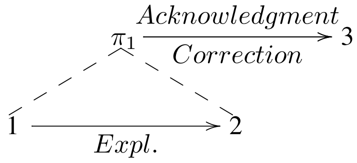


This is a very interesting example. As the discourse analysis shows the undirected cycle that is produced between EDUs 3, 4 and 5 enables the second speaker to explain why she disagrees with the whole of the initial statement.

Disputing the argumentative orientation of a reason

In the final mode of counter-argumentation that Apothéloz (1989) has proposed the second speaker does not dispute neither the reason nor the conclusion. Instead she argues that the reason corroborates towards the opposite of the conclusion. This can be illustrated with the following example.

(9) — [Running a marathon is exhausting.]$_1$ [The whole body undergoes too much stress.]$_2$
— [That's precisely what makes it nice!]$_3$



5 Discussion and Future Work

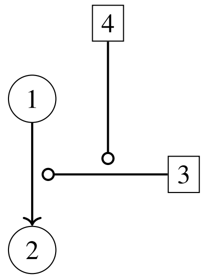
In the previous section we have showed via the use of a case study how the use of a discourse representation theory can help us represent in fine detail the phenomena that take place during argumentation—in this particular case, counter argumentation during a dialogue. In order to represent discourse we have chosen to use the Segmented Discourse Representation Theory (SDRT) of Asher and Lascarides (2003). This choice was made after careful consideration of the phenomena present during argumentation as well as the expressive power of other discourse representation theories.

Take for example the Rhetorical Structure Theory (RST, Mann and Thompson (1988)), which is the most widely cited and used discourse representation theory currently. In RST, as in SDRT, the basic units are the same, namely EDUs.[4] In RST *adjacent* EDUs can be linked together with rhetorical relations in order to form what in RST's jargon are called *spans*. Spans can be linked with rhetorical relations either with other *adjacent* EDUs or *adjacent* spans. We keep on emphasizing the word "adjacent" since this constitutes in our opinion (but see also (Peldszus and Stede, 2013)) a limitation of RST since it does not allow this theory to have *long distance dependencies*, a crucial phenomenon in argumentation. SDRT does not have this limitation. Consider example (7). In this simple example the $Correction$ relation—which, incidentally, is the backbone of the second speaker's attack—holds between non-adjacent EDUs. Even if the first speaker's argument was much longer, or if the second speaker elaborated on the fact that Clara did not work hard (and thus we had many EDUs intervening between π_1 and 4) it wouldn't influence the fact that the complex segment π_1 would be attached to EDU 4. Such long distance attachments are impossible with SDRT which requires that each EDU or span is attached to an adjacent EDU or span.

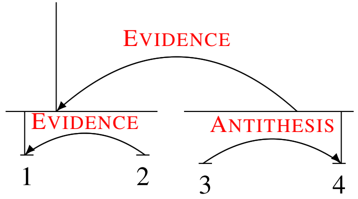
The second problem that RST has as far as the representation of argumentative structures is concerned, is that it cannot correctly represent rebuttals. This is problem that is also reported by Peldszus and Stede (2013) so we are using their example, slightly modified in order to illustrate this point. Consider the following dialogue:

(10) — [We should tear the building down.]$_1$ [It is full of asbestos.]$_2$
— [It is possible to clean it up.]$_3$
— [But that would be forbiddingly expensive!]$_4$

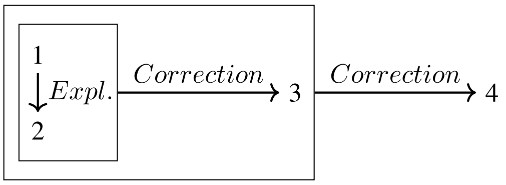
The argumentation graph that results from this dialogue, according to the scheme proposed in (Peldszus and Stede, 2013) is the following:



where edges with arrows denote support relations and edges with circles denote undercuts. The RST graph for the above dialogue is the following:



As we can see, the structural properties of those two graphs are completely different and the use of RST for argumentative analysis does not seem to be a promising path to follow. On the other hand, SDRT neatly follows the argumentation graph (we have used the box representation of SDRT here) making it thus more appropriate for use in argumentative analysis.



At this point we would like to say a few words on the computational extraction of discourse structures. Most of the published work currently is using the RST framework. This is due to two facts. Firstly there are more annotated data available for RST and secondly the problem is computationally less demanding since decisions are always made locally (attachments can be either left or right of a given span) which renders this framework more simple and thus more attractive to researchers. Of course, this implies that all long distance attachements are completely lost, an aspect which is crucial, as we have seen, for argumentation.

Muller et al. (2012) have recently attempted extraction of SDRT structures using data from the ANNODIS corpus (Afantenos et al., 2012), annotated with SDRT structures, with state of the art results. The authors attack the problem of predicting SDRT discourse structures by making some simplifications to the objects that they need to predict, namely they eliminate CDUs by making the assumption that, semantically speaking, attachment to a CDU amounts to attaching to its head—that is the uppermost and leftmost EDU. They have thus structures reminiscent of dependency graphs in syntactic analysis.

The authors perform structured prediction on the dependency graphs they produced which can be broken

[4] There is a big difference as far as EDUs are concerned between the two theories. In SDRT EDUs can be embedded the one within the other whilst RST does not allow it.

down into two steps. Initially they learn local probability distributions for attaching and labeling EDUs, based on naive bayes and logistic regression models. They effectively thus create a complete graph where each node represents an EDU and each arc a probability of attachment. The authors then move on to the decoding phase where the goal is to extract the graph that approaches the reference object. They use two decoding approaches based on $A*$ and Maximum Spanning Tree (MST) algorithms.

Closing this paper we would like to state that one of the main reasons that extraction of argumentative structures has not been more widely explored by the computational linguistics community is due to the fact that few annotated corpora exist. We believe that a project with the goal of jointly annotating argumentative and discourse structures is crucial for the advancement of this field, as well as other fields such as automatic summarization (Afantenos et al., 2008), question answering, etc.

References

Stergos D. Afantenos, Vangelis Karkaletsis, Panagiotis Stamatopoulos, and Constantin Halatsis. 2008. Using Synchronic and Diachronic Relations for Summarizing Multiple Documents Describing Evolving Events. *Journal of Intelligent Information Systems*, 30(3):183–226, June.

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cecile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paul Pery-Woodley, Laurent Prevot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret, and Laure Vieu. 2012. An empirical resource for discovering cognitive principles of discourse organisation: the annodis corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Leila Amgoud and Henri Prade. 2012. Can ai models capture natural language argumentation? *International Journal of Cognitive Informatics and Natural Intelligence*, 6(3):19–32, July.

Denis Apothéloz. 1989. Esquisse d'un catalogue des formes de la contre-argumentation. *Travaux du Centre de Recherches Sémiologiques*, 57:69–86.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK.

Nicholas Asher, Antoine Venant, Philippe Muller, and Stergos D. Afantenos. 2011. Complex discourse units and their semantics. In *Contstraints in Discourse (CID 2011)*, Agay-Roches Rouges, France.

Elena Cabrio and Serena Villata. 2012a. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea, July. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2012b. Natural Language Arguments: A Combined Approach. In *20th European Conference on Artificial Intelligence (ECAI 2012)*, Montpellier, France.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jacob Glazer and Ariel Rubinstein. 2004. On optimal rules of persuasion. *Econometrica*, 72(6):1715–1736, November.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Towards a Functional Theory of Text Organization. *Text*, 8(3):243–281.

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of COLING 2012*, pages 1883–1900, Mumbai, India, December. The COLING 2012 Organizing Committee.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Simone Teufel and Marc Moens. 2002. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, 28(4):409–445, December.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Articles*. PhD Thesis, University of Edinburgh.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Applying Argument Extraction to Improve Legal Information Retrieval

Kevin D. Ashley
University of Pittsburgh School of Law
Pittsburgh, Pennsylvania, USA 15260
ashley@pitt.edu

Abstract

Argument extraction techniques can likely improve legal information retrieval. Any effort to achieve that goal should take into account key features of legal reasoning such as the importance of legal rules and concepts, support and attack relations among claims, and citation of authoritative sources. Annotation types reflecting these key features will help identify the roles of textual elements in retrieved legal cases in order to better inform assessments of relevance for users' queries. As a result, legal argument models and argument schemes will likely play a central part in the text annotation type system.

1 Introduction

With improved prospects for automatically extracting arguments from text, we are investigating whether and how argument extraction can improve legal information retrieval (IR). An immediate question in that regard is the role that argument models and argument schemes will play in achieving this goal.

For some time, researchers in Artificial Intelligence and Law have developed argument models, formal and dialectical process models to describe arguments and their relations. They have also implemented these models in computer programs that construct legal arguments. Some of these models employ argument schemes to provide semantics and describe reasonable arguments. Each scheme corresponds to a typical domain-specific inference sanctioned by the argument, a kind of *prima facie* reason for believing the argument's conclusion. See (Prakken, 2005, p. 234).

By and large, however, these argument models and schemes and their computational implementations have not had much of a practical effect on legal practice. A primary reason for this is the well-known bottleneck in representing knowledge from the legal texts (e.g., statutes, regulations, and cases) that play such an important role in legal practice in a form so that the the computational implementations can reason with them.

Meanwhile, legal information retrieval systems have proven to be highly functional. They provide legal practitioners with convenient access to millions of legal texts without relying on argument models or schemes, relying instead on Bayesian statistical inference based on term frequency. Users of legal information systems can submit queries in the form of a natural language description of a desired fact pattern and retrieve numerous relevant cases.

Useful as they are, however, legal information retrieval systems do not provide all of the functionality that practitioners could employ. What IR system users often want "is not merely IR, but AR", that is, "argument retrieval: not merely sentences with highlighted terms, but arguments and argument-related information. For example, users want to know what legal or factual issues the court decided, what evidence it considered relevant, what outcomes it reached, and what reasons it gave." (Ashley and Walker, 2013a).

Recently, IBM announced its Debater project, an argument construction engine which, given a corpus of unstructured text like Wikipedia, can automatically construct a set of relevant pro/con arguments phrased in natural language. Built upon the foundation of IBM's Jeopardy-game-winning Watson question answering system, the advent of Debater raises some interesting related questions. A central hypothesis of the Watson project was to answer questions based on shallow syntactic knowledge and its implied semantics. This was preferred to formally represented deep semantic knowledge, the acquisition of which is difficult and expensive (Fan et al., 2012). If Debater is applied to legal domains (*See, e.g.,*(Beck, 2014)), one wonders to what extent the same will be true of Debater. In particular, to what extent will explicit argumentation models and their schemes for the legal domain be necessary or useful for the effort to extract legal arguments? And, can techniques in Debater be adapted to improve legal IR?

2 Related Work

The seminal work on extracting arguments and argument-related information from legal case decisions is (Mochales and Moens, 2011). Operationally, the authors defined an argument as "a set of propositions, all of which are premises except, at most, one, which is a conclusion. Any argument follows an argumentation scheme. . . ." Using machine learning based on manually classified sentences from the Araucaria corpus, including court reports, they achieved good performance on classifying sentences as propositions in arguments or not and classifying argumentative propositions as premises or conclusions. Given a limited set of documents, their manually-constructed rule-based argument grammar also generated argument tree structures (Mochales and Moens, 2011).

In identifying argumentative propositions, Mochales and Moens achieved accuracies of 73% and 80% on two corpora, employing domain-general features (including, e.g., each word, pairs of words, pairs and triples of successive words, parts of speech including adverbs, verbs, modal auxiliaries, punctuation, keywords indicating argumentation, parse tree depth and number of subclauses, and certain text statistics.) For classifying argumentative propositions as premises or conclusions, their features included the sentence's length and position in the document, tense and type of main verb, previous and successive sentences' categories, a preprocessing classification as argumentative or not, and the type of rhetorical patterns occurring in the sentence and surrounding sentences (i.e., Support, Against, Conclusion, Other or None). Additional features, more particular to the legal domain included whether the sentence referred to or defined a legal article, the presence of certain argumentative patterns (e.g. "see", "*mutatis mutandis*", "having reached this conclusion", "by a majority") and whether the agent of the sentence is the plaintiff, the defendant, the court or other (Mochales and Moens, 2011).

Factors, stereotypical fact patterns that strengthen or weaken a side's argument in a legal claim, have been identified in text automatically. Using a HYPO-style CBR program and an IR system relevance feedback module, the SPIRE program retrieved legal cases from a text corpus and highlighted passages relevant to bankruptcy law factors (Daniels and Rissland, 1997). The SMILE+IBP program learned to classify case summaries in terms of applicable trade secret law factors (Ashley and Brüninghaus, 2009), analyzed automatically classified squibs of new cases, predicted outcomes, and explained the predictions. (Wyner and Peters, 2010) presents a scheme for annotating 39 trade secret case texts with GATE in terms of finer grained components (i.e., factoroids) of a selection of factors.

Using an argument model to assist in representing cases for conceptual legal information retrieval was explored in (Dick and Hirst, 1991). More recently, other researchers have addressed automatic semantic processing of case decision texts for legal IR, achieving some success in automatically:

- assigning rhetorical roles to case sentences based on 200 manually annotated Indian decisions (Saravanan and Ravindran, 2010),
- categorizing legal cases by abstract Westlaw categories (e.g., bankruptcy, finance and banking) (Thompson, 2001) or general topics (e.g., exceptional services pension, retirement) (Gonçalves and Quaresma, 2005),
- extracting treatment history (e.g., "affirmed", "reversed in part") (Jackson et al., 2003),
- determining the role of a sentence in the legal case (e.g., as describing the applicable law or the facts) (Hachey and Grover, 2006),
- extracting offenses raised and legal principles applied from criminal cases to generate summaries (Uyttendaele et al., 1998),
- extracting case holdings (McCarty, 2007), and
- extracting argument schemes from the Araucaria corpus such as argument from example and argument from cause to effect (Feng and Hirst, 2011).

We aim to develop and evaluate an integrated approach using both semantic and pragmatic (contextual) information to retrieve arguments from legal texts in order to improve legal information retrieval. We are working with an underlying argumentation model and its schemes, the Default Logic Framework (DLF), and a corpus of U.S. Federal Claims Court cases (Walker et al., 2011; Walker et al., 2014; Ashley and Walker, 2013a). Like (Mochales and Moens, 2011) and (Sergeant, 2013), we plan to:

1. Train an annotator to automatically identify propositions in unseen legal case texts,
2. Distinguish argumentative from non-argumentative propositions and classify them as premises or conclusions,
3. Employ rule-based or machine learning models to construct argument trees from unseen cases based on a manually annotated training corpus, but also to
4. Use argument trees to improve legal information retrieval reflecting the uses of propositions in arguments.

Before sketching our approach for the legal domain, however, we note that IBM appears to have developed more domain independent techniques for identifying propositions in documents and classifying them as premises in its Debater system.[1]

On any topic, the Debater's task is to "detect relevant claims" and return its "top predictions for pro claims and con claims." On inputting the topic, "The sale of violent videogames to minors should be banned," for example, Debater:

(1) scanned 4 million Wikipedia articles,

(2) returned the 10 most relevant articles,

(3) scanned the 3000 sentences in those 10 articles,

(4) detected those sentences that contained "candidate claims",

(5) "identified borders of candidate claims",

(6) "assessed pro and con polarity of candidate claims",

(7) "constructed a demo speech with top claim predictions", and

(8) was then "ready to deliver!"

Figure 1 shows an argument diagram constructed manually from the video recording of Debater's oral output for the example topic.

3 Key Elements of Legal Argument

Debater's argument regarding banning violent video games is meaningful but compare it to the *legal* argument concerning a similar topic in Figure 2. The Court in Video Software Dealers Assoc. v. Schwarzenegger, 556 F. 3d 950 (9th Cir. 2009), addressed the issue of whether California (CA) Civil Code sections 1746-1746.5 (the "Act"), which restrict sale or rental of "violent video games" to minors, were unconstitutional under the 1st and 14th Amendments of the U.S. Constitution. The Court held the Act unconstitutional. As a presumptively invalid content-based restriction on speech, the Act is subject to strict scrutiny and the State has not demonstrated a compelling interest.

In particular, the Court held that CA had not demonstrated a compelling government interest that "the sale of violent video games to minors should be banned." Figure 2 shows excerpts from the portion of the opinion in which the Court justifies this conclusion. The nodes contain propositions from that portion and the arcs reflect the explicit or implied relations among those propositions based on a fair reading of the text.

The callout boxes in Figure 2 highlight some key features of legal argument illustrated in the Court's argument:

1. Legal rules and concepts govern a court's decision of an issue.
2. Standards of proof govern a court's assessment of evidence.
3. Claims have support / attack relations.
4. Authorities are cited (e.g., cases, statutes).
5. Attribution information signals or affects judgments about belief in an argument (e.g., "the State relies").
6. Candidate claims in a legal document have different plausibility.

[1] *See, e.g.,* `http://finance.yahoo.com/blogs/the-exchange/ibm-unveils-a-computer-than-can-argue-181228620.html`. A demo appears at the 45 minute mark: `http://io9.com/ibms-watson-can-now-debate-its-opponents-1571837847`.

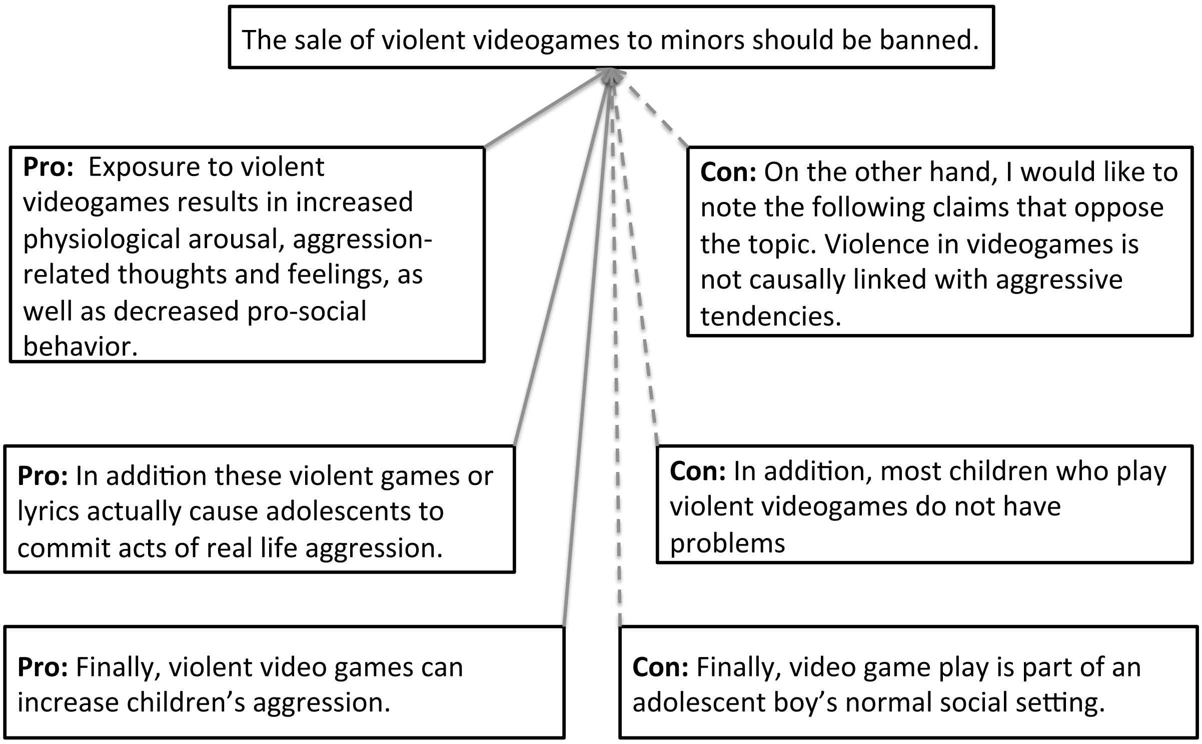

The sale of violent videogames to minors should be banned.
Pro: Exposure to violent videogames results in increased physiological arousal, aggression-related thoughts and feelings, as well as decreased pro-social behavior.
Con: On the other hand, I would like to note the following claims that oppose the topic. Violence in videogames is not causally linked with aggressive tendencies.
Pro: In addition these violent games or lyrics actually cause adolescents to commit acts of real life aggression.
Con: In addition, most children who play violent videogames do not have problems
Pro: Finally, violent video games can increase children's aggression.
Con: Finally, video game play is part of an adolescent boy's normal social setting.

Figure 1: Argument Diagram of IBM Debater's Output for Violent Video Games Topic (root node)

Although the argument diagrams in Figures 1 and 2 address nearly the same topic and share similar propositions, the former obviously lacks these features that would be important in legal argument (and, as argued later, important in using extracted arguments to improve legal IR). Of course, on one level this is not surprising; the Debater argument is *not* and does not purport to be a legal argument.

On the other hand, given the possibility of applying Debater to legal applications and argumentation, it would seem essential that it be able to extract such key information. In that case, the question is the extent to which explicit argument models and argument schemes of legal reasoning would be useful in order to assist with the extraction of the concepts, relationships, and information enumerated above and illustrated in Figure 2.

4 Default-Logic Framework

Vern Walker's Default Logic Framework (DLF) is an argument model plus schemes for evidence-based legal arguments concerning compliance with legal rules. At the Research Laboratory for Law, Logic and Technology (LLT Lab) at Hofstra University, researchers have applied the DLF to model legal decisions by Court of Federal Claims "Special Masters" concerning whether claimants' compensation claims comply with the requirements of a federal statute establishing the National Vaccine Injury Compensation Program. Under the Act, a claimant may obtain compensation if and only if the vaccine caused the injury.

In order to establish causation under the rule of Althen v. Secr. of Health and Human Services, 418 F.3d 1274 (Fed.Cir. 2005), the petitioner must establish by a preponderance of the evidence that: (1) a "medical theory causally connects" the type of vaccine with the type of injury, (2) there was a "logical sequence of cause and effect" between the particular vaccination and the particular injury, and (3) a "proximate temporal relationship" existed between the vaccination and the injury. Walker's corpus comprises all decisions in a 2-year period applying the *Althen* test of causation-in-fact (35 decision texts, 15-40 pages per decision). In these cases, the Special Masters decide which evidence is relevant to which issues of fact, evaluate the plausibility of evidence in the legal record, organize evidence and draw reasonable inferences, and make findings of fact.

The DLF model of a single case "integrates numerous units of reasoning" each "consisting of one

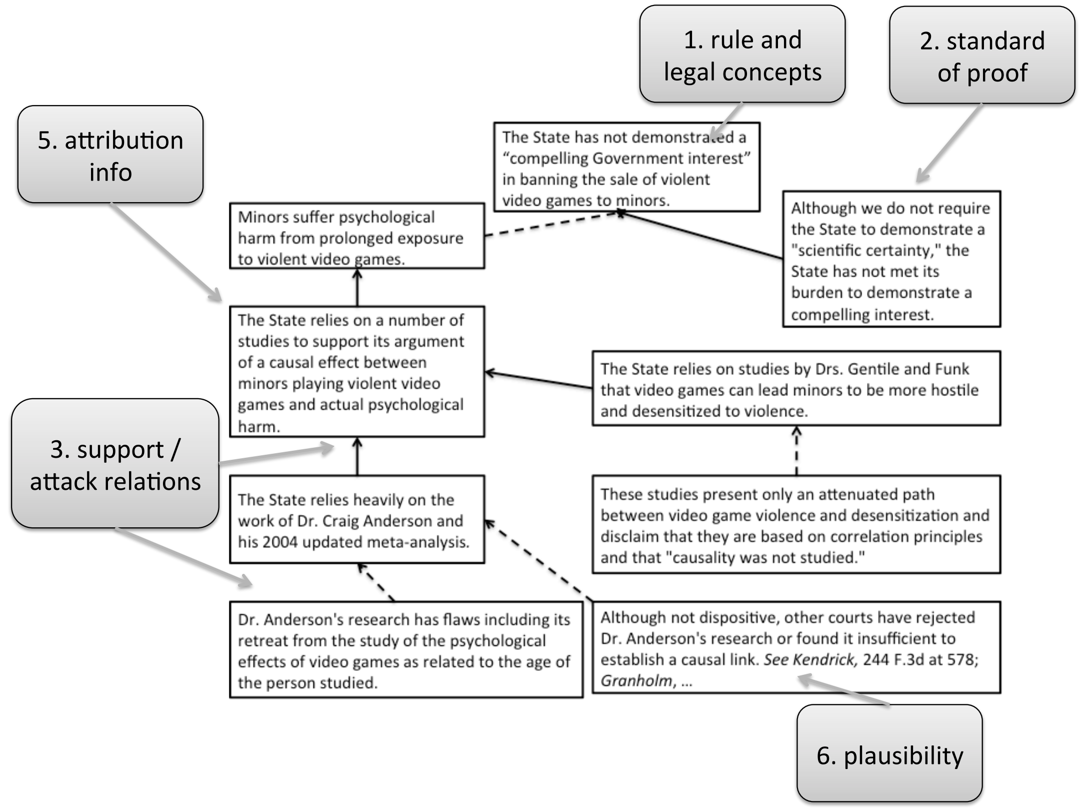

1. rule and legal concepts
2. standard of proof
5. attribution info
The State has not demonstrated a "compelling Government interest" in banning the sale of violent video games to minors.
Minors suffer psychological harm from prolonged exposure to violent video games.
Although we do not require the State to demonstrate a "scientific certainty," the State has not met its burden to demonstrate a compelling interest.
The State relies on a number of studies to support its argument of a causal effect between minors playing violent video games and actual psychological harm.
The State relies on studies by Drs. Gentile and Funk that video games can lead minors to be more hostile and desensitized to violence.
3. support / attack relations
The State relies heavily on the work of Dr. Craig Anderson and his 2004 updated meta-analysis.
These studies present only an attenuated path between video game violence and desensitization and disclaim that they are based on correlation principles and that "causality was not studied."
Dr. Anderson's research has flaws including its retreat from the study of the psychological effects of video games as related to the age of the person studied.
Although not dispositive, other courts have rejected Dr. Anderson's research or found it insufficient to establish a causal link. See Kendrick, 244 F.3d at 578; Granholm, ...
6. plausibility

Figure 2: Diagram Representing Realistic Legal Argument Involving Violent Video Games Topic

conclusion and one or more immediately supporting reasons (premises)" and employing four types of connectives (min (and), max (or), evidence factors, and rebut) (Walker et al., 2014). For example, Figure 3 shows an argument diagram representing the excerpt of the the DLF model of the special master's finding in the case of Cusati v. Secretary of Health and Human Services, No. 99-0492V (Office of Special Masters, United States Court of Federal Claims, September 22, 2005) concerning whether the first *Althen* condition for showing causation-in-fact is satisfied.

The main point is that the DLF model of a legal argument and its argument schemes represent the above-enumerated key features of legal argument. As illustrated in the callout boxes of Figure 3, the model indicates: (1) the 1st *Althen* rule and causation-in-fact concept that govern the decision of the causation issue, (2) the preponderance of evidence standard of proof governing the court's assessment, (3) support relations among the propositions, the Special Master having recorded no counterarguments, (4) citation to the statute, 42 USC 300aa-11(c)(1)(C)(ii)), and to the *Althen* and *Shyface* case authorities, (5) some attribution information that signals judgments about the Special Master's belief in an argument (e.g., "Dr. Kinsbourne and Dr. Kohrman agree"), and (6) four factors that increase plausibility of the claim of causation.

5 Legal Argument and Legal IR

Legal decisions contain propositions and arguments how to "prove" them. Prior cases provide examples of how to make particular arguments in support of similar hypotheses and of kinds of arguments that have succeeded, or failed, in the past. Consider a simple query discussed in (Ashley and Walker, 2013a): Q1: "MMR vaccine can cause intractable seizure disorder and death."

An attorney/user in a new case where an injury followed an MMR vaccination might employ this query to search for cases where such propositions had been addressed. Relevant cases would add confidence that the propositions and accompany-

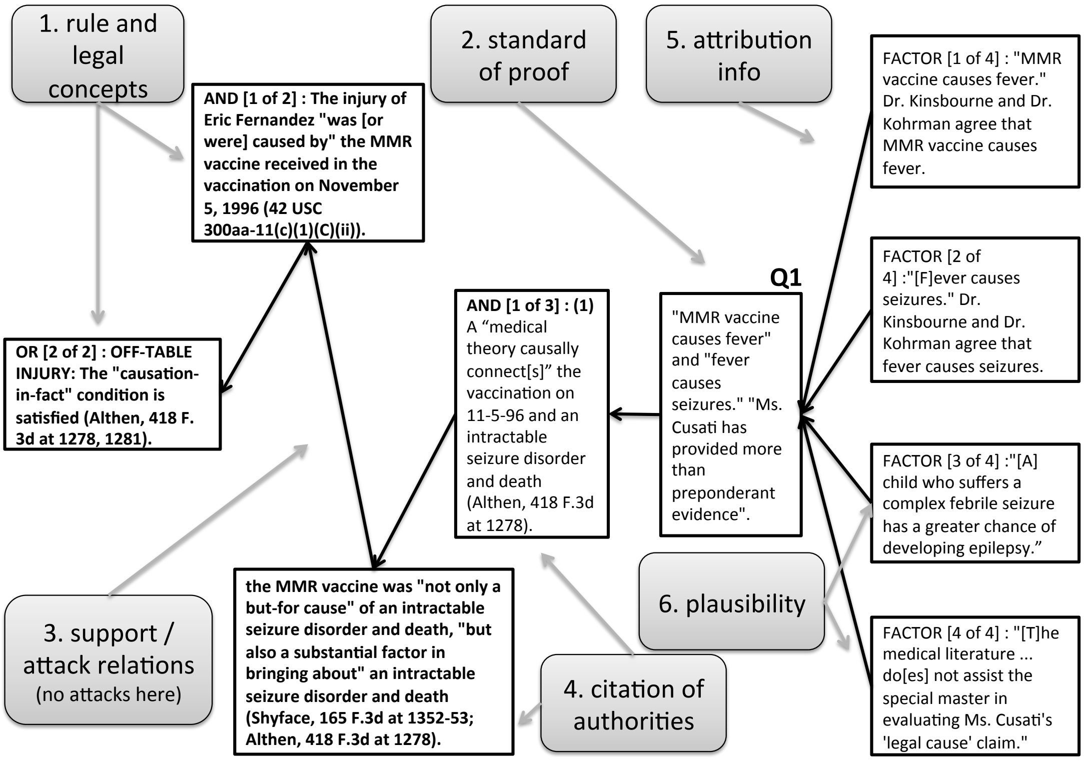

1. rule and legal concepts
2. standard of proof
5. attribution info
AND [1 of 2] : The injury of Eric Fernandez "was [or were] caused by" the MMR vaccine received in the vaccination on November 5, 1996 (42 USC 300aa-11(c)(1)(C)(ii)).
FACTOR [1 of 4] : "MMR vaccine causes fever." Dr. Kinsbourne and Dr. Kohrman agree that MMR vaccine causes fever.
Q1
FACTOR [2 of 4] :"[F]ever causes seizures." Dr. Kinsbourne and Dr. Kohrman agree that fever causes seizures.
AND [1 of 3] : (1) A "medical theory causally connect[s]" the vaccination on 11-5-96 and an intractable seizure disorder and death (Althen, 418 F.3d at 1278).
"MMR vaccine causes fever" and "fever causes seizures." "Ms. Cusati has provided more than preponderant evidence".
OR [2 of 2] : OFF-TABLE INJURY: The "causation-in-fact" condition is satisfied (Althen, 418 F. 3d at 1278, 1281).
FACTOR [3 of 4] :"[A] child who suffers a complex febrile seizure has a greater chance of developing epilepsy."
6. plausibility
the MMR vaccine was "not only a but-for cause" of an intractable seizure disorder and death, "but also a substantial factor in bringing about" an intractable seizure disorder and death (Shyface, 165 F.3d at 1352-53; Althen, 418 F.3d at 1278).
3. support / attack relations (no attacks here)
4. citation of authorities
FACTOR [4 of 4] : "[T]he medical literature ... do[es] not assist the special master in evaluating Ms. Cusati's 'legal cause' claim."

Figure 3: Diagram of DLF Model of Special Master's Finding in *Cusati* Case re 1st *Althen* Condition

ing arguments were reasonable and had been successful.

Importantly, the cases retrieved will be more relevant to the extent that the proposition is used in a similar argument. That is, they will be more relevant to the extent that the proposition plays roles in the case arguments similar to the role in which the attorney intends to use it in an argument about the current case.

An argument diagram like that of Figure 3 can illustrate the effect of the six key elements of legal reasoning illustrated above on how relevant a retrieved case is to a user's query. The diagram shows a legal argument in which the proposition corresponding to Q1 plays a role in the *Cusati* case as an evidence-based finding of the Special Master, namely, that "MMR vaccine causes fever" and "fever causes seizures."

Such diagrams have a "legal rule-oriented" direction (i.e., to the left in Figure 3) and an "evidentiary factors-oriented" direction (i.e., to the right in this diagram). For instance, an attorney whose client sustained seizures after receiving the MMR vaccine probably knows that he/she will have to satisfy a requirement of causation. The attorney may not know, however, what legal standard defines the relevant concept of causation or what legal authority may be cited as an authoritative source of the standard. In that situation, retrieved cases will likely be more relevant to the extent that that they fill in the legal rule-oriented direction, relative to a proposition similar to the one marked "Q1", with *legal rules* about the *concept* of causation and *citations* to their *authoritative sources*.

If the attorney is unsure of the kinds of evidence that an advocate should employ in convincing a Special Master to make the finding of fact on causation or of the relevant standard of proof for assessing that evidence of causation, retrieved cases will be more relevant to the extent that they fill in the evidentiary factors-oriented direction, relative to a proposition similar to the one marked "Q1", with evidentiary factors and an identification of the *standard of proof*.

The attorney may be interested in better understanding how to improve the *plausibility* of a proposition about causation as an evidence-based finding. Cases will be more relevant to the extent that they contain evidentiary factors that *support* such a finding. An attorney interested in attacking the *plausibility* of the evidence-based finding might be especially interested in seeing cases involving examples of evidentiary factors that *attack* such a finding.

Finally, the cases will be more relevant to the extent that the proposition similar to the one marked "Q1" concerning MMR vaccine's causing injury is *attributable* to the Special Master as opposed merely to some expert witness's statement.

6 Specifying/Determining Propositions' Argument Roles

The importance of a proposition's argument role in matching retrieved cases to users' queries raises two questions: (1) How does the user specify the target propositions and their argumentative roles in which he is interested? (2) How does a program determine the roles that propositions play in retrieved case arguments?

An argument diagram like that of Figure 3 may play a role in enabling users to specify the arguments and propositions in which they are interested. One can imagine a user's inputting a query by employing a more abstract version of such a diagram. For instance, in the Query Input Diagram of Figure 4, the nodes are labeled with, or refer to, argument roles. These roles include:

Legal Rule: sentences that state a legal rule in the abstract, without applying the rule to the particular case being litigated

Ruling/Holding: sentences that apply a legal rule to decide issues presented in the particular case being litigated

Evidence-Based Finding: sentences that report a trier-of-fact's ultimate findings regarding facts material to the particular case being litigated

Evidence-Based Reasoning: sentences that report the trier-of-fact's reasoning in assessing the relevant evidence and reaching findings regarding facts material to the particular case being litigated (e.g., evidentiary factors)

Evidence: sentences that describe any type of evidence legally produced in the particular case being litigated, as part of the proof intended to persuade the trier-of-fact of alleged facts material to the case (e.g., oral testimony of witnesses, including experts on technical matters; documents, public records, depositions; objects and photographs)

Citation: sentences that credit and refer to authoritative documents and sources (e.g., court decisions (cases), statutes, regulations, government documents, treaties, scholarly writing, evidentiary documents)

In the "text", "concept", and "citation" slots of the appropriate nodes of the query input diagram, Figure 4, users could specify the propositions, concepts, or citations that they know or assume and check the targeted nodes in the directions (rule-oriented or evidentiary-factors-oriented) or ranges that they hope to fill through searching for cases whose texts satisfy the diagram's argument-related constraints. In effect, the diagram will guide the IR system in ranking the retrieved cases for relevance and in highlighting their relevant parts.

Regarding the second question, concerning how a program will determine propositions' argument roles in case texts, that is the third task that Mochales and Moens addressed with a rule-based grammar applied to a small set of documents. While their rules employed some features particular to legal argument, (e.g., whether a sentence referred to a legal article) one imagines that additional features would be needed, pertaining to legal argument or to the regulated domain of interest. These features would become the predicates of additional grammar rules or be annotated in training cases for purposes of machine learning.

The legal argument roles listed above are a first cut at a more comprehensive enumeration of the types of legal argument features with which to annotate legal case texts in an Unstructured Information Management Architecture (UIMA) annotation pipeline for purposes of extracting argument information and improving legal IR.

UIMA, an open-source Apache framework, has been deployed in several large-scale government-sponsored and commercial text processing applications, most notably, IBM's Watson question answering system (Epstein et al., 2012). A UIMA

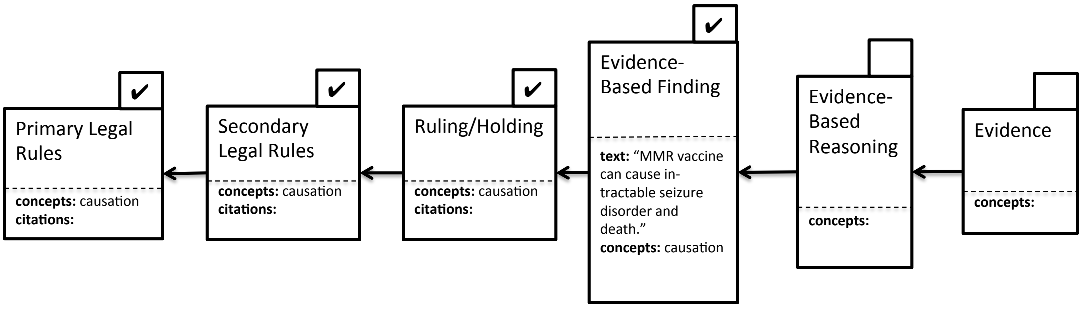


Figure 4: Sample Query Input Diagram

pipeline is an assemblage of integrated text annotators. The annotators are "a scalable set of cooperating software programs, . . . , which assign semantics to some region of text" (Ferrucci, 2012), and "analyze text and produce annotations or assertions about the text" (Ferrucci et al., 2010, p. 74).

A coordinated type system serves as the basis of communication among these annotators; a type system embodies a formalization of the annotators' analysis input and output data (Epstein et al., 2012, p. 3). In (Ashley and Walker, 2013b) and (Ashley and Walker, 2013a) the authors elaborate three additional bases for annotations, which, with further refinement, may serve as a conceptual substrate for the annotation types listed above:

1. DLF annotations, as suggested in Figure 3, capture "(i) the applicable statutory and regulatory requirements as a tree of authoritative rule conditions (i.e., a "rule tree") and (ii) the chains of reasoning in the legal decision that connect evidentiary assertions to the special master's findings of fact on those rule conditions (Walker et al., 2011)."

2. Annotations in terms of presuppositional information that "identifies entities (e.g., types of vaccines or injuries), events (e.g., date of vaccination or onset of symptoms) and relations among them used in vaccine decisions to state testimony about causation, assessments of probative value, and findings of fact." (Ashley and Walker, 2013a).

3. Annotations of of argument patterns based on: inference type (e.g., deductive or statistical), evidence type (e.g., legal precedent, policy, fact testimony), or type of weighing of source credibility to resolve evidentiary discrepancies (e.g., in terms of expert vs. expert or of adequacy of explanation) (Walker et al., 2014) .

If we succeed in designing a system of coordinated legal annotation types and operationalizing a UIMA annotation pipeline, we envision adding a module to a full-text legal IR system. At *retrieval time* it would extract semantic / pragmatic legal information from the top *n* cases returned by a traditional IR search and re-rank returned cases to reflect the user's diagrammatically specified argument need. The module would also summarize highly ranked cases and highlight argument-related information (Ashley and Walker, 2013a). Since the module processes the texts of cases returned by the information retrieval system, no special knowledge representation of the cases in the IR system database is required; the knowledge representation bottleneck will have been circumvented.

7 Conclusion

According to Wittgenstein, meaning lies in the way knowledge is used. Legal argument models and argument schemes can specify roles for legal propositions to play (and, interestingly, Stephen Toulmin was a student of Wittgenstein.) Thus, researchers can enable machines to search for and use legal knowledge intelligently in order, among other things, to improve legal information retrieval.

Although IBM Debater may identify argument propositions (e.g., claims), legal argument schemes could help it to address legal rules and concepts, standards of proof, internal support and attack relations, citation of statutory and case authorities, attribution, and plausibility. Open questions include the extent to which legal expert knowledge will be needed in order to operationalize argument schemes to extract arguments from legal case texts.

Acknowledgments

My colleagues Vern Walker, Matthias Grabmair, and Eric Nyberg make this work possible.

References

K. Ashley and S. Brüninghaus. 2009. Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, pages 125–165.

K. Ashley and V. Walker. 2013a. From information retrieval (IR) to argument retrieval (AR) for legal cases: Report on a baseline study. In K. Ashley, editor, *JURIX*, volume 259 of *Frontiers in Artificial Intelligence and Applications*, pages 29–38. IOS Press.

K. Ashley and V. Walker. 2013b. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In *Proc. 14th Int'l Conf. on Artificial Intelligence and Law*, ICAIL '13, pages 176–180, New York, NY, USA. ACM.

S. Beck. 2014. Emerging technology shapes future of law. `http://www.americanlawyer.com/id=1202664266769/Emerging-Technology-Shapes-Future-of-Law`. Accessed: 2014-09-20.

J. Daniels and E. Rissland. 1997. Finding legally relevant passages in case opinions. In *ICAIL*, pages 39–46.

J. Dick and G. Hirst. 1991. A case-based representation of legal text for conceptual retrieval. In *Proceedings, Workshop on Language and Information Processing*, American Society for Information Science, pages 93–102.

EA Epstein, MI Schor, BS Iyer, A. Lally, EW Brown, and J. Cwiklik. 2012. Making Watson fast. *IBM J. Res. and Dev.*, 56(3.4):15–1.

J. Fan, A. Kalyanpur, DC Gondek, and DA Ferrucci. 2012. Automatic knowledge extraction from documents. *IBM J. Res. and Dev.*, 56(3.4):5–1.

V. Feng and G. Hirst. 2011. Classifying arguments by scheme. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *ACL*, pages 987–996. The Association for Computer Linguistics.

D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. 2010. Building Watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79.

D. Ferrucci. 2012. Introduction to "This is Watson". *IBM J. Res. and Dev.*, 56(3.4):1–1.

T. Gonçalves and P. Quaresma. 2005. Is linguistic information relevant for the classification of legal texts? In *Proc. 10th Int'l Conf. on AI and Law*, ICAIL '05, pages 168–176, NY, NY. ACM.

B. Hachey and C. Grover. 2006. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345.

P. Jackson, K. Al-Kofahi, A. Tyrrell, and A. Vachher. 2003. Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, 150(1-2):239–290, November.

L.T. McCarty. 2007. Deep semantic interpretations of legal texts. In *Proc. 11th Int'l Conf. on AI and Law*, ICAIL '07, pages 217–224, NY, NY. ACM.

R. Mochales and M.-F. Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

H. Prakken. 2005. AI & Law, logic and argument schemes. *Argumentation*, 19(3):303–320.

M. Saravanan and B. Ravindran. 2010. Identification of rhetorical roles for segmentation and summarization of a legal judgment. *Artificial Intelligence and Law*, 18(1):45–76.

A. Sergeant. 2013. Automatic argumentation extraction. In et al. P. Cimiano, editor, *ESWC*, volume 7882 of *Lecture Notes in Computer Science*, pages 656–660. Springer.

P. Thompson. 2001. Automatic categorization of case law. In *Proc. 8th Int'l Conf. on AI and Law*, ICAIL '01, pages 70–77, NY, NY. ACM.

C. Uyttendaele, M.-F. Moens, and J. Dumortier. 1998. Salomon: Automatic abstracting of legal cases for effective access to court decisions. *Artificial Intelligence and Law*, 6(1):59–79.

V. Walker, N. Carie, C. DeWitt, and E. Lesh. 2011. A framework for the extraction and modeling of fact-finding reasoning from legal decisions: Lessons from the vaccine/injury project corpus. *Artificial Intelligence and Law*, pages 291–331.

V. Walker, K. Vazirova, and C. Sanford. 2014. Annotating patterns of reasoning about medical theories of causation in vaccine cases: Toward a type system for arguments. In *Proc. 1st Workshop on Argumentation Mining, ACL 2014*.

A. Wyner and W. Peters. 2010. Lexical semantics and expert legal knowledge towards the identification of legal case factors. In *Proc. 23d Conf. on Legal Knowledge and Information Systems: JURIX 2010*, pages 127–136, Amsterdam. IOS Press.

Argumentation Mining on the Web from Information Seeking Perspective

Ivan Habernal[†‡], **Judith Eckle-Kohler**[†‡], **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract

In this paper, we argue that an annotation scheme for argumentation mining is a function of the task requirements and the corpus properties. There is no one-size-fits-all argumentation theory to be applied to realistic data on the Web. In two annotation studies, we experiment with 80 German newspaper editorials from the Web and about one thousand English documents from forums, comments, and blogs. Our example topics are taken from the educational domain.

To formalize the problem of annotating arguments, in the first case, we apply a Claim-Premise scheme, and in the second case, we modify Toulmin's scheme. We find that the choice of the argument components to be annotated strongly depends on the register, the length of the document, and inherently on the literary devices and structures used for expressing argumentation. We hope that these findings will facilitate the creation of reliably annotated argumentation corpora for a wide range of tasks and corpus types and will help to bridge the gap between argumentation theories and actual application needs.

1 Introduction

Argumentation mining apparently represents an emerging field in Natural Language Processing (NLP) with publications appearing at mainstream conferences, such as ACL (Cabrio and Villata, 2012; Feng and Hirst, 2011; Madnani et al., 2012) or COLING (Stab and Gurevych, 2014; Levy et al., 2014; Wachsmuth et al., 2014a). In particular, there is an increasing need for tools capable of understanding argumentation on the large scale, because in the current information overload, humans cannot feasibly process such massive amounts of data in order to reveal argumentation. Unfortunately, even current Web technologies (such as search engines or opinion mining services) are not suitable for such a task. This drives the research field to the next challenge – argumentation mining on the Web. The abundance of freely available (yet unstructured, textual) data and possible applications of such tools makes this task very appealing.

Our research into argumentation mining is motivated by the *information seeking perspective*. The key sources are discussions (debates) about controversies (contentions) targeted at a particular topic which is of the user's interest. The scope is not limited to a particular media type as the source types can range from the on-line newspapers' editorials to user-generated discourse in social media, such as blogs and forum posts, covering different aspects of the issues. Understanding positions and argumentation in on-line debates helps users to form their opinions on controversial issues and also fosters personal and group decision making (Freeley and Steinberg, 2008, p. 9). The main task would be to identify and extract the core argumentation (its formal aspects will be discussed later) and present this new knowledge to users. By utilizing argumentation mining methods, users can be provided with the most relevant information (arguments) regarding the controversy under investigation.

Although argumentation mining on the Web has already been partly outlined in the literature (Schneider et al., 2012; Sergeant, 2013), the requirements and use-case scenarios differ substantially. Various tasks are being solved, most of them depending on the domain, e.g., product reviews or political contentions. As a result, different interpretations of arguments and argumentation have been developed in NLP, and therefore, most of the existing researches are not directly adaptable.

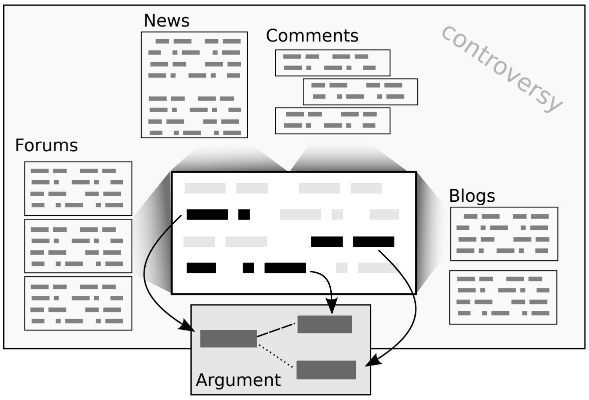


Figure 1: Schematic overview of argumentation mining on the Web

Morover, not all of the related research works are tightly connected to argumentation theories (de Moor et al., 2004; Villalba and Saint-Dizier, 2012; Cabrio et al., 2013b; Llewellyn et al., 2014). However, we feel that it is vital to ground NLP research in argumentation mining in existing work on argumentation.

In this article, we will particularly focus on bridging the gap between argumentation theories and actual application needs that has not been targeted in the relevant literature. We will support our findings by comprehensively surveying existing works and presenting results from two extensive annotation studies.

Our main findings and suggestions can be summarized as follows: First, the use-case of any research in argumentation mining must be clearly stated (i.e., in terms of expected outcomes). Second, properties of the data under investigation must be taken into account, given the variety of genres and registers (Biber and Conrad, 2009). Third, an appropriate argumentation model must be chosen according to the requirements. Therefore, we claim that it is not possible to formulate a single argumentation mining perspective that would be applicable to the Web data in general.

2 Relation to Argumentation Theories

Research on argumentation is widely interdisciplinary, as it spreads across philosophy and rhetoric (Aristotle and Kennedy (translator), 1991; Perelman and Olbrechts-Tyteca, 1991; Walton et al., 2008), informal and formal logic (Dung, 1995; Henkemans, 2000; Stoianovici, 2009; Schneider et al., 2013; Hunter, 2013), educational research (Weinberger and Fischer, 2006; Noroozi et al., 2013), pragmatics (Xu and Wu, 2014), psychology (Larson et al., 2004), and many others. Given so many different perspectives on investigating argumentation, there is a plethora of possible interpretations of argumentation mining. Thus, finding a common understanding of this evolving field is a fundamental challenge.

For NLP research, this overwhelming amount of related works brings many theoretical and practical issues. In particular, there is no one-size-fits-all argumentation theory. Even argumentation researchers disagree on any widely-accepted ultimate concept. For example, Luque (2011) criticizes the major existing approaches in order to establish a new theory which is later again severely criticized by other in-field researches (Andone, 2012; Xie, 2012). Given this diversity of perspectives, NLP research cannot simply adopt one particular approach without investigating its theoretical background as well as its suitability for the particular task.

2.1 What we do not tackle

Given the breath of argumentation mining just outlined, we would also like to discuss aspects that do not fit into our approach to argumentation mining, namely macro argumentation and evaluation using formal frameworks.

First, we treat argumentation as a product (micro argumentation or monological models), not as a process (macro argumentation or dialogical models). While dialogical models highlight the process of argumentation in a dialogue structure, monological models emphasize the structure of the argument itself (Bentahar et al., 2010, p. 215). Therefore, we examine the relationships between the different components of a given argument, not a relationship that can exist between arguments.[1] Exploring how argumentation evolves between parties in time remains out of our scope.

Second, we do not tackle any logical reasoning, defeasibility of reasoning, or evaluating argumentation with formal frameworks in general. Although this is an established field in informal logic (Prakken, 2010; Hunter, 2013; Hunter, 2014), such an approach might not be suitable directly for Web data as it assumes that argumentation is logical (such a strong assumption cannot be guar-

[1] For further discussion see, e.g., (Blair, 2004; Johnson, 2000; Reed and Walton, 2003) or Micheli (2011) who summarizes the distinction between the process (at a pragmatic level) and the product (at a more textual level).

anteed). Furthermore, acceptability of arguments also touches the fundamental problem of the target audience of the argument, as different groups have different perceptions. Crosswhite et al. (2004) point out that "one of the key premises from which the study of rhetoric proceeds is that influencing real audiences is not simply a matter of presenting a set of rational, deductive arguments."

2.2 Common terminology

Let us set up a common terminology. *Claim* is "the conclusion we seek to establish by our arguments" (Freeley and Steinberg, 2008, p. 153) or "the assertion put forward publicly for general acceptance" (Toulmin et al., 1984, p. 29). *Premises* are "connected series of sentences, statements, or propositions that are intended to give reasons of some kind for the claim" (Freeley and Steinberg, 2008, p. 3).

3 Related Work

3.1 Opinion mining perspective

In existing works on argumentation mining of the Web data, the connection is often made to opinion mining (Liu, 2012). From the users' point of view, opinion mining applications reveal *what people think about something.* The key question which brings argumentation on the scene is *why do they think so?* – in other words, explaining the reasons behind opinions.

Villalba and Saint-Dizier (2012) approach aspect-based sentiment of product reviews by classifying discourse relations conveying arguments (such as justification, reformulation, illustration, and others). They build upon Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) and argue that rhetorical elements related to explanation behave as argument supports.

For modeling argumentation in social media, Schneider et al. (2012) suggest using Dung's framework (Dung, 1995) with Walton schemes (Walton et al., 2008), but do not provide evidence for such a decision. They admit that "It is far from clear how an argument [...] can be transformed into a formal argumentation scheme so that it can be reasoned in an argumentation framework" (Schneider et al., 2012, p. 22).

Schneider and Wyner (2012) focus on the product reviews domain and develops a number of argumentation schemes (inspired by (Walton et al., 2008)) based on manual inspection of their corpus. Appropriateness of such an approach remains questionable. On the one hand, Walton's argumentation schemes are claimed to be general and domain independent. On the other hand, evidence from the field shows that schemes might not be the best means for analyzing user-generated argumentation. In examining real-world political argumentation from (Walton, 2005), Walton (2012) found out that 37.1% of the arguments collected did not fit any of the fourteen schemes they chose so they created new schemes ad-hoc. Cabrio et al. (2013a) select five argumentation schemes from Walton and map these patterns to discourse relation categories in the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), but later they define two new schemes that they discovered in PDTB. These findings confirm that the schemes lack coverage for dealing with real argumentation in natural language texts.

3.2 Previous works on annotation

Table 1 summarizes the previous research on annotating argumentation. Not only it covers related work from the NLP community but also studies from general discourse analysis (Newman and Marshall, 1991; Walton, 2012) and road-maps or position papers (Schneider and Wyner, 2012; Peldszus and Stede, 2013a; Sergeant, 2013). The heterogeneity of used argumentation models and the domains under investigation demonstrates the breath of the argumentation mining field. We identified the following research gaps.

- Most studies dealing with Web data use some kind of proprietary model without relation to any argumentation theory (Bal and Saint-Dizier, 2010; Rosenthal and McKeown, 2012; Conrad et al., 2012; Schneider and Wyner, 2012; Villalba and Saint-Dizier, 2012; Florou et al., 2013; Sergeant, 2013; Wachsmuth et al., 2014b; Llewellyn et al., 2014).

- Inter-annotation agreement (IAA) that reflects reliability of the annotated data is either not reported (Feng and Hirst, 2011; Mochales and Moens, 2011; Walton, 2012; Florou et al., 2013; Villalba and Saint-Dizier, 2012), or is not based on a chance-corrected measure (Llewellyn et al., 2014).

This motivates our research into annotating Web data relying on a model based on a theoretical

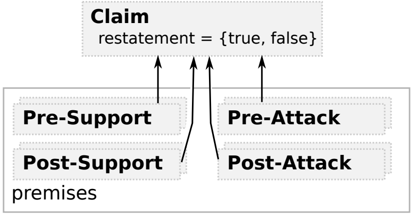


Figure 2: Claim-Premise scheme. Note that the relations (arrows) are only illustrative; they are implicitly encoded in the roles of the particular argument components.

background in argumentation and reporting IAA that would confirm suitability of the model and reliability of the annotated data.

4 Annotating argumentation in Web data

Up until now, we have used the terms argumentation and argument in their common meaning without any particular formal definition. We will now elaborate on annotation schemes and discuss their suitability and reliability for the Web data.

4.1 Annotation Schemes

Because of the lack of a single general-purpose argumentation model (cf. discussion in §1), we present here two different schemes.[2] Both are built upon foundations in argumentation theories, but they differ in their granularity, expression power, and other properties.

4.1.1 Claim-Premises scheme

The Claim-Premises scheme is widely used in previous work on argumentation mining, e.g., (Palau and Moens, 2009; Florou et al., 2013; Peldszus and Stede, 2013b). It defines an argument as consisting of a (possibly empty) set of premises and a single claim; premises either support or attack the claim (Besnard and Hunter, 2008). We adopted this general scheme for the purpose of annotating arguments in long Web documents (Kluge, 2014). According to this adopted version of the scheme, claims, restatements and premises are subsumed under the term argument component; a restatement of a claim is also considered as claim and is part of the same argument. The scheme is depicted in Figure 2.

Premises either support or attack a claim, i.e., there is a support or attack relation between each premise and a claim. The simplest way to represent the support and attack relations is to attach labels to adjacent argument components, which indicate their argumentative role. The span of argument components is left unspecified, allowing for argument components spanning a clause or one to several sentences. Using the six labels *claim, restatement, pre-claim support, post-claim support, pre-claim attack and post-claim attack*, a linear sequence of non-nested arguments can be represented.

While graph structures where nodes stand for argument components, and edges for support or attack relations are a more general way to represent arguments (equivalent to, i.e., (Dung, 1995) or (Freeman, 1991)), it is unclear which additional benefits such a more fine-grained annotation of arguments brings for the annotation of Web documents. In a pre-study performed by Kluge (2014), the possibility to annotate nested arguments turned out to be a drawback, rather than an advantage, because the inter-annotator agreement dropped considerably.

Suitability of the scheme The main advantage of the Claim-Premises scheme is its simplicity. Therefore, it is particularly suited for annotating arguments in long Web documents, such as news articles, editorials or blog posts. Kluge (2014) found that most documents of these text types consist of three major parts: an introductory part, summarizing the document content in one or two paragraphs, the main part, presenting a linear sequence of arguments, and an optional concluding part summarizing the main arguments.

The Claim-Premise scheme can be used to provide an overview of the claims and their supporting or attacking premises presented in a long Web document. From an information seeking perspective, arguments could be clustered by similar claims or similar premises, and then ranked in the context of a specific information need by a user. In a similar way, this scheme could be used for automatic summarization.

However, the Claim-Premises scheme does not allow to distinguish between different kinds of premises supporting the claim. Hence, fine-grained distinctions of premises into specific factual evidence versus any kind of common ground can not be captured.

[2] An exhaustive overview of various argumentation models, their taxonomy, and properties can be found in (Bentahar et al., 2010).

Source	**Arg. Model**	**Domain**	**Size**	**IAA**
Newman and Marshall (1991)	Toulmin	legal domain (People vs. Carney, U.S. Supreme Court)	qualitative	N/A
Bal and Saint-Dizier (2010)	proprietary	socio-political newspaper editorials	56 documents	Cohen's κ (0.80)
Feng and Hirst (2011)	Walton (top 5 schemes)	legal domain (Araucaria corpus, 61% subset annotated with Walton scheme)	$\approx$ 400 arguments	not reported claimed to be small
Georgila et al. (2011)	proprietary	general discussions (negotiations between florists)	21 dialogues	Krippendorf's α (0.37-0.56)
Mochales and Moens (2011)	Claim-Premise based on Freeman	legal domain (Aracuraria corpus, European Human Rights Council)	641 documents w/ 641 arguments (Aracuraria) 67 documents w/ 257 arguments (EHRC)	not reported
Walton (2012)	Walton (14 schemes)	political argumentation	256 arguments	not reported
Rosenthal and McKeown (2012)	opinionated claim, sentence level	blogposts, Wikipedia discussions	4000 sentences	Cohen's κ (0.50-0.57)
Conrad et al. (2012)	proprietary (spans of arguing subjectivity)	editorials and blogpost about Obama Care	84 documents	Cohen's κ (0.68) on 10 documents
Schneider and Wyner (2012)	proprietary, argumentation schemes	camera reviews	N/A (proposal/position paper)	N/A
Schneider et al. (2012)	Dung + Walton	unspecified social media	N/A (proposal/position paper)	N/A
Villalba and Saint-Dizier (2012)	proprietary, RST	hotel reviews, hi-fi products, political campaign	50 documents	not reported
Peldszus and Stede (2013a)	Freeman + RST	Potsdam Commentary Corpus	N/A (proposal/position paper)	N/A
Florou et al. (2013)	none	public policy making	69 argumentative segments / 322 non-argumentative segments	not reported
Peldszus and Stede (2013b)	based on Freeman	not reported, artificial documents created for the study	23 short documents	Fleiss' κ multiple results
Sergeant (2013)	N/A	Car Review Corpus (CRC)	N/A (proposal/position paper)	N/A
Wachsmuth et al. (2014b)	none	hotel reviews	2100 reviews	Fleiss' κ (0.67)
Llewellyn et al. (2014)	proprietary, no argumentation theory	Riot Twitter Corpus	7729 tweets	only percentage agreement reported
Stab and Gurevych (2014)	Claim-Premise based on Freeman	student essays	90 documents	Krippendorf's α_U (0.72) Krippendorf's α (0.81)

Table 1: Previous works on annotating argumentation. IAA = Inter-annotation agreement; N/A = not applicable.

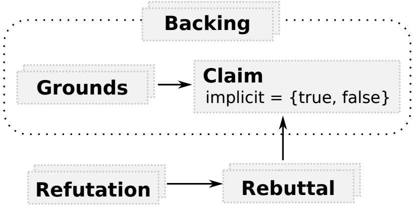


Figure 3: Extended Toulmin's scheme. Note that the relations (arrows) are only illustrative; they are implicitly encoded in the roles of the particular argument components.

4.1.2 Toulmin's scheme

The Toulmin's model (Toulmin, 1958) is a conceptual model of argumentation, in which different components play distinct roles. In the original form, it consists of six components: *claim*, *data (grounds)*, *warrant*, *backing*, *qualifier*, and *rebuttal*.

The roles of *claim* and *grounds* correspond to the definitions introduced earlier (*claim* and *premises*, respectively). The role of *warrant* is to justify a logical inference from *grounds* to *claim*. To assure the trustworthiness of the *warrant*, *backing* provides further set of information. *Qualifier* limits the degree of certainty under which the argument should be accepted and *rebuttal* presents a situation in which the *claim* might be defeated. For examples of arguments based on Toulmin's original model see, e.g., (Freeley and Steinberg, 2008, Chap. 8).

Based on our experiments during annotation pre-studies, we propose an extension of the Toulmin's model by means of (1) omitting the *qualifier* for stating modality, as people usually do not state the degree of cogency, (2) omitting the *warrant* as reasoning for justifying the move from grounds to claims is not usually explained, (3) extending the role of *backing* so it provides additional set of information to back-up the argument as a whole but is not directly bound to the *claim* as the *grounds* are, and (4) adding *refutation* which attacks the *rebuttal* (attacking the attack). The scheme is depicted in Figure 3.

Suitability of the scheme As pointed out by Bentahar et al. (2010), many argumentation systems make no distinction between their premises, despite the fact that in arguments expressed in natural language we can typically observe premises playing different roles. Toulmins' scheme allows such a distinction using the set of different components (roles). “By identifying these roles, we can present the arguments in a more readily understandable fashion, and also identify the various ways in which the argument may be accepted or attacked” (Bentahar et al., 2010, p. 216).

Toulmin's model, as a general framework for modeling static monological argumentation (Bentahar et al., 2010), has been used in works on annotating argumentative discourse (Newman and Marshall, 1991; Chambliss, 1995; Simosi, 2003; Weinberger and Fischer, 2006). However, its complexity and the fact that the description of the components is informal and sometimes ambiguous, poses challenges for an application of the model on real-world data, especially user-generated discourse on the Web. Moreover, some of the components are usually left implicit in argumentation, such as the warrant or even the claim (Newman and Marshall, 1991).

5 Preliminary results of annotation studies

In order to examine the proposed approaches, we conducted two extensive independent annotation studies. The central controversial topics were related to education. One distinguishing feature of educational topics is their breadth, as they attract researchers, practitioners, parents, or policymakers. Since the detailed studies are being published elsewhere, we summarize only the main results and outcomes in this paper.

In the first study, we used the Claim-Premises scheme for annotating a dataset of web documents consisting of 80 documents from six current topics related to the German educational system (e.g., mainstreaming, staying down at school), which is described in (Kluge, 2014). The dataset contains (newspaper) articles, blog posts, and interviews. It was created by Vovk (2013) who manually selected documents obtained from a focused crawler and the top 100 search engine hits (per topic).

In the second study, the annotation was split into two stages. In the first stage, we annotated 990 English comments to articles and forums posts with their argumentativeness (persuasiveness). The source sites were identified using a standard search engine and the content was extracted manually; we chose the documents randomly without any pre-filtering. In the second stage, we applied the extended Toulmin's scheme on 294 argumentative English comments to articles and forums posts and 57 English newspaper editorials and blog posts. The topics cover, e.g., mainstreaming,[3] single-sex schools, or home-schooling, among others.

Measuring inter-annotator agreement For any real large-scale annotation attempt, measuring inter-annotator agreement (IAA) is crucial in order to estimate the reliability of annotations and the feasibility of the task itself. Both annotation approaches share one common sub-task: labeling spans of tokens with their corresponding argumentation concept, the boundaries of the spans are not known beforehand. Therefore, the most appropriate measure here is the unitized Krippendorf's α_U as the annotators identify and label the units in the same text (Krippendorff, 2013). Other measures, such as Cohen's κ or Fleiss' π, expect the units (boundaries of the argument component) to be known beforehand, which is not the case here.

5.1 Outcomes of annotating with Claim-Premises scheme

During an annotation study of 6 weeks, three annotators (one inexperienced annotator and two experts) annotated 80 documents belonging to six topics. On average, each annotator needed 23 hours to annotate the 3863 sentences. The annotators marked 5126 argument components (53% premises, 47% claims) and 2349 arguments, which is 2.2 argument components per argument. On average, 74% of the tokens in the dataset are covered by an argument component indicates that the documents are in fact highly argumentative. An average claim spans 1.1 sentences, whereas an average premise spans 2.2 sentences.

While the IAA scores appeared to be non-substantial, ranging from α_U=34.6 (distinguishing all 6 annotation classes and non-argumentative) to α_U=42.4 (distinguishing between premises, claims and non-argumentative), they are in line with previous results: Peldszus and Stede (2013b) report α_U=42.5 for their sentence-level annotation study.

By analysing typical patterns of argument components used in arguments, Kluge (2014) found that almost three quarters of arguments (72.4%) consist of one claim and one premise. In 59.5% of these arguments, the support follows the claim, whereas only in 11.6% of the arguments, the support precedes the claim. The corresponding patterns consisting of attack and claim are significantly less frequent: only 3.4% of the arguments consist of a claim and an attack.

Argument Component	**Comments, Forums**	**Blogs**	**Articles**
Claim	0.57	0.17	0.23
Grounds	0.64	0.32	0.11
Backing	0.41	-0.16	0.28
Rebuttal	0.33	-0.02	0.00
Refutation	0.06	0.35	0.00

Table 2: IAA scores (Krippendorf's α_U) from annotations using the Toulmin's scheme.

Annotated examples can be found in §A.1.

5.2 Outcomes of annotating with Toulmin's scheme

In the first stage, three independent annotators labeled 524 out of 990 documents as argumentative/persuasive on the given topic. Total size of this dataset was 130,085 tokens (mean 131, std. dev. 139) and 6,371 sentences (mean 6.44, std. dev. 6.53). Agreement on the first sub-set of this dataset of 300 documents was 0.51 (Fleiss' π, three annotators per document), the second sub-set (690 documents) was then annotated by two annotators with agreement 0.59 (Cohen's κ). This stage took in total about 17 hours per annotator.

In the second phase that took about 33 hours per annotator, a collection of comments and forum posts (294 documents) was randomly chosen from the previously labeled argumentative documents from the previous stage together with 49 blog posts and 8 newspaper articles. The total size of this dataset was 345 documents, containing 87,286 tokens (mean 253.00, std. dev. 262.90) and 3,996 sentences (mean 11.58, std. dev. 11.72). Three independent annotators annotated the whole dataset in multiple phases. After each phase, they discussed discrepancies, resolved issues and updated the annotation guidelines. The inter-annotator agreement was measured on the last phase containing 93 comments and forum posts, 8 blogs, and 6 articles. During the annotations, 2 articles and 4 forum posts/comments were also discarded as non-argumentative.

Agreement (Krippendorf's α_U) varies significantly given different argumentation components

[3] Discussion about benefits or disadvantages of including children with special needs into regular classes.

and registers, as shown in Table 2. Given these results, we formulate the following conclusions.

This scheme seems to fit well short documents (forum posts and comments) as they tend to bring up one central *claim* with a support (*grounds*). Its suitability for longer documents (blogposts and editorials) is doubtful. We examined the annotation errors and found that in well-structured documents, the annotators were able to identify the concepts reliably. However, if the discussion of the controversy is complex (many sub-aspects are discussed) or follows a dialogical manner, application the Toulmin's scheme is all but straightforward.

Furthermore, the distinction between *grounds* and *backing* also allows to capture different kinds of evidence. Authors purposely use *grounds* to explicitly support their *claim*, while *backing* mostly serves as an additional information (i.e., author's personal experience, referring to studies, etc.) and the argument can be still acceptable without it. However, boundaries between these two components are still fuzzy and caused many disagreements.

We show few annotation examples (as agreed by all annotators after the study) in §A.2.

6 Observations

In this section, we would like to summarize some important findings from our annotation studies.

6.1 Data heterogeneity

Variety or registers There exist many on-line registers that carry argumentation to topics under investigation, such as newspaper reports (i.e., events), editorials (opinions), interviews (single party, multiple parties), blogposts,[4] comments to articles and blogs (threaded allowing explicit discussion, linear with implicit discussion by quoting and referencing), discussion forums, Twitter, etc.

Short versus long documents Different document lengths affect the style of argumentation. Short documents (i.e., Tweets in the extreme case) have to focus on the core of the argument. By contrast, long documents, such as blog posts or editorials, may elaborate various aspects of the topic and usually employ many literary devices, such as narratives, quotations from sources, or direct and indirect speech.

Well structured newspaper articles versus poorly structured user-generated content Producing a well-understandable argument is actually a human skill that can be acquired by learning; many textbooks are available on that topic, e.g., (Sinnott-Armstrong and Fogelin, 2009; Weston, 2008; Schiappa and Nordin, 2013). Thus, it is very likely that, for example, trained journalists in editorials and lay people in social media will produce very different argumentation, in terms of structure, language, etc.

6.2 Properties of argumentation in user-generated discourse

Non-argumentative texts Distinguishing argumentative from non-argumentative discourse is a necessary step that has to be undertaken before annotating argument components. While in newspaper editorials some parts (such as paragraphs) may be ignored during argument annotation (Kluge, 2014), in comments and forum posts we had to perform an additional step to filter documents that do not convey any argumentation or persuasion (cf. §5.2 or Example 4 in §A.2).

Implicit argumentation components in Toulmin's model As already reported by Newman and Marshall (1991), some argument components are not explicitly expressed. This is mostly the case of *warrant* in the original Toulmin's model; we also discarded this component from our extension. However, even the *claim* is often not stated explicitly, as seen in example 3 (§A.2). The claim reflects the author's stance and can be understood (inferred) by readers, but is left implicit.

Other rhetorical dimensions of argument All the models for argumentation discussed so far focus solely on the *logos* part of the argument. However, rhetorical power of argumentation also involves other dimensions, namely *pathos*, *ethos*, and *kairos* (Aristotle and Kennedy (translator), 1991; Schiappa and Nordin, 2013). These have never been tackled in computational approaches to modeling argumentation. Furthermore, figurative langauge, fallacies, or narratives (see example 3 in §A.2) are prevalent in argumentation on the Web.

[4] In contrast to traditional publisher, bloggers do not have to comply with strict guidelines or the use of formal language (Santos et al., 2012).

6.3 Recommendations

Based on the experience from the annotation studies, we would like to conclude with the following recommendations: (1) selection of argumentation model should be based on the data at hand and the desired application; our experiments show that Toulmin's model is more expressive than the Claim-Premise model but is not suitable for long documents, (2) annotating argumentation is time-demanding and error-prone endeavor; annotators thus have to be provided with detailed and elaborated annotation guidelines and be extensively trained (our experiments with crowdsourcing were not successful).

7 Follow-up use cases

Understanding argumentation in user-generated content can foster future research in many areas. Here we present two concrete applications.

7.1 Understanding argumentative discourse in education

Computer-supported argumentation has been a very active research field, as shown by Scheuer et al. (2010) in their recent survey of various models and argumentation formalisms from the educational perspective. Many studies on computer-supported collaboration and argumentation (Noroozi et al., 2013; Weinberger and Fischer, 2006; Stegmann et al., 2007) can directly benefit from NLP techniques for automatic argument detection, classification, and summarization. Instead of relying on scripts (Dillenbourg and Hong, 2008; Scheuer et al., 2010; Fischer et al., 2013) or explicit argument diagramming (Scheuer et al., 2014), collaborative platforms can further provide scholars with a summary of the whole argumentation to the topic, reveal the main argumentative patterns, provide the weaknesses of other's arguments, as well as identify shortcomings that need to be improved in the argumentative knowledge construction. Automatic analysis of micro-arguments can also help to overcome the existing trade-off between freedom (free-text option) and guidance (scripts) (Dillenbourg and Hong, 2008).

7.2 Automatic summarization of argumentative discourse

When summarizing argumentative discourse, knowledge of the underlying structure of the argument is a valuable source. Previous work in this area includes, e.g., opinion-based summarization of blogposts (a pilot task in TAC 2008[5]). Carenini and Cheung (2008) compared extractive and abstractive summaries in controversial documents and found out that a high degree of controversiality improved performance of their system. Similarly, presenting argumentation in a condensed form (the large concepts of the argument are compressed or summarized) may improve argument comprehension. This approach would mainly utilize tools for document compression (Qian and Liu, 2013).

8 Conclusions

In this article, we formulated our view on argumentation mining on the Web and identified various use-case scenarios and expected outcomes. We thoroughly reviewed related work with focus on Web data and annotation approaches. We proposed two different annotation schemes based on their theoretical counterparts in argumentation research and evaluated their suitability and reliability for Web data in two extensive independent annotation studies. Finally, we outlined challenges and gaps in current argumentation mining on the Web.

Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806, and by the German Institute for Educational Research (DIPF).

A Annotated examples

A.1 News articles using Claim-Premises scheme

Example 1

[*claim:* „Die Umstellung zu G8 war schwierig“, sagt Diana.] [*support:* In den Sommerferien nach dem Sitzenbleiben holte sie das nach, was ihr die G8er voraus hatten: Lateinvokabeln, Stochastik, Grammatik. „Den Vorteil, durch das Wiederholen den Stoff noch mal zu machen, hatte ich nicht.“]

[*claim:* "The change [to G8] was difficult," says Diana.] [*support:* (Since) After staying down, she had to catch up with the G8 students during her summer holiday, studying Latin vocabulary,

[5] http://www.nist.gov/tac/publications/2008/papers.html

stochastics, and grammar. "I did not have the advantage of reviewing previous material."]

Example 2

[*claim:* Lehrer wird man, weil das ein sicherer Beruf ist.] [*support:* So denken noch immer viele junge Leute, die sich für eine Pädagogenlaufbahn entscheiden. Gut acht von zehn Erstsemestern, die 2009 mit einem Lehramtsstudium anfingen, war dieser Aspekt ihres künftigen Berufs wichtig oder sogar sehr wichtig. Keine andere Studentengruppe, die die Hochschul-Informations-System GmbH HIS befragte, legt so viel Wert auf Sicherheit.]

[*claim:* People become teachers because it is a safe job.] [*support:* This is what more and more young people who decide to become a teacher think. Well over eight of 10 freshman students who started to study to become teachers in 2009 considered this an important or very important aspect. No other group of students interviewed by the HIS set that much value on safeness.]

Example 3

[*claim:* Für die Unis sind Doktoranden günstige Arbeitskräfte.] [*support:* Eine Bekannte hatte mit ihrem Doktorvater zu kämpfen, der versuchte, sie noch am Institut zu halten, als ihre Arbeit längst fertig war. Er hatte immer neue Ausreden, weshalb er noch keine Note geben konnte. Als sie dann auch ohne Note einen guten Job bekam, auerhalb der Uni, spielte sich eine Art Rosenkrieg zwischen den beiden ab. Bis heute verlangt er von ihr noch Nacharbeiten an der Dissertation. Sie schuftet jetzt spätabends und am Wochenende für ihren Ex-Prof, der natürlich immer nur an ihrem Fortkommen interessiert war.]

[*claim:* At university, graduate students are cheap employees.] [*support:* An acquaintance struggled with her Ph.D. supervisor, who tried to keep her in his group at any rate, even though she had already completed her thesis. He pled more and more excuses for not yet grading her work. When she finally found a good job outside university even without a final grade a martial strife arose. Still today, he asks her to rework her dissertation. Now, she is drudging for her ex-supervisor, who always only wanted the best for her, late in the evening or on the weekend.]

A.2 Forum posts using extended Toulmin's scheme

Example 1

[*backing:* I'm a regular education teacher. I have students mainstreamed into my class every year.] [*grounds:* My opinion is that it needs to be done far more judiciously than it is done now- if six exceptional children are put in my class, that is the equivalent of putting an entire special ed classroom into my regular class.] [*grounds:* I personally feel like these kids are shortchanged- some of them are good kids who need an adult close by and able to give more focused attention. In a class of 30+, this isn't going to happen consistently.] [*grounds:* And some of the ones who come to me have legally imposed modifications, some of which have little or no bearing on what I teach, so I am not allowed to handle my class in a way I think it should be done. That impairs my efficiency as an educator.] [*grounds:* Also, some have so many modifications that for all intents and purposes they are merely taking a special ed class whose physical location just happens to be in a regular classroom.] [*claim:* From my point of view, mainstreaming is not a terrible idea, but it is lamentable in its execution, and because of that, damaging in its results.]

Comments Quite a good argument with an explicit claim, few grounds and some backing.

Example 2

tara_mommy:
I agree with you too, which is why I said:
[*rebuttal:* There are obviously cases where this isn't going to work. Extreme behavioral trouble, kids that just aren't able to keep up with what they're learning in average classes, etc.] [*claim:* But on the whole, I like mainstreaming.]

Comments Only claim and rebuttal; no supporting grounds.

Example 3

l think as parents of the child you have to be certain and confident that your child is ready to mainstream. lf not, it can backfire on the child. [*backing:* My child was in "preschool handicapped" from age 2-5. We tried to mainstream him in kindergarten, but he had a hard time adjusting. So the school got him a one on one para and it helped a bit. 2 grades later, he still has

a one on one aide but doing EXCELLENT.]
Our goal is for him to not have a one on one by middle school. We took him off meds and we have a strong behavior plan, he sees therapists, and it is hourly teaching and redirecting with him. Truth be told College may not be in his future, but we will do everything in our power to try to get him there.

Comments The claim is implicit, the author is slightly against mainstreaming. Mainly storytelling, which is not considered as grounds but as backing. The typos (using ‘l’ instead of ‘I’) are kept uncorrected.

Example 4

My lo has mild autism, he has only just been diagnosed, he is delayed in some areas (but not others), he goes to ms school, and has some one to one (this should increase now, I hope). There is one TA and a full time TA who supports another child with autism. It’s a smallish school.
He isn’t disruptive (well he sometimes doesn’t do as asked and can be a little akward), he has never been aggressive in anyway, he is very happy.
I am worried about his future (high school)after reading this.
Sarah x

Comments Not an argumentative/persuasive text.

References

Corina Andone. 2012. Bermejo-Luque, Lilian. Giving Reasons. A Linguistic-Pragmatic Approach to Argumentation Theory. *Argumentation*, 26(2):291–296.

Aristotle and George Kennedy (translator). 1991. *On Rhetoric: A Theory of Civil Discourse*. Oxford University Press.

Bal Krishna Bal and Patrick Saint-Dizier. 2010. Towards Building Annotated Resources for Analyzing Opinions and Argumentation in News Editorials. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, pages 1152–1158. European Language Resources Association (ELRA).

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33:211–259.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge Textbooks in Linguistics. Cambridge University Press.

J. Anthony Blair. 2004. Argument and its uses. *Informal Logic*, 24:137151.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212, Jeju Island, Korea, July. Association for Computational Linguistics.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013a. From Discourse Analysis to Argumentation Schemes and Back: Relations and Differences. In João Leite, Tran Cao Son, Paolo Torroni, Leon Torre, and Stefan Woltran, editors, *Proceedings of 14th International Workshop on Computational Logic in Multi-Agent Systems*, volume 8143 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg.

Elena Cabrio, Serena Villata, and Fabien Gandon. 2013b. A support framework for argumentative discussions management in the web. In Philipp Cimiano, Oscar Corcho, Valentina Presutti, Laura Hollink, and Sebastian Rudolph, editors, *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 412–426. Springer Berlin Heidelberg.

Giuseppe Carenini and Jackie Chi Kit Cheung. 2008. Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG ’08, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marilyn J. Chambliss. 1995. Text cues and strategies successful readers use to construct the gist of lengthy written arguments. *Reading Research Quarterly*, 30(4):778–807.

Alexander Conrad, Janyce Wiebe, and Rebecca Hwa. 2012. Recognizing arguing subjectivity and argument tags. In Roser Morante and Caroline Sporleder, editors, *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88, Jeju Island, Korea. Association for Computational Linguistics.

Jim Crosswhite, John Fox, Chris Reed, Theodore Scaltsas, and Simone Stumpf. 2004. Computational models of rhetorical argument. In Chris Reed and Timothy J. Norman, editors, *Argumentation Machines*, volume 9 of *Argumentation Library*, pages 175–209. Springer Netherlands.

Aldo de Moor, Lilia Efimova, and Aldo De Moor. 2004. An Argumentation Analysis of Weblog Conversations. In *Proceedings of the 9th International Working Conference on the Language-Action Perspective on Communication Modelling (LAP 2004)*, volume 197, pages 1–16.

Pierre Dillenbourg and Fabrice Hong. 2008. The mechanics of CSCL macro scripts. *International Journal of Computer-Supported Collaborative Learning*, 3(1):5–23.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321 – 357.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 987–996, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frank Fischer, Ingo Kollar, Karsten Stegmann, and Christof Wecker. 2013. Toward a script theory of guidance in computer-supported collaborative learning. *Educational Psychologist*, 48(1):56–66.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria. ACL.

Austin J. Freeley and David L. Steinberg. 2008. *Argumentation and Debate*. Cengage Learning, Stamford, CT, USA, 12th edition.

James B Freeman. 1991. *Dialectics and the macrostructure of arguments: A theory of argument structure*, volume 10 of *Trends in Linguistics*. De Gruyter.

Kallirroi Georgila, Ron Artstein, Angela Nazarian, Michael Rushforth, David Traum, and Katia Sycara. 2011. An annotation scheme for cross-cultural argumentation and persuasion dialogues. In *Proceedings of the SIGDIAL 2011 Conference: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 272–278, Portland, Oregon. Association for Computational Linguistics.

A. Francisca Snoeck Henkemans. 2000. State-of-the-art: The structure of argumentation. *Argumentation*, 14(4):447–473.

Anthony Hunter. 2013. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81, January.

Anthony Hunter. 2014. Probabilistic qualification of attack in abstract argumentation. *International Journal of Approximate Reasoning*, 55(2):607–638, January.

Ralph H Johnson. 2000. *Manifest rationality: A pragmatic theory of argument*. Routledge.

Roland Kluge. 2014. Automatic Analysis of Arguments about Controversial Educational Topics in Web Documents, Master Thesis, Ubiquitious Knowledge Processing Lab, TU Darmstadt.

Klaus Krippendorff. 2013. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications, 3rd edition.

Meredith Larson, M. Annae Britt, and Aaron Larson. 2004. Disfluencies in comprehending argumentative texts. *Reading Psychology*, 25:205–224.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, August. To appear.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Clare Llewellyn, Claire Grover, Jon Oberlander, and Ewan Klein. 2014. Re-using an Argument Corpus to Aid in the Curation of Social Media Collections. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 462–468.

Lilian Bermejo Luque. 2011. *Giving Reasons: A Linguistic-Pragmatic Approach to Argumentation Theory*, volume 20 of *Argumentation Library*. Springer Netherlands.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 20–28, Stroudsburg, PA, USA. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical report, Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA.

Raphaël Micheli. 2011. Arguing Without Trying to Persuade? Elements for a Non-Persuasive Definition of Argumentation. *Argumentation*, 26(1):115–126, September.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, April.

S. Newman and C. Marshall. 1991. Pushing Toulmin Too Far: Learning From an Argument Representation Scheme. Technical report, Xerox Palo Alto Research Center 3333 Coyote Hill Road, Palo Alto, CA 94034.

Omid Noroozi, Armin Weinberger, Harm J.a. Biemans, Martin Mulder, and Mohammad Chizari. 2013. Facilitating argumentative knowledge construction through a transactive discussion script in CSCL. *Computers & Education*, 61:59–76, February.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107, New York, NY, USA. ACM.

Andreas Peldszus and Manfred Stede. 2013a. From Argument Diagrams to Argumentation Mining in Texts:. *International Journal of Cognitive Informatics and Natural Intelligence*, 7(1):1–31, January.

Andreas Peldszus and Manfred Stede. 2013b. Ranking the annotators : An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperabilty with Discourse*, pages 196–204. Association for Computational Linguistics.

Chaim Perelman and Lucie Olbrechts-Tyteca. 1991. *The New Rhetoric*. University of Notre Dame Press.

Henry Prakken. 2010. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, June.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1–4. European Language Resources Association (ELRA).

Xian Qian and Yang Liu. 2013. Fast joint compression and summarization via graph cuts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1502, Seattle, Washington, USA, October. Association for Computational Linguistics.

Chris Reed and Douglas Walton. 2003. Argumentation schemes in argument-as-process and argument-as-product. In *Proceedings of the conference celebrating informal Logic*, volume 25.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting Opinionated Claims in Online Discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37. IEEE, September.

Rodrygo LT Santos, Craig Macdonald, Richard MC McCreadie, Iadh Ounis, Ian Soboroff, et al. 2012. Information retrieval on the blogosphere. *Foundations and Trends in Information Retrieval*, 6(1):1–125.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M. McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-Supported Collaborative Learning*, 5(1):43–102.

Oliver Scheuer, BruceM. McLaren, Armin Weinberger, and Sabine Niebuhr. 2014. Promoting critical, elaborative discussions through a collaboration script and argument diagrams. *Instructional Science*, 42(2):127–157.

Edward Schiappa and John P. Nordin. 2013. *Argumentation: Keeping Faith with Reason*. Pearson UK, 1st edition.

Jodi Schneider and Adam Wyner. 2012. Identifying Consumers' Arguments in Text. In Diana Maynard, Marieke van Erp, and Brian Davis, editors, *Semantic Web and Information Extraction SWAIE 2012*, pages 31–42.

Jodi Schneider, B Davis, and Adam Wyner. 2012. Dimensions of argumentation in social media. In *Lecture Notes in Computer Science*, volume 7603, pages 21–25. Springer Berlin Heidelberg.

Jodi Schneider, Tudor Groza, and Alexandre Passant. 2013. A review of argumentation for the social semantic web. *Semantic Web*, 4(2):159–218.

Alan Sergeant. 2013. Automatic Argumentation Extraction. In *ESWC 2013*, pages 656–660. Springer-Verlag Berlin Heidelberg.

Maria Simosi. 2003. Using Toulmin's framework for the analysis of everyday argumentation: Some methodological considerations. *Argumentation*, 17:185–202.

Walter Sinnott-Armstrong and Robert J. Fogelin. 2009. *Understanding Arguments: An Introduction to Informal Logic*. Cengage Learning, 8 edition.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, August. To appear.

Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2007. Facilitating argumentative knowledge construction with computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, 2(4):421–447.

Dragan Stoianovici. 2009. Formal Logic vs. Philosophical Argument. *Argumentation*, 24(1):125–133, January.

Stephen Toulmin, Richard Rieke, and Allan Janik. 1984. *An Introduction to Reasoning*. Macmillan, 2nd edition.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some Facets of Argument Mining for Opinion Analysis. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *Proceedings of Fourth International Conference on Computational Models of Argument, COMMA 2012*.

Artem Vovk. 2013. Discovery and Analysis of Public Opinions on Controversial Topics in the Educational Domain, Master Thesis, Ubiquitious Knowledge Processing Lab, TU Darmstadt.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014a. Modeling Review Argumentation for Robust Sentiment Analysis. In *Proceedings of the 25th International Conference on Computational Linguistics COLING 2014*, page To appear, Dublin, Ireland.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014b. A Review Corpus for Argumentation Analysis. In Alexander Gelbukh, editor, *15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 14)*, pages 115–127. Springer.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Douglas Walton. 2005. *Fundamentals of Critical Argumentation*. Critical Reasoning and Argumentation. Cambridge University Press, 1 edition.

Douglas Walton. 2012. Using Argumentation Schemes for Argument Extraction: A Bottom-Up Method. *International Journal of Cognitive Informatics and Natural Intelligence*, 6(3):33–61.

Armin Weinberger and Frank Fischer. 2006. A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46(1):71–95, January.

Anthony Weston. 2008. *A Rulebook for Arguments*. Hackett Pub Co., 4 edition.

Yun Xie. 2012. Review of Giving Reasons. *Informal Logic*, 32(4).

Cihua Xu and Yicheng Wu. 2014. Metaphors in the perspective of argumentation. *Journal of Pragmatics*, 62:68–76, February.

Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective

Christian Stab[†], **Christian Kirschner**†‡, **Judith Eckle-Kohler**†‡ **and Iryna Gurevych**[†‡]

[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
`www.ukp.tu-darmstadt.de`

Abstract

In this paper, we analyze and discuss approaches to argumentation mining from the discourse structure perspective. We chose persuasive essays and scientific articles as our example domains. By analyzing several example arguments and providing an overview of previous work on argumentation mining, we derive important tasks that are currently not addressed by existing argumentation mining systems, most importantly, the identification of argumentation structures. We discuss the relation of this task to automated discourse analysis and describe preliminary results of two annotation studies focusing on the annotation of argumentation structure. Based on our findings, we derive three challenges for encouraging future research on argumentation mining.

1 Introduction

Argumentation mining is a recent research area which promises novel opportunities not only for information retrieval, educational applications or automated assessment tools but also aims at improving current legal information systems or policy modeling platforms. It focuses on automatically identifying and evaluating arguments in text documents and includes a variety of subtasks like identifying argument components, finding accepted arguments and discovering argumentation structures. Researchers have already investigated argumentation mining in several domains. For instance, Teufel (1999) aims at identifying rhetorical roles of sentences in scientific articles and Mochales-Palau and Moens (2011) identify arguments in legal documents. Also, Feng and Hirst (2011) investigated argumentation schemes in newspapers and court cases and Florou et al. (2013) applied argumentation mining in policy modeling.

However, current approaches mainly focus on the identification of arguments and their components and largely neglect the identification of argumentation structures although an argument consists not only of a set of propositions but also exhibits a certain structure constituted by argumentative relations (Peldszus and Stede, 2013; Sergeant, 2013). We argue in this paper that identifying argumentative relations and the argumentation structure respectively is an important task for argumentation mining. First, identifying argumentative relations between argument components enables the identification of additional reasons for a given claim and thus allows the creation of valuable knowledge bases e.g. for establishing new information retrieval platforms. Second, it is important to recognize which premises belong to a claim, since it is not possible to evaluate arguments without knowing which premises belong to it. Third, automatically identifying the structure of arguments enables novel features of applications, such as providing feedback in *computer-assisted writing* (e.g., recommending reasonable usage of discourse markers, suggesting rearrangements of argument components) or extracting argumentation structures from scientific publications for *automated summarization* systems.

In this paper, we analyze several examples of argumentative discourse from the discourse structure perspective.[1] We outline existing approaches on argumentation mining and discourse analysis and provide an overview of our current work on argumentation structure annotation in scientific articles and persuasive essays. We conclude this paper with a list of challenges for encouraging future

[1] The examples are taken from persuasive essays which are either collected from the writing feedback section of `http://www.essayforum.com` or from the corpus compiled by Stab and Gurevych (2014)

research on argumentation mining.

2 Background

Philosophy and Logic proposed a vast amount of argumentation theories (e.g. Toulmin (1958), Walton et al. (2008), Freeman (2011)).[2] The majority of these theories generally agree that an *argument* consists of several *argument components* which can either be a premise or a claim. The simplest form of an argument includes one claim that is supported by at least one premise (figure 1).

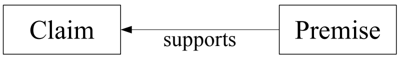


Figure 1: Illustration of a simple argument

The *claim*[3] is the central component of an argument that can either be true or false. Thus, the claim is a statement that should not be accepted by the reader without additional reasons. The second component of an argument, the *premise*[4], underpins the plausibility of the claim. It is usually provided by the proponent (writer) for convincing the reader of the claim. Examples (1) and (2) illustrate two simple arguments, each containing a claim (in bold face) and a single premise (underlined):

> (1) *“**It is more convenient to learn about historical or art items online.** With Internet, people do not need to travel long distances to have a real look at a painting or a sculpture, which probably takes a lot of time and travel fees.”*
>
> (2) *“**Locker checks should be made mandatory and done frequently** because they assure security in schools, make students healthy, and will make students obey school policies.”*

These examples illustrate that there exist argument components both on the sentence level and on the clause level.

Argumentative relations are usually directed relations between two argument components and represent the *argumentation structure*. There exist different types like *support* or *attack* (Peldszus and Stede, 2013) which indicate that the source argument component is a reason or a refutation for the target component. For instance, in both of the examples above, an argumentative support relation holds from the premise to the claim. The following example illustrates a more complex argument including one claim and three premises:

> (3) *“**Everybody should study abroad**$_a$. It’s an irreplaceable experience if you learn standing on your own feet$_b$ since you learn living without depending on anyone else$_c$. But one who is living overseas will of course struggle with loneliness, living away from family and friends$_d$.”*

Figure 2 shows the structure of the argument in (3). In this example, premise$_b$ supports the claim$_a$ whereas premise$_d$ attacks the claim$_a$.

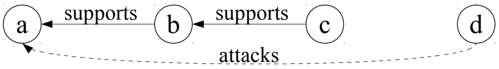


Figure 2: Argumentation structure of example (3).

This example illustrates three important properties of argumentation structures:

1. Argumentative relations can hold between non-adjacent sentence/clauses, e.g. the argumentative attack relation from premise$_d$ to the claim$_a$.

2. Some argumentative relations are signaled by indicators, whereas others are not. For instance, the argumentative attack relation from premise$_d$ to the claim$_a$ is indicated by the discourse marker *‘but’*, whereas the argumentative support relation from premise$_b$ to claim$_a$ is not indicated by a discourse marker.

3. Argumentative discourse might exhibit reasoning chains, e.g. the chain constituted between argument components a, b, and c.

3 Argumentation Mining

Previous approaches on argumentation mining cover several subtasks including the separation of argumentative from non-argumentative text units (Moens et al., 2007; Florou et al., 2013), the classification of argument components (with different component classes) (Rooney et al., 2012;

[2] A review of argumentation theory is beyond the scope of this paper. A survey can be found in Bentahar et al. (2010)

[3] also called conclusion (Mochales-Palau and Moens, 2009)

[4] sometimes called support (Besnard and Hunter, 2008) or reason (Anne Britt and Larson, 2003)

Mochales-Palau and Moens, 2009; Teufel, 1999; Feng and Hirst, 2011), and the identification of argumentation structures (Mochales-Palau and Moens, 2009; Wyner et al., 2010).

3.1 Separation of Argumentative from Non-argumentative Text Units

The first step of an argumentation mining pipeline typically focuses on the identification of argumentative text units before analyzing the components or the structure of arguments. This task is usually considered as a binary classification task that labels a given text unit as argumentative or non-argumentative. One of the first approaches was proposed by (Moens et al., 2007). They focus on the identification of argumentative text units in newspaper editorials and legal documents included in the Araucaria corpus (Reed et al., 2008). The annotation scheme utilized in Araucaria is based on a domain-independent argumentation theory proposed by Walton (1996). A similar approach is reported by Florou et al. (2013). In their experiments, they classify text segments crawled with a focused crawler as either containing an argument or not. They focus on the identification of arguments in the policy modeling domain for facilitating decision making. For that purpose, they utilize several discourse markers and features extracted from the tense and mood of verbs.

Although the separation of argumentative from non-argumentative text units is an important step in argumentation mining, it merely enables the detection of text units relevant for argumentation and does not reveal the argumentative role of argument components.

3.2 Classification of Argument Components

The classification of argument components aims at identifying the *argumentative role* (e.g. claims and premises) of argument components.

One of the first approaches to identify argument components is *Argumentative Zoning* proposed by (Teufel, 1999). Each sentence is classified as one of seven rhetorical roles including e.g. claim, result or purpose using structural, lexical and syntactic features. The underlying assumption of this work is that argument components extracted from a scientific article provide a good summary of its content. Rooney et al. (2012) also focus on the identification of argument components but in contrast to the work of Teufel (1999) their scheme is not tailored to a particular genre. In their experiments, they identify claims, premises and non-argumentative text units in the Araucaria corpus. Feng and Hirst (2011) also use the Araucaria corpus for their experiments, but focus on the identification of *argumentation schemes* (Walton, 1996) which are templates for arguments (e.g. argument from example or argument from position to know). Since their approach is based on features extracted from mutual information of claims and premises, it requires that the argument components are reliably identified in advance. Mochales-Palau and Moens (2009) report several experiments for classifying argument components. They solely focus on the legal domain and in particular on legal court cases from the European Court of Human Rights (ECHR). They consider the classification of argument components as two consecutive steps. They utilize a maximum entropy model for identifying argumentative text units before identifying the argumentative role (claim and premise) of the identified components using a Support Vector Machine.

3.3 Identification of Argumentation Structures

Currently, there are only few approaches aiming at the identification of argumentation structures. For instance, the approach proposed by Mochales-Palau and Moens (2011) relies on a manually created context-free grammar (CFG) and on the presence of discourse markers for identifying a tree-like structure between argument components. However, the approach relies on the presence of discourse markers and exploits manually created rules. Therefore, it does not accommodate ill-formatted arguments (Wyner et al., 2010) and is not capable of identifying implicit argumentation structures which are common in argumentative discourse. Indeed, Marcu and Echihabi (2002) found that only 26% of the evidence relations in the RST Discourse Treebank (Carlson et al., 2001) include discourse markers.

Another approach was presented by Cabrio and Villata (2012). They identify relations between arguments of an online debate platform for identifying accepted arguments and to support the interactions in online debates. In contrast to the work of Mochales-Palau and Moens (2011), this approach aims at identifying relations between arguments (macro-level) and not between argument components (micro-level).

4 Argumentation and Discourse Analysis

Discourse analysis aims at identifying discourse relations that hold between adjacent text units with text units being sentences, clauses or nominalizations (Webber et al., 2012). Since text units might be argument components and discourse relations are often closely related to argumentative relations, previous work in automated discourse analysis is highly relevant for argumentation mining.

4.1 Discourse Relations and Argumentative Relations

Most previous work in automated discourse analysis is based on corpora annotated with general discourse relations, most notably the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and the Rhetorical Structure Theory (RST) Discourse Treebank (Carlson et al., 2003). Whereas RST represents the discourse structure as a tree, the PDTB allows more general graph structure. For the annotation of discourse relations in the PDTB, two different types of discourse relations were distinguished: implicit and explicit relations. Whereas *explicit discourse relations* are indicated by discourse markers, *implicit discourse relations* are not indicated by discourse markers and the identification of those relations requires more sophisticated methods.

Take as an example the argumentation structure discussed in section 2.

> *"Everybody should study abroad*$_a$*. It's an irreplaceable experience if you learn standing on your own feet*$_b$ ***since*** *you learn living without depending on anyone else*$_c$*.* ***But*** *one who is living overseas will of course struggle with loneliness, living away from family and friends*$_d$*."*

Whereas the argument components b and c, as well as c and d are related through the discourse marker *'since'* (signalling an explicit CAUSE relation) and *'but'* (signalling an explicit CONTRAST relation), the discourse relation JUSTIFY between a and b is an implicit relation.

Existing approaches of discourse analysis proposed different sets of discourse relations, and there is currently no consensus in the literature about the 'right' set of discourse relations. For instance, the RST (Mann and Thompson, 1988) uses a different set of discourse relations than the PDTB (Prasad et al., 2008).

It is still an open question how the proposed discourse relations relate to argumentative relations. Although, there are preliminary findings that indicate that there are certain similarities (Cabrio et al., 2013), approaches like RST and PDTB aim at identifying general discourse structures and are not tailored to argumentative discourse.

The difference of the relations is best illustrated by the work of Biran and Rambow (2011), which is to the best of our knowledge the only approach that focuses on the identification of distinct argumentative relations. The authors argue that existing definitions of discourse relations are only usable as a building block for argumentation mining and that there are no distinct argumentative relations included in existing approaches. Therefore, they combine 12 relations from the RST Discourse Treebank (Carlson et al., 2001) to a single argumentative support relation for identifying justifications in online discussions.

4.2 Discourse Markers and Indicators of Argumentative Relations

There is a large body of previous research in linguistics on the role of *discourse markers*, signalling discourse relations (e.g.*'because'*, *'therefore'*, *'since'*, etc.) in discourse analysis. Most previous investigations of discourse markers are based on the PDTB (Prasad et al., 2008) and on the RST Discourse Treebank (Carlson et al., 2003).

However, a critically discussed question in this context is the definition of discourse markers. Are discourse markers in the sense of indicators marking discourse relations just words like *'because'*, *'therefore'*, *'since'*? Taboada (2006) investigates the role of discourse markers in corpora annotated with discourse relations according to the RST. In her discussion of related work on discourse markers in linguistics, she concludes that there are many lexical and linguistic devices signalling discourse relations beyond discourse markers, such as the mood (e.g. indicative or conjunctive) or the modality (e.g. possibility, necessity) of a sentence.

In particular, for *argumentative* discourse, the role of indicators, such as discourse markers, is not well-understood yet, which is due to the lack of corpora annotated with argumentation structures. Recently, Tseronis (2011) summarized intermediate results of a corpus-based analysis of argumentative moves, aiming at the identification of linguistic surface cues that act as *argumentative markers*. According to Tseronis (2011), *any* single or complex lexical expression can act as an argumentative marker, and it can either mark an argumentative relation (i.e., connecting two arguments or argument components) or signal a certain argumentative role, such as a claim or a premise. Moreover, he observed that also sequential patterns of argumentative markers indicate particular argumentative moves, for instance, first stating the common ground (e.g., using the marker *it is understandable ...*) and then presenting an attack to this common ground (e.g., using a marker such as *nevertheless*).

5 Argumentation Structure Annotation

Our research in argumentation mining is motivated by the (1) information access and (2) computer-assisted writing perspective. Currently, we are conducting two annotation studies, focused on analyzing argumentation structures in scientific articles and persuasive essays. In the following subsections we provide an overview of the (preliminary) results.

5.1 Argumentation Structures in Scientific Articles

One of the main goals of any scientific publication is to present new research results to an expert audience. In order to emphasize the novelty and importance of the research findings, scientists usually build up an argumentation structure that provides numerous arguments in favor of their results. The goal of this annotation study is to automatically identify those argumentation structures on a fine-grained level in scientific publications in the educational domain and thereby to improve information access. A potential use case could be an automated summarization system creating a summary of important arguments presented in a scientific article.

Up to now only coarse-grained approaches like Argumentative Zoning (Teufel et al., 2009; Liakata et al., 2012; Yepes et al., 2013) have been developed for argumentation mining in scientific publications. These approaches classify argument components according to their argumentative contribution to the document (see section 3.2) but they do not consider any relations between the argument components. To the best of our knowledge, there is no prior work on identifying argumentation structures on a fine-grained level in scientific full-texts yet (see section 3.3).

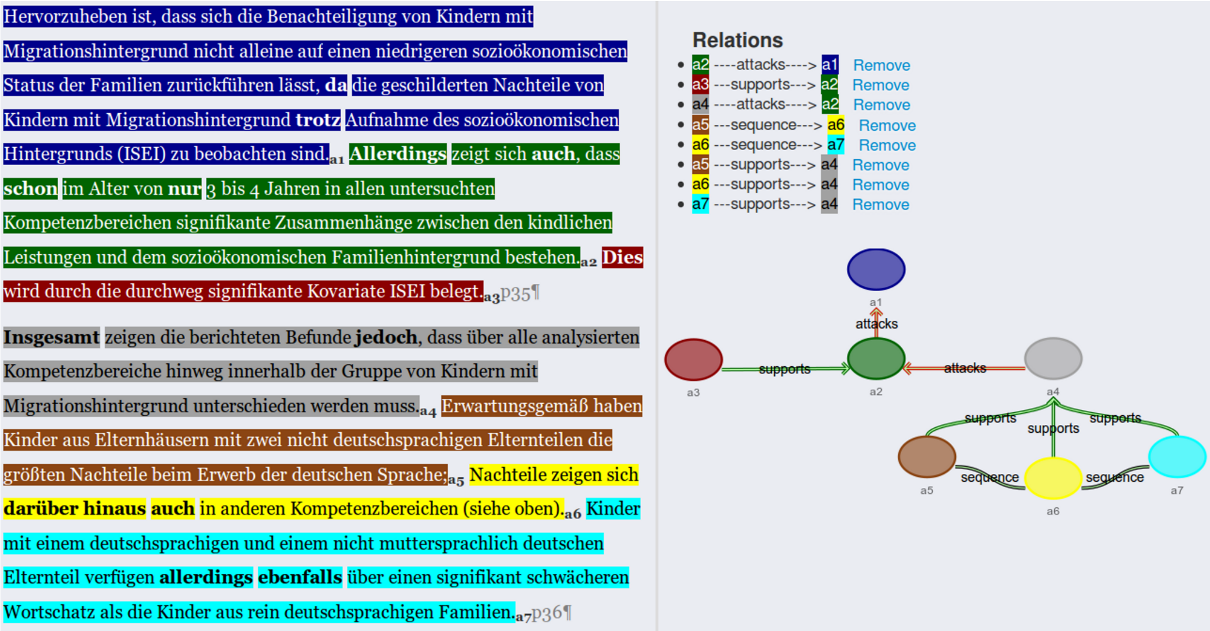
Due to the lack of evaluation datasets, we are performing an annotation study with four annotators, two domain experts and two annotators who developed the annotation guidelines. Our dataset consists of about 20 scientific full-texts from the educational domain. For the annotation study, we developed our own Web-based annotation tool (see figure 3 for a screenshot). The annotation tool allows to label argument components directly in the text with different colors and to add different relations (like support or attack) between argument components. The resulting argumentation structure is visualized as a graph (see figure 3).

Next, we plan to develop weakly supervised machine learning methods to automatically annotate scientific publications with argument components and the relations between them. The first step will be to distinguish non-argumentative parts (for example descriptions of the document structure) from argumentative parts (see section 3.1). The second step will be to identify support and attack relations between the argument components. In particular, we will explore lexical features, such as discourse markers (for example *'hence'*, *'so'*, *'for that reason'*, *'but'*, *'however'*, see section 4), and semantic features, such as text similarity or textual entailment.

5.2 Identifying Argumentation Structures for Computer-Assisted Writing

The goal of computer-assisted writing is to provide feedback about written language in order to improve text quality and writing skills of authors respectively. Common approaches are for instance focused on providing feedback about spelling and grammar, whereas more sophisticated approaches also provide feedback about discourse structures (Burstein et al., 2003), readability (Pitler and Nenkova, 2008), style (Burstein and Wolska, 2003) or aim at facilitating second language writing (Chen et al., 2012; Huang et al., 2012).

Argumentative Writing Support is a particular type of computer-assisted writing that aims at providing feedback about argumentation and thus postulates methods for reliably identifying arguments. Besides the recognition of argument components, the identification of the argumentation



Hervorzuheben ist, dass sich die Benachteiligung von Kindern mit Migrationshintergrund nicht alleine auf einen niedrigeren sozioökonomischen Status der Familien zurückführen lässt, **da** die geschilderten Nachteile von Kindern mit Migrationshintergrund **trotz** Aufnahme des sozioökonomischen Hintergrunds (ISEI) zu beobachten sind.$_{a1}$ **Allerdings** zeigt sich **auch**, dass **schon** im Alter von **nur** 3 bis 4 Jahren in allen untersuchten Kompetenzbereichen signifikante Zusammenhänge zwischen den kindlichen Leistungen und dem sozioökonomischen Familienhintergrund bestehen.$_{a2}$ **Dies** wird durch die durchweg signifikante Kovariate ISEI belegt.$_{a3}$P35¶

Insgesamt zeigen die berichteten Befunde **jedoch**, dass über alle analysierten Kompetenzbereiche hinweg innerhalb der Gruppe von Kindern mit Migrationshintergrund unterschieden werden muss.$_{a4}$ Erwartungsgemäß haben Kinder aus Elternhäusern mit zwei nicht deutschsprachigen Elternteilen die größten Nachteile beim Erwerb der deutschen Sprache;$_{a5}$ Nachteile zeigen sich **darüber hinaus** **auch** in anderen Kompetenzbereichen (siehe oben).$_{a6}$ Kinder mit einem deutschsprachigen und einem nicht muttersprachlich deutschen Elternteil verfügen **allerdings** **ebenfalls** über einen signifikant schwächeren Wortschatz als die Kinder aus rein deutschsprachigen Familien.$_{a7}$P36¶

Relations

- a2 ----attacks----> a1 Remove
- a3 ---supports---> a2 Remove
- a4 ----attacks----> a2 Remove
- a5 ---sequence---> a6 Remove
- a6 ---sequence---> a7 Remove
- a5 ---supports---> a4 Remove
- a6 ---supports---> a4 Remove
- a7 ---supports---> a4 Remove

Figure 3: Screenshot of the annotation tool for argumentation structure annotation in scientific full-texts: The left side includes the text of a scientific article and the argument components marked with different colors and labels (a1-a7). The graph visualization on the right side illustrates the argumentation structure. Each node represents an argument component connected with several relations (*'support'*, *'attack'*, *'sequence'*).

structure is crucial for argumentative writing support, since it would open novel possibilities for providing formative feedback about argumentation. On the one hand, an analysis of the argumentation structure would enable the recommendation of more meaningful arrangements of argument components and a reasonable usage of discourse markers. Both have been shown to increase argument comprehension and recall, and thus the quality of the text (Anne Britt and Larson, 2003). On the other hand, by identifying which premises belong to a claim, it would be possible to advice the author to add additional support in her/his argumentation to improve the persuasiveness.

Following this vision, we conducted an annotation study with three annotators to model argument components and the argumentation structure in persuasive essays at the clause-level. The corpus includes 90 persuasive essays which we selected from *essayforum.com*. Our annotation scheme includes three argument components (*major claim*, *claim* and *premise*) and two argumentative relations (*support* and *attack*). For defining the annotation guidelines and the annotation process we conducted a preliminary study on a corpus of 14 short text snippets with five non-trained annotators and found that information about the topic and the author's stance is crucial for annotating arguments. According to these findings, we defined a top-down annotation process starting with the major claim and drilling-down to the claims and the premises so that the annotators are aware of the author's stance and the topic before annotating other components. Using this strategy, we achieved an inter-rater agreement of $\alpha_U = 0.72$[5] for argument components and $\alpha = 0.81$ for argumentative relations indicating that the proposed scheme and annotation process successfully guides annotators to substantial agreement. For more details about this annotation study, we refer the interested reader to (Stab and Gurevych, 2014), which includes a detailed description of the annotation scheme, an analysis of inter-annotator agreements on different granularities and an error analysis. The corpus as well as the annotation guidelines are freely available to encourage future research.[6]

[5] We used Krippendorff's α_U (Krippendorff, 2004) for measuring the agreement since there are no predefined marbles in our study and annotators had also to identify the boundaries of argument components.

[6] `http://www.ukp.tu-darmstadt.de/data/argumentation-mining`

6 Challenges

Existing approaches of argumentation mining mainly focus on the identification of argument components (section 3). Based on the examples analyzed in section 2 and on the experience gained in our annotation studies (section 5), we identified the following challenges for future research in argumentation mining that have not been addressed adequately by previous work.

Segmentation: Most of the existing approaches are based on the sentence-level. However, for analyzing arguments, a more fine-grained segmentation is needed (Sergeant, 2013). Apart from the sentence level, in real world data argument components exist on the clause level or can spread over several sentences. For instance, example (4) illustrates that a single sentence can contain multiple argument components (claim in bold face and premise underlined) (see also example (2) in section 2). In example (5) the premise consists of two sentences, because both sentences are needed to represent and support the "different opinions" in the claim.

> (4) *"**Eating apples is healthy** which has to do with substrates which prevent cancer and other diseases."*
>
> (5) *"**There are different opinions about coffee.** Some people say they need it to stay awake. Other people think it's unhealthy."*

It is an open question if existing segmentation approaches can be used for reliably identifying the boundaries of argument components. In example (4) we find two times the word "which". This makes it hard for a segmenter to split the sentence correctly in only two parts. On the other hand, the combination of sentences (example (5)) also requires more elaborated techniques that are able to identify sentences that are related and only form in combination the support of a particular claim.

Context Dependence: The context is crucial for identifying arguments, their components and argumentation structures. As illustrated by Stab and Gurevych (2014), it is even a hard task for human annotators to distinguish claims and premises without being aware of the context. For instance, the following three argument components constitute a reasoning chain in which *c* is a premise for *b* and *b* a premise for *a*:

> (6) *"Random locker checks should be made obligatory.*$_a$ *Locker checks help students stay both physically and mentally healthy.*$_b$ *It discourages students from bringing firearms and especially drugs.*$_c$*"*

In this argumentation structure, *a* can be classified as a claim. However, without being aware of the argument component *a*, *b* becomes a claim which is supported by premise *c*. The same situation can be found in example (3) in section 2. If we look at the argument components *b* and *c* in isolation, we can classify *b* as claim. However, looking at the whole example, the argument component *a* is the claim, supported by the premise *b*. The same holds for the argument components *c* and *a* which would be connected by a support relation if they are considered in isolation. Both examples illustrate that the context is crucial for classifying argument components as claims or premises and for identifying the argumentation structure. Although, Stab and Gurevych (2014) proposed an annotation process that facilitates these decisions in manual annotation studies of persuasive essays, it is still an open issue how to model the context in order to improve the performance of automatic argumentation mining methods.

Ambiguity of Argumentation Structures: The most important challenge for identifying argumentation structures is ambiguity, since there are often several possible interpretations of argumentation structures which makes it hard or even impossible to identify one correct interpretation. In previous examples, we have already seen that the classification of argument components depends on the context and the considered argument components respectively. However, even if we consider all components of an argument, there might be several reasonable interpretations of its structure. For instance, the structure of example (6) can be interpreted in three different ways (figure 4). In the first interpretation, the argument component *c* supports argument component *b* and argument component *b* supports argument component *a*, whereas in the second interpretation argument components *b* and *c* both support argument component *a*. The third interpretation contains all possible argumentative relations from the first and second interpretation combined, and thus represents a graph structure (in contrast to a tree structure).

The ambiguity of argumentation structures rep-

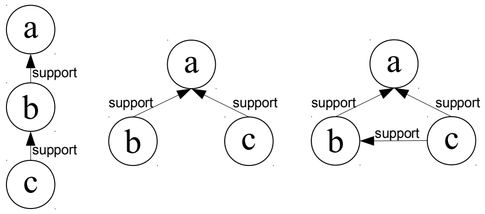


Figure 4: Several interpretations of the argumentation structure of example (6).

resents a major challenge for argument annotation studies and consequently the creation of reliable gold standards for argumentation mining. In all annotation studies we know, exactly one annotation is considered to be correct which means that other possibly correct interpretations are considered as incorrect and therefore downgrade the results for the inter annotator agreement and the performance of automatic classifiers. Consequently, it might be interesting to explore different evaluation methods. For instance, evaluation schemes used in automatic text summarization could be considered as an alternative. In text summarization, inter annotator agreement for human-generated summaries is particularly low, and hence, each human-generated summary is considered valid for evaluating an automatic summarization system (Nenkova and McKeown, 2012).

7 Conclusion

In this paper, we showed that existing approaches to argumentation mining mainly focus on the identification of argument components and largely neglect the identification of argumentation structures, although this task is crucial for many promising applications, e.g., for building novel argument related knowledge bases. By examining several examples, we derived characteristic properties of argumentation structures. We discussed the relation of discourse analysis and argumentation structure and showed that previous works in discourse analysis are not capable of identifying argumentation structures, because discourse relations do not cover all argumentative relations and are limited to relations between adjacent text units. Based on our observations, we derived three challenges for encouraging future research, i.e., (i) identifying the boundaries of argument components, (ii) modeling the context of argument components and argumentative relations, and (iii) addressing the problem of ambiguous argumentation structures. In particular, the ambiguity of argumentation structure poses an important issue for future work.

References

M. Anne Britt and Aaron A. Larson. 2003. Constructing representations of arguments. *Journal of Memory and Language*, 48(4):794–810.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*, volume 47. MIT press Cambridge.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 05(04):363–381.

Jill Burstein and Magdalena Wolska. 2003. Toward Evaluation of Writing Style: Finding Overly Repetitive Word Use in Student Essays. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 35–42, Budapest, Hungary.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18(1):32–39.

Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI '12, pages 205–210, Montpellier, France.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In João Leite, TranCao Son, Paolo Torroni, Leon Torre, and Stefan Woltran, editors, *Computational Logic in Multi-Agent Systems*, volume 8143 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10, Aalborg, Denmark.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski, 2003. *Building a discourse-tagged corpus in the framework of rhetorical structure theory*, chapter 5, pages 85–112. Springer.

Mei-Hua Chen, Shih-Ting Huang, Hung-Ting Hsieh, Ting-Hui Kao, and Jason S. Chang. 2012. FLOW: A First-language-oriented Writing Assistant System. In *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, pages 157–162, Jeju Island, Korea.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 987–996, Stroudsburg, PA, USA. Association for Computational Linguistics.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. 2013. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, Sofia, Bulgaria, August. Association for Computational Linguistics.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*, volume 18 of *Argumentation Library*. Springer.

Chung-chi Huang, Ping-che Yang, Mei-hua Chen, Hung-ting Hsieh, Ting-hui Kao, and Jason S. Chang. 2012. TransAhead: A Writing Assistant for CAT and CALL. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 16–19, Avignon, France.

Klaus Krippendorff. 2004. Measuring the Reliability of Qualitative Text Analysis Data. *Quality & Quantity*, 38(6):787–800.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 368–375.

Raquel Mochales-Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

Raquel Mochales-Palau and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA. ACM.

Ani Nenkova and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining Text Data*, pages 43–76.

Andreas Peldszus and Manfred Stede. 2013. From Argument Diagrams to Argumentation Mining in Texts: A Survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 186–195.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 28–30, Marrakech, Morocco.

Chris Reed, Raquel Mochales-Palau, Glenn Rowe, and Marie-Francine Moens. 2008. Language resources for studying argument. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, LREC '08, pages 2613–2618, Marrakech, Morocco.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, FLAIRS '12, pages 272–275, Marco Island, FL, USA.

Alan Sergeant. 2013. Automatic argumentation extraction. In *Proceedings of the 10th European Semantic Web Conference*, ESWC '13, pages 656–660, Montpellier, France.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, page (to appear), Dublin, Ireland, August.

Maite Taboada. 2006. Discourse Markers as Signals (or Not) of Rhetorical Relationsteu. *Journal of Pragmatics*, 38(4):567–592.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1493–1502.

Simone Teufel. 1999. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, University of Edinburgh.

Stephen E. Toulmin. 1958. *The uses of Argument*. Cambridge University Press.

Assimakis Tseronis. 2011. From connectives to argumentative markers: A quest for markers of argumentative moves and of related aspects of argumentative discourse. *Argumentation*, 25(4):427–447.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Douglas N Walton. 1996. *Argumentation schemes for presumptive reasoning*. Routledge.

Bonnie Webber, Mark Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18:437–490, 10.

Adam Wyner, Raquel Mochales Palau, Marie-Francine Moens, and David Milward. 2010. Approaches to text mining arguments from legal cases. In *Semantic Processing of Legal Texts*, volume 6036 of *Lecture Notes in Computer Science*, pages 60–79. Springer.

Antonio Jimeno Yepes, James G. Mork, and Alan R. Aronson. 2013. Using the argumentative structure of scientific literature to improve information access. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 102–110.

Argumentation, Ideology, and Issue Framing in Parliamentary Discourse

Graeme Hirst,* Vanessa Wei Feng,* Christopher Cochrane,† and Nona Naderi*
*Department of Computer Science and †Department of Political Science
University of Toronto, Toronto, Ontario, Canada
*{gh,weifeng,nona}@cs.toronto.edu
†christopher.cochrane@utoronto.ca

Abstract

In argumentative political speech, the way an issue is framed may indicate the unstated assumptions of the argument and hence the ideological position of the speaker. Our goal is to use and extend our prior work on discourse parsing and the identification of argumentation schemes to identify specific instances of issue framing and, more generally, ideological positions as they are expressed in text. We are using annotated historical and contemporary proceedings of the British, Canadian, and Dutch parliaments, looking in particular at speech on the topic of immigration.

1 Introduction

A key aspect of any argument is the unstated assumptions and beliefs that underlie it. At bottom, all naturally occurring arguments are enthymematic. Our research in argumentation has the long-term goal of identifying these unstated elements, both at the micro level — the specific unstated premises of an argument — and at the macro level — the belief system or ideology within which the entire argument is constructed, which may in turn contribute to its unstated premises (and also to any unstated conclusions).

Our past research has concerned analysis of argumentation, and the related issue of determining the rhetorical structure of discourse, at the micro level. In this paper, we briefly describe this work. We then describe our present and planned research on ideology-based argumentation, including, in particular, the identification of specific kinds of issue framing and their role in ideological disagreement.

Our research is part of the project Digging Into Linked Parliamentary Data ("Dilipad"), an interdisciplinary tri-national project that is collecting and richly annotating historical and contemporary parliamentary proceedings of the U.K., Canada, and the Netherlands for use in studies in political science, political history, and other areas of social science and linguistics.[1] The project includes two case studies on the identification of ideology, ideological frameworks, and argumentation in the data, which we will describe below.

2 Argumentation analysis

The context for our initial research on argumentation (presented in detail by Feng and Hirst (2011)) was the early work of Mochales and Moens (2008; 2009a; 2009b), who focused on automatic detection of arguments in legal texts. With each sentence represented as a vector of shallow features, they trained a multinomial naïve Bayes classifier and a maximum entropy model on the Araucaria corpus. In their follow-up work, they trained a support vector machine to further classify each argumentative clause into a premise or a conclusion. In addition, they developed a context-free grammar for argumentation structure parsing. Our work is "downstream" from that of Mochales and Moens. Assuming the eventual success of their, or others', research program on detecting and classifying the components of an argument, we sought to determine how the pieces fit together as an instance of an argumentation scheme. This, in turn, would be used, in future work, to understand the argument and recover the unstated assumptions. Figure 1 shows the structure of a complete posited system, with our work addressing the part inside the red dashed line.

Of Walton's set of 65 argumentation schemes (Walton *et al.*, 2008), we focused on the five that are most frequent in the Araucaria dataset (Reed and Rowe, 2004; Rowe and Reed, 2008): ar-

[1] For more details of the project, including the other participating institutions and researchers, see http://dilipad.history.ac.uk

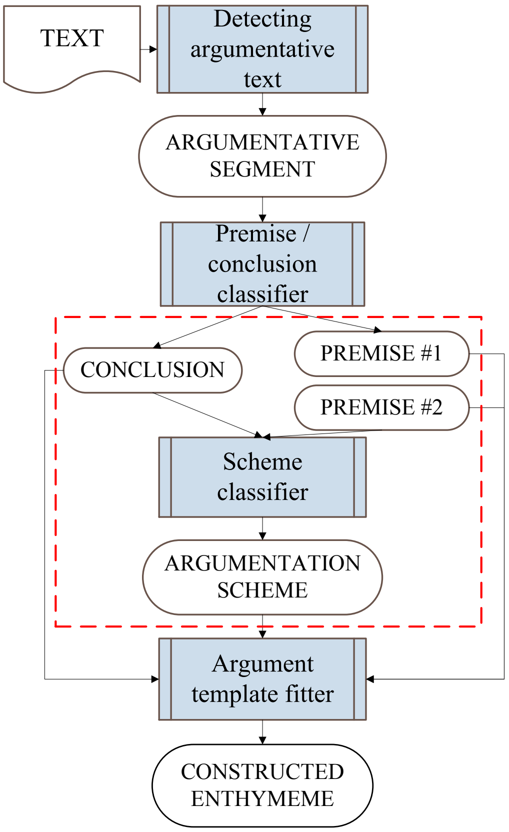


Figure 1: Overall framework of our research on argumentation schemes.

gument from example, argument from cause to effect, practical reasoning, argument from consequences, and argument from verbal classification. Casting the problem as one of text classification, we built a pruned C4.5 decision tree (Quinlan, 1993) for both one-against-others classification of each scheme and for pairwise classification of each possible pairing of schemes. We used a variety of textual features, some of them specific to a particular argument scheme and others identical across schemes. They ranged from specific keywords and phrases to word-pair similarity between the premise and the conclusion, the starting point of the premise or conclusion in its sentence, and various syntactic dependency relations. Additionally, we used one feature that cannot at present be automatically derived from text, but which we assume may be determined by cues such as discourse relations: whether the argument is linked or convergent; that is, whether or all just one of the premises suffice for the conclusion.

Using Araucaria for both training and testing, we achieved high accuracy in one-against-others classification for argument from example and practical reasoning: 90.6% and 90.8% (baseline is 50%). The accuracy of classification of argument from cause to effect was just over 70%. However, with the other two schemes (argument from consequences and argument from verbal classification), accuracy was only in the low 60s. This is probably due at least partly to the fact that these schemes do not have such obvious cue phrases or patterns as the other three schemes, and therefore may require more world knowledge, and also because the available training data for each in Araucaria was relatively small (44 and 41 instances, respectively). In pairwise classification, we were able to correctly differentiate between most of the scheme pairs, with accuracies as high as 98% (baseline is again 50%). Performance was poor (64.0%) only for argument from consequences against argument from verbal classification — perhaps not coincidentally the two schemes for which performance was poorest in the one-against-others task.

3 Discourse analysis for argumentation analysis

The rhetorical or discourse structure of an argumentative text contributes to (or is, in part, determined by) the structure of the argument that it expresses. Consequently, much of our recent work has focused on **discourse parsing**, that is, determining the hierarchical rhetorical structure of the text: the logical relationships between sentences. Following the tenets of **Rhetorical Structure Theory (RST)** (Mann and Thompson, 1988), this is a tree structure that covers the text whose leaves are the elementary discourse units (EDUs) of text (roughly speaking, clauses and clause-like constituents) and whose edges are the RST relations that hold between EDUs or spans of related text. The set of relations include many that are pertinent to the structure of argumentation, such as CONTRAST, CAUSE, SUMMARY and ENABLEMENT. Also, as we noted above, an analysis of discourse structure may help us to discriminate convergent from linked arguments. So while an RST structure is not an argumentation structure per se, it clearly contains information that contributes to building an argumentation structure.

Our research on discourse parsing has three facets: improving the initial segmentation of text into EDUs (Feng and Hirst, 2014b); improving the parsing itself by using rich linguistic features (Feng and Hirst, 2012); and technically improving the parser both in accuracy and in efficiency by separating the parsing of intra-sentence and multi-sentence structures into separate processes (following Joty *et al.* (2013)), and adding a post-editing pass to each process (Feng and Hirst, 2014a). Bringing the improvements together, and training and testing in the RST Discourse Treebank (Carlson *et al.*, 2001), we achieved an F_1 score of 92.6% on discourse segmentation, and an accuracy of 58.2% (against a baseline of 29.6%)[2] on recognizing discourse relations on a gold-standard segmentation.

Our next task will be to combine our discourse parser with our earlier work on identifying argumentation schemes. We will augment our classifier with new features derived from the discourse structure in order to improve its accuracy. We will also use discourse structure features to improve the upstream classification that feeds into the argumentation scheme classifier, and to begin the task of further downstream analysis. In particular, this will include analysis of arguments to determine the underlying ideology of a text.

4 Ideology and issue framing

Social scientists usually define ideology as a belief system: "a configuration of ideas and attitudes in which the elements are bound together by some form of constraint or functional interdependence" (Converse, 1964, p. 207). The **left / right** political divide is a systematic and enduring ideological cleavage that divides "the world of political thought and action" in democratic countries (Bobbio, 1996). Systematic left / right differences appear in the voting records of politicians in legislative assemblies (Hix *et al.*, 2006), in the election platforms of political parties (Budge *et al.*, 2001; Klingemann *et al.*, 2006), and in the patterns of public opinion (Jost, 2006). The left / right divide is so pervasive and enduring that many now wonder whether these political differences are manifestations of deeply rooted, and perhaps heritable, psychological traits (Alford *et al.*, 2005; Carney *et al.*, 2008; Haidt, 2012).

Several computational studies have looked at the question of whether a political speaker's ideological position on the left / right spectrum can be determined just from a quantitative analysis of the vocabulary that they use — both from the way they talk about particular topics and (in some contexts) from the topics that they tend to talk about (Lin *et al.*, 2006; Mullen and Malouf, 2006; Yu *et al.*, 2008; Diermeier *et al.*, 2012; Zirn, 2014). Typically, these studies attempt to induce a classifier from word-frequency vectors. Results have been mixed; for example, extreme positions in the U.S. Congress can be distinguished from those of the other side — sometimes by the use of topic-dependent shibboleths such as *gay* (liberal Democrat) or *homosexual* (conservative Republican) — but more-moderate positions cannot be (Yu *et al.*, 2008).

In our earlier work (Hirst *et al.*, 2010; Hirst *et al.*, 2014), we showed that the U.S. results do not apply to the Canadian Parliament. On one hand, we were able to classify party membership more reliably overall than the U.S. research did, but on the other hand we also showed that distinctions in the vocabulary of the speakers depend far more upon whether their party was in government or in opposition than upon their ideological position. The differences reflect primarily defence (government) and attack (opposition), a feature inherent to parliamentary governments in general, and especially to the Canadian parliament where party discipline is particularly strict (Savoie, 1999). When we applied classification methods based on word-frequency to the proceedings of the European Parliament, in which the factor of government–opposition status is absent, we achieved a more-accurate ideological classification of speakers from the five major parties across the left / right spectrum (Hirst *et al.*, 2014). This confounding role of institutions on left / right differences align with what others have recently uncovered in cross-national analysis of legislative voting patterns (Hix and Noury, 2013).

Casual observers of politics recognize left / right differences when they see them, but even experts struggle to define these terms. The root of the problem is the effort to define left and right by reducing each side to a single idea or "essential core". The morphology of left and right is inconsistent with such a specification. Rather, left and right describe "family resemblances" between the systems of political ideas that actors on each side advance on the questions of political disagreement (Cochrane, 2014). Although no single idea de-

[2] This is the majority baseline of always labeling the resulting subtree with the relation ELABORATION with the current span as the nucleus and the next span as the satellite.

fines the left or the right, ideas are more or less central to one of these resemblances to the extent that they are more common among the belief systems of actors that are inside each category than they are among the beliefs systems of actors that are outside each category. From this vantage point, the central ideas on the political left are commitments to equality, pacifism, and, more recently, the environment. The distinguishing ideas on the right are support for capitalist economic orthodoxy, law and order, and patriotic militarism (Cochrane, 2014). The differences between political parties in their support for these ideas explain more than two-thirds of the variation in how citizens and experts position the parties on a left / right dimension (Cochrane, 2014).

The "content" of a belief system is the set of preferences that an actor harbours about political issues. The "structure" of a belief system is the way in which an actor puts different political issues together into bundles of constrained preferences. Actors that think about politics from the vantage point of altogether different ideas not only disagree in their positions on issues, they also disagree in their views of how different issues fit together logically in the political world around them. Thus, the content and the structure of belief systems varies on the left and the right (Cochrane, 2013).

Because of these differences, individuals from different ideological positions will often **frame** things differently in argumentation on any particular issue. For example, on the issue of how much immigration should be allowed into their country, one person might frame the argument as one of economic benefit or detriment, a second person as one of the benefits or problems of multiculturalism, and a third person as one of social justice.[3] These differences will be reflected in the vocabulary that each of these people uses, which accounts for the results presented above on identifying ideology based on vocabulary alone; in the absence of confounding factors, as we saw most clearly in the case of the European Parliament, vocabulary is a strong indicator all by itself.

So we see that the framing of an issue by a speaker in an argumentative text is not, ultimately, a linguistic entity; it's an ideological viewpoint or perspective: a set of beliefs, assumptions and precompiled arguments.[4] Nonetheless, for automatic text analysis, quantifiable semantic characteristics of the speaker's presentation of a position are indicators or proxies of the framing, which can then be interpreted qualitatively (by a human). In a simple analysis, this might be a statistical analysis of the key concepts of the text, as denoted by content words, significant collocations of words, and syntactic structures, much as in the simple text-classification–based ideology studies mentioned above, or a topic-model–based analysis, as in the work of Nguyen *et al.* (2013).

In our research, however, we are also proposing a novel, more-sophisticated analysis in which we also look at the actual argumentation structures and discourse relationships of the text and how the concepts adduced by the lower-level linguistic components are used in these structures. We will describe these proposals in the next section.

5 Argumentation and issue framing in parliamentary speech

Left / right speech is a subset of ideological speech more generally. Ideological speech is a subset of political speech more generally. As we noted above, previous analyses of political speech attempt to induce left / right classifiers from analyses of vocabulary across all of the many topics of discussion in a dataset. But this approach disregards the results of an extensive body of political science research that analyzes left / right ideological disagreement in legislative voting records (Poole and Rosenthal, 2007; Hix and Noury, 2013), party election manifestos (Budge *et al.*, 2001; Klingemann *et al.*, 2006), and opinions (Jost, 2006). A key finding from these studies concerns the varying centrality of specific actors, ideas, and topics to left / right political disagreement. Some actors are more central to the left or to the right than are other actors. Some ideas are more central to the left or to the right than are other ideas. Left / right disagreements implicate some political issues and not others. This provides an informative prior for models that seek to uncover left / right differences from the patterns of vocabu-

[3] Immigration is in fact the particular topic on which we will conduct our case study on the framing of arguments; see section 5 below.

[4] A fortiori, framing is a political action: "Framing essentially involves selection and salience. To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described" (Entman, 1993). But here, we focus on the linguistic and argumentative aspects of framing.

lary and argumentation in political text. The likelihood that speech conveys information about left / right argumentation is a function of the speaker and the topic.

Thus, the goal of our work, broadly speaking, is to develop computational models for the automatic analysis of ideology and issue framing in political speech that are better informed than the simple vocabulary-based models and that draw on automatic discourse parsing and automatic analysis of argumentation as their primary mechanism. We would like to look more narrowly and more deeply at argumentation on specific issues by individuals across the left / right spectrum, and develop automatic methods of analysis that will identify, or help analysts to identify, different frames and ideological positions. Our "help to" hedge reflects the difficulty of the goal and the context of our research as part of a much-larger project that is building datasets and tools to assist political scientists and political historians in their analyses.

The primary data for our work is the annotated parliamentary proceedings, from the present back to the mid-1800s or earlier, that are being produced by the Dilipad project (see section 1 above), from which we will draw speech[5] on specific topics for diachronic and cross-national analysis of argumentation and framing. Immigration is a topic of special interest here, as it has been an important and recurring issue since the nineteenth century in all three participating countries. We hope to identify national and temporal differences and similarities in the frames used to discuss the issue.

In our models, we will bring together, and extend, the work on discourse parsing and argumentation scheme identification described in sections 2 and 3 above. Although these techniques are far from perfect, we hypothesize that typical political speech contains a sufficiently well-cued discourse structure that the analyses that we can achieve, although still quite imperfect, will be usefully indicative of issue framing and other ideological signals, and will be more immune to confounding factors, such as the attack-and-defence dynamics of parliamentary debates, than simple vocabulary classification. In particular, we will use features from discourse units and rhetorical relations to find claims and analyze the reasoning structure that is used to justify, support, and derive the claims. In addition, we will take into account how the concepts adduced by lower-level linguistic components — phrases, syntactic dependency structures — are used in the actual argumentation structures and discourse relationships of the text. We hope to be able to recognize instances of known frames in the text, and possibly even discover new ones. Because we will be developing deeper and hence more tentative methods of computational linguistic analysis, we do not expect to provide a complete automated analysis of text in the first instance, but rather to provide data that can then be interpreted by a human analyst.

In parallel with this approach, we will also develop text-classification methods for identifying ideological positions in speech that will look beyond vocabulary and also take into consideration frequent collocations and lexicalized syntactic dependency structures as features. This will allow us to include differences in the way that particular words are used (even where speakers use the word with the same frequency) as a feature of the classification. This will provide a new, higher baseline against which the results of the discourse- and argumentation-based analysis can be evaluated. It may also provide information that can itself be a component of that analysis. In addition, the words, collocations, and dependency structures that are most informative for classification will, as with our other methods, be available for human interpretation.

6 Conclusion

Our work focuses on the structure of discourse and arguments to better understand ideological positions and issue framing through their linguistic realizations. By applying discourse parsing and the analysis of argumentation to parliamentary debates, we hope to determine how speakers with various ideologies argue on a range of issues. Ideologies are manifested not only by the vocabularies used, but also by how the differing beliefs of political speakers lead to different framing of issues. Ideology detection can therefore benefit from argumentation and discourse analysis techniques.

[5] Although we refer to *political* and *parliamentary speech* and *speakers*, as is conventional, we are working only with the published textual transcriptions of the parliamentary debates. We are not using audio data or any kind of automatic speech recognition.

Acknowledgements

The Digging Into Linked Parliamentary Data project is funded through the Digging Into Data Challenge. The Canadian arm of the project is funded by the Natural Sciences and Engineering Research Council of Canada through its Discovery Frontiers program and by the Social Sciences and Humanities Research Council. We are grateful to Kaspar Beelen for helpful comments on an earlier draft of this paper.

References

Alford, John R.; Funk, Carolyn L.; and Hibbing, John R. (2005). *American Political Science Review*, 99(2), 153–167.

Bobbio, Noberto (1996). *Left and Right: The Significance of a Political Distinction*. Cambridge, UK: Polity Press.

Budge, Ian; Klingemann, Hans-Dieter; Volkens, Andrea; and Bara, Judith (2006). *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945–1998*. Oxford, UK: Oxford University Press.

Carlson, Lynn; Marcu, Daniel; and Okurowski, Mary Ellen (2001). Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. *Proceedings of Second SIGDial Workshop on Discourse and Dialogue (SIGDial 2001)*, Aalborg, 1–10.

Carney, Dana R.; Jost, John T.; Gosling, Samuel D.; and Potter, Jeff (2012). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, 29(6), 807–840.

Cochrane, Christopher (2013). The asymmetrical structure of left / right disagreement: Left-wing coherence and right-wing fragmentation in comparative party policy. *Party Politics*, 19(1), 104–121.

Cochrane, Christopher (2014). *Left and Right: The Small World of Political Ideas*. MS under review.

Converse, Philip E. (1964). The nature of belief systems in mass publics. In David E. Apter, ed. *Ideology and Discontent*. London, UK: Collier-MacMillan, 206–261.

Diermeier, Daniel; Godbout, Jean-François; Yu, Bei; and Kaufmann, Stefan (2012). Language and ideology in Congress. *British Journal of Political Science*, 42(1), 31–55.

Entman, Robert M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4), 51–58.

Feng, Vanessa Wei and Hirst, Graeme (2011). Classifying arguments by scheme. *Proceedings, 49th Annual Meeting of the Association for Computational Linguistics*, Portland, Oregon, 978–996.

Feng, Vanessa Wei and Hirst, Graeme (2012). Text-level discourse parsing with rich linguistic features. *Proceedings, 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Korea, 60–68.

Feng, Vanessa Wei and Hirst, Graeme (2014a). A linear-time bottom-up discourse parser with constraints and post-editing. *Proceedings, 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore.

Feng, Vanessa Wei and Hirst, Graeme (2014b). Two-pass discourse segmentation with pairing and global features. Submitted.

Haidt, Jonathan (2012). *The Righteous Mind: Why Good People are Divided by Politics and Religion*. New York, NY: Pantheon Books.

Hirst, Graeme; Riabinin, Yaroslav; and Graham, Jory (2010). Party status as a confound in the automatic classification of political speech by ideology. *Proceedings, 10th International Conference on Statistical Analysis of Textual Data / 10es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2010)*, Rome, 731–742.

Hirst, Graeme; Riabinin, Yaroslav; Graham, Jory; Boizot-Roche, Magali; and Morris, Colin (2014). Text to ideology or text to party status? In: Kaal, Bertie; Maks, E. Isa; and van Elfrinkhof, Annemarie M.E. (editors), *From Text to Political Positions: Text analysis across disciplines,* Amsterdam: John Benjamins, 93–115.

Hix, Simon; and Noury, Abdul (2013). Government-opposition or left-right? The institutional determinants of voting in legislatures. Working paper, retrieved from `http://personal.lse.ac.uk/hix/Research.HTM`, 2014-06-14.

Hix, Simon; Noury, Abdul; and Roland, Gérare (2006). Democratic politics in the European Parliament. *American Journal of Political Science*, 50(2), 494–520.

Jost, John T. (2006). The end of the end of ideology. *American Psychologist*, 61(7), 651–670.

Joty, Shafiq; Carenini, Giuseppe; Ng, Raymond; and Mehdad, Yashar (2013). Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, Sofia, Bulgaria, 486–496.

Klingemann, Hans-Dieter; Volkens, Andrea; Bara, Judith; Budge, Ian; and McDonald, Michael (2006). *Mapping Policy Preferences II: Estimates for Parties, Electors, and Governments in Eastern Europe, European Union, and OECD 1990–2003*. Oxford, UK: Oxford University Press.

Lin, Wei-Hao; Wilson, Theresa; Wiebe, Janyce; and Hauptmann, Alexander (2006). Which side are you on? Identifying perspectives at the document and sentence levels. *Proceedings of the 10th Conference on Natural Language Learning (CoNLL-X)*, 109–116.

Mann, William and Thompson, Sandra (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3), 243–281.

Mochales, Raquel and Moens, Marie-Francine (2008). Study on the structure of argumentation in case law. *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems*, Amsterdam: IOS Press, 11–20.

Mochales, Raquel and Moens, Marie-Francine (2009a). Argumentation mining: the detection, classification and structure of arguments in text. *ICAIL '09: Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ACM, 98–107.

Mochales, Raquel and Moens, Marie-Francine (2009b). Automatic argumentation detection and its role in law and the Semantic Web. *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web*, Amsterdam: IOS Press, 115–129.

Mullen, Tony and Malouf, Robert (2006). A preliminary investigation into sentiment analysis of informal political discourse. *Proceedings of the AAAI Symposium on Computational Approaches to Analyzing Weblogs*, 159–162.

Nguyen, Viet-An; Boyd-Graber, Jordan; and Resnik, Philip (2013). Lexical and hierarchical topic regression. *Proceedings of Advances in Neural Information Processing Systems 26*, Lake Tahoe, NV, 1106–1114.

Poole, Keith T. and Rosenthal, Howard L. (2007). *Ideology and Congress*. New Brunswick, NJ: Transaction Publishers.

Quinlan, J. Ross (1993). C4.5: Programs for machine learning, *Machine Learning*, 16(3), 235–240.

Savoie, Donald (1999). *Governing from the Centre: The Concentration of Power in Canadian Politics*. Toronto, ON: University of Toronto Press.

Reed, Chris and Rowe, Glenn (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal of Artificial Intelligence Tools*, 14, 961–980.

Rowe, Glenn and Reed, Chris (2008). Argument diagramming: The Araucaria project. In *Knowledge Cartography: Software tools and mapping techniques*, edited by Alexandra Okada, Simon J. Buckingham Shum, and Tony Sherborne, London: Springer-Verlag, 163–181.

Walton, Douglas; Reed, Chris; and Macagno, Fabrizio (2008). *Argumentation Schemes*. Cambridge University Press.

Yu, Bei; Kaufmann, Stefan; and Diermeier, Daniel (2008). Classifying party affiliation from political speech. *Journal of Information Technology in Politics*, 5(1), 33–48.

Zirn, Cäcilia (2014). Analyzing positions and topics in political discussions of the German Bundestag. *Proceedings of the ACL 2014 Student Research Workshop*, Baltimore, 26–33.

Argumentation Theory in the Field: An Empirical Study of Fundamental Notions

Ariel Rosenfeld
Bar-Ilan University, Ramat-Gan, Israel
rosenfa5@cs.biu.ac.il

Sarit Kraus
Bar-Ilan University, Ramat-Gan, Israel
sarit@cs.biu.ac.il

Abstract

Argumentation Theory provides a very powerful set of principles, ideas and models. Yet, in this paper we will show that its fundamental principles unsatisfactorily explain real-world human argumentation and should be adapted. We will present an extensive empirical study on the incompatibility of abstract argumentation and human argumentative behavior, followed by practical expansion of existing models.

1 Introduction

Argumentation Theory has developed rapidly since Dung's seminal work (Dung, 1995). There has been extensive work extending Dung's framework and semantics; Value Argumentation Framework (VAF) (Bench-Capon et al., 2002), Bipolar Argumentation Framework (BAF) (Cayrol and Lagasquie-Schiex, 2005) and Weighted Argumentation Framework (WAF) (Dunne et al., 2011) to name a few. All reasonable frameworks and semantics rely on the same fundamental notions: *Conflict Freedom*, *Acceptability*, *Extensions* from (Dung, 1995), and expand upon them in some way. One more notion, which was not addressed in (Dung, 1995), *Support*, has been increasingly gaining attention (Boella et al., 2010). Overall, the same principals and ideas have prevailed for many years.

All of these models and semantics try to provide a *normative* approach to argumentation, i.e, how argumentation should work from a logical standard. From a *descriptive* point of view, the study of (Rahwan et al., 2010), where the authors investigated the reinstatement principle in behavioral experiments, is the only experimental study, as far as we know, that tested argumentation in the field. Nevertheless, many argumentative tools have been developed over time; MIT's delibrium (Klein, 2011), Araucaria (Reed and Rowe, 2004), ArgTrust (Tang et al., 2012) and Web-Based Intelligent Collaborative System (Liu et al., 2007), that try to provide systems where people can handle argumentative situations in a coherent and valid way. We believe that these argumentative tools and others, as efficient and attractive as they might be, have a difficult time attracting users outside the academia due to the gap between the Argumentation Theory and the human argumentative behavior, which, as previously stated, has not been addressed in the context of Argumentation Theory thus far.

In order to further develop argumentative applications and agents, we conducted a novel empirical study, with hundreds of human subjects, showing the incompatibility between some of the fundamental ideas, stated above, and human argumentation. In an attempt to mimic and understand the human argumentative process, these inconsistencies, which appear even in the weakest argumentative requirements as conflict freedom, pose a large concern for theoreticians and practitioners alike. Our findings indicate that the fundamental notions are not good predictive features of people's actions. A possible solution is also presented which provided better results in explaining people's arguments than the existing theory. This solution, which we call *Relevance*, captures a perceptual distance between arguments. That is, how one argument affects another and how this affect is comprehended by a reasoner. Relevance also holds a predicatory value as shown in recent work (Rosenfeld and Kraus, 2014).

This article's main contribution is in showing that the Argumentation Theory has difficulties in explaining a big part of the human argumentative behavior, in an extensive human study. Secondly, the proposed notion of relevance could in turn provide the argumentation community with an additional tool to investigate the existing theory and semantics.

2 Dung's Fundamental Notions

Argumentation is the process of supporting claims with grounds and defending them against attacks. Without explicitly specifying the underlying language (natural language, first order logic...), argument structure or attack/support relations, Dung has designed an abstract argumentation framework (Dung, 1995). This framework, combined with proposed semantics (reasoning rules), enables a reasoner to cope and reach conclusions in an environment of arguments that may conflict, support and interact with each other. These arguments may vary in their grounds and validity.

Definition 1. A Dungian Argumentation Framework (AF) is a pair $< A, R >$, where A is a set of arguments and R is an attack relation over $A \times A$.
Conflict-Free: A set of arguments S is conflict-free if there are no arguments a and b in S such that aRb holds.
Acceptable: An argument $a \in A$ is considered acceptable w.r.t a set of arguments S iff $\forall b.bRa \rightarrow \exists c \in S.cRb$.
Admissible: A set S is considered admissible iff it is conflict-free, and each argument in S is acceptable with respect to S.

Dung also defined several semantics by which, given an AF, one can derive the sets of arguments that should be considered *Justified* (to some extent). These sets are called *Extensions*. The different extenstions capture different notions of justification where some are more strict than others.

Definition 2. An extension $S \subseteq A$ is a set of arguments that satisfies some rules of reasoning.
Complete Extension: E is a complete extension of A iff it is an admissible set and every acceptable argument with respect to E belongs to E.
Preferred Extension: E is a preferred-extension in A iff it is a maximal (with respect to set inclusion) admissible set of arguments.
Stable Extension: E is a stable-extension in A iff it is a conflict-free set that attacks every argument that does not belong in E. Formally, $\forall a \in A \backslash E, \exists b \in S$ such that bRa.
Grounded Extension: E is the (unique) grounded extension of A iff it is the smallest element (with respect to the inclusion) among the complete extensions of A.

Definition 3. Similar to the attack relation R, one can consider a separate relation S which indicates *Support* (Amgoud et al., 2008). A supporting argument can also be viewed as a part of another argument internal structure. These two options only differ in the AF structure; the reasoning outcome is not influenced. The support relation was introduced in order to better represent realistic knowledge.

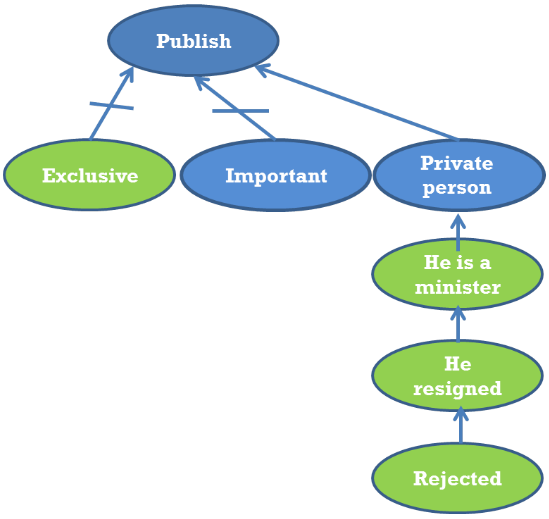


Figure 1: An example of a Bipolar Argumentation Framework; nodes are arguments, arrows indicate attacks and arrows with diagonal lines indicate support.

Let us consider the following example;
Example.
During a discussion between reporters, R_1 and R_2, about the publication of information I concerning person X, the following arguments are presented:
R_1: I is important information, thus we must publish it.
R_2: I concerns the person X, where X is a private person and we cannot publish information about a private person without his consent.
If you were R_1, what would you say next?
A. X is a minister, so X is a public person, not a private person.
B. X has resigned, so X is no longer a minister.
C. His resignation has been refused by the chief of the government.
D. This piece is exclusive to us; If we publish it we can attain a great deal of appreciation from our readers.
See Figure 1 for a graphical representation.

In this example, all mentioned semantics agree on a single (unique) extension which consists of all arguments except "Resigned" (option B) and "Private Person" (R_2's argument). Thus, all arguments except ”Resigned” and ”Private person” should be considered *Justified*, regardless of the choice of semantics.

Argumentation Theory consists of many more ideas and notions, yet the very fundamental ones stated above are the focus of this work.

3 Real Dialogs Experiment

To get a deeper understanding of the relations between people’s behaviour in argumentation and the stated notions, we used real argumentative conversations from Penn Treebank Corpus (1995) (Marcus et al., 1993) of transcribed telephone calls and a large number of chats collected toward this aim. The Penn Treebank Corpus consists of transcribed phone calls on various topics, among them some controversial topics such as ”Should the death penalty be implemented?” and ”Should a trial be decided by a judge or jury?”, with which we chose to begin. We went through all 33 dialogs on ”Capital Punishment” and 31 dialogs on ”Trial by Jury” to identify the arguments used in them and cleared all irrelevant sentences (i.e, greetings, unrelated talk etc.). The shortest deliberation consisted of 3 arguments and the longest one comprised of 15 arguments (a mean of 7). To these dialogs we added another 157 online chats on *”Would you get an influenza vaccination this winter?”* collected from Israeli students, ages ranging from 19 to 32 (mean=24), using a chat interface we implemented. We constructed 3 BAFs, similar to the one in Figure 1, using the arguments extracted from 5 randomly selected conversations. Each conversation which was not selected for the BAF construction was then annotated using the arguments in the BAFs. All in all, we had 64 phone conversations and 157 online chats, totaling 221, all of which are of argumentative nature.

Every conversation provided us with 2 argument sets A_1 and A_2, both subsets of A. We tested every A_i ($i = 1, 2$) such that $|A_i| \geq 3$ in order to avoid almost completely trivial sets.

Participants were not expected to be aware of *all* arguments in the BAF, as they were not presented to them. Thus, in testing the *Admissibility* of A_i and whether A_i is a part of some *Extension*, we examined both the original BAF and the *restricted* BAF induced by $A_1 \cup A_2$. That is, the argumentation framework in which $A = A_1 \cup A_2$ and the attack and support relations are defined over $A_1 \cup A_2 \times A_1 \cup A_2$, denoted as $AF\downarrow_{A_1 \cup A_2}$.

3.1 Results

The first property we tested was *Conflict-Freedom*, which is probably the weakest requirement of a set of arguments. We had anticipated that all A_i would have this property, yet only 78% of the deliberants used a conflict-free set A_i. Namely, that 22% of the deliberants used at least 2 conflicting arguments, i.e, one attacks the other. From a purely logical point of view, the use of conflicting arguments is very grating. Yet, we know that some people try to portray themselves as balanced and unbiased, and as such use contradictory arguments to show that they can consider both ends of the argument and can act as good arbitrators. When we examined *Acceptability*, we tested if every argument $a \in A_i$ is acceptable w.r.t $A_i \setminus \{a\}$. We found that 58% of the deliberants followed this rule. *Admissibility* was tested according to both the original framework and the restricted framework. Merely 28% of the A_is used are considered admissible w.r.t the original framework, while more than 49% qualify when considering the restricted BAF. We can see that people usually do not make the extra effort to ensure that their argument-set is admissible. A possible explanation can be values (norms and morals), as described in (Bench-Capon et al., 2002). Given a set of values, a reasoner may not recognize the attacking arguments as defeating arguments as they advocate a weaker value. As such, the reasoner considers his set admissible. A similar explanation is provided in (Dunne et al., 2011), where a reasoner can assign a small weight to the attacking arguments and as such still consider his set admissible. These explanations can also partially account for the disheartening results in the test of *Extensions*. When examining the original framework, less than 30% of A_is used were a part of some extension, with Preferred, Grounded and Stable performing very similarly (28%, 30%, 25%). When considering the restricted framework, 49%, 50% and 37% of the deliberants used A_is that were part of some extension prescribed by Preferred, Grounded and Stable (respectively) under the restricted BAF. As for *Support*, 27% of the arguments selected were supporting arguments, i.e, arguments which do not attack any other argument in the framework. Although they cannot change the reasoning outcomes, people naturally consider the supporting arguments, which traditionally are not considered ”powerful”.

To strengthen our findings we performed yet another experiment. We tested the notions in a controlled and structured environment, where the participant is aware of all arguments in the framework.

4 Structured Argumentative Scenarios

We collected 6 fictional scenarios, based on known argumentative examples from the literature (Walton, 2005; Liu et al., 2007; Cayrol and Lagasquie-Schiex, 2005; Amgoud et al., 2008; Tang et al., 2012).

Two groups of subjects took part in this study; the first consisted of 64 US citizens, all of whom are workers of Amazon Mechanical Turk, ages ranging from 19 to 69 (mean=38, s.d=13.7) with varying demographics. The second consisted of 78 computer science B.Sc. students from Bar-Ilan University (Israel), ages ranging from 18 to 37 (mean=25, s.d=3.7) with similar demographics.

Each subject was presented with the 6 scenarios. Each scenario was presented in a short textual dialog between 2 participants, similar to the journalists' example above. The subject was instructed to place himself in one of the deliberants' roles, given the partial conversation, and to choose the next argument he would use from the four available arguments. We instructed the subject to consider only the arguments in the dialog and the proposed ones, and refrain from assuming any other information or possible arguments in the dialog's context.

The following example, based on (Liu et al., 2007), was presented to the subjects;

Example.
A couple is discussing whether or not to buy an SUV.
Spouse number 1 (S_1): "We should buy an SUV; it's the right choice for us".
Spouse number 2 (S_2): "But we can't afford an SUV, it's too expensive".
The participant was then asked to put himself in S_1's shoes and choose the next argument to use in the conversation. The options were: A. "Good car loan programs are available from a bank", B. "The interest rates on car loans will be high"', C. "SUVs are very safe, safety is very important to us", D. "There are high taxes on SUVs".
See Figure 2 for a graphical representation of the aforementioned framework.

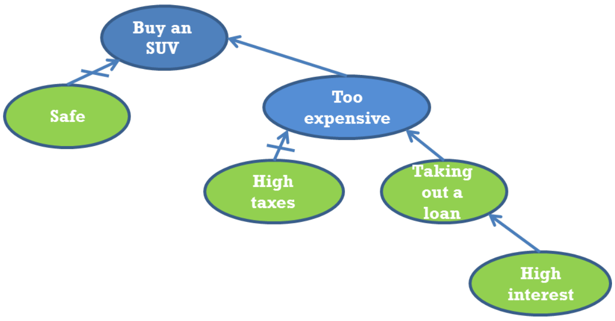


Figure 2: SUV example of BAF

The distribution of selections in the above example was as follows; A.35%, B.24%, C.8%, D. 33%. There is only one (unique) extension in this scenario which includes "High interest" and "high taxes". Especially when considering "Taking out a loan", it should be considered overruled (unjustified/invalid), or at least very weak, as it is attacked by an undisputed argument. As we can see, only slightly over half of the subjects choose an argument from the extension, i.e, a somewhat *Justified* argument.

4.1 Results

The distribution of selections, in all scenarios, suggests that there could be different factors in play, which differ from one subject to another. Thus, there is no decisive answer to what a person would say next. Unfortunately, testing *Conflict Freedom* and *Admissibility* is inapplicable here. None of the subjects was offered an argument that conflicts with its previous one and could not choose more than one argument to construct an admissible set. When examining *Extensions*, all scenarios which were presented to the subject are *Well Founded* (that is to say, there exists no infinite sequence $a_0, a_1, \dots, a_n, \dots$ such that $\forall i.(a_i, a_{i+1}) \in R$). As such, all mentioned semantics coincide - only one extension is Grounded, Stable and Preferred. Of the 6 scenarios, 5 had suggested 2 justified arguments and 2 overruled arguments (arguments which are not part of any extension) to the subject. In these 5 scenarios, 67.3% of the time a justified argument was selected (on average). This result is disappointing since 50% is achieved by randomly selecting arguments. As for *Support*, 49.4% of the arguments selected were supporting arguments, i.e, arguments which do not attack any other argument in the framework. Even more interesting is that 80% of the time people chose (directly or indirectly) an argument supporting their first argument. This phenomenon can be regarded as a *Confirmation Bias*, which is recorded in many fields (Nickerson, 1998). Confirmation bias is a phenomenon wherein people have been shown to actively seek and assign more weight to evidence that confirms their beliefs, and ignore or underweigh evidence that could disconfirm their beliefs. Confirmation Bias can also explain the persistence of discredited beliefs, i.e, why people continue to consider an argument valid/invalid despite its logical argumentative status. Here it is extremely interesting since the subjects only played a role and it was not really their original argument. There is a strong tension between the *Confirmation Bias* and *Extensions*. In some scenarios the subject is given a situation in which he "already used" an overruled argument, and therefore had a problem advocating it by using a supporting argument.

We had anticipated that in finite and simple argumentative frameworks people would naturally choose the "right" arguments, yet we again see that the argumentative principals unsatisfactorily explain people's argumentative selections. This is not a complete surprise, since we have many examples in the literature where people do not adhere to the optimal, monolithic strategies that can be derived analytically (Camerer, 2003).

We have shown here, in two separate experiments, that a similar phenomenon occurs in the context of argumentation - people do not choose "ideal" arguments according to the Argumentation Theory.

5 Relevance

It is well known that human cognition is limited, as seen in many examples in (Faust, 1984) and others. In chess for example, it is common to think that a beginner can consider about 3 moves ahead and a master about 6. If we consider the argumentation process as a *game* (McBurney and Parsons, 2009), a player (an arguer) cannot fully comprehend *all* possible moves (arguments) and their utility (justification status) before selecting a move (argument to use) when the game (framework) is complex. The depth and branching factor limitations of the search algorithms are of course personal. For example, we would expect an educated adult to be able to better consider her arguments than a small child.

Definition 4. Let a,b be arguments in some AF. $Rel : A \rightarrow P(A)$ is a personal relevance function which given argument $a \in A$ (for evaluation) returns a set of arguments $A' \subseteq A$ which are, given the reasoner's cognitive limitations and knowledge, relevant to a. Using Rel, we can distinguish between relevant and irrelevant arguments w.r.t a given argument, yet we gain additional strength in incorporating the reasoner's limitation and biases.

We denote the restriction of AF to arguments relevant to a as $AF{\downarrow_{Rel(a)}} \equiv < A', R' >$ where $A' = Rel(a)$ and $R' = A' \times A' \cap R$.
On $AF{\downarrow_{Rel(a)}}$ one can deploy any semantic of choice.

The simplest way to instantiate the *Rel* is $Rel(\cdot) = A$, meaning that all arguments in the AF are relevant to the given argument. This instantiation is the way the classic frameworks address the reasoner's limitations, simply by saying – there are none. As shown in (Liao and Huang, 2013), it is not necessary to discover the status of all arguments in order to evaluate a specific argument/set of arguments. Thus, considering $Rel(a)$ as the maximal set of *affecting arguments* (arguments in which their status affects the status of a) is another natural way to consider relevance, yet without considering cognitive limitations.

We suggest the following instantiation, which we examined empirically.

Definition 5. Let $D(a, b)$ be a distance function, which given arguments a, b returns the directed distance from argument a to b in AF's graph.

Given a distance measurement D we can define an edge-relevance function as follows:

Definition 6. $Rel_D(a) = \{b | D(b, a) \leq k\}$ where k is a non-negative constant.

Naturally, when setting k to 0, every argument a is considered justified in $AF{\downarrow_{Rel_D(a)}}$ (under any semantics). k can be thought of as a depth limitation for the search algorithm used by the reasoner. Of course, if $k = \infty$, $AF{\downarrow_{Rel_D(a)}} = \{$All affecting arguments on $a\}$.

5.1 Empirical Testing

We used several D functions in our work on predicting arguments given a partial conversation (Rosenfeld and Kraus, 2014). When $k = 0$, as stated above all arguments should be considered justified. Analyzing the free-form dialogs using Grounded semantics with $k = 2$ resulted in 72% of the arguments used being part of some extension, whereas without relevance a little less than 50% was part of some extension.

Relevance provides a way to rationally justify every argument within an AF to some extent. Unlike VAF (Bench-Capon et al., 2002) and WAF (Dunne et al., 2011), which rely on exogenous knowledge about values and weights from the reasoner, relevance can be instantiated without any prior knowledge on the reasoner and still offer a better explanatory analysis of the framework.

6 Conclusions

We presented an empirical study, with over 400 human subjects and 250 annotated dialogs. Our results, based on both free-form human deliberations and structured experiments, show that the fundamental principles of Argumentation Theory cannot explain a large part of the human argumentative behavior. Thus, Argumentation Theory, as it stands, should not be assumed to have descriptive or predicatory qualities when it is implemented with people.

Our relevance notion provides a new way to rationalize arguments without prior knowledge about the reasoner. Relevance, as well as other psychological and social aspects, should be explored to better fit the Argumentation Theory to human behavior. This required step is crucial to the integration of argumentation in different human domains.

References

Leila Amgoud, Claudette Cayrol, Marie-Christine Lagasquie-Schiex, and Pierre Livet. 2008. On bipolarity in argumentation frameworks. *International Journal of Intelligent Systems*, 23(10):1062–1093.

Trevor JM Bench-Capon, Sylvie Doutre, and Paul E Dunne. 2002. Value-based argumentation frameworks. In *Artificial Intelligence*.

Guido Boella, Dov M Gabbay, Leendert WN van der Torre, and Serena Villata. 2010. Support in abstract argumentation. In *COMMA*, pages 111–122.

Colin Camerer. 2003. *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Claudette Cayrol and Marie-Christine Lagasquie-Schiex. 2005. On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and quantitative approaches to reasoning with uncertainty*, pages 378–389. Springer.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Paul E Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge. 2011. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486.

David Faust. 1984. *The limits of scientific reasoning*. U of Minnesota Press.

Mark Klein. 2011. How to harvest collective wisdom on complex problems: An introduction to the mit deliberatorium. *Center for Collective Intelligence working paper*.

Beishui Liao and Huaxin Huang. 2013. Partial semantics of argumentation: basic properties and empirical. *Journal of Logic and Computation*, 23(3):541–562.

Xiaoqing Frank Liu, Samir Raorane, and Ming C Leu. 2007. A web-based intelligent collaborative system for engineering design. In *Collaborative product design and manufacturing methodologies and applications*, pages 37–58. Springer.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Peter McBurney and Simon Parsons. 2009. Dialogue games for agent argumentation. In *Argumentation in artificial intelligence*, pages 261–280. Springer.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175.

Iyad Rahwan, Mohammed I Madakkatel, Jean-François Bonnefon, Ruqiyabi N Awan, and Sherief Abdallah. 2010. Behavioral experiments for assessing the abstract argumentation semantics of reinstatement. *Cognitive Science*, 34(8):1483–1502.

Chris Reed and Glenn Rowe. 2004. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

Ariel Rosenfeld and Sarit Kraus. 2014. Providing arguments in discussions based on the prediction of human argumentative behavior. Unpublished manuscript.

Yuqing Tang, Elizabeth Sklar, and Simon Parsons. 2012. An argumentation engine: Argtrust. In *Ninth International Workshop on Argumentation in Multiagent Systems*.

Douglas N Walton. 2005. *Argumentation methods for artificial intelligence in law*. Springer.

Legal Argumentation Concerning Almost Identical Expressions (AIE) In Statutory Texts

Michał Araszkiewicz
Department of Legal Theory
Jagiellonian University
Bracka 12, 31-005 Kraków, Poland
michal.araszkiewicz@uj.edu.pl

Agata Łopatkiewicz
Institute of Education
Jagiellonian University
Stefana Batorego 12, 31-135 Kraków, Poland
agata.lopatkiewicz@uj.edu.pl

Abstract

This paper deals with the problem of reasoning with synonymic expressions in the domain of statutory law. It is shown that, even in cases of strong lexical synonymy (what is referred to here as 'Almost Identical Expressions'), it is necessary to engage in complicated argumentative structures in order to obtain justified conclusions concerning the mutual interreplaceability of legal terms. This result has implications for the methods adopted in research on the automated analysis of the corpora of legal texts.

1 Introduction

The aim of this paper is to analyse the phenomenon of legal argumentation that makes use of almost identical expressions extracted from statutory texts. It is often the case that a lawmaker makes use of expressions A and B in statutory rules, where A and B have such a similar meaning that they would be presumably treated as strictly synonymic by a native speaker of a language. Therefore, a native speaker of a language would be inclined to assign identical consequences to states of affairs designated by expressions A and B. The similarity between the mentioned expressions seems to constitute grounds for the application of arguments based on plain meaning and analogy. However, there exist certain rationality assumptions behind the making of laws leading to the conclusion that, if two expressions are not strictly identical, they should be treated as different by the addressee of the statutory regulation. These two argumentative stances point out inconsistent solutions and, therefore, may cause divergent opinions concerning the rights and obligations of addressees of the law. Therefore, the investigation of this phenomenon is important for the sake of legal policy matters. However, the analysis of argumentation encompassing almost identical expressions is also of crucial importance for the development of legal knowledge-based systems. Such systems should take into account that the relation of synonymy between linguistic expressions should be treated more carefully than in less formal contexts of discourse in order to avoid oversimplifications and potentially wrong suggestions to the user.

2 The Notion of Synonymy

Synonymy has always been considered one of the most basic semantic relations between linguistic expressions (for instance, Murphy 2003). The relation is also useful in contemporary research on Natural Language Processing[1] (see also Hirst 2004). Although synonymy is generally accounted for as similarity of meaning, in specialised contexts, this account is insufficient because of notorious problems concerning the understanding of notions regarding 'similarity' and 'meaning'.

Due to these problems, the relation of synonymy has been a subject of interest for linguistic philosophers. A classic contribution to the debate is a paper by Goodman (1949), in which he argues than no two non-identical words can have the same meaning. Instead of the theory of synonymy

[1] The WordNet project involves the notion of synsets: sets of cognitive synonyms that represent certain concepts. http://wordnet.princeton.edu, accessed on September 24, 2014.

as ‘sameness of meaning’, he advocated a theory of ‘likeness of meaning’, according to which two names of predicates may be treated as synonyms if their meaning is similar enough to warrant the thesis of their ‘sameness’, or mutual interreplaceability, in certain contexts of discourse. The criteria used here may vary from one context to another (Goodman 1949, 7).

As with any philosophical thesis, Goodman’s proposal remained controversial in the literature of the subject (for a relatively recent contribution, see Heydrich 1993). The philosophical discussion of synonymy is deeply connected with such topics as analyticity and necessity. For obvious reasons, we cannot investigate these extremely complicated issues here (see Soames 2003). However, we claim that Goodman’s thesis captures an important insight into the pragmatic dimensions of synonymy: two linguistic expressions, A and B, may be seen as mutually interreplaceable in the context of discourse C_1 while they could be assessed as different (and, therefore, not mutual substitutions) in the context of discourse C_2. The relation of synonymy depends on the context of assessment regarding this relation.

Philosophical controversies notwithstanding, the notion of synonymy is widely used in lexicography, and the existence of thesauri and dictionaries of synonyms is obvious evidence for the usefulness of this relation for language users. The words ‘synonym’ and ‘synonymy’ are actually used by the speakers of languages, and the corpora of conversational material are investigated in order to establish their actual understanding of the term. Murphy (2013) notes the following accounts of the word ‘synonym’ as found in the analysed corpora:

1) synonymy as ‘sameness’ or ‘near sameness’ of meaning,

2) synonymy as the possibility of substituting one word for another,

3) synonymy as the co-extensional character of two scientific names (in biology).

There are more specific understandings of the word ‘synonym’ in computer science (Murphy 2013, 281), but they are not relevant to the discussion of the present paper. Interestingly, the relation of synonymy is also found in translational contexts: the words that are mutual translations in different languages are also seen as synonyms (Murphy 2013, 282).

It is easy to note that the use of the term ‘synonymy’ in descriptive lexicography tends to avoid the discussion of philosophical problems of this linguistic phenomenon. Generally, the people interested in finding synonyms to certain words are interested in substitutability of these words without changing their meaning (as regards both denotative, connotative and social meaning, Murphy 2013, 302). These empirical findings are compatible with Goodman’s thesis referred to above.

3 Almost Identical Expressions (AIE) in Statutory Language

The texts of statutes consist of linguistic expressions. Generally speaking, a lawmaker intends to indicate certain states of affairs and to assign legal consequences to them. The lawmaker indicates these states of affairs by means of linguistic expressions. The language of law shares many features with plain language, such as indeterminacy and vagueness (Bix, Endicott); however, although it is often presumed that statutory texts should be understood with regard to the ‘plain, natural meaning’ (Interpreting Statutes), often special, legal meaning should be ascribed to the used terms (for a recent elaboration of this subject, see Araszkiewicz 2014).

It is often the case that the lawmaker chooses similar, yet not identical, terms to refer to certain states of affairs that are assigned to legal consequences. In such contexts, there is a situation of doubt whether the lawmaker intended to refer to the same, or to different (sets of) states of affairs. The pragmatic context of interpreting such statutory language expressions is set out by the adversarial character of legal proceedings. Each party is interested in persuading the judge to ascribe such meaning to a statutory term that leads to the legal consequences desired by this party. Consequently, a party to the dispute may be interested in treating similar expressions alike with respect to their legal result; another party may be interested in strict differentiation between the meanings of slightly different expressions.

There are different approaches to the indicated problem in different jurisdictions. Sometimes, even the lawmaker gives explicit guidelines to show how similar expressions should be interpreted. For instance, the Australian *Acts Interpretation Act 1901*[2] contains a provision, 15AC, according to which, ‘when an Act has expressed an

[2] http://www.comlaw.gov.au/Series/C1901A00002, last accessed on September 24, 2014.

idea in a particular form of words and a later Act appears to have expressed the same idea in a different form of words for the purpose of using a clearer style, the ideas shall not be taken to be different merely because different forms of words were used.'[3] However, typically, the lawmaker will be reluctant to give the addressees of legal texts such explicit suggestions. Thus, the dilemma concerning the ascription of identical or non-identical meaning to slightly different linguistic expressions will remain an open issue.

This dilemma is particularly visible with regard to the class of expressions we refer to as Almost Identical Expressions (AIE). By definition, the linguistic expressions E_1 and E_2 in language L belong to the set of AIE if and only if:

1) they stem from the same lexical root,

2) they are not identical from the syntactic point of view,

3) they would be considered as natural mutual substitutions by a competent native speaker of language L (in a relevant context of discourse C).

The point 3) is the most important one: AIE create a strong inclination in the native speakers of the language to treat them interchangeably in the relevant context of discourse. But point 2) creates the possibility for the construction of arguments to the contrary. The next two sections are devoted to the discussion of an exemplary legal question encompassing the use of AIE.

4 The Legal Research Problem

The legal research problem that focused our attention on the argumentation concerning AIE is as follows: what are the legal consequences of non-compliance of subjects of law with the requirement of concluding contracts and making other statements in writing? The Polish Civil Code (Act of 23 April 1964, consolidated version: Journal of Laws 2014.121, hereafter referred to as the PCC) contains approximately 100 instances of expressions lexically cognate with the word 'writing', most of which are parts of provisions specifying requirements of the form of contracts and other statements. There are three types of these expressions, forming a set of AIE:

1) 'in written form' (PL: *w formie pisemnej*),

2) 'in writing' (PL: *na piśmie*) and

3) 'stated in writing' (PL: *stwierdzone pismem*).

All these expressions would be treated as mutual substitutions in the majority of contexts of discourse by a native speaker of the Polish language; interestingly, lawyers are also often inclined to see these three expressions as interreplaceable ones. However, this contention does not lead to any immediate answers concerning both the content of requirements that are expressed by the expressions listed above and the consequences of non-compliance with these requirements.

For the sake of clarity regarding the following investigations, it is necessary to delineate the legal context concerning the 'written form' requirement in Polish civil law. The basic rules dealing with this issue are in art. 73 § 1 of the PCC:

If the law stipulates that a legal act be made in written form, an act made without observing the stipulated form is invalid only if the law provides for a nullity clause.

and article 74 § 1 of the PCC:

The stipulation of written form without a nullity clause leads, if the stipulated form is not observed in litigation, to witness evidence or evidence in the form of declarations of the parties concerning the performance of the act being inadmissible.[4]

The legal consequences stemming from the quoted rules are straightforward. If a given act should be made in written form and the law prescribes for the pain of nullity, in the case of failure to fulfil the requirement, the act is not valid. Conversely, if the pain of nullity is not mentioned in the law (or in the statement of the parties), the act cannot be invalid in the case of non-compliance with the written form requirement. This consequence is uncontroversial. The legal results provided by the latter of the quoted provisions are more nuanced: if a written form is required for an act and it is not complied with, the act is still valid. However, certain types of evidence are not admissible to prove that such act has taken place. Let us refer to this legal consequence as the consequence of evidentiary difficulties. In the following analyses, we will focus on this latter legal consequence only. The consequence of invalidity is an easy topic from the point of view of argument

[3] We are grateful to Graeme Hirst for pointing out this interesting regulation during the BiCi seminar on Frontiers and Connections between Argumentation Theory and Natural Language Processing in Bertinoro (July 20-24th, 2014).

[4] The translations of the provisions are taken from the commercial Legalis system provided by the C.H. Beck publishing house, with certain modifications by the authors.

mining and natural language processing of statutory texts: an act is invalid only if there is an explicit clause providing for such consequence. In the absence of such a clause, the consequence of the failure to meet the requirement of a 'written form' should lead to evidentiary difficulties. This contention is, again, uncontroversial, with regard to the requirement of 'written form' as indicated in the latter of the quoted provisions. The question is, first, whether the requirements provided by the law should be understood identically where the law speaks about 'written form', 'in writing' and 'stated in writing', respectively. Second, what are the legal consequences of the failure to meet the requirements referred to as 'in writing' and 'stated in writing'?

Let us present the existing controversy in a more explicit manner. Let us assume that a legal provision of the PCC has the following scheme:

(X) Legal act X should be performed in written form.

The quoted art. 74 § 1 of the PCC enables us to derive the following conclusion from (X):

(X-con) If the legal act X is not performed in written form, then the consequence of evidentiary difficulties shall apply as regards the legal act X.

Let us recall the expression 'in written form' forms an AIE set with the expressions 'in writing' and 'stated in writing'. This enables us to present the two following schemes of provisions (actually often instantiated in the PCC):

(Y) Legal act Y should be performed in writing.
(Z) Legal act Z should be stated in writing.

The precise formulation of the legal research questions goes as follows: (Q1) Is the meaning of X, Y and Z identical? (Q2) Is it the case that Y and Z lead to the formulation of Y-con and Z-con rules analogous to the X-con rule?

In order to establish valuable answers to these questions, a corpus of judgments (>30 cases) and legal doctrinal works (5 sources) were examined. The results are reported in the following section.

5 Analysis of Actual Arguments as Found in the Corpora

The analysis of the existing material led to the following answers to the questions outlined above: Q1: undecided (there are authoritative sources that tend to give positive and negative answers to the question) and Q2: positive (but the interpretation of the answer depends on the chosen answer to Q1).

Theoretically, several argumentation schemes can play their role is justifying different answers to Q1. For instance, the argument from plain natural meaning would support a positive answer to Q1. The argument would run as follows.

Premise 1. Statutory terms should be interpreted in accordance with their plain natural meaning.

Premise 2. According to plain natural meaning, the expressions 'in written form', 'in writing' and 'stated in writing' should be treated as (strict) synonyms.

Conclusion. The meaning of X, Y and Z is identical (positive answer to Q1).

Let us note that this argument could be further backed by analogous reasoning: Premise 2 could be refined to relax the assumption of strict synonymy in favour of the claim that, in the context if legal discourse, these AIE should be treated as carrying the same meaning (because the differences between them could be reasonably ignored).

Actually, a refined version of this argument scheme was used by one of the most influential legal scholars in Poland, Zbigniew Radwański (Radwański 2002, 134). The remaining analysed doctrinal sources also adopt this view. Let us reconstruct his argument:

Premise 1. If differences between the terms used by the legislator are a matter of style only, then the terms should be treated as (strict) synonyms.

Premise 2. 'In written form', 'in writing' and 'stated in writing' are terms that differ with respect to style only.

Conclusion. The meanings of X, Y and Z are identical (positive answer to Q1).

Let us note that a positive answer to Q1 implies, as a matter of logic, a positive answer to Q2.

However, it is also possible to formulate arguments to the contrary. According to the rationality postulates concerning legislative process, if the legislator intends to indicate the same state of affairs in different parts of regulation, he uses one and the same term. If he uses (even slightly) different terms instead, this means that his intent was to designate different states of affairs. This argumentative pattern is often referred to as the prohibition of synonymic interpretation:

Premise 1. The terms used in the statute should not be assigned with an identical meaning unless they are syntactically identical.

Premise 2. 'In written form', 'in writing' and 'stated in writing' are not syntactically identical.

Conclusion. The meanings of X, Y and Z are not identical (negative answer to Q1).

Note that a negative answer to Q1 does not logically imply a negative answer to Q2. A negative answer to Q1 consists only of holding that the 'written form' requirement is something other than 'in writing' or 'stated in writing'. Let us add in this connection that, uncontroversially, the 'written form' requirement is satisfied only if a statement is manually[5] undersigned by a person.

Consequently, the controversy between a positive and negative answer to Q1 boils down to the set of sufficient conditions to satisfy a given requirement. Undoubtedly, if a legal provision is based on the scheme (X) presented above, the requirement is not met unless the statement encompassing the content of legal act X is manually undersigned by a person. The question (Q3) is whether this sufficient condition should also be met for the satisfaction of requirements formulated in schemes (Y) and (Z). As a matter of course, a positive answer to Q1 implies a positive answer to Q3, while a negative answer to Q1 implies a negative answer to Q3.

Interestingly, the judicial opinions reviewed in the research tend to adopt a rather negative answer to Q1 (unlike doctrinal sources quoted above). This may be caused by the fact that judicial authorities are closer to legal practice and they do not intend to impose unnecessary burdens on the addressees of the provisions. This is particularly visible in the context of the interpretation of the following provision (art. 514 of the PCC) related to the institution of a claim assignment:

If a claim is stated in writing, a contractual stipulation that assignment cannot be made without the debtor's consent is effective towards the assignee only when the document contains a mention of the stipulation unless the assignee knew of the stipulation at the time of assignment.

The courts tend to adopt a negative answer to Q1 in this context. For instance, in the Resolution of 6 July 2005 (III CZP 40/05), the Supreme Court stated that:

Stating of the claim in writing in the understanding of the art. 514 of the PCC is satisfied also in case the creditor issues a document (e.g. an invoice) that confirms the performance of an obligation and the debtor accepts the document.

Thus, the Supreme Court ruled that the requirements for satisfying the 'stated in writing' requirement are less severe than 'in written form'. The satisfaction of the latter implies the satisfaction of the former, but not the other way around.

The reconstruction of an argument justifying this conclusion from the wording of the Resolution is a non-trivial task due to the highly complex structure of the analysed sentences. The proposal of the argument's structuration would be as follows:

Premise 1. There is no need to delimit the types of documents that may be used for the identification and confirmation of legal facts (wrt art. 514 of the PCC).

Premise 2. Adoption of a positive answer to Q1 would amount to the undue delimitation of the types of documents used for the identification and confirmation of legal facts.

Conclusion. Q1 should be answered negatively.

The argument formulated by the Supreme Court is enthymematic, especially with regard to the premise 1: the court seems to assume that the possibility of identification and confirmation of legal facts is a worthwhile value, which should be realised at the expense of more firm protection of debtors. This stems from the contention of the Supreme Court, according to which an invoice issued by the creditor but not accepted by the debtor would be insufficient to fulfil the condition of 'being stated in writing', because the protection of the debtor would be too weak if a broader interpretation were accepted. This value judgment can be reconstructed from the text only by a person who possesses at least basic legal training. However, this does alter the conclusion that the Su-

[5] For the sake of brevity, we leave the problems of electronic signatures aside.

preme Court rejects the thesis concerning the mutual interreplaceability of expressions ‘in written form’ and ‘stated in writing’.

It is worth emphasising that the same interpretation has been accepted by the courts with regard to the interpretation of art. 511 of the PCC:

If a claim is stated in writing, its assignment should also be stated in writing.

For instance, in the Judgment of the Appellate Court in Katowice of 8 March 2005, I ACa 1516/04, the negative answer to the Q1 was advanced on the basis of a literal reading of the statute: If the legislator speaks about ‘stating in writing’, this means that he does not intend to introduce a requirement of ‘written form’, simply because these expressions are not identical.

Let us note that the answer to the Q2 may remain positive even if Q1 is answered negatively. However, different situations will have to be considered as regards the satisfaction of ‘written form’ and ‘stated in writing’ requirements.

6 Conclusion

The investigations of this paper lead to the formulation of the following conclusions. The peculiarities of statutory text make the NLP analyses related to this material very difficult. In particular, such ubiquitous semantic relations as synonymy have to be dealt with in a non-standard manner as regards the statutory text. Even in the case of AIE that seem to be very close, or even perfect synonyms in other contexts of discourse, establishing the interreplaceability relations between terms is a problematic issue. Reaching a justified conclusion as regards this relation in legal contexts is a complicated process, also due to the fact that lawyers disagree about the existence or non-existence of synonymy relations between the analysed terms. This process involves the reconstruction of legal arguments used in different authoritative sources. The reconstruction is not an easy task due to the complicated structure of sentences present in judicial opinions and doctrinal theories as well as posing hypotheses about enthymematic premises. The latter activity involves a vast amount of professional legal knowledge. Therefore, the corpora of legal texts should be annotated by legal professionals (or at least legal students) in the process of argumentation mining rather than by laymen in order to avoid misunderstandings generated by a lack of legal knowledge.

Even in the case of AIE, which seem to be (near) synonyms on purely linguistic grounds, as it was shown, the discussion of their interreplaceability involves the use of not only linguistic arguments, but also teleological arguments possessing a complicated structure. The obtained conclusions are contextual and perhaps defeasible, as is often the case in the context of legal discourse.

The most important conclusion stemming from the investigations above is that, in the context of an NLP analysis of the corpora of legal texts (aiming at the creation of intelligent databases of legal knowledge), one should be very cautious as regards the use of any databases of synonyms. Moreover, the corpora of statutory texts should not be analysed apart from the legal doctrine and (most importantly) databases of legal cases. These sources should serve for the reconstruction of arguments used to determine the meaning and scope of statutory expressions.

References

Graeme Hirst. 2004. Ontology and the Lexicon. In S. Staub, R. Studer (eds.). *Handbook on Ontologies*, 269-292. Springer, Berlin Heidelberg.

M. Lynne Murphy. 2013. What we talk about when we talk about synonyms (and what it can tell us about thesauruses). *International Journal of Lexicography* 26(3): 279-304.

M. Lynne Murphy. 2003. *Semantic Relations and the Lexicon*. Cambridge University Press, Cambridge, UK.

Michał Araszkiewicz. 2014. Legal Interpretation: Intensional and Extensional Dimensions of Statutory Terms. In E. Schweighofer, M. Handstanger, H. Hoffmann, F. Kummer, E. Primosch, G. Schefbeck, G. Withalm (eds.). *Zeichen und Zauber des Rechts. Festschrift für Friedrich Lachmayer*, 496-492. Weblaw, Switzerland.

Nelson Goodman. 1949. On Likeness of Meaning. *Analysis* 10(1): 1-7.

Wolfgang Heydrich. 1993. A Reconception of Meaning. *Synthese* 95(1): 77-94.

Scott Soames. 2003. *Philosophical Analysis in the 20th Century. Vol. 2: The Age of Meaning*. Princeton University Press: Princeton and Oxford.

Zbigniew Radwański (ed.). 2002. System prawa prywatnego. Tom 2. Prawo cywilne – część ogólna. /The System of Private Law. Vol. 2. Civil Law – the General Part/. C.H. Beck, Instytut Nauk Prawnych PAN (The Institute of Legal Sciences of the Polish Academy of Sciences), Warszawa, Poland.

Extracting and Understanding Arguments about Motives from Stories

Floris Bex
Department of Information and Computing Systems
Utrecht University
The Netherlands
f.j.bex@uu.nl

Trevor Bench-Capon
Department of Computer Science
University of Liverpool
United Kingdom
tbc@liverpool.ac.uk

Abstract

In this paper, we discuss how Value-based Argumentation can be used as a tool in human and computer *story understanding*, especially where understanding the story requires understanding of the motives of its characters. It is shown how arguments about motives can be extracted from stories, and how dialogues about these arguments can aid in story understanding.

1 Introduction

In this paper, a short version of which was published as (Bex and Bench-Capon, 2014), we discuss the important connections between narratives, or stories, and argumentation. We often persuade not by imparting facts and rules, but by providing an interesting narrative, particularly when trying to convince others to adopt particular values and attitudes. Presentation of an argument as a story engages our natural reaction to a story, to attempt to understand it. Thus, the story form fosters engagement, encouraging the right choices by appealing to common values rather than by imposing a rule that is to be followed.

A central concept in the research on story understanding is that of *scripts* (Schank and Abelson, 1977), coherent scenarios about common situations such as visiting a restaurant. Despite the apparent failure of scripts to deliver the promised advances in computational linguistics, they still play an important part in computational and cognitive approaches of story understanding (Mueller, 2004), and they are widely applied in, for example, case-based reasoning (Gentner and Forbus, 2011), scenario-based evidence analysis (Vlek et al., 2013) and narrative generation (Gervas et al., 2005).

In our opinion, purely script-based approaches to story interpretation are not suited to understanding persuasive stories concerning values, such as parables. Scripts represent the way in which we expect typical situations to play out: the more a story adheres to a familiar script, the more plausible a story is considered to be. However, many memorable stories such as parables depend on a twist in the story, something which is out of the ordinary and which challenges conventional attitudes (Govier and Ayers, 2012). For example, noone expects a father to organise a feast for a son who has spent all of his money on wild living (*The Prodigal Son*[1]). Furthermore, the most interesting stories are often those with conflicting attitudes (Wilensky, 1982). For example, in the *Prodigal Son*, the son's older brother wants to turn away his sibling: why welcome a sinner? The father, however, forgives and welcomes his son. In models based on scripts, in which stories are rendered only as causal sequences, these conflicts between characters' values remain largely implicit and unexplained.

For a computational model of story understanding, we need to add a more fine-grained psychological dimension to the causal narrative, in which conflicts between characters' attitudes and challenges to common attitudes can be modelled. This gives us an internal perspective that allows us to represent the deliberations of the characters involved, which allows for a much more subtle analysis of character motive and attitude than we can perform with the external causal perspective. This in turn allows us to show how the relevant stories can influence the audience's attitudes or, in other words, how these stories can persuade an audience to adopt a different attitude.

Recently, we have proposed a model for story understanding (Bex et al., 2014a)(Bex and Bench-Capon, 2014), which draws from value-based practical reasoning (Atkinson and Bench-Capon, 2007). Stories can be represented as (causal)

[1] Luke 15:11-32. We use the World English Bible translation available at http://www.ebible.org/

state transition diagrams, where the transitions represent possible actions by the characters in the story. Character motives are represented by indicating which values are promoted or demoted by the actions in the story. We can then extract practical reasoning arguments of the form *I should perform Action because it promotes Value* and *I should not perform Action because it demotes Value* from the diagram. If we also have separate arguments denoting the characters' attitudes (value orderings), we can construct an Extended Argumentation Framework (EAF) with values (Modgil, 2009), a set of (possibly conflicting) arguments representing character choices and attitudes. Given an EAF, we can then infer attitudes given the choices made in the story. In section 4.1 we show how a particular story interpreted by means of an EAF can be used as an argument in a particular dialogical context, using (Modgil and Bench-Capon, 2008)'s extended TPI-protocol for argumentative dialogue to argue for a change in value preferences in a dialogical setting.

2 Motivating example: The Good Samaritan

Stories can be a powerful vehicle of persuasion. A story does not persuade by imparting explicit rules, but by exposing a coherent narrative aimed at changing or reinforcing attitudes, so that the stories exemplify various group cultural norms. Many folktales are of this type, as are parables, both secular and biblical. As an example of a well-known parable, we will consider *The Good Samaritan*. Since we will be discussing this parable throughout the paper, we will quote it in full. The context is established in Luke 10:25-27:

> Behold, a certain lawyer stood up and tested him, saying, "Teacher, what shall I do to inherit eternal life?"
>
> He said to him, "What is written in the law? How do you read it?"
>
> He answered, "You shall love the Lord your God with all your heart, with all your soul, with all your strength, with all your mind, [Deuteronomy 6:5]; and your neighbour as yourself [Leviticus 19:18]."
>
> He said to him, "You have answered correctly. Do this, and you will live."
>
> But he, desiring to justify himself, asked Jesus, "Who is my neighbour?"

Thus the lawyer asks two questions. The first, "what shall I do to inherit eternal life?", receives an answer justified by scriptural authority. But the second, "Who is my neighbour?", is met simply by a story.

> Jesus answered, "A certain man was going down from Jerusalem to Jericho, and he fell among robbers, who both stripped him and beat him, and departed, leaving him half dead. By chance a certain priest was going down that way. When he saw him, he passed by on the other side. In the same way a Levite also, when he came to the place, and saw him, passed by on the other side. But a certain Samaritan, as he travelled, came where he was. When he saw him, he was moved with compassion, came to him, and bound up his wounds, pouring on oil and wine. He set him on his own animal, and brought him to an inn, and took care of him. On the next day, when he departed, he took out two denarii, and gave them to the host, and said to him, 'Take care of him. Whatever you spend beyond that, I will repay you when I return.' Now which of these three do you think seemed to be a neighbour to him who fell among the robbers?"
>
> He said, "He who showed mercy on him."
>
> Then Jesus said to him, "Go and do likewise."

This provides a very clear example of a story being used as an argument to justify a particular answer to a question, "Who is my neighbour?". However, it is not meant as a theoretical argument: the aim is not that the lawyer should believe that the Samaritan is his neighbour (nor, since the one in the story is a fictional character, that all Samaritans are his neighbour). Nor is the lawyer intended to set out to assist wounded travellers on the road from Jerusalem to Jericho. Unlike practical reasoning proper, there is no specific situation, with a specific choice of actions to resolve. Rather the argument is intended to convince the lawyer (and ultimately of course the reader) to *become* a different person, the sort of person who will enjoy eternal life.

So how exactly does the story convince its audience to change their ways? Govier and Ayers (Govier and Ayers, 2012) have recently explored this question in detail. They specifically address the relation between parables and argument using the *Good Samaritan* as one of their examples. They reconstruct the *Good Samaritan* as the following argument (italicised statements are said in (Govier and Ayers, 2012) to be implicit):

1. *If supposedly holy people (the priest and the Levite) were to ignore an unknown and needy person on a road, they would not treat that person as a neighbour.*

2. If a person who was of no special status and did not know an unknown and needy person on a road were to treat him with mercy and kindness, that person would treat the needy person as a neighbour.
   So

3. What matters about being a neighbour is not one's status or one's prior knowledge of a person.

4. What matters about being a neighbour is treating another with mercy and kindness when that person is needy and one encounters him.

5. *It is good to treat a needy stranger as a neighbour if one encounters him.*
   Therefore

6. *One should treat other people, when they are in need and one encounters them, as one's neighbours with mercy and kindness.*

Statements 1 and 2, which both can be said to follow from the story in some way[2], lead to conclusions 3 and 4. These two conclusions together with the value judgement contained in 5 then lead to the final conclusion 6. The addition of 5 and 6 is, in our opinion, somewhat contentious because it transforms the argument into an argument with a normative conclusion, advocating particular behaviour. This is perhaps justified by the comment 'Go and do likewise' made by Jesus, since this shows that the intention in telling the parable is to affect future actions. However, we would contend that the intention of the parable should not be of the form *in certain situations you should do this* - a norm, but rather an invitation to adopt different attitudes, to be like the Samaritan and recognise that duties between people arise from their common humanity rather than any social or religious ties (statements 3 and 4). To enable a story to have this effect we need a detailed account of the reasoning of the Samaritan, the Priest and the Levite, since otherwise we cannot articulate the differences in attitude between the three characters, and so cannot identify the attitudes we are being urged to abandon and adopt.

3 Understanding stories using value-based argumentation

The computational model for story understanding we propose is based on (Atkinson and Bench-Capon, 2007)'s framework for value-based practical reasoning. We previously used this model to capture abductive reasoning in which stories served as explanations for particular evidence (Bex et al., 2009). The model contains three main elements: (i) *Action-Based Alternating Transition Systems with Values* (AATS+V) for encapsulating stories; (ii) arguments based on the *Practical Reasoning Argumentation Scheme* (PRAS), to generate arguments concerning the individual choices a story character can make; and (iii) Value-based Argumentation Frameworks (VAF), representing the set of arguments and counterarguments a story character uses to make his individual choices on the basis of his preferences and attitudes. Because we want to be able to explicitly reason about characters' value orderings, we use (Modgil, 2009)'s Extended Argumentation Frameworks (EAF) instead of the original VAFs. Below, we will discuss each of these elements by means of our example.

3.1 Stories as AATS+V

Structuralist accounts of narrative argue that actions that represent transitions between states are the basic building blocks of stories. It is for this reason that we choose the mechanism of Action-based Alternating Transition Systems with Values (AATS+V) as our basic formalization method for stories. An AATS consists of a set of states and transitions between them, with the transitions labelled with joint actions, that is actions compris-

[2] It is unclear why (Govier and Ayers, 2012) consider 1 to be implicit and 2 not.

ing an action of each of the agents concerned. In an AATS+V, the transitions are labelled with the values that motivate the characters in the story. A basic version of the parable of the Good Samaritan can be rendered as the AATS+V in Figure 1.

At the beginning of the story q_0, the condition of the traveller is *wounded*. In q_4, the traveller's wounds have been bandaged and he is in a stable condition. In addition to the actions taken by the characters in the story (j_1, j_3, j_6), we have also included the hypothetical actions the characters could have performed: for example, the Priest could also have helped the traveller (j_2). Action choice in parables is often more or less binary ($help$ or $\neg help$, $accept$ or $\neg accept$ in the *Prodigal Son*), so modelling these extra actions does not require much extra information besides the original story text. The values that are promoted by each action are included in the AATS+V: Religious Duty (+RD), Religious Law (+RL), National Solidarity (+NS), Racial Solidarity (+RS), Compassion (+C), Prudence (+P), Convenience (+Cv) and Revenge (+R). Adding the values requires more background knowledge. For example, we need to know that the traveller and the Levite were of the same race, and that Samaritans were a common enemy for the Jewish people. Nowadays, this background information can be gained from Biblical texts, or from the many varied accounts on how parables should be interpreted, but it would have been well-known to the original audience. The values in figure 1 are a selection that the authors have heard from a variety of sources over the years.

3.2 Arguments based on the story

The idea of arguments based on stories is that we look for arguments that instantiate the Practical Reasoning Argumentation Scheme (PRAS). Such arguments are of the following form.

1. In the current circumstances R
2. We should perform action A
3. Which will result in new circumstances S
4. Which will promote some value V

Now, given an AATS+V, we can construct these arguments for the different characters. The basic idea expressed in (Atkinson and Bench-Capon, 2007) is that the AATS+V serves as a formal grounding for arguments that instantiate the Practical Reasoning Argumentation Scheme (PRAS), as follows (where the line numbers in the above PRAS scheme correspond to the line numbers in the formal rendering below).

1. In the initial state $q_0 = q_x \in Q$
2. Agent $i \in Ag$ should participate in joint action $j_n \in J_{Ag}$, where $j_n^i = \alpha_i$
3. Such that $\tau(q_x j_n)$ is q_y
4. Such that for some $v_u \in Av_i$, $\delta(q_x, q_y, v_u)$ is $+$

Here, Q is a finite, non-empty set of *states*, $Ag = \{1, \ldots, n\}$ is a finite, non-empty set of *agents*, α_i defines the set of states from which action α may be executed by agent i, τ is a partial *system transition function*, which defines the state $\tau(q, j)$ that would result from the performance of action j in state q, Av_i is a finite, non-empty set of values for agent i and δ is a *valuation function* which defines the status (promoted (+), demoted (–) or neutral (=)) of a value ascribed to the transition between two states.

Given this mapping of PRAS on an AATS+V, we can generate the arguments from the AATS+V, noting that arguments for different actions attack each other because the actions are mutually exclusive, *i.e.*, one cannot help and not help someone at the same time. First, there are the two arguments that might apply to the priest.

- A_1: I should help the man because I have a religious duty to do so. This will promote Religious Duty (+RD)
- A_2: I should not help the man because I risk uncleanliness through contact with his blood. This will promote Religious Law (+RL).

The following apply to the Levite.

- A_3: I should help the man because he is a fellow countryman. This will promote National Solidarity (+NS).
- A_4: I should help the man because he is of my race. This will promote Racial Solidarity (+RS).

None of the above arguments apply to the Samaritan. The following arguments apply to all three characters.

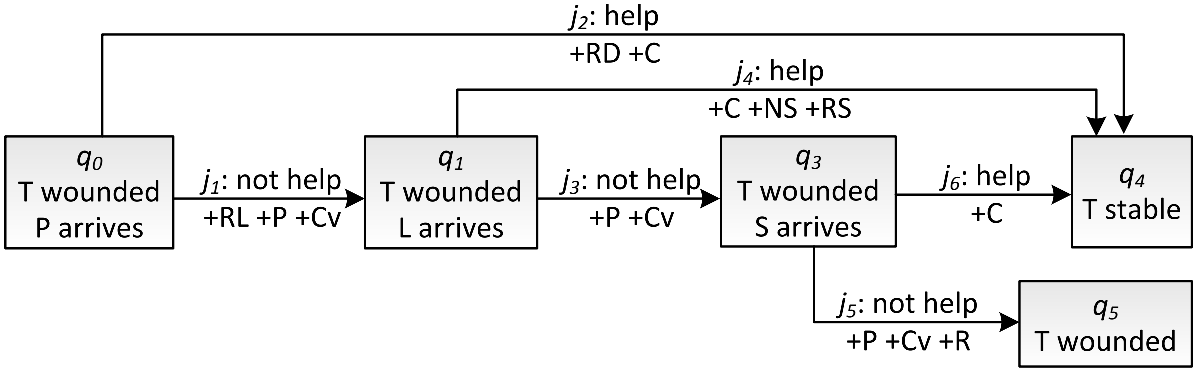


Figure 1: AATS+V for the Good Samaritan

- A_5: I should help the man because he is a fellow human being. This will promote Compassion (+C).
- A_6: I should not help the man because it may be trap and I may be robbed. This will promote Prudence (+P).
- A_7: I should not help the man because it will interrupt my journey. This will promote Convenience (+Cv).

Finally there is an argument that applies only to the Samaritan:

- A_8: I should not help this man, because his people have quarrelled with mine. This will promote Revenge (+R).

All of the arguments A_1-A_4 relate to duties of one sort or another, arising from religious law or duty, or one form or another of social relationship (nation, race). A_5-A_8 all arise from natural human instincts, unconnected with any social institution.

3.3 Constructing an Argumentation Framework

From these arguments, we can construct a Value-based Argumentation Framework (VAF). A VAF is based on (Dung, 1995)'s standard Argumentation Frameworks. An Argumentation Framework $AF = (Args, \mathcal{R})$, where $Args$ is a set of arguments, and $\mathcal{R} \subseteq (Args \times Args)$ is a binary attack relation between pair of arguments. The attack relations between arguments A_1-A_8 are straightforward: arguments concluding *help* attack and are attacked by those concluding *do not help*. A VAF also contains a set of values, and a mapping that associates a value with each argument. Furthermore, a VAF has associated audiences, each of which represents a total ordering of these values.

The purpose of building a VAF is to find a subset of the arguments which is at once conflict free (i.e. no two arguments in the subset attack one another), and collectively able to defend itself (i.e. any attacker of an argument in the subset is itself attacked by an argument in the subset). The maximal such subset is called a preferred extension, and represents a maximal consistent position given the arguments presented. The key feature of VAFs is that they allow a distinction to be made between successful attacks (defeats) and unsuccessful attacks, on the basis of the values associated with the arguments: attacks succeed only if the value associated with the attacking argument is ranked by the audience as equal to, or higher than, the argument it attacks. The VAF thus accounts for elements of subjectivity in that the arguments that are acceptable are dependent upon the audience's ranking of the values involved in the scenario.

We now attempt to explain the actions of the three characters by considering different value orderings, different audiences. Suppose that the Priest puts religion before all else (*i.e.*, Religious Duty and Religious Law are preferred to Convenience, Compassion and Prudence). He then has a conflict between A_1, which argues he should help to promote RD, and A_2, which argues he should not help to promote RL. In the story, he chooses to observe of the law, which applies specifically to himself because of his special role, over the vaguer practical obligation to serve others. This ranking of strict observance of the law over more human concerns is criticised elsewhere in the Gospels, e.g. Mark 2:27 (Then Jesus said to them, "The

Sabbath was made to meet the needs of people, and not people to meet the requirements of the Sabbath.").

The Levite must be supposed to act on either A_6 or A_7, overriding the specific duties of A_3 and A_4 as well as A_5. But because we can assume to have a type of a morally respectable man, it must be assumed that we are being invited to conclude that these preferences are acceptable in the eyes of the current moral climate: that it is morally acceptable for prudence and/or personal convenience to override obligations arising from country or race, let alone from natural feelings of compassion.

The Samaritan, in contrast has no duties prompting him help the man, and must balance his compassion against the other natural human instincts. That he helps the man (A_5), can only be explained in terms of him putting compassion before all other values, individually and in combination, and this is what we are invited to conclude is what being a neighbour really is. The context supplied in the coda quoted above invites the hearer to adopt these value preferences, to become a person who places compassion above creed, country and convenience and to act in accordance with these priorities in future.

4 Stories as arguments in a dialogical context

In the previous section, we discussed how stories, the characters in them and these characters' motivations can be understood using VAFs. We can now use the story as an argument, using exactly this interpretation of the story. The conclusion the audience is invited to draw from the story depends on the context in which the story is told. In the case of *Good Samaritan*, this context is provided by the exchange between Jesus and the lawyer, and specifically the lawyer's question "Who is my neighbour?". As we have argued in section 2, the actual question is something like "what does it mean to love your neighbour like yourself?", and the answer is not the literal "The Samaritan is my neigbour" but rather an understanding of why the Samaritan acts as he does, which encourages one to adopt similar attitudes to the Samaritan.

4.1 Extended Argumentation about Values

In our model, the audience of the story should identify the value-based arguments in the story and then reason about which values will explain the behaviour of the Samaritan. In (Atkinson and Bench-Capon, 2007) the value orderings themselves cannot be reasoned with or about, as they are not represented in the object language. We therefore use the machinery of (Modgil, 2009) to represent statements about value orderings as arguments in an Extended Argumentation Framework (EAF). In addition to a set of arguments $Args$ and attacks between arguments $\mathcal{R}$, EAFs also contain a set $\mathcal{D} \subseteq (Args \times R)$ of *attacks on attacks*. The idea is that arguments about preferences attack some attack between arguments and thus influence the preferred extension. For example, if argument A attacks argument B and vice versa, there are normally two preferred extensions, $\{A\}$ and $\{B\}$. However, if we add the argument $A > B$ (expressing that A is preferred to B), which attacks and defeats the attack from B on A, there is only one preferred extension namely $\{A, A > B\}$.

In the EAF for the Samaritan there potentially two value-preference arguments for each pair of values, for example:

AV1 Prudence is preferred to Compassion ($P > C$).

AV2 Compassion is preferred to Prudence ($C > P$).

These pairs will mutually attack, but more importantly they will attack the attack from the argument motivated by the less preferred value on arguments motivated by the other value. The complete EAF for the parable will now contain all the base arguments A_1-A_8 and a value preference argument for each attack between these original arguments. Furthermore, we introduce arguments for the various characters: AC_1 (Character is a priest), AC_2 (Character is a Levite) and AC_3 (Character is a Samaritan). This will enable us to eliminate arguments which do not apply to particular characters from consideration: thus AC_1 will attack A_3, A_4 and A_8, AC_2 will attack A_1, A_2 and A_8, and AC_3 will attack A_1, A_2, A_3 and A_4. Adding AC_3 to the AF that contains all characters' arguments A_1 - A_8 then produces the EAF applicable to just the Samaritan, as shown in Figure 2. Similarly, we can introduce AC_2 to get the EAF applicable to the Levite and AC_1 to get the EAF applicable to the priest.

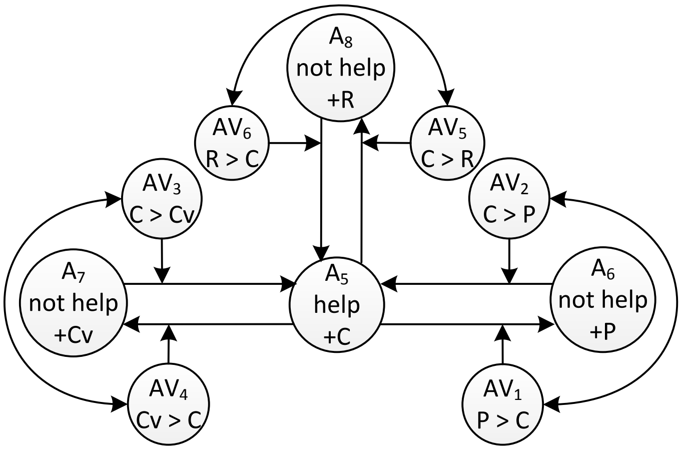


Figure 2: EAF for the Samaritan

4.2 The dialogical argument of the Good Samaritan

Now that we have established appropriate EAFs for the various characters, we need to evaluate them to explain the choices they make in the story. Thus, in the case of the Samaritan, we need to construct an admissible set containing an argument to justify helping the traveller, and then to consider what value preferences it contains. One method of constructing admissible sets from Dung style AFs is to use a dialogue game, such as the TPI (Two Party Immediate Response) Game of (Dunne and Bench-Capon, 2003). As was shown in (Modgil and Bench-Capon, 2008) this can be adapted to EAFs as follows. First, we rewrite the object level arguments of the EAF as meta level statements. This is a purely mechanical process: each pair of arguments in an attack relation is replaced by four arguments and their attack relations. Thus, for example,A_6 attacks A_5 is rewritten as: A_5 *holds*, which is attacked by A_6 *defeats* A_5, which is attacked by A_6 *does not hold* which is attacked by A_6 *holds*. Note that A_5 holds and A_6 *holds* do not directly attack one another, and so are not in conflict. Where A_5 and A_6 are value based arguments, we can reject A_6 *defeats* A_5 not only because we reject A_6, but also because we prefer the value of A_5 to the value of A_6. Thus A_6 *defeats* A_5 is attacked by (in our example) *compassion is preferred to prudence*, which is itself attacked by *prudence is preferred to compassion*. Each pair of attacking arguments is thus rewritten as a regular AF; figure 3 shows the new, regular AF, structure for the pair of arguments A_5 and A_6.

A TPI game proceeds by the proposer playing an argument, the opponent playing an attacker, the proposer playing an attacker of that argument and so on, until one player cannot move. At this point a player can back up to a choice point and play a different attacker. This continues until no moves are possible (note that arguments under attack cannot be played). At this point we will have an admissible set containing the arguments played by the last player to move. If this was the proposer is will contain the original argument and this will have been shown to be acceptable. Because it is the Samaritan's preference we are trying to determine, we use the EAF in figure 2, rewritten as a regular AF. The dialogue then proceeds as follows:

> **Samaritan**: A_5 holds. *This is an argument justifying what the Samaritan did in the story*: current position is {A_5 holds}.
>
> **Opponent**: A_6 defeats A_5. *Opponent chooses a way to attack* A_5.
>
> **Samaritan**: AV_2 C > P. The preference argument is played: the alternative would eventually require A_5 holds to be played, but this is under attack. Current position is {A_5 holds, C > P}.
>
> **Opponent**: A_7 defeats A_5. *Opponent cannot play P > C, because it is under attack, and so backs up and chooses another line of attack.*
>
> **Samaritan**: AV_3 C > Cv. Current position is {A_5 holds, C > P, C > Cv}.
>
> **Opponent**: A_8 defeats A_5. *Again the opponent must back up since Cv > C is under attack.*

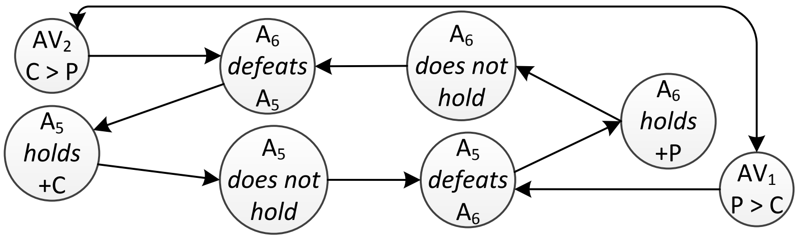


Figure 3: Regular AF for the A_5 - A_6 part if the EAF in figure 2

> **Samaritan**: AV_5 C > R. Current position is $\{A_5$ holds, C > P, C > Cv, C > R$\}$.

At this point the opponent must stop, since there are no further lines of attack. The Samaritan's position, $\{AC_3, A_5, AV_2, AV_3, AV_5\}$, comprises an argument justifying his action A_5, and the three value preferences required to defend that argument AV_2, AV_3 and AV_5. It is exactly this position that the audience is being urged to adopt, since it provides the answer to the lawyer's question "what does it mean to love your neighbour like yourself?".

In our opinion, the argument that the story of the *Good Samaritan* presents is accurately captured by the above dialogue. In contrast to (Govier and Ayers, 2012)'s traditional, more syllogistic analysis of the argument presented by the story (section 2), in the case of the dialogue no explicit norm or course of action is being advocated. This is exactly the way it should be: instead of advocating norms, stories (especially parables) convince by having the audience consider a character's motives by, as it were, engaging in an internal dialogue with the character.

5 Implementing our model

Generating arguments from stories and presenting the different possible extensions based on the value orderings allows one to gain insight into the point of the story: why did the characters act as they did, and which attitudes are advocated in the story? Whilst this is interesting as a theoretical exercise, one additional aim is to implement a system that allows people to explore the stories and character motives in an interactive and intuitive way. One option is to allow humans to engage in a dialogue akin to the ones in section 4.2, thus allowing users to for example, interrogate an agent representing the Samaritan about his motives, and thus gain a better understanding of the story. This can then be used for educational purposes, for example, schoolchildren learning about values through stories.

For such a system, the following separate elements need to be implemented.

1. *Construct initial AATS+V on the basis of a story.*

2. *Include additional hypothetical transitions: 'what could the characters have done and why?'.*

3. *Generate a VAF of arguments and critiques based on AATS+V.*

4. *Execute a dialogue based on the VAF.*

Elements 1 and 2 have been done manually for a few stories: the fable of the *Ant and the Grasshopper* and the Parables of the *Prodigal Son* and the *Good Samaritan*. Ideally part of this process is automated if we want to build a more substantial corpus. For element 1, we can first automatically extract the characters and events from stories, especially from fairly short and simple stories such as fables. This is certainly not trivial but very well possible (see e.g. (Hogenboom et al., 2011)). However, as was discussed earlier, the values expressed by the story depend on the cultural background of the reader: the same story may have different interpretations. Furthermore, element 2 is also hard to fully automate as additional hypothetical transitions are often implicit in the stories, so for elements 1 and 2 human annotation will have to be used, based on skeleton AATS+V's that are constructed using event extraction.

For element 3, currently, Prolog and PHP implementations[3] exist (Wyner et al., 2012),(Wardeh et al., 2013). The PHP tool is based on (Atkinson

[3] The PHP application can be used at http://cgi.csc.liv.ac.uk/∼maya/ACT/. A Prolog program that represents the AATS in Figure 1 and systematically generates the full suite of arguments and objections based on that structure is included in Appendix A.

and Bench-Capon, 2007) and so does not include arguments based on look ahead.

Once the arguments are available, it becomes possible to reason with them in a dialogue. Recently a dialogue game for arguing about the motives found in fables and parables was proposed (Bex and Bench-Capon, 2014). This protocol can be implemented in a dialogue game execution engine (Bex et al., 2014b), which allows for mixed initiative dialogues between software agents and humans through a simple interface (see (Bex et al., 2013)), making it possible to reason with the agents in a story in a similar way as shown in section 4.2. Furthermore, users can input new, value-based arguments about what they think the characters' choices in the story were. These arguments can then relatively easily be inserted as a new transition in the AATS+V (cf. (Reed et al., 2010)), using the mapping given in this article. Thus, the interface may also serve as a knowledge ellicitation tool to find different interpretations of the stories.

6 Conclusion

In this paper we have shown two important connections between computational models of narrative and computational models of argumentation: how argumentation can be used to understand stories, in terms of the motives and attitudes of the characters, and how stories can themselves be used to present arguments, especially arguments designed to persuade the audience to adopt particular attitudes. We have argued that parables can be interpreted as arguments of this sort, and illustrated our views with the famous parable of the *Good Samaritan*. We have identified several advantages of using stories in this way.

Using stories enables the consideration of hypothetical choices, so that the choice can be made clear and memorable, allowing us to benefit from the vividness of the concrete, without needing to have had any particular experience. Moreover using stories excludes irrelevant considerations: we need not consider facts and actions not mentioned in the story; this simplifies the construction of the AATS, and disbars irrelevant counter arguments, allowing for focus to be kept on the main point at issue. Stories are intended to reinforce or change attitudes: this is preferred to presenting a specific set of norms, since attitudes tend to produce an instinctive, and hence more immediate, response and can be applied to numerous, as yet unforeseen, situations. Moreover, they go deeper and so are more to be relied on. This is why soldiers are taught the history of their regiments: the tales of heroism and derring-do can inspire the loyalty and camaraderie required to bind them into an effective unit in a way in which standing orders cannot hope to do. Often there is no objective argument for an attitude or a norm, and so we need to rely on an emotional reaction, which is more easily produced by a story, especially one which allows the hearers to draw the conclusion for themselves (as does the good Samaritan parable, where the conclusion is stated by the addressee, not in the parable itself).

Engaging in a dialogue about a story further draws out the message of the story, and thus dialogue can act as an aid for story understanding. Our model, when combined with an application for argumentative dialogue, makes these dialogues about stories possible. Users can engage in meaningful discussions about a story not just with each other but also with the characters in a story which, when asked, will explain their motives and thus clarify the point of the story. In this way, our model comprises not just a theoretical discussion of understanding and arguing with stories, but also provides a first step towards a promising applications that can be used in, for example, educational settings.

References

K. Atkinson and T. Bench-Capon. 2007. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10-15):855–874.

F. Bex and T. Bench-Capon. 2014. Understanding narratives with argumentation. In *Computational Models of Argument: Proceedings of COMMA 2014*, pages 11 – 18. IOS Press.

F. Bex, T. Bench-Capon, and K. Atkinson. 2009. Did he jump or was he pushed? abductive practical reasoning. *Artificial Intelligence and Law*, 17:79–99.

F.J. Bex, J. Lawrence, M. Snaith, and C. Reed. 2013. Implementing the argument web. *Communications of the ACM*, 56(10):66–73.

F. Bex, K. Atkinson, and T. Bench-Capon. 2014a. Arguments as a new perspective on character motive in stories. *Literary and Linguistic Computing*. to appear.

F.J. Bex, J. Lawrence, and C.A. Reed. 2014b. Generalising argument dialogue with the dialogue game execution platform. In *Proceedings of COMMA 2014*. IOS Press. to appear.

P.M. Dung. 1995. On the acceptability of arguments and its fundamental rolein nonmonotonic reasoning, logic programming, and *n*–person games. *Artificial Intelligence*, 77:321–357.

Paul E. Dunne and Trevor J. M. Bench-Capon. 2003. Two party immediate response disputes: Properties and efficiency. *Artificial Intelligence*, 149(2):221–250.

D. Gentner and K.D. Forbus. 2011. Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2:266–276.

P. Gervas, B. Diaz-Agudo, F. Peinado, and R. Hervas. 2005. Story plot generation based on cbr. *Knowledge-Based Systems*, (4-5):235–242.

T. Govier and L. Ayers. 2012. Logic and parables: Do these narratives provide arguments? *Informal Logic*, 32(2):161–189.

F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong. 2011. An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (ISWC 2011)*, volume 779, pages 48–57.

Sanjay Modgil and Trevor J. M. Bench-Capon. 2008. Integrating object and meta-level value based argumentation. In *Computational Models of Argument: Proceedings of COMMA 2008*, pages 240–251. IOS Press.

S. Modgil. 2009. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173:901–934.

E.T. Mueller. 2004. Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research*, (4):307–340.

C.A. Reed, S. Wells, K. Budzynska, and J. Devereux. 2010. Building arguments with argumentation : the role of illocutionary force in computational models of argument. In *Proceedings of COMMA 2010*.

R.C. Schank and R.P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. Lawrence Erlbaum, Hillsdale, NJ.

C.S. Vlek, H. Prakken, S. Renooij, and B. Verheij. 2013. Modeling crime scenarios in a bayesian network. In *Proceedings of the 14th International Conference on Artificial Intelligence and Law*, pages 150 –159, Rome, Italy.

M. Wardeh, A. Wyner, T. Bench-Caopn, and K. Atkinson. 2013. Argumentation based tools for policy-making. In *Proceedings of the 14th International Conference on Artificial Intelligence and Law (ICAIL 2013)*, pages 249–250, New York (NY). ACM.

R. Wilensky, 1982. *Points: A Theory of the Structure of Stories in Memory*. Erlbaum, Hillsdale, NJ.

A. Wyner, K. Atkinson, and T. Bench-Capon. 2012. Critiquing justifications for action using a semantic model. In *Proceedings of COMMA 2012*.

Encompassing uncertainty in argumentation schemes

Pietro Baroni and **Massimiliano Giacomin**
Università degli Studi di Brescia
{pietro.baroni, massimiliano.giacomin}@ing.unibs.it

Beishui Liao[*]
Zhejiang University
baiseliao@zju.edu.cn

Leon van der Torre
University of Luxembourg
leon.vandertorre@uni.lu

Abstract

In existing literature, little attention has been paid to the problems of how the uncertainty reflected by natural language text (e.g. verbal and linguistic uncertainty) can be explicitly formulated in argumentation schemes, and how argumentation schemes enriched with various types of uncertainty can be exploited to support argumentation mining and evaluation. In this paper, we focus on the first problem, and introduce some preliminary ideas about how to classify and encompass uncertainty in argumentation schemes.

1 Introduction

Mining and evaluating arguments from natural language text (Green et al., 2014) is a relatively new research direction with applications in several areas ranging from legal reasoning (Palau and Moens, 2011) to product evaluation (Wyner et al., 2012). Argumentation schemes (Walton et al., 2008) are commonly adopted in this context as a first modeling tool: it is assumed that natural arguments adhere to a set of paradigmatic schemes, so that these schemes can be used both to drive the identification of the arguments present in the text and, after that, to support their formal representation. As a further step, the assessment of argument justification status requires to identify the relations among them and to apply a formal method, called *argumentation semantics* to derive the status from these relations. For instance, the well known[1] Dung's theory of abstract argumentation (Dung, 1995) focuses on the relation of attack between arguments and provides a rich variety of alternative semantics (Baroni et al., 2011) for argument evaluation on this basis.

When dealing with natural language sources, one of the challenging problems is to handle the uncertainty of arguments. In fact, natural language statements typically include several kinds of uncertainty. This calls for the need to encompass uncertainty in the formalisms which are meant to provide a representation of natural arguments, first of all in argumentation schemes, in order to avoid that some useful information carried by the text source is lost in the first modelling step.

To illustrate this problem, let us consider a simple example concerning two conflicting natural language excerpts E_1 and E_2, possibly taken from some medical publications:

E_1: According to [Smith 98], drug X often causes the side effect Y.

E_2: According to recent experimental trials, it is highly likely that drug X does not increase the probability of the side effect Y.

In order to identify argument structures in these texts, one may resort to specific argumentation schemes. Referring to the classification proposed in (Walton et al., 2008), E_1 can be represented by an argument A_1 which is an instance of the scheme *Argument from Expert Opinion*, while E_2 by an argument A_2 which is an instance of the scheme *Argument From Falsification*.

After A_1 and A_2 are identified, it may be noted that (though expressed with different linguistic nuances) their conclusions are in conflict: briefly, A_1 leads to the claim that X causes Y, while A_2 to the claim that X does not cause Y. As a consequence, a mutual attack relation between A_1 and A_2 can be identified. Then, the arguments and their attacks can be formalized as an abstract argumentation framework $AF = (\{A_1, A_2\}, \{(A_1, A_2), (A_2, A_1)\})$ and the status of arguments in AF can be evaluated according to a given argumentation semantics. For instance, under grounded semantics, both A_1 and A_2 are not accepted. It must be noted however that such a modelling approach

[*] Corresponding author

[1] Due to space limitations, we assume knowledge of Dung's theory in the following.

(and the relevant outcome in terms of argument evaluation) overlooks some information which is (implicitly or explicitly) carried by the text and that may lead, in particular, to have one of the arguments prevailing over the other. For instance, as considered in (Bex et al., 2013), one may have a preference relation over argument schemes so that, for instance, the scheme *Argument From Falsification* is preferred to the scheme *Argument from Expert Opinion*. Accordingly, A_2 would be preferred to A_1, and the attack relation would not be mutual, due to the inability of A_1 to attack A_2 (see the notion of *preference-dependent* attack in (Bex et al., 2013)). In this case, we would get a different argumentation framework $AF' = (\{A_1, A_2\}, \{(A_2, A_1)\})$. Then, under grounded semantics, A_1 is rejected, while A_2 is accepted.

However, a static preference relation on the adopted scheme appears too rigid: in most cases the preference for an argument over another one is not simply based on their structure but, rather, on their content. To exemplify, in this case, one may have different opinions on the reliability of the source [Smith 98], mentioned in E_1, and of the experimental trials mentioned in E_2. Moreover, the two excerpts include several terms expressing vagueness and/or uncertainty, like *often*, *highly likely*, *the probability of*, that may be taken into account in the preference ranking of arguments. However, this is not possible in the approach sketched above, since the argument schemes adopted in the formalization do not encompass these forms of uncertainty and the relevant information carried by the text is lost in the first modelling step.

Given the pervasiveness of vagueness and uncertainty in natural language this appears to be a severe limitation for the use of argumentation schemes in argument mining from texts. To overcome this problem we envisage the study of argumentation schemes extended with uncertainty in the context of the process sketched in Figure 1. Here argumentation schemes with uncertainty are used to extract arguments from texts, keeping explicit the relevant uncertainties that can then be used in the step of argument evaluation using suitable abstract formalisms and semantics with uncertainty. As to the latter step, the study of extensions of Dung's framework with explicit uncertainty representation is receiving increasing attention in recent years (Li et al., 2011a; Thimm, 2012; Hunter, 2013a; Hunter, 2014) while, to the best of our knowledge, lesser work has been devoted to encompassing uncertainty in argumentation schemes.

This long-term research goal involves several basic questions including:

1) How the uncertainty reflected by natural language text can be explicitly formulated in argumentation schemes?

2) How argumentation schemes enriched with various types of uncertainty can be exploited to support argument mining and evaluation?

3) Which is (are) the most appropriate abstract formalism(s) for the evaluation of arguments with uncertainty?

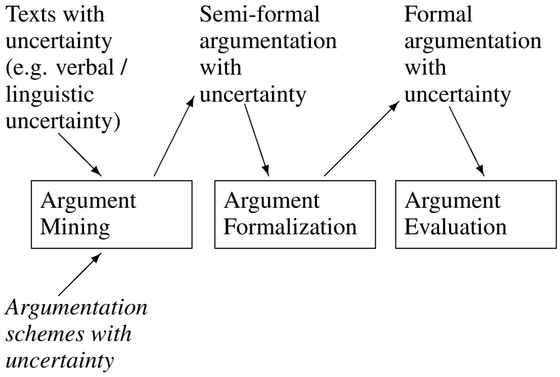

Texts with uncertainty (e.g. verbal / linguistic uncertainty)
Semi-formal argumentation with uncertainty
Formal argumentation with uncertainty
Argument Mining
Argument Formalization
Argument Evaluation
Argumentation schemes with uncertainty

Figure 1: From natural language to argument evaluation: a schematic process

By focusing on the first question, this paper presents some preliminary ideas for encompassing uncertainty in argumentation schemes.

The paper is organised as follows. We review some examples of uncertainty classifications in natural language texts in Section 2 and analyze the non-uniformity of uncertainty representation in existing argumentation schemes in Section 3. Then, in Section 4 we exemplify and discuss a preliminary approach for encompassing uncertainty in argumentation schemes. Finally, Section 5 concludes the paper.

2 Classifying uncertainty types in natural language texts

In natural language texts different types of uncertainty can be identified. To give a brief account of the richness and complexity of this topic and of the research activities that are being carried out in this area, we quickly recall some examples of uncertainty classifications considered in the literature.

In the context of scientific discourse, de Waard and Maat (2012) distinguish knowledge evaluation (also called epistemic modality) from knowledge attribution (also called evidentiality). The former basically concerns the degree of commitment with respect to a given statement, while the latter concerns the attribution of a piece of knowledge to a source. Accordingly, different kinds of uncertainty can be identified.

For instance, according to de Waard and Maat (2012), sources of knowledge may be distinguished into the following categories:

1) Explicit source of knowledge: the knowledge evaluation can be explicitly owned by the author ('*We therefore conclude that ...*') or by a named referent ('*Vijh et al. [28] demonstrated that ...*').

2) Implicit source of knowledge: if there is no explicit source named, knowledge can implicitly still be attributed to the author (' *these results suggest ...*') or an external source ('*It is generally believed that ...*').

3) No source of knowledge: the source of knowledge can be absent entirely, e.g. in factual statements, such as '*transcription factors are the final common pathway driving differentiation*'.

Since different sources may have different degrees of credibility, this leads to identify a first type of uncertainty, namely the (possibly implicit) *source uncertainty*.

As to knowledge evaluation, de Waard and Maat (2012), following Wilbur et al. (2006), distinguish four *levels of certainty* in the degree of commitment of a subject to a statement: 1) Doxastic (firm belief in truth), 2) Dubitative (some doubt about the truth exists), 3) Hypothetical (the truth value is only proposed), and 4) Lack of knowledge.

This kind of evaluation, called *uncertainty about statements* in the following, is typically expressed through suitable linguistic modifiers.

Actually linguistic modifiers have a quite generic nature and have been the subject of specific studies by themselves: Clark (1990) provides an extensive review of experimental studies concerning the use of *linguistic uncertainty* expressions, such as *possible*, *probable*, *likely*, *very likely*, *highly likely*, etc., and their numerical representation. Linguistic uncertainty is pervasive in natural language communication. On the one hand, it can be regarded as a form of *uncertainty expression* (alternative to, e.g., numerical or implicit uncertainty expressions) rather than as a distinct uncertainty type. On the other hand, linguistic uncertainty may be regarded as a generic type of uncertainty, of which other more specific forms of uncertainty are subtypes. This generic type can be associated to those natural language statements to which a more specific uncertainty type can not be applied. For the sake of the preliminary analysis carried out in this paper, we will adopt the latter view.

Regan et al. (2002) distinguish between epistemic uncertainty (uncertainty in determinate facts) and linguistic uncertainty (uncertainty in language) and claims that the latter has received by far less attention in uncertainty classifications in the fields of ecology and biology. Linguistic uncertainty is in turn classified into five distinct types: vagueness, context dependence, ambiguity, indeterminacy of theoretical terms, and underspecificity, with vagueness being claimed to be the most important for practical purposes. In fact, all of them refer in some way to the problem that some natural language expressions admit alternative interpretations. Hence this classification is focused on a specific form of uncertainty and the use of the term linguistic uncertainty here is rather restricted with respect to other works.

Taking into account the discussion above, in this paper we consider, as a starting point, three uncertainty types:

1) *Source uncertainty*, denoted in the following as U_1, concerning the fact that to evaluate the credibility of different statements one may take into account the credibility of their sources;

2) *Uncertainty about a statement*, denoted as U_2, arising in situations where a subject making a statement expresses a partial degree of commitment to the statement itself;

3)*Linguistic uncertainty or uncertainty inside a statement*, denoted as U_3, namely uncertainty generically present in natural language statements, with no further more precise meaning specified.

For instance in the sentences "*According to [Smith 98]*, Drug X causes headache" and "*According to recent experimental trials*, Drug X causes headache", one may identify U_1 since they refer the statement "Drug X causes headache" to a source (a paper and clinical trials, respectively).

On the other hand, the sentence "It is *likely* that Drug X causes headache" provides an example of U_2 since the statement "Drug X causes headache" is not regarded as certain.

Finally, a sentence like "Drug X *sometimes* causes *severe* headache" provides an example of U_3.

For a more articulated example including several uncertainty types, let us consider the following text, taken from (Swenson, 2014): "*. . . , the Mg inhibition of the actin-activated ATPase activity observed in class II myosins is likely the result of Mg-dependent alterations in actin binding. Overall, our results suggest that Mg reduces the ADP release rate constant and rate of attachment to actin in both high and low duty ratio myosins.* "

Here, some expressions (*likely* and *suggest that*) indicate a partial commitment of authors to the corresponding statements (U_2), and the knowledge source is made explicit by the citation of (Swenson, 2014) (U_1). Further, the vague terms (*high* and *low*) correspond to a form of generic linguistic uncertainty inside the relevant statement (U_3).

3 Non-uniformity of uncertainty representation in existing schemes

Given that uncertainty pervades natural language texts and argumentation schemes appear as suitable formal tool for argumentation mining from texts, the question of how to capture uncertainty in argumentation schemes naturally arises. This appears to be an open research question, as the state-of-the-art formulation of argumentation schemes (Walton et al., 2008) does not consider uncertainty explicitly, and, more critically, does not seem to deal with uncertainty in a systematic way, though somehow recognizing its presence. To exemplify this problem let us compare two argumentation schemes[2] from (Walton et al., 2008).

The first scheme we consider, called *Argument from Position to Know* (APK), is defined as follows:

Major Premise: Source a is in a position to know about things in a certain subject domain S containing proposition A.
Minor Premise: a asserts that A (in domain S) is true (false).
Conclusion: A is true (false).
CQ1: Is a in a position to know whether A is true (false)?
CQ2: Is a an honest (trustworthy, reliable) source?
CQ3: Did a assert that A is true?

In this scheme, no explicit uncertainty is included, but the critical questions correspond to several forms of uncertainty that may affect it.

The second scheme, called *Argument from Cause to Effect* (ACE), is defined as follows:

Major Premise: Generally, if A occurs, then B will (might) occur.
Minor Premise: In this case, A occurs (might occur).
Conclusion: Therefore, in this case, B will (might) occur.
CQ1: How strong is the causal generalization?
CQ2: Is the evidence cited (if there is any) strong enough to warrant the casual generalization?
CQ3: Are there other causal factors that could interfere with the production of the effect in the given case?

In this case, in addition to the implicit uncertainty corresponding to critical questions, explicit expressions of uncertainty are included, namely the modifier *Generally* and the *might* specifications in the parentheses.

Clearly the representation of uncertainty in the two schemes is not uniform (since the second scheme encompasses explicit uncertainty in the premises and the conclusion, while the first does not) but it is not clear whether this non-uniformity is based on some underlying difference between the schemes or is just accidental in the natural language formulation of the schemes. Indeed, it seems possible to reformulate these schemes in a dual manner (adding explicit uncertainty mentions to the first one, removing them from the second one) while not affecting their meaning, as follows:

APK with explicit uncertainty:
Major Premise: Source a is (possibly) in a position to know about things in a certain subject domain S containing proposition A.
Minor Premise: a asserts that A (in domain S) is (might be) true (false).
Conclusion: A is (might be) true (false).

ACE without explicit uncertainty:
Major Premise: If A occurs, then B will occur.
Minor Premise: In this case, A occurs.
Conclusion: Therefore, in this case, B will

[2] Recall that an argument scheme basically consists of a set of premises, a conclusion defeasibly derivable from the premises according to the scheme, and a set of critical questions (CQs) that can be used to challenge arguments built on the basis of the scheme.

occur.

The above-mentioned non-uniformity suggests that a more systematic treatment of uncertainty in argument schemes is needed in order to face the challenges posed by the representation of natural language arguments.

Indeed, a recent work (Tang et al., 2013) addresses the relationships between uncertainty and argument schemes in a related but complementary research direction. While the work described in the present paper aims at enriching argumentation schemes proposed in the literature with explicit uncertainty representation in a systematic way, Tang et al. (2013) introduce several novel argument schemes concerning reasoning *about* uncertainty. This is done using Dempster-Shafer theory of evidence in the context of a formalism for the representation of *evidence arguments*. Different schemes basically differ in the choice of the rule for (numerical) evidence combination among the many alternative combination rules available in the literature, and the critical questions in each scheme refer to the applicability conditions of the relevant rule (e.g. Is each piece of evidence independent?). Investigating the possible reuse of some of the specific ideas presented by Tang et al. (2013) in the context of our broader modelling approach is an interesting direction of future work.

4 Encompassing uncertainty in argumentation schemes

Devising a systematic approach to encompass natural language uncertainty in argumentation schemes is a long term research goal, posing many conceptual and technical questions and challenges, partly evidenced in the previous sections. We suggest that such an approach should include the following “ingredients”:

1) a classification of uncertainty types;

2) a characterization of the uncertainty types relevant to each argumentation scheme;

3) a formalism for the representation of uncertainty evaluations (of various types) in actual arguments, i.e. in instances of argument schemes;

4) a mechanism to derive an uncertainty evaluation for the conclusion of an argument from the evaluations concerning the premises and the applied scheme.

While each of the items listed above is, by itself, a large and open research question, we provide here some preliminary examples of point 2, using for point 1 the simple classification introduced in Section 2. In particular we suggest that the scheme specification should be accompanied by an explicit account of the types of uncertainty it may involve, while the use of linguistic uncertainty expressions in the scheme (like in ACE above) should be avoided within the natural language description of the scheme itself. This approach prevents the non-uniformities pointed out in Section 3 and enforces the adoption of clear modelling choices about uncertainty at the moment of definition of the scheme. In particular, as evidenced below, it may point out some ambiguities in the definition of the scheme itself.

In the following examples, we explicitly associate uncertainty types with the premises of the considered schemes (that may affected by them) and with the critical questions (that point out the potential uncertainty affecting the premises). Analysing the uncertainty possibly affecting the scheme itself or its applicability (that may also be expressed by some critical questions) is left to future work (and requires a richer classification of uncertainty types), while, according to point 4 above, the uncertainty about the conclusion is regarded as a derived notion and, for the sake of the present analysis, is considered as derived uncertainty, denoted as DU. The syntax we use to associate uncertainty types with parts of argument schemes is as follows: $\{\ldots\}[\mathrm{U_x},\ldots]$, where the part of the scheme (possibly) affected by uncertainty is enclosed in braces and is followed by the relevant uncertainty type(s) enclosed in brackets.

First, let us consider the APK scheme. Here, the major premise explicitly refers to a source a, so it can be associated with $\mathrm{U_1}$ (as evidenced by the critical questions CQ1 and CQ2). Further one may consider that the inclusion of proposition A in domain S and the proposition A itself can be specified with some linguistic uncertainty ($\mathrm{U_3}$). As to the minor premise, since it refers explicitly to a given assertion, it can be associated with uncertainty about assertions ($\mathrm{U_2}$). Actually, the critical question CQ3 refers to the minor premise and its statement “Did a assert that A is true?” is, in fact, ambiguous as far as the type of uncertainty is concerned. On the one hand it might raise a doubt about the fact that a did actually make any assertion about A, on the other hand it might raise a doubt about the contents of the assertion made by

a. For instance, a might have made a weaker assertion, like "A is probably true", or a completely different assertion like "A is false". The three alternatives mentioned above are rather different and involve different uncertainty types. The possibility that a made a weaker assertion is a case of U_2, while if a made a completely different assertion (or no assertion at all) about A, the entire minor premise is challenged, and this amounts to be uncertain about the credibility of the (implicit) source from which we learned that "a asserted that A is true", hence a case of U_1. As this ambiguity is present in the current formulation of the scheme, we leave it unresolved and indicate both types of uncertainty for the minor premise and CQ3.

This leads to reformulate APK as follows:

Major Premise: $\{$Source a is in a position to know about things in a certain subject domain $S\}[\mathrm{U}_1]$ $\{$containing proposition $A\}[\mathrm{U}_3]$.

Minor Premise: $\{a$ asserts that A (in domain S) is true (false)$\}[\mathrm{U}_1, \mathrm{U}_2]$.

Conclusion: $\{A$ is true (false)$\}$[DU].

CQ1: $\{$Is a in a position to know whether A is true (false)?$\}[\mathrm{U}_1]$

CQ2: $\{$Is a an honest (trustworthy, reliable) source?$\}[\mathrm{U}_1]$

CQ3: $\{$Did a assert that A is true?$\}[\mathrm{U}_2, \mathrm{U}_1]$.

Let us now consider the ACE scheme. Its first premise is a causal generalization, which, as suggested by the use of *(might)* in its original formulation, is not always valid. In our simple classification this can be regarded as a form of linguistic uncertainty inside the statement (U_3). This kind of uncertainty may also affect the actual formulation of the statements A and B in the instantiations of the scheme. The major premise is challenged by CQ1 and CQ2. While their interpretation allows some overlap, CQ1 seems to concerns the "strength" of the causal generalization as it is formulated, while CQ2 refers to the implicit evidential source of knowledge supporting the causal generalization. Accordingly, CQ1 may be referred to U_3, while CQ2 to U_1.

The minor premise concerns the observation of a fact (the occurrence of A), that might involve linguistic uncertainty U_3. Indeed, also the observation of the occurrence of A might have a source, so that, in principle, the second premise might be affected by U_1, and one might have an additional critical question CQ+ like "Does A actually occur?", which would turn out very similar in nature to CQ3 in the APK scheme. The fact that a question like CQ+ is not considered in this scheme, points out a further non-uniformity in the formulation of argument schemes: one may wonder why a sort of explicit confirmation of the minor premise is required by a critical question in the APK scheme, while the same kind of confirmation is not required in the ACE scheme. While one might answer that similar questions may have a different importance in different schemes, we suggest that a further analysis is needed to address these issues in a systematic way and that a classification of uncertainty types can be very useful in this respect. To point out this, we add CQ+ in the revised version of the ACE scheme, with the relevant uncertainty type U_1 associated with the minor premise. Finally, CQ3 raises the question about possible other factors interfering with the causal relation between A and B, i.e. suggests the presence of possible exceptions in the application of the scheme. This kind of uncertainty is not encompassed in our simplistic preliminary classification, hence we let it unspecified (denoted as [??]), as a pointer to future developments. This leads to reformulate ACE as follows:

Major Premise:$\{$If A occurs, then B will occur$\}$ $[\mathrm{U}_1, \mathrm{U}_3]$.

Minor Premise: $\{$In this case, A occurs$\}$ $[\mathrm{U}_1, \mathrm{U}_3]$.

Conclusion:$\{$Therefore, in this case, B will occur$\}$ $[DU]$.

CQ1: $\{$How strong is the causal generalization?$\}[\mathrm{U}_3]$

CQ2: $\{$Is the evidence cited (if there is any) strong enough to warrant the casual generalization?$\}[\mathrm{U}_1]$

CQ+: $\{$Does A actually occur?$\}[\mathrm{U}_1]$

CQ3: $\{$Are there other causal factors that could interfere with the production of the effect in the given case?$\}[??]$

5 Conclusions

In recent years, the issue of combining explicit uncertainty representation and argumentation has received increasing attention, with several works dealing in particular with probabilistic argumentation (Dung and Thang, 2010; Hunter, 2012; Hunter, 2013b; Li et al., 2011b). These works are based on formal argumentation theories like Dung's abstract argumentation frameworks (Dung, 1995) or logic-based argumentation (Hunter, 2013b). This paper suggests that these investigations on the formal side should be complemented by efforts on the conceptual and semi-formal side, with particular reference to the argumentation schemes model. Argumentation schemes provide a very intuitive semi-formal representation approach for natural arguments and are indeed adopted in several works as a first level modelling tool to identify and extract arguments from natural language texts. However, as evidenced in this paper, argumentation schemes need to be enriched and extended in order to capture the various kinds of uncertainty typically present in natural language arguments. The present work provides a preliminary contribution to this research line, by pointing out some problems and providing some simple examples of how they might be tackled. Future work directions are huge and include an extensive review of the uncertainty types considered in the literature, with special attention to works in the area of argumentation mining, and a systematic analysis of the various ways argument schemes may be affected by different uncertainty types.

Acknowledgment

The research reported in this paper was partially supported by the National Science Foundation of China (No.61175058), and Zhejiang Provincial Natural Science Foundation of China (No. LY14F030014).

References

P. Baroni, M. Caminada, and M. Giacomin. 2011. An introduction to argumentation semantics. *Knowledge Eng. Review*, 26(4):365–410.

F. Bex, S. Modgil, H. Prakken, and C. Reed. 2013. On logical specifications of the argument interchange format. *J. Log. Comput.*, 23(5):951–989.

D. A. Clark. 1990. Verbal uncertainty expressions: A critical review of two decades of research. *Current Psychology: Research & Reviews*, 9(3):203–235.

A. de Waard and H. P. Maat. 2012. Epistemic modality and knowledge attribution in scientific discourse: A taxonomy of types and overview of features. In *Proc. of ACL 2012*, pages 47–55.

P. M. Dung and P. M. Thang. 2010. Towards (probabilistic) argumentation for jury-based dispute resolution. In *Proc. of COMMA 2010*, pages 171–182.

P. M. Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence*, 77(2):321–357.

N. Green, K. Ashley, D. Litman, C. Reed, and V. Walker, editors. 2014. *ACL 2014 Proceedings of the First Workshop on Argumentation Mining, Baltimore, Maryland, USA*.

A. Hunter. 2012. Some foundations for probabilistic abstract argumentation. In *Proc. of COMMA 2012*, pages 117–128.

A. Hunter. 2013a. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81.

A. Hunter. 2013b. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81.

A. Hunter. 2014. Probabilistic qualification of attack in abstract argumentation. *International Journal of Approximate Reasoning*, 55(2):607 – 638.

H. Li, N. Oren, and T. J. Norman. 2011a. Probabilistic argumentation frameworks. In *Proc. of TAFA 2011*, pages 1–16.

H. Li, N. Oren, and T. J. Norman. 2011b. Probabilistic argumentation frameworks. In *Proc. of TAFA 2011*, pages 1–16.

R. Mochales Palau and M.-F. Moens. 2011. Argumentation mining. *Artif. Intell. Law*, 19(1):1–22.

H. M. Regan, M. Colyvan, and M. A. Burgman. 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, 12(2):618–628.

A. M. Swenson. 2014. Magnesium modulates actin binding and adp release in myosin motors. *J. Biol. Chem.*

Y. Tang, N. Oren, S. Parsons, and K. Sycara. 2013. Dempster-shafer argument schemes. In *Proc. of ArgMAS 2013*.

M. Thimm. 2012. A probabilistic semantics for abstract argumentation. In *Proc. of ECAI 2012*, pages 750–755.

D. Walton, C. Reed, and F. Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

W. J. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356.

A. Wyner, J. Schneider, K. Atkinson, and T. J. M. Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In *Proc. of COMMA 2012*, pages 43–50.

An Informatics Perspective on Argumentation Mining

Jodi Schneider[*]
INRIA Sophia Antipolis France
jschneider@pobox.com

Abstract

It is time to develop a community research agenda in argumentation mining. I suggest some questions to drive a joint community research agenda and then explain how my research in argumentation, on support tools and knowledge representations, advances argumentation mining.

1 Time for a community research agenda

This year, argumentation mining is receiving significant attention. Five different events from April to July 2014 focus on topics such as arguing on the Web, argumentation theory and natural language processing, and argumentation mining. A coordinated research agenda could help advance this work in a systematic way.

We have not yet agreed on the most fundamental issues:

Q1 What counts as 'argumentation', in the context of the argumentation mining task?

Q2 How do we measure the success of an argumentation mining task? (e.g. corpora & gold standards)

> *"Argumentation mining, is a relatively new challenge in corpus-based discourse analysis that involves automatically identifying argumentative structures within a document, e.g., the premises, conclusion, and argumentation scheme of each argument, as well as argument-subargument and argument-counterargument relationships between pairs of arguments in the document."*[1] *(Green et al., 2014)*

An informatics perspective (i.e. concerned with supporting human activity) could help us understanding how we will apply argumentation mining; this should sharpen the definition of the argumentation mining task(s). Given such an operationalization, we can then use the standard natural language processing approach: define a corpus of interest, make a gold standard annotation, test algorithms, iterate...

For instance, to operationalize the definition of argumentation mining (Q1), we need to know:

Q1a How do we plan to use the results of argumentation mining?

Q1b What domain(s) and human tasks are to be supported?

Q1c What is the appropriate level of granularity of argument structures in a given context? Which models of argumentation are most appropriate?

This can be challenging because argumentation has a variety of meanings and uses, in fields from philosophy to rhetoric to law; some of the purposes for using argumentation are shown in Figure 1.Understanding how we will use the results of argumentation mining can help address important questions related to Q2, such as measuring the success of algorithms and support tools for identifying arguments. In particular:

Q2a How accurate does argumentation mining need to be?

Q2b In which applications are algorithms for automatically extracting argumentation most appropriate?

Q2c In which applications are support tools for semi-automatically extracting argumentation more appropriate?

In my work I have tried to bring applications of argumentation mining to the forefront. My work falls into three main areas: supporting human argumentation with computer tools (CSCW), rep-

[*] This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n^{o} 246016.

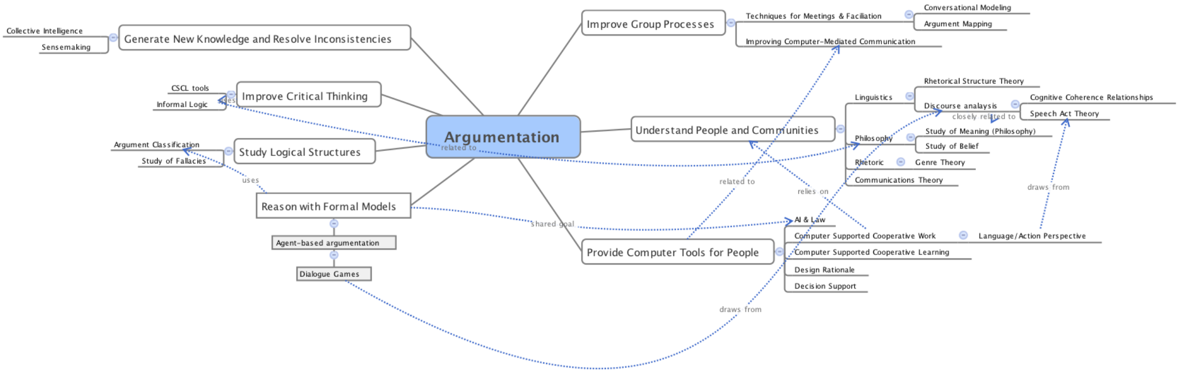

Collective Intelligence
Sensemaking
Generate New Knowledge and Resolve Inconsistencies
CSCL tools
Informal Logic
uses
Improve Critical Thinking
Argument Classification
Study of Fallacies
Study Logical Structures
uses
Argumentation
related to
Reason with Formal Models
Agent-based argumentation
Dialogue Games
shared goal
Improve Group Processes
Techniques for Meetings & Faciliation
Conversational Modeling
Argument Mapping
Improving Computer-Mediated Communication
Understand People and Communities
Linguistics
Rhetorical Structure Theory
Discourse analaysis
Cognitive Coherence Relationships
Speech Act Theory
closely related to
Philosophy
Study of Meaning (Philosophy)
Study of Belief
Rhetoric
Genre Theory
Communications Theory
related to
relies on
draws from
AI & Law
Computer Supported Cooperative Work
Language/Action Perspective
Provide Computer Tools for People
Computer Supported Cooperative Learning
Design Rationale
Decision Support
draws from

Figure 1: Argumentation can be used for many purposes.
Download an editable version of this figure from FigShare DOI `http://dx.doi.org/10.6084/m9.figshare.1149925`

resenting argumentation in ontologies (knowledge representation), and mining arguments from social media (information extraction using argumentation theory).

2 Computer-Supported Collaborative Work

Arguing appears throughout human activity, to support reasoning and decision-making. The application area determines the particular genres and subgenres of language that should be investigated (Q1b). The appropriate level of granularity (Lawrence et al., 2014) depends on whether we are in a literary work or a law case or a social media discussion (Q1c). The acceptable error rate (Q2a) follows from human tolerances, which we expect to depend on the area; this in turn determines whether we completely automate argumentation mining (Q2b) or merely provide semi-automatic support (Q2c). This is why I emphasize looking at application areas to determine which problems to focus our attention on, for argument mining.

My thesis described a general, informatics approach to supporting argumentation in collaborative online decision-making (Schneider, 2014b):

1. Analyze requirements for argumentation support in a given situation, context, or community.
2. Consider which argumentation models to use; test their suitability, using features such as the appropriate level of granularity and the tasks to be supported.
3. Build a prototype support tool, using a model of argumentation structures.
4. Evaluate and iterate.

In this approach, argumentation mining supports scalability, by providing automatic or semi-automatic identification of the relevant arguments.

I have applied this methodology to Wikipedia information quality debates, which are used to determine whether to delete a given topic from the encyclopedia (Schneider, 2014b). We tested two argumentation models: Walton's argumentation schemes (Schneider et al., 2013) and the theory of factors/dimensions (Schneider et al., 2012c), and our annotated data is available online.[2] Whereas Walton's argumentation schemes could have provided support for writing arguments, we instead chose to use domain-specific decision factors to filter the overall debate in the prototype support tool we built. One difference is that Walton's argumentation schemes are at the micro-level—structuring the premises and conclusions of a given argument—whereas decision factors are at the macro-level, identifying the topics important to discuss; this distinction may be relevant for argumentation mining (Schneider, 2014a).

3 Knowledge Representation

Argumentation mining assumes a way to package arguments so that they can be exchanged and shared. Structured representations of arguments allow "evaluating, comparing and identifying the relationships between arguments" (Rahwan et al., 2011). And the knowledge representations most commonly used for the Web are ontologies.

To investigate the existing ontologies for structuring arguments on the social web, we wrote "A Review of Argumentation for the Social Semantic Web" (Schneider et al., 2012b).

[2] `http://purl.org/jsphd`

The review compares:

- 13 theoretical models for capturing argument structure (Toulmin, IBIS, Walton, Dung, Value-based Arg. Frameworks, Speech Act Theory, Language/Action Perspective, Pragma-dialectic, Metadiscourse, RST, Coherence, and Cognitive Coherence Relations).
- Applications of these theoretical models.
- Ontologies incorporating argumentation (including AIF, LKIF, IBIS and many others).
- 37 collaborative Web-based tools with argumentative discussion components (drawn from Social Web practice as well as from academic researchers).

Thus the argumentation community can choose from a number of existing approaches for structuring argumentation on the Web.

Still, new approaches continue to be suggested. Peldszus and Stede have suggested a promising proposal for annotating arguments using Freeman's argumentation macrostructure (Peldszus and Stede, 2013). And for biomedical communications, Clark et al have proposed a micropublications ontology based on Toulmin's model for pay-as-you-go construction of claim-argument networks from scientific papers (Clark et al., 2014). We are using this ontology—the micropublications ontology[3]—to model evidence about pharmacokinetic drug interactions (Schneider et al., 2014a) in a joint project organized by Richard Boyce.

We have also developed two ontologies related to argumentation. First, WD, the Wiki Discussion ontology[4] (Schneider, 2014b) was alluded to in Section 2: WD is used for argumentation support for decision-making discussions in ad-hoc online collaboration, applying factors/dimensions theory. Second, ORCA is an Ontology of Reasoning, Certainty and Attribution[5] (de Waard and Schneider, 2012). Based on a taxonomy by de Waard, ORCA is motivated by scientific argument. ORCA allows distinguishing completely verified facts from hypotheses: it records the certainty of knowledge (lack of knowledge; hypothetical; dubitative; doxastic) as well as its basis (reasoning, data, unidentified) and source (author or other, explicitly or implicitly; or none).

4 Mining from Social Media

The third strand of our research is in mining arguments from social media.

4.1 Characteristics of social media

To identify arguments in social media, we need to know where to look. The intention of the author might be relevant, for instance we can expect different types of argument in messages, depending on whether they are recreation, information, instruction, discussion, and recommendation (Schneider et al., 2014b). In (Schneider et al., 2012a), we suggested that relevant features for argumentation in social media may include the genre, metadata, properties of users, goals of a particular dialogue, context and certainty, informal and indirect speech, implicit information, sentiment and subjectivity.

4.2 Information extraction based on argumentation schemes

In a corpus of camera reviews, we examine the argument that consumers give in reviews, focusing on rationales about camera properties and consumer values.

In collaboration with Liverpool researchers including Adam Wyner (Wyner et al., 2012), we describe the argumentation mining task in consumer reviews as an information extraction task, where we fill slots in a predetermined argumentation scheme, such as:

Consumer Argumentation Scheme:
Premise: Camera *X* has property *P*.
Premise: Property *P* promotes value *V* for agent*A*.
Conclusion: Agent *A* should *Action1* camera *X*.

Further details of the information extraction are given in (Schneider and Wyner, 2012). In particular, we developed gazetteers for the camera domain and user domain, and selected appropriate discourse indicators and sentiment terminology. These form part of an NLP pipeline in the General Architecture for Text Engineering framework. Resulting annotations can be viewed on a document or searched with a corpus indexing and querying tool, informing an argument analyst who wishes to construct instances of the consumer argumentation scheme.

We have also presented additional argumentation schemes that model evaluative expressions in reviews, focusing in (Wyner and Schneider, 2012) on user models within a context of hotel reviews.

[3] http://purl.org/mp/

[4] http://purl.org/wd/

[5] http://vocab.deri.ie/orca

5 Conclusions

We have described our work related to argumentation mining, which uses CSCW, knowledge representation, argumentation theory and information extraction. As we noted, different approaches are appropriate for identifying and modeling arguments in online debates (Schneider, 2014b) versus scientific papers (Schneider et al., 2014a), so different application areas need to be considered. We hope that our questions about argumentation mining—starting with *What counts as 'argumentation', in the context of the argumentation mining task?* and *How do we measure the success of an argumentation mining task?*—drive the community towards establishing shared tasks. Shared corpora and well-defined tasks are needed to propel argumentation mining beyond a highly discussed area into an agreed upon research challenge.

References

Tim Clark, Paolo N. Ciccarese, and Carole A. Goble. 2014. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics*, 5(27), July.

Anita de Waard and Jodi Schneider. 2012. Formalising uncertainty: An Ontology of Reasoning, Certainty and Attribution (ORCA). In *Joint Workshop on Semantic Technologies Applied to Biomedical Informatics and Individualized Medicine (SATBI+SWIM 2012) at International Semantic Web Conference (ISWC 2012)*.

Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker. 2014. Workshop description, first workshop on argumentation mining at the association for computational linguistics.

John Lawrence, Chris Reed, Colin Allen, Simon McAlister, and Andrew Ravenscroft. 2014. Mining arguments from 19th century philosophical texts using topic based modelling. In *Proceedings of the First Workshop on Argumentation Mining*, pages 79–87. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Iyad Rahwan, Bita Banihashemi, Chris Reed, Douglas Walton, and Sherief Abdallah. 2011. Representing and classifying arguments on the Semantic Web. *The Knowledge Engineering Review*, 26(04):487–511.

Jodi Schneider and Adam Wyner. 2012. Identifying consumers' arguments in text. In *SWAIE 2012: Semantic Web and Information Extraction. In conjunction with EKAW 2012*.

Jodi Schneider, Brian Davis, and Adam Wyner. 2012a. Dimensions of argumentation in social media. *Knowledge Engineering and Knowledge Management*, pages 21–25.

Jodi Schneider, Tudor Groza, and Alexandre Passant. 2012b. A review of argumentation for the Social Semantic Web. *Semantic Web-Interoperability, Usability, Applicability*, 4(2):159–218.

Jodi Schneider, Alexandre Passant, and Stefan Decker. 2012c. Deletion discussions in Wikipedia: Decision factors and outcomes. In *Proceedings of the International Symposium on Wikis and Open Collaboration*, WikiSym 2012, pages 17:1–17:10.

Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the ACM conference on Computer Supported Cooperative Work*, CSCW 2013, pages 1069–1080.

Jodi Schneider, Carol Collins, Lisa Hines, John R Horn, and Richard Boyce. 2014a. Modeling arguments in scientific papers. In *The 12th ArgDiaP Conference: From Real Data to Argument Mining*.

Jodi Schneider, Serena Villata, and Elena Cabrio. 2014b. Why did they post that argument? Communicative intentions of Web 2.0 arguments. In *Arguing on the Web 2.0 at the 8th ISSA Conference on Argumentation*.

Jodi Schneider. 2014a. Automated argumentation mining to the rescue? Envisioning argumentation and decision-making support for debates in open online collaboration communities. In *Proceedings of the First Workshop on Argumentation Mining*, pages 59–63, Baltimore, Maryland, June. Association for Computational Linguistics.

Jodi Schneider. 2014b. *Identifying, Annotating, and Filtering Arguments and Opinions in Open Collaboration Systems*. Ph.D. dissertation, Digital Enterprise Research Institute, National University of Ireland, Galway. Corpus and supplementary material also available online at `http://purl.org/jsphd`.

Adam Wyner and Jodi Schneider. 2012. Arguing from a point of view. In *First International Conference on Agreement Technologies*.

Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In *Computational models of argument: Proceedings of COMMA 2012*.

Mining Fine-grained Argument Elements

Adam Wyner
Department of Computing Science
University of Aberdeen
Meston Building, Meston Walk
Aberdeen, AB24 3UK, Scotland
azwyner@abdn.ac.uk

Abstract

The paper discusses the architecture and development of an Argument Workbench, which supports an analyst in reconstructing arguments from across textual sources. The workbench takes a semi-automated, interactive approach searching in a corpus for fine-grained argument elements, which are concepts and conceptual patterns in expressions that are associated with argumentation schemes. The expressions can then be extracted from a corpus and reconstituted into instantiated argumentation schemes for and against a given conclusion. Such arguments can then be input to an argument evaluation tool.

1 Introduction

We have large corpora of unstructured textual information such as in consumer websites (Amazon), newspapers (BBC's "Have Your Say", or in policy responses to public consultations. The information is *complex*, *high volume*, *fragmentary*, and either *linearly* (Amazon or BBC) or *alinearly* (policy responses) presented as a series of comments or statements. Given the lack of structure of the corpora, the cumulative *argumentative* meaning of the texts is obscurely distributed across texts. In order to make coherent sense of the information, the content must be extracted, analysed, and restructured into a form suitable for further formal and automated reasoning (e.g. ASPARTIX (Egly et al., 2008) that is grounded in Argumentation Frameworks (Dung, 1995)). There remains a significant *knowledge acquisition bottleneck* (Forsythe and Buchanan, 1993) between the textual source and formal representation.

Argumentation text is rich, multi-dimensional, and fine-grained, consisting of (among others): a range of (explicit and implicit) *discourse relations between statements in the corpus*, including indicators for conclusions and a premises; speech acts and propositional attitudes; contrasting sentiment terminology; and domain terminology that is represented in the verbs, nouns, and modifiers of sentences. Moreover, linguistic expression is various, given alternative syntactic or lexical forms for related semantic meaning. It is difficult for people to reconstruct argument from text, and even moreso for a computer.

Yet, the presentation of argumentation in text is not a random or arbitrary combination of such elements, but is somewhat structured into reasoning patterns, e.g. defeasible argumentation schemes (Walton, 1996). Furthermore, the scope of linguistic variation is not unlimited, nor unconstrained: diathesis alternations (related syntactic forms) appear in systematic patterns (Levin, 1993); a thesarus is a finite compendium of lexical semantic relationships (Fellbaum, 1998); discourse relations (Webber et al., 2011) and speech acts (Searle and Vanderveken, 1985) (by and large) signal systematic semantic relations between sentences or between sentences and contexts; and the expressivity of contrast and sentiment is scoped (Horn, 2001; Pang and Lee, 2008). A more open-ended aspect of argumentation in text is domain knowledge that appears as terminology. Yet here too, in a given corpus on a selected topic, discussants demonstrate a high degree of topical coherence, signalling that similar or related conceptual domain models are being deployed. Though argumentation text is complex and coherence is obscured, taken together it is also underlyingly highly organised; after all, people do argue, which is meaningful only where there is some understanding about what is being argued about and how the meaning of their arguments is linguistically conveyed. Without such underlying organisation, we could not successfully reconstruction and evaluate arguments from source materials, which is contrary to what is accomplished in argument analysis.

The paper proposes that the elements and structures of the lexicon, syntax, discourse, argumentation, and domain terminology can be deployed to support the identification and extraction of relevant fine-grained textual passages from across complex, distributed texts. The passages can then be reconstituted into instantiated argumentation schemes. It discusses an argument workbench that takes a semi-automated, interactive approach, using a text mining development environment, to flexibly query for concepts (i.e. semantically annotated) and patterns of concepts within sentences, where the concepts and patterns are associated with argumentation schemes. The concepts and patterns are based on the linguistic and domain information. The results of the queries are extracted from a corpus and interactively reconstituted into instantiated argumentation schemes for and against a given conclusion. Such arguments can then be input to an argument evaluation tool. From such an approach, a "grammar" for arguments can be developed and resources (e.g. gold corpora) provided.

The paper presents a sample use case, elements and structures, tool components, and outputs of queries. Broadly, the approach builds on (Wyner et al., 2013; Wyner et al., 2014; Wyner et al., 2012). The approach is contrasted against statistical/machine learning, high level approaches that specify a grammar, and tasks to annotate single passages of argument.

2 Tool Development and Use

In this section, some of the main elements of the tool and how it is used are briefly outlined.

2.1 Use Case and Materials

The sample use case is based on Amazon consumer reviews about purchasing a camera. Consumer reviews can be construed as presenting arguments concerning a decision about what to buy based on various factors. Consumers argue in such reviews about what features a camera has, the relative advantages, experiences, and sources of misinformation. These are qualitative, linguistically expressed arguments.

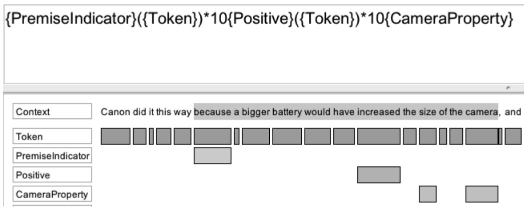
2.2 Components of Analysis

The analysis has several subcomponents: a consumer argumentation scheme, discourse indicators, sentiment terminology, and a domain model. The consumer argumentation scheme (CAS) is derived from the value-based practical reasoning argumentation scheme (Atkinson and Bench-Capon, 2007); it represents the arguments for or against buying the consumer item relative to preferences and values. A range of explicit discourse indicators (Webber et al., 2011) are automatically annotated, such as those signalling premise, e.g. *because*, conclusion e.g. *therefore*, or contrast and exception, e.g. *not, except*. Sentiment terminology (Nielsen, 2011) is signalled by lexical semantic contrast: *The flash worked poorly* is the semantic negation of *The flash worked flawlessly*, where *poorly* is a negative sentiment and *flawlessly* is a positive sentiment. Contrast indicators can similarly be used. Domain terminology specifies the objects and properties that are relevant to the users. To some extent the terminology can be automatically acquired (term frequency) or manually derived and structured into an ontology, e.g from consumer report magazines or available ontologies. Given the modular nature of the analysis as well as the tool, auxilary components can be added such as speech act verbs, propositional attitude verbs, sentence conjunctions to split sentences, etc. Each such component adds a further dimension to the analysis of the corpus.

2.3 Components of the Tool

To recognise the textual elements of Section 2.2, we use the GATE framework (Cunningham et al., 2002) for language engineering applications. It is an open source desktop application written in Java that provides a user interface for professional linguists and text engineers to bring together a wide variety of natural language processing tools in a pipeline and apply them to a set of documents. Our approach to GATE tool development follows (Wyner and Peters, 2011). Once a GATE pipeline has been applied to a corpus, we can view the annotations of a text either *in situ* or extracted using GATE's ANNIC (ANNotations In Context) corpus indexing and querying tool.

In GATE, the gazetteers associate textual passages in the corpus that match terms on the lists with an annotation. The annotations introduced by gazetteers are used by JAPE rules, creating anno-



```
{PremiseIndicator}({Token})*10{Positive}({Token})*10{CameraProperty}
```

Context: Canon did it this way because a bigger battery would have increased the size of the camera, and

Figure 1: Query and Sample Result

tations that are visible as highlighted text, can be reused to construct higher level annotations, and are easily searchable in ANNIC. Querying for an annotation or a pattern of annotations, we retrieve all the terms with the annotation.

2.4 Output and Queries

The ANNIC tool indexes the annotated text and supports semantic querying. Searching in the corpus for single or complex patterns of annotations returns all those strings that are annotated with pattern along with their context and source document. Complex queries can also be formed. A query and a sample result appear in Figure 1, where the query finds all sequences where the first string is annotated with *PremiseIndicator*, followed some tokens, then a string annotated with *Positive* sentiment, some tokens, and finally ending with a string that is annotated as *CameraProperty*. The search returned a range of candidate structures that can be further scrutinised; the query can be iteratively refined to zero on in other relevant passages. The example can be taken as part of a positive justification for buying the camera. The query language (the language of the annotations) facilitates complex search for any of the annotations in the corpus, enabling exploration of the statements in the corpus.

2.5 Analysis of Arguments and their Evaluation

The objective of the tool is to find specific patterns of terminology in the text that can be used to instantiate the CAS argumentation scheme both for and against purchase of a particular model of camera. We iteratively search the corpus for properties, instantiate the argumentation scheme, and identify attacks. Once we have instantiated arguments in attack relations, we may evaluate the argumentation framework. Our focus in this paper is the identification of arguments and attacks from the source material rather than evaluation. It is important to emphasise that we provide an analyst's *support tool*, so some degree of judgement is required.

From the results of queries on the corpus, we have identified the following premises bearing on *image quality*, where we paraphrase the source and infer the values from context. Agents are also left implicit, assuming that a single agent does not make contradictory statements. The premises instantiate the CAS in a positive form, where *A1* is an argument for buying the camera; similarly, we can identify statements and instantiated argumentation schemes against buying the camera.

A1. P1: The pictures are perfectly exposed.
P2: The pictures are well-focused.
V1: These properties promote image quality.
C1: Therefore, you (the reader) should by the Canon SX220.

Searching in the corpus we can find statements contrary to the premises in A1, constituting an attack on A1. To defeat these attacks and maintain A1, we would have to search further in the corpus for contraries to the attacks. Searching for such statements and counterstatements is facilitated by the query tool.

3 Discussion

The paper presents an outline of an implemented, semi-automatic, interactive rule-based text analytic tool to support analysts in identifying fine-grained, relevant textual passages that can be reconstructed into argumentation schemes and attacks. As such, it is not evaluated with respect to *recall* and *precision* (Mitkof, 2003) in comparison to a gold standard, but in comparison to user facilitation (i.e. analysts qualitative evaluation of using the tool or not), a work that remains to be done. The tool is an advance over graphically-based argument extraction tools that rely on the analysts' unstructured, implicit, non-operationalised knowledge of discourse indicators and content (van Gelder, 2007; Rowe and Reed, 2008; Liddo and Shum, 2010; Bex et al., 2014). There are logic programming approaches that automatically annotate argumentative texts: (Pallotta and Delmonte, 2011) classify statements according to rhetorical roles using full sentence parsing and semantic translation; (Saint-Dizier, 2012) provides a rule-oriented approach to process specific, highly structured argumentative texts. (Moens et al., 2007) manually annotates legal texts then constructs a grammar that is tailored to automatically annotated the passages. Such rule-oriented approaches do not use argumentation schemes or domain models; they do not straightforwardly provide for complex annotation querying; and they are stand-alone tools that are not integrated with other NLP tools.

The interactive, incremental, semi-automatic approach taken here is in contrast to statistical/machine learning approaches. Such approaches rely on prior creation of *gold standard corpora* that are annotated manually and adjudicated (considering interannotator agreement). The gold standard corpora are then used to induce a model that (if succesful) annotates corpora comparably well to the human annotation. For example, where sentences in a corpora are annotated as premise or conclusion, the model ought also to annotate the sentences similarly; in effect, what a person uses to classify a sentence as premise or conclusion can be acquired by the computer. Statistical approaches yield a probability that some element is classified one way or the other; the *justification*, such as found in a rule-based system, for the classification cannot be given. Moreover, refinement of results in statistical approaches rely on enlarging the training data. Importantly, the rule-based approach outlined here could be used to support the creation of gold standard corpora on which statistical models can be trained. Finally, we are not aware of statistical models that support the extraction of the fine-grained information that appears to be required for extracting argument elements.

We should emphasis an important aspect of this tool in relation to the intended use on corpora. The tool is designed to apply to reconstruct or construct arguments that are identified in *complex*, *high volume*, *fragmentary*, and *alinearly* presented comments or statements. This is in contrast to many approaches that, by and large, follow the structure of arguments within a particular (large and complex) document, e.g. the BBC's Moral Maze (Bex et al., 2014), manuals (Saint-Dizier, 2012), and legal texts (Moens et al., 2007). In addition, the main focus of our tool is not just the premise-claim relationship, but rich conceptual patterns that indicate the content of expressions and are essential in instantiating argumentation schemes.

The development of the tool can proceed modularly, adding argumentation schemes, developing more articulated domain models, disambiguating discourse indicators (Webber et al., 2011), introducing auxilary linguistic indicators such as other verb classes, and other parts of speech that distinguish sentence components. The tool will be applied to more extensive corpora and have output that is associated with argument graphing tools. More elaborate query patterns could be executed to derive more specific results. In general, the openness and lexibility of the tool provide a platform for future, detailed solutions to a range of argumentation related issues.

References

[Atkinson and Bench-Capon2007] Katie Atkinson and Trevor J. M. Bench-Capon. 2007. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10-15):855–874.

[Bex et al.2014] Floris Bex, Mark Snaith, John Lawrence, and Chris Reed. 2014. Argublogging: An application for the argument web. *J. Web Sem.*, 25:9–15.

[Cunningham et al.2002] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, pages 168–175.

[Dung1995] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358.

[Egly et al.2008] Uwe Egly, Sarah Alice Gaggl, and Stefan Woltran. 2008. Answer-set programming encodings for argumentation frameworks. *Argument and Computation*, 1(2):147–177.

[Fellbaum1998] Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

[Forsythe and Buchanan1993] Diana E. Forsythe and Bruce G. Buchanan. 1993. Knowledge acquisition for expert systems: some pitfalls and suggestions. In *Readings in knowledge acquisition and learning: automating the construction and improvement of expert systems*, pages 117–124. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Horn2001] Laurence Horn. 2001. *A Natural History of Negation*. CSLI Publications.

[Levin1993] Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

[Liddo and Shum2010] Anna De Liddo and Simon Buckingham Shum. 2010. Cohere: A prototype for contested collective intelligence. In *ACM Computer Supported Cooperative Work (CSCW 2010) - Workshop: Collective Intelligence In Organizations - Toward a Research Agenda*, Savannah, Georgia, USA, February.

[Mitkof2003] Ruslan Mitkof, editor. 2003. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

[Moens et al.2007] Marie-Francine Moens, Erik Boiy, Raquel Mochales-Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law (ICAIL '07)*, pages 225–230, New York, NY, USA. ACM Press.

[Nielsen2011] Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

[Pallotta and Delmonte2011] Vincenzo Pallotta and Rodolfo Delmonte. 2011. Automatic argumentative analysis for interaction mining. *Argument and Computation*, 2(2-3):77–106.

[Pang and Lee2008] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January.

[Rowe and Reed2008] Glenn Rowe and Chris Reed. 2008. Argument diagramming: The Araucaria Project. In Alexandra Okada, Simon Buckingham Shum, and Tony Sherborne, editors, *Knowledge Cartography: Software Tools and Mapping Techniques*, pages 163–181. Springer.

[Saint-Dizier2012] Patrick Saint-Dizier. 2012. Processing natural language arguments with the <TextCoop> platform. *Argument & Computation*, 3(1):49–82.

[Searle and Vanderveken1985] John Searle and Daniel Vanderveken. 1985. *Foundations of Illocutionary Logic*. Cambridge University Press.

[van Gelder2007] Tim van Gelder. 2007. The rationale for Rationale. *Law, Probability and Risk*, 6(1-4):23–42.

[Walton1996] Douglas Walton. 1996. *Argumentation Schemes for Presumptive Reasoning*. Erlbaum, Mahwah, N.J.

[Webber et al.2011] Bonnie Webber, Markus Egg, and Valia Kordoni. 2011. Discourse structure and language technology. *Natural Language Engineering*, December. Online first.

[Wyner and Peters2011] Adam Wyner and Wim Peters. 2011. On rule extraction from regulations. In Katie Atkinson, editor, *Legal Knowledge and Information Systems - JURIX 2011: The Twenty-Fourth Annual Conference*, pages 113–122. IOS Press.

[Wyner et al.2012] Adam Wyner, Jodi Schneider, Katie Atkinson, and Trevor Bench-Capon. 2012. Semi-automated argumentative analysis of online product reviews. In *Proceedings of the 4th International Conference on Computational Models of Argument (COMMA 2012)*, pages 43–50. IOS Press.

[Wyner et al.2013] Adam Wyner, Tom van Engers, and Anthony Hunter. 2013. Working on the argument pipeline: Through flow issues between natural language argument, instantiated arguments, and argumentation frameworks. In ??, editor, *Proceedings of the Workshop on Computational Models of Natural Argument*, volume LNCS, pages ??–?? Springer. To appear.

[Wyner et al.2014] Adam Wyner, Katie Atkinson, and Trevor Bench-Capon. 2014. A functional perspective on argumentation schemes. In Peter McBurney, Simon Parsons, and Iyad Rahwan, editors, *Post-Proceedings of the 9th International Workshop on Argumentation in Multi-Agent Systems (ArgMAS 2013)*, pages ??–?? To appear.

Genre distinctions and discourse modes: Text types differ in their situation type distributions

Alexis Palmer and Annemarie Friedrich
Department of Computational Linguistics
Saarland University, Saarbrücken, Germany
{apalmer,afried}@coli.uni-saarland.de

Abstract

In this paper we explore the relationship between the genre of a text and the types of situations introduced by the clauses of the text, working from the perspective of the theory of discourse modes (Smith, 2003). The typology of situation types distinguishes between, for example, events, states, generic statements, and speech acts. We analyze texts of different genres from two English text corpora, the Penn Discourse TreeBank (PDTB) and the Manually Annotated SubCorpus (MASC) of the Open American National Corpus. Texts of different types – genres in the PDTB and subcorpora in MASC – are segmented into clauses, and each clause is labeled with the type of situation it introduces to the discourse. We then compare the distribution of situation types across different text types, finding systematic differences across genres. Our findings support predictions of the discourse modes theory and offer new insights into the relationship between text types and situation type distributions.

1 Introduction

Language is not a unitary phenomenon, and patterns of language use change according to the type of text under investigation. In natural language processing, furthermore, it has been shown that there are strong effects from both the domain and the genre of texts on the performance of systems performing automatic analysis. These effects are relevant at nearly all levels of analysis, from part-of-speech tagging to discourse parsing, yet they are in some ways poorly understood. For example, there is no single agreed-upon set of text types that suits all levels of analysis, nor are we aware of systematic guidelines for sorting texts into genre categories; this process often relies on human intuition and the claim that "I know [a document of type X] when I see one."

Rather than conceptualizing text type purely as a document-level characteristic, in this study we take inspiration from a theory which targets *text passages* as an intermediate level of representation. The idea is that most texts are in fact a mix of passages of different types. For example, a news story may begin with a short narrative passage which focuses on one individual's reaction to the newsworthy event and then proceed with a more informative discussion of the topic at hand. Smith (2003) identifies five different types of text passages, or **discourse modes**, each of which is associated with certain linguistic characteristics of the text passage. (See Sec. 2 for more on the modes and the linguistic characteristics.) This study investigates how closely the predicted linguistic characteristics of certain text types are reflected in a body of naturally occurring texts.

We focus on genre differences at the level of the clause, considering the types of situations introduced to the discourse by clauses of text. According to Smith, the situation (or **situation entity**) types presented in a text are an important characteristic for distinguishing between the different types of text passages. Using two sets of documents (see Sec. 3) with genre labels, we investigate the distributions of situation types (see Sec. 2.1 for the inventory of situation types) for the different text types. We find systematic differences between news/jokes texts on the one hand and essay/persuasive texts on the other, as the theory predicts. In the final section of the paper, we briefly discuss potential applications of these findings to argumentation mining.

Mode	Distribution of SEs	Progression
NARRATIVE	mostly Event, State	SEs relate to one another; dynamic events advance narrative time
REPORT	mostly Event, State, General Stative	SEs related to Speech Time; time progresses forward & backward from that time
DESCRIPTION	mostly Event, State, ongoing Event	Time is static; text progresses in spatial terms through the scene described
INFORMATION	mostly General Stative	atemporal; progressing on a metaphoric path through the domain of the text
ARGUMENT / COMMENTARY	mostly General Stative, Fact, Proposition	atemporal; progressing on a metaphoric path through the domain of the text

Table 1: Discourse modes and their linguistic correlates according to Smith (2005).

2 Discourse modes: a theory of text passages and their types

Smith (2003) proposes to analyze discourse at the level of the text passage, viewing each individual text as a mixture of text passages. These passages are contiguous regions of text, generally one or more paragraphs, with particular discourse functions. Each passage belongs to one of five discourse modes: NARRATIVE, REPORT, DESCRIPTION, INFORMATION, ARGUMENT/COMMENTARY. Importantly, the modes can be characterized according to two broad classes of linguistic correlates: the mode of progression through the text passage (roughly temporal or atemporal), and the distribution of situation entity types. The modes and their correlates appear in Table 1.

2.1 Situation entities

In this work we are directly concerned with the second type of linguistic correlate: the situation entities. A situation entity (SE) can be thought of as the abstract object introduced to the discourse by a clause of text. The type of the SE introduced by a clause depends on, among other things, the internal temporal properties of the verb and its arguments. The interpretation of the verb constellation may of course by influenced by adverbials and other linguistic factors. We are primarily interested in finite clauses, for the most part assuming that each clause introduces one SE.[1]

The SE types fall into four broad categories. **Eventualities** describe particular situations such as Events (1) or States (2).

(1) The tour guide pointed to the mosaic. (EVENT)

(2) The view from the castle is spectacular. (STATE)

The class of **General Statives** includes Generalizing Sentences (3), which report regularities, and Generic Sentences (4), which make statements about kinds or classes.

(3) Silke often feeds my cats. (GENERALIZING SENTENCE)

(4) The male cardinal has a black beak. (GENERIC SENTENCE)

The third class of SE types are **Abstract Entities**, which differ from the other SE types in how they relate to the world: Eventualities and General Statives are located spatially and temporally in the world, but Abstract Entities are not. Facts (5) are objects of knowledge, and Propositions (6) are objects of belief. In the following examples, the underlined clauses introduce Abstract Entities to the discourse.

(5) I know that his plane arrived at 11:00. (FACT)

(6) I believe that his plane arrived at 11:00. (PROPOSITION)

Finally, we introduce the category **Speech Acts** for clauses whose main function is performative: namely, Questions (7) and Imperatives (8).

(7) Why is it so? (QUESTION)

[1] For a more detailed discussion of situation entities, please see Friedrich and Palmer (2014b). For even more information, see our project page (`http:\\sitent.coli.uni-saarland.de`) and the references cited there, including a detailed annotation manual.

(8) Please sign and return to the sender. (Imperative)

2.2 Linking situation types and discourse modes: what does the theory predict?

The broad aim of this study is to compare the predictions of the theory to evidence from text corpora, in particular with respect to the distributions of SEs across different text types. We focus on two modes: Report and Argument/Commentary. For the Report mode, the expectation is that text passages should be made up primarily of Eventualities (Events and States) with some General Statives. The most frequent SE types in the Arg/Comm mode, on the other hand, should be primarily Abstract Entities (Facts and Propositions) and General Statives.

To date there is no large body of data annotated with discourse modes. Therefore, we instead look directly at the distributions of SEs within text passages for which we have annotated data (Friedrich and Palmer, 2014b), taking the genre category assigned within our text corpora as a proxy for discourse mode. We do this under the assumption that some genres are associated with a certain predominant discourse mode. From that assumption, we consider the average SE distributions per text type to reflect the distributions expected from the predominant mode. Specifically, we map texts from the genres news and jokes to the Report mode, and essays and fundraising letters to the Arg/Comm mode.

3 Data for corpus study

We test the predictions of the theory on sets of texts extracted from two different corpora, described below. These corpora were chosen in large part because they both group their texts according to genre. Although the two corpora use a different set of genre labels, both cover the two broad categories we are interested in. Annotation and analysis of the two data sets are described in Sec. 4.

3.1 Penn Discourse TreeBank

The Penn Discourse TreeBank (PDTB) (Prasad et al., 2008) provides annotations of discourse structure over a collection of texts from the Wall Street Journal; these texts are from the Penn TreeBank (Marcus et al., 1993), one of the most widely-used annotated corpora in natural language processing. In addition to discourse structure annotations, PDTB texts are hand-labeled with part-of-speech tags, syntactic structure, and, as of relatively recently, genre designations. Webber (2009) found that the texts in PDTB belong to a number of different categories and, further, that the discourse relations marked in the texts pattern according to the genre of the text. In fact, Webber (2009) inspired the current study, raising the question of whether the SE type distributions found in texts similarly reflect the genre of the text.

PDTB	*news*	790
	essays	1723
MASC	*news*	2563
	jokes	3453
	essays	2404
	letters	1850

Table 2: Number of SE-bearing clauses analyzed per corpus, per genre.

The PDTB texts are predominantly from the *news* genre (roughly 1900 texts), with much smaller numbers of texts from four other genres: *essays* (roughly 170 texts), *letters* (roughly 60 texts), *highlights* (roughly 40 texts), and *errata* (25 texts). From these, we extract 20 news texts and 20 essay texts to be used in our study.

3.2 Manually Annotated Sub-Corpus

The second corpus used in this study is MASC (Ide et al., 2008), the Manually Annotated Sub-Corpus of the Open American National Corpus.[2] Overall, MASC contains roughly 500,000 words of text (both written text and transcribed speech), balanced over 19 text types. In addition to manually-checked annotations of sentence and word boundaries, part-of-speech tags, named entities, and both shallow and deeper syntactic structure, some portions of MASC have been annotated for a number of semantic and pragmatic phenomena. For this study, though, we use only the genre labels and our own SE annotations (see Sec. 4).

For our study, we extract texts from the written part of MASC. We use the texts from four of the genres: *news*, *jokes*, *essays*, and *letters*. The letters fall into two sub-categories (*philanthropic-fundraising* and *solicitation-brochures*), though all of the letters have the same general goal of soliciting donations, whether of money, time, or goods.

[2] `http://www.anc.org/data/masc`

4 Corpus study

In this section we describe the segmentation and annotation of the data, the situation type inventories reflected in the analysis, and the methodology used for computing results. We then present and discuss our findings.[3]

4.1 Segmentation and annotation

Having selected texts for analysis, we next segmented them into clauses, again following the assumption of one SE per clause (with a few exceptional cases). The PDTB texts were segmented manually by the annotator, and the MASC texts using SPADE (Soricut and Marcu, 2003) with some heuristic post-processing. Each clause was then manually labeled with its SE type.

The PDTB annotations were performed by one paid annotator with extensive background in linguistics, with ample training time but only a minimal annotation manual.

The MASC annotations are part of a large ongoing annotation project with multiple paid annotators, an extensive manual, and a structured training phase. In the latter, we take a feature-driven approach to annotation which improves the quality of the annotations, leading to substantial inter-annotator agreement (see Table 3). In addition to the SE type label, annotators mark each clause with three relevant linguistic features, which are not used in the current study, but which guide the annotators to find the best-fitting SE type label. These are inherent lexical aspect of the verb (Friedrich and Palmer, 2014a), genericity of the main referent, and habituality of the event described. Details regarding the annotation scheme and the benefits of feature-driven annotation appear in Friedrich and Palmer (2014b).

genre	clauses	Kappa
news	2563	0.667
jokes	3453	0.756
essays	2404	0.493
letters	1850	0.612

Table 3: Number of clauses, inter-annotator agreement (Cohen's Kappa) for MASC subcorpora.

4.2 SE inventories

Each of the two analyses uses a slightly different set of SE types. The main difference between the two is that for the PDTB data annotations were done mostly at a coarse-grained level, and the MASC annotations are more fine-grained.

The PDTB analysis remains close to the inventory of SE types presented in Sec. 2.1, with the modification that three of the four coarse-grained categories (i.e. General Statives, Abstract Entities, and Speech Acts) are treated as SE types. In other words, for each of these categories, we conflate its subtypes into a single higher-level type. States and Events are treated as separate categories. The coarse-grained analysis still captures the relevant distinctions yet allows us to make useful generalizations over the relatively small amount of data.

For MASC, we return to a fine-grained analysis. General Statives and Speech Acts are counted at the fine-grained level, and Abstract Entities do not appear in the analysis at all. We add the REPORT type of situation entity, which is a subtype of EVENTS, designed to capture cases like (9).

(9) ..., said the President of the Squash Association. (REPORT)

4.3 Method

For both data sets, we compute the distributions of SE types per genre. For each genre, we collect the counts of situation entity types assigned and then compute the corresponding percentages. For the PDTB data (Figure 2), this is a straightforward analysis, as there was only one annotator.

For MASC (Figure 1), we use the annotations of two annotators to compute the distributions. Annotators are allowed to mark a segment with multiple situation types; we simply use all markings of types to compute the percentages. When annotators disagree, we do not adjudicate but rather count both annotations; when they do agree, we counts two instance of the agreed-upon label. Hence, the statistics presented in Figure 1 present an average over the two annotator's assignments. The distributions shown in Figure 1 all differ significantly ($p < 0.01$) from each other according to a χ^2-test, which means that the SE type distributions of the genres are all significantly different from each other: text types differ in their situation type distributions.

[3] Results from the PDTB portion of the analysis were first presented at the 2009 Texas Linguistics Society conference in Austin, Texas.

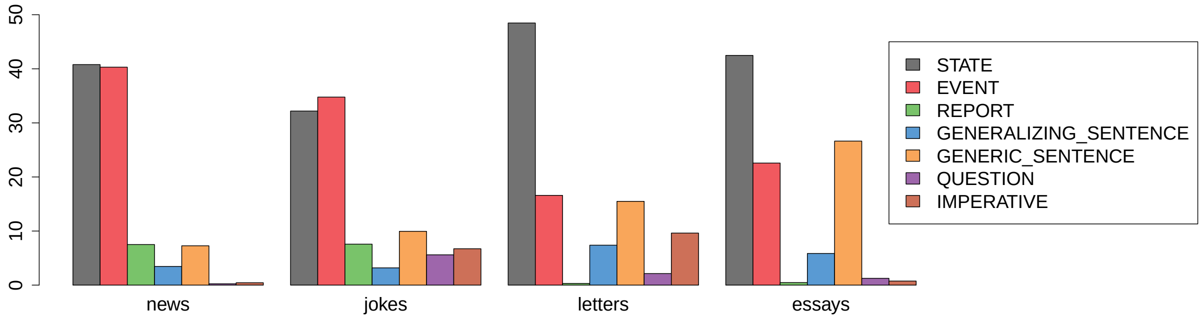


Figure 1: Distributions of situation entity types in four MASC genres.

4.4 Findings

The broad finding is that General Statives play a predominant role for texts associated with the ARGUMENT/COMMENTARY mode, and Events and States for texts associated with the REPORT mode. With these results, we begin to replace the vague distributional statements in Table 1 with more precise characterizations of SE type distributions.

We first compare the two genres shared across both data sets: *news* and *essays*. For both data sets, we see that the proportion of Eventualities is highest for the news genre, and that within Eventualities, Events are more frequent than States.[4] This supports the theoretical claim that passages in REPORT mode predominantly consist of Events and States. Smith (2005) also predicts a significant number of General Statives for REPORT passages; in our study we observe these types in the news texts, but less frequently than Eventualities.[5]

We see more General Statives in essays than in news. The predominance of General Statives is not surprising, given that arguments are frequently built from generalizations and statements about classes or kinds. An interesting result that is not predicted by the theory is that in essays, States are much more frequent than Events. Together with the higher prevelance of General Statives, this suggests that essays rely heavily on describing and discussing states of affairs rather than particular actions or events.

Now we turn to the two additional genres in MASC: *jokes* and *letters*. First it should be noted that it's not clear whether a distinction should be made between (persuasive) essays and the persuasive letters that appear in MASC. Second, we can see that the predominance of State-type SEs is even stronger for letters than it is for essays. In addition, we see that letters use more generalizing statements and fewer generics, and a rather high proportion of Imperatives. The expected distribution of Imperatives is not explicitly treated by the theory, but one can easily imagine the sorts of Imperative statements that would appear in fundraising and solicitation letters: e.g. "Send a check now! Don't delay! Save the whales!"

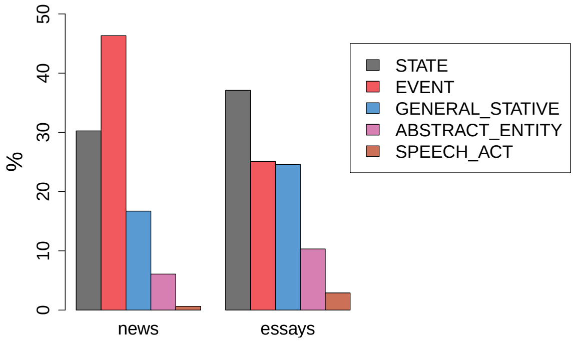


Figure 2: Distributions of situation entity types in two PDTB genres.

Jokes are interesting in that they pattern quite similarly to news texts, but with a higher proportion of Speech Act types. The latter can be attributed to the fact that jokes contain more direct and reported speech than news.

5 Discussion and conclusion

The corpus study described above investigates, across two different datasets of written English text, the relationship between situation entities and text type on the basis of the available data. In

[4] For MASC this second result comes from conflating the categories of Event and Report.

[5] It would be interesting to compare this distribution to texts from another mode (e.g. NARRATIVE) for which Smith (2005) does not predict many General Statives in order to determine the relative importance of General Statives in the REPORT mode.

both cases, and taking genre as a proxy for discourse mode, we find support for Smith's theoretical prediction that different types of text show different characteristic distributions of the types of SEs introduced by the clauses of the text. We find this specifically for two broad text types: news/jokes (mapped to the REPORT mode of discourse) and essays/persuasive texts (mapped to the ARGUMENT/COMMENTARY mode of discourse). The current study analyzes SE distributions over collections of texts; a logical next step is to do this analysis in a more fine-grained fashion, associating SE distributions with text passages labeled with discourse modes. This would remove the need for the genre-as-proxy assumption and move us even further toward a clearer understanding of how discourse modes and situation entity types pattern together.

In future work, we plan to create automatic methods to label clauses with their SE type, which could then be used to automatically identify the types of text passages present in documents.

Relevance for argumentation mining

Some current research in argumentation mining investigates the question of whether performance for automatically extracting argument components from text improves when a system can first narrow down the search space to the argumentative regions of the document. (For example, see Stab and Gurevych (2014) and Levy et al. (2014).) Our finding that essays and persuasive texts show a different distribution of SE types than news texts suggests one way to approach the challenge of finding the argumentative portions of texts.

So far work in argumentation mining has focused predominantly on finding arguments in argumentative texts: opinion pieces, argumentative essays, editorials, and the like. This is to some extent a limiting assumption, as texts from a wide range of genres can in fact contain argumentative passages. A method for finding argumentative passages could extend the range of texts available for argumentation mining.

Acknowledgments

For the PDTB case study, we gratefully acknowledge Caroline Sporleder for much interesting and insightful discussion, as well as Todd Shore both for his annotation work and for discussions arising from that work. This study has also benefitted from discussions with Bonnie Webber and Manfred Pinkal. Finally, huge thanks to the participants of the Bertinoro symposium on the intersection of Argumentation Theory and Natural Language Processing for a highly engaging and intellectually stimulating week. This research was supported in part by the MMCI Cluster of Excellence, and the second author is supported by an IBM PhD Fellowship.

References

Annemarie Friedrich and Alexis Palmer. 2014a. Automatic prediction of aspectual class of verbs in context. In *Proceedings of ACL 2014*.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. In *Proceedings of The Linguistic Annotation Workshop*.

Nancy Ide, Collin Baker, Christiane Fellbaum, and Charles Fillmore. 2008. MASC: The manually annotated sub-corpus of American English.

Ran Levy, Yonatan Bilu, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of LREC 2008*.

Carlota S Smith. 2003. *Modes of discourse: The local structure of texts*. Cambridge University Press.

Carlota S Smith. 2005. Aspectual entities and tense in discourse. In *Aspectual Inquiries*, pages 223–237. Springer.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings ACL-HLT 2003*.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014*.

Bonnie Webber. 2009. Genre distinctions for discourse in the Penn TreeBank. In *Proceedings of ACL 2009*.

Scientific argumentation detection as limited-domain intention recognition

Simone Teufel
Computer Laboratory
University of Cambridge
15 JJ Thomson Avenue, Cambridge, UK
Simone.Teufel@cl.cam.ac.uk

Abstract

We describe the task of intention-based text understanding for scientific argumentation. The model of scientific argumentation presented here is based on the recognition of 28 concrete rhetorical moves in text. These moves can in turn be associated with higher-level intentions. The intentions we aim to model operate in the limited domain of scientific argumentation and justification; it is the limitation of the domain which makes our intentions predictable and enumerable, unlike general intentions.

We explain how rhetorical moves relate to higher-level intentions. We also discuss work in progress towards a corpus annotated with limited-domain intentions, and speculate about the design of an automatic recognition system, for which many components already exist today.

1 Introduction

Automatically recognising the structure of an argument is an attractive and challenging task, which has received interest for a long time from the AI as well as the natural language processing community, and recently from both communities together in a joint effort. Because arguments are global text structuring devices, argument recognition has the potential to advance text understanding and the many real-life tasks that could profit from it.

There are various definitions of what an argument is (Toulmin, 1958; Cohen, 1984; Dung, 1995; Brüninghaus and Ashley, 2005; Besnard and Hunter, 2008; Walton et al., 2008; Green, 2014). We are here interested in a definition close to discourse structure, and concentrate in particular on the recognition of prototypical argumentation steps in scientific exposition. We posit that these argumentation steps can be defined at an abstract level so that world knowledge is not required for their recognition.

There is a clear connection between our goal and intention recognition. Fully understanding every aspect of an author's argumentation requires the recognition of all of their intentions, which in turn means that we would have to model, generalise over, and do inference with general world knowledge. This is of course an AI-hard task fraught with many theoretical and practical problems; consider the symbolic AI work on this and closely related problems (e.g., Schank and Abelson, 1977; Pollack, 1986, 1990; Norvig, 1989; Cohen et al., 1990 and Carberry, 1990).

We will propose instead to reframe argumentation detection as a *limited-domain* intention recognition task. The basic building blocks of our model of an argument are instances of higher-level intentions which the authors are likely to have had when they were writing their paper. The representation we suggest for intentions does not contain any propositional content based on arbitrary world knowledge. Instead, our intentions are represented as generalised propositions such as “Our solution is better than the competition's”. Such speech acts realise parts of the author's intention of persuading the reader that the work described in the paper is novel and significant. When during processing we encounter the sentence

To our knowledge, our system is the first one *aimed at building semantic lexicons from raw text without using any additional semantic knowledge.*

(9706013, S-171)

our representation only registers the author's intention of staking a novelty claim for their new work. The proposition is generalised in that the propositional content of the novelty, i.e., the fact that the authors built the first lexicon from raw text without any additional semantic knowledge, is not encoded. This detail is not important at the level of abstraction we have in mind.

The simplification of argument recognition into a limited-domain intention recognition problem is possible because of the high degree of conventionalisation of scientific argumentation. Following Swales (1990), we call explicit statements such as the above novelty claim "rhetorical moves". Rhetorical moves are well-documented in various disciplines: they occur frequently, and they can be enumerated and classified, as applied linguists have done in some detail for several disciplines (e.g., Myers, 1992; Hyland, 1998; Salager-Meyer, 1992).

Swales also coined the expression "research space" – a cognitive construct consisting of scientific problems, methods and research acts that authors use when they locate their research with respect to historical approaches and current trends.

When we faced the decision of which types of semantic participants to encode in our representation of rhetorical moves, we tried to achieve as much generalisation as possible, in line with the Knowledge Claim Discourse Model (KCDM, Teufel, 2010). In fact, the core semantic participants in rhetorical moves can be reduced to just two sets – US (the paper's authors) and THEM (everybody else who has ever published).

When it comes to the states and events expressed in rhetorical moves, we maximally generalise again and end up with four classes of predicates, where the classes are defined based on the number of participants in the logical act expressed in the move. We differentiate statements about the authors' own work (US); statements about others' previous work (THEM); statements about the connection between the authors' work with previous work (US and THEM); and finally statements about the research space and the authors' position in it. Another relevant observation is that rhetorical moves often contain sentiment, in the form of "good" vs. "bad" situations, as well as successful vs. failed problem solving acts.

As far as the representation of time in the events and states described in rhetorical moves is concerned, another simplification is possible: it suffices to model three points in time, the time before the authors' research activity begins (t_0), and the times during (t_1) and after (t_2) their research activity. Of course, the real actions by the authors that gave rise to the research in the paper are spread in time in far more complex ways, but a scientific paper is a social construct (Bazerman, 1985). The telling of "the story" follows the convention that all research acts associated with the paper happen simultaneously, and that they transform an earlier state of the world into a new (better) one.

These simplifications allow us to define the 28 rhetorical moves in Figure 1[1]. We also give some examples of rhetorical moves from the chemistry, computational linguistics and agriculture literature, which were sourced from our annotated corpora.

The overall argumentation structure we propose concerns the author's argument that their research was worthy of publication, and all of its subarguments – which, at its heart, is always the same argument. Argument recognition then corresponds to a guess as to which strategy the author pursued in making this argument. This process will have to be driven by a bottom-up recognition of rhetorical moves, as these are the only explicitly expressed parts of the argument. This will trigger a simple form of inference as to which higher-level intention might have been present during the writing of the paper.

In previous work, we have used a robust classification model called Argumentative Zoning (AZ; Teufel, 2000, 2010; Teufel et al. 2009, O'Seaghdha and Teufel 2014), that turns some aspects of the more general argumentation recognition model of the KCDM into a simple sentence classification task. In AZ, rhetorical moves with a similar function were bundled together into 7 (in later versions 15 or 6) flat classes or zones, and each sentence was classified into one of these on the basis of surface features,

[1] An earlier version of the list of moves appears in Teufel (1998).

	I. Properties of research space
R-1	Problem addressed is a problem
R-2	New goal/problem is new
R-3	New goal/problem is hard
R-4	New goal/problem is important/interesting
R-5	Solution to new problem is desirable
R-6	No solution to new problem exists
	II. Properties of new solution (US)
R-7	New solution solves problem
R-8	New solution avoids problems
R-9	New solution necessary to achieve goal
R-10	New solution is advantageous
R-11	New solution has limitations
R-12	Future work follows from new solution
	III. Properties of existing solution (THEM)
H-1	Existing solution is flawed
H-2	Existing solution does not solve problem
H-3	Existing solution introduces new problem
H-4	Existing solution solves problem
H-5	Existing solution is advantageous
	IV. Relationships between existing and new solutions (US and THEM)
H-6	New solution is better than existing solution
H-7	New solution avoids problems (when existing does not)
H-8	New goal/problem/solution is different from existing
H-9	New goal/problem is harder than existing goal/problem
H-10	New result is different from existing result
H-11	New claim is different from/clashes with existing claim
H-12	Agreement/support between existing and new claim
H-13	Existing solution provides basis for new solution
H-14	Existing solution provides part of new solution
H-15	Existing solution (adapted) provides part of new solution
H-16	Existing solution is similar to new solution

Recently, **R-4** *the use of imines as starting materials in the synthesis of nitrogen-containing compounds has attracted a lot of interest from synthetic chemists.*$^{(1)}$ (b200198e)

H-4 *This account makes reasonably good empirical predictions, though* **H-2** *it does fail for the following examples: . . .* (9503014, S-75)

H-12 *Greater survival of tillers under irrigated conditions agrees with other reports in barley [4,28] and wheat [10,13,26].* (A027)

Figure 1: Rhetorical moves; some examples

including sequence information. This way of phrasing the problem allows for tractable recognition and evaluation. AZ classification has been shown to lead to stable and reliable annotation on several scientific disciplines, and it is also demonstrably useful for a set of applications such as the detection of new ideas in a large scientific area, summarisation, search, and writing assistance.

Nevertheless, AZ is only a flat approximation of a larger argumentation model of scientific justification. The work presented here is a departure from AZ in that it aims to model the stages of scientific argumentation in a more informative, finer-grained way.

2 The role of citations in the argument

The reader may have noticed that the rhetorical moves in parts III and IV of Fig. 1, which are concerned with statements about THEM (i.e., other published authors), are closely connected to citation function[2]. In fact, we have in the past attempted the recognition of some of the H-moves as an isolated task, in the form of citation function classification (CFC; Teufel et al., 2006); others (Garzone and Mercer, 2000; Cohen et al., 2006) have used other schemes for similar citation classification tasks.

Where, how often, and how authors cite previous work is an important aspect of their overall scientific argument. For instance, the authors might choose one of the possible articles types (review, research paper, pioneer work etc) to support a particular point in their overall argument. The choice of a particular pioneer paper might signal their intellectual heritage. They might tell us who their rivals are, and who uses similar methods for a different goal (i.e., not rivals), whose infrastructure they borrow, and whose work supports theirs and vice versa. These questions will crucially influence where in the text (physically and logically in terms of the argumentation) a given citation will occur.

As a result of all this, it is often possible to determine some citations as being particularly central to the authors' paper. This information, if it could be automatically determined from text in a reliable

[2] These 16 moves also follow a different naming scheme, where the move name starts with the letter "H" – historically, such moves were called "hinge" moves, as opposed to the "R" ("rhetorical") moves in parts I and II of Fig 1.

way, would vastly improve bibliographic search. It also has the potential to improve bibliometric assessments of a piece of work's impact, e.g. in the sense of Borgman and Furner (2002), White (2004), and Boyak and Klavans (2010).

3 Higher-level intentions

There are some rhetorical moves that at first glance seem to make litte sense. Stating H-5, praise of other people's work, might comparatively weaken the author's own knowledge claim. Similarly, stating H-9, the fact that the author's research goal is harder than other people's goal, might prompt the criticism that the authors have simply chosen their goal badly – had they chosen an easier goal, the solution might have been easier, or achieved better results.

However, rhetorical moves must be interpreted as part of the larger picture of the overall scientific argument. Scientific writing can be seen as one big game where an author's overall goal is to successfully manoeuvre their paper past the peer review, so that it can be published.

According to the conventions of peer review, there is a small set of criteria for acceptance – the authors need to show that the problem they address is justified (High-Level-Goal 1 or HLG-1 for short), that their knowledge claim is significant (HLG-2) and novel (HLG-3), and that the research methodology they use is sound (HLG-4). If valid evidence for the fulfilment of these criteria is presented, the peer review cannot justifiably reject the paper.

Fig. 2 spells out how the overall argument for validity is put together from high- and medium-level intentions and rhetorical moves[3]. Rhetorical moves in Fig. 2 appear in shaded boxes (H- and R-type moves in different shades of grey). Above the rhetorical moves, we see a simple representation of the intentions posited in the model. For simplicity and readability, Fig. 3 repeats the same network without rhetorical moves. The arrows in both figures express the "supports" relationship in argumentation theory. For instance, in order to argue for the novelty of one's work, a state-of-the-art comparison may or may not be necessary – this depends on whether one describes the research goal as new or not. For new research goals, one may simply show that no other work is similar enough to one's goal: new goals (created at t_1) cannot be compared to existing state-of-the-art, which is frozen in time at t_0. (Novelty is a rare example of a high-level intention which can be left to the reader to infer, or alternatively stated explicitly as move R-2 or R-6.)

Note that each citation that has an H-type rhetorical move associated with it automatically strengthens the claim that the authors are knowledgeable in the field (one of the important subgoals of HLG-4, soundness). Under our model, citations without any associated H-move are not contributing to this goal, as a knowledgeable author must be able to state the relationship of the current work to earlier work. (A simple statement of similarity with somebody else's work should barely count, but has been given a "weak" move, H-16, because we encountered it so frequently in our corpus studies.)

From Fig. 2 we can now see why stating H-5 can be a good strategic move even though it praises other people's work – it supports HLG-4 (soundness of methodology) via the sub-argument that by including praise-worthy existing work, the authors make sure they use the best methods currently available. Similarly, the statement that one's goal is harder than somebody else's motivates that the authors' chosen problem is justified (HLG-1) and significant (HLG-2), and additionally strengthens HLG-4 (via the claim that the authors know their field well). This illustrates that a rhetorical move can support more than one high-level intention.

4 Knowledge representation of moves and intentions

What has been said so far raises the question of which knowledge representation is most suited for modelling intentions and rhetorical moves. Designing a propositional logic that expresses the full semantics of rhetorical moves and of higher-level intentions is a task that goes far beyond the current paper; it requires a thorough design of the semantics of objects and events/states in this limited domain, as well as an appropriate type of inference. Nevertheless, we will sketch some of the principles of what might be usefully encoded.

The THEM entities would need to be grounded to

[3] An earlier version of this diagram appears as Fig.3.1.7 in Teufel (2000, p.105).

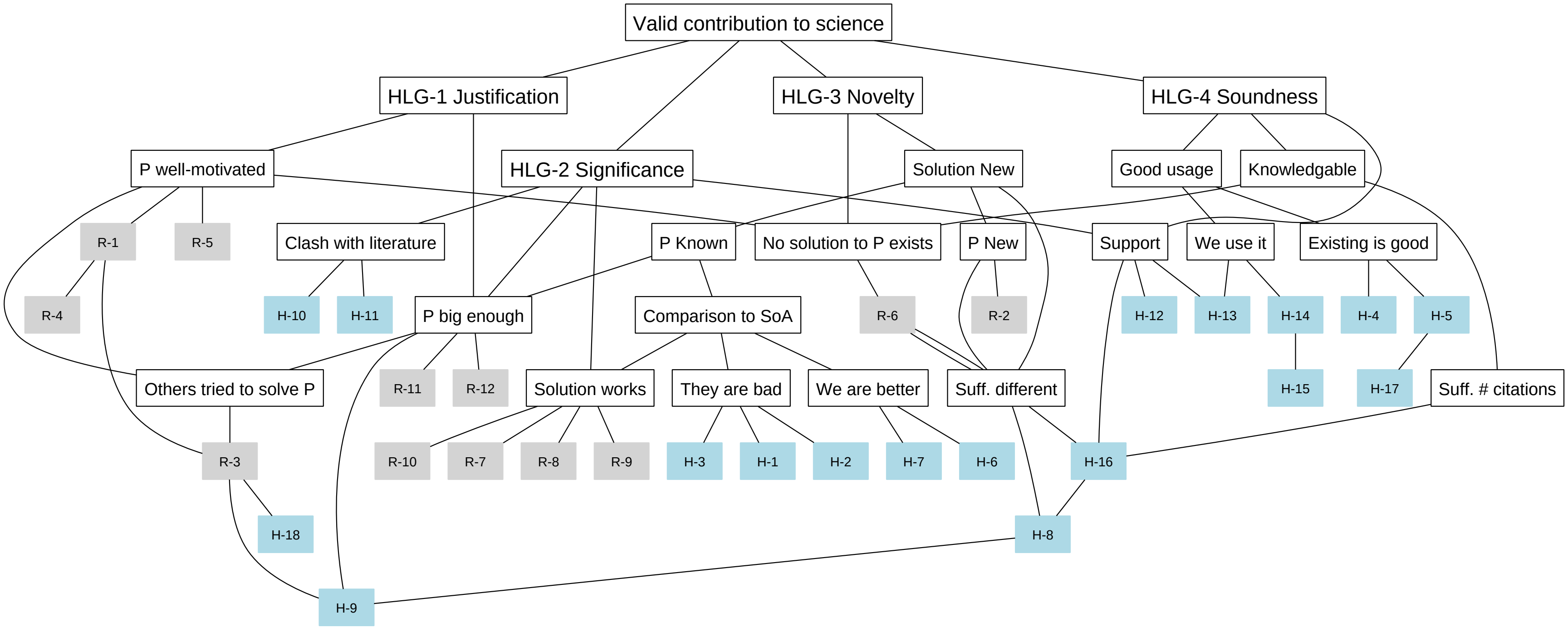


Figure 2: Argumentation network (including rhetorical moves).

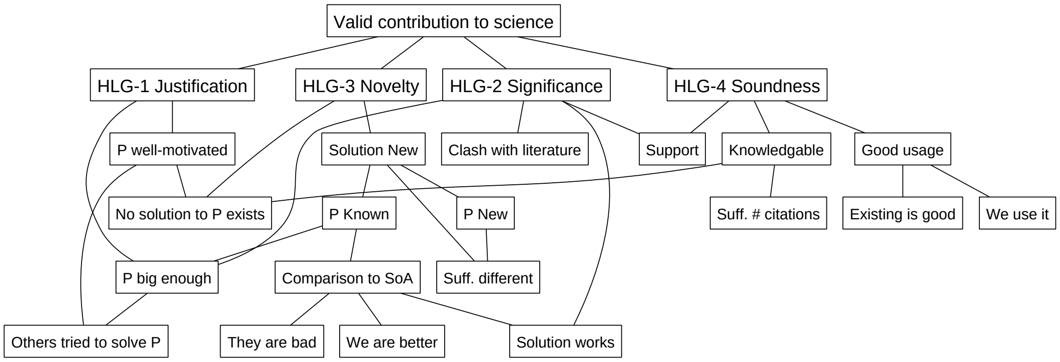


Figure 3: Argumentation network (excluding rhetorical moves).

citations, possibly also to more general entities such as "many linguists in the 1970s". Entities would need to be tracked throughout the paper, for instance by performing co-reference. We would also need to represent problems, solutions and goals as atomic types, i.e., the fact that they are considered problems, solutions and goals, rather than their content. (The system should keep pointers to the textual strings that express this content, so that down-stream processing or human users can gain access to this information.)

The exact representation of a proposition is open to speculation at this point, but moves would likely be decomposed into atomic clauses. Events and properties in the limited domain (such as changing a solution into another one, or the fact that one solution is better than another) would be associated with a time; for instance all actions that logically happen during the research act presented in the paper would be associated with t_1.

Inference could be performed by a theorem prover, which could inhibit or further activate the potentially possible "supports" relationships given in Fig. 1, by taking the plausibility of a particular inference into account, in the light of the textual evidence encountered.

Axioms could directly encode some of the rules of the scientific publication game, such that the existence of a problem is a bad state, that of a solution is a good state, but that a solution *needing* something else is a bad state again. Temporal inference could require axioms such as things that persist at a certain time also persist in later times, unless they are changed.

R-5	solution(s) $\wedge$ solve(s, p, t_1) $\wedge$ good(a, t_2) $\wedge$ aspect(a, s) $\wedge$ problem(p) $\wedge$ address(US$, p$)
R-12	problem(p_1) $\wedge$ cause(s, p_1, t_1) $\wedge$ solution(s) $\wedge$ solve(s, p) $\wedge$ problem(p) $\wedge$ address(US$, p$
H-1	solution(s_1) $\wedge$ own(THEM$, s_1$) $\wedge$ bad(a, t_0) $\wedge$ aspect(a, s) $\wedge$ solve(s_1, p) $\wedge$ problem(p) $\wedge$ address(US$, p$
H-7	solution(s_1) $\wedge$ own(THEM$, s_1$) $\wedge$ solution(s) $\wedge$ own(US$, s$) $\wedge$ $/$ solve(s_1, p, t_0) ($\wedge$ solves(s, p, t_1)
H-15	own(THEM$, s_1$) $\wedge$ solution (s_1) $\wedge$ solution (s_2) $\wedge$ change(US$, s_1, s_2, t_1$) $\wedge$ use(US$, s_2, t_1$)

Figure 4: Sketch of knowledge representation for selected rhetorical moves

As an example of what the representation might look like, Fig. 4 expresses five moves in a simple prepositional logic. Here, ownership of solutions (by US or THEM) is expressed directly, as are simple relationships between solutions, problems, results and claims. Consider move H-15, for instance – adapting somebody else's solution means taking it, changing it into something else, and then using the changed solution. Some moves, such as R-6 and R-9, look like they might require quantification, which exceeds the expressivity of simple predicate logic.

Several aspects of the moves' semantics are not explicitly expressed in text; they could even be modelled as presuppositions. For instance, R-7 states that a rival's solution does not solve one's problem, which presupposes that the author's solution does, otherwise it would not be a relevant statement. R-7 thereby implicitly invokes a comparison between the author's approach and the rivals', which is won by the authors. Crucially, whether or not the authors' successful problem-solving is explicitly mentioned in the text or not is optional. Another example is the need to know whether a problem mentioned in a certain rhetorical move is actually the problem that the authors address in the current paper. This is often decisive, because the knowledge claim of the paper is connected exclusively to this particular problem. In some part of the paper, the authors give us the information which problem it is that they address, but they will typically not repeat this elsewhere.

It is the discourse model's job to accumulate the information about the identity of important problems in its knowledge representation. This can be done either via coreference or via some other mechanism that infers that the discourse is still concerned with the same problem. This may seem a very hard task, but at least it is not doomed in principle: in earlier work we managed to train non-experts in performing similar inferences and judgements during AZ annotation, using no world knowledge, only discourse cues.

5 Design of a recogniser

How could all this be recognised in unlimited text? The recognition of rhetorical moves would drive recognition with this model; as the only visible parts of the argument, rhetorical moves correspond to the bottom-up element. In contrast, high-level intentions form the top-down, *a priori* expectations. They can only ever be inferred, because the authors typically leave them implicit, so their recognition will never be made with absolute certainty.

A hybrid statistical-symbolic recogniser of scientific argumentation could instantiate the network in Fig. 2 on the fly for each new incoming paper, and keep a knowledge base of propositions derived during recognition. Whenever one of the moves is detected, the activation of its associated box is triggered. Statistically trained recognisers based on superficial features and evidence from tens of thousands of analysed papers provide a confidence value for the recognition of each move, which is translated into the strength of activation.The symbolic part of the recogniser keeps track of the logic representation accumulated up to that point in processing, and performs inference as to which higher-level intention is supported by currently activated rhetorical moves.

The output of such an analysis would be a partially activated network expressing the overall argument likely to be followed in the paper, where each node in the network is annotated with a more or less instantiated knowledge representation. The activated network can be considered as an automatically-derived explanation for the place in the research space where the authors situate themselves.

Newly-derived, intermediate levels of information should be additionally available from such an analysis, as a side-effect of this hybrid style of recognition. For instance, coreference resolution is an important aspect of analysis and contributes to the superficial features. It could also feed into a mechanism that determines which of the cited previous approaches is central to the argumentation in the paper, which of these the authors present as their main rivals or collaborators, and which aspects of existing work they criticise or praise.

It is quite obvious that a solution to this task would be immediately useful for a host of applications in search, summarisation and the teaching of scientific writing. As the system would be able to associate textual statements with the corresponding likely intentions it recognised, it could produce a justification for its overall analysis of the argument. Operating as a text critiquer, such a system could point out badly-expressed instances of well-known argumentation patterns, e.g. missing or weak evidence for particular high-level intentions.

Appealing though such applications are, the main point of the analysis laid out here is the development of a theory of text understanding of naturally occurring arguments in scientific text. Given the state of current NLP technology, some of the intermediate levels of recognition necessary for this seem to us to be within reach in the near future.

6 Conclusions

This paper promotes robust text understanding of scientific articles in a deeper manner than is currently practiced, as this would lead to more informative, symbolic representations of argument structuring. Mature technologies exist for determining specific scientific entities such as gene names (cf. the review by Campos et al., 2014) and specific events such as protein–gene interactions (e.g., Rebholz et al., 2005). In contrast to our work, such approaches are domain-specific and only recognise a small part of the entities or relationships modelled here. A different line of research associates text pieces with the research phase or information structure a given statement belongs to, where information structure is defined in terms of methods, results, conclusions etc, as in the work of Liakata et al. (2010), Guo et al (2013) and Hirohata et al. (2008). A related task, hedge detection in science, has been established and competitively evaluated (see Farkas et al. (2010) for an overview of the respective CoNLL shared task). While these two approaches (information structuring and hedge recognition) are domain-independent like ours, the analysis presented here aims at a deeper, more informative representation of relationships between general entities in the research space.

At the other end of the spectrum, we are aware of at least one deeper analysis of argument structure in science than ours, which is manual and takes world-knowledge into account, namely Green (2014); our approach differs from hers in that we opt to model argumentation in a domain- and discipline-independent manner, which is automatic but necessarily at a far shallower level.

Our claims in this paper include that a logical scientific argument structure exists and can be interpreted by a human reader, even in light of ambiguity and although only some steps of the argumentation are explicitly stated. We have also claimed that this type of analysis holds for all disciplines in principle, but certainly for all empirical sciences. We further claim that a substantial part of the argumentation in a well-written paper is recognisable to a reader even if they do not have any domain knowledge. These are rather strong claims: It is not even clear whether humans can recognise the explicit argumentation parts, let alone the inferred ones. We therefore need to substantiate the claims with annotation experiments.

In our work to date, we have made empirical observations about argumentation structure in synthetic chemistry, computer science, computational linguistics, and agriculture, but many of these are confined to the level of AZ or CFC. We are now in the process of corroborating the argumentation-level observations by corpus annotation of rhetorical moves. This initially takes the form of adding information to already existing AZ- and CFC-level annotation, with the aim of constructing a full-scale rhetorical move annotation. Higher-level goals will then be annotated as a second step.

Practical work also concerns building the recognisers of rhetorical moves. Several such recognisers already exist and will be refined in future work. It will be interesting to study exactly when inference about higher-level intentions becomes necessary, and which kinds of constraints can be derived from the argumentation network and the knowledge representation so as to usefully guide the inference mechanism.

References

Charles Bazerman. 1985. Physicists reading physics, schema-laden purposes and purpose-laden schema. *Written Communication*, 2(1):3–23.

Philippe Besnard and Anthony Hunter. 2008. *Elements of argumentation*. MIT Press.

Christine L. Borgman and Jonathan Furner. 2002. Scholarly communication and bibliometrics. In *Annual review of information science and technology: Vol. 36*, pages 3–72. Information Today, Medford, NJ.

Kevin W. Boyack and Richard Klavans. 2010. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404.

Stefanie Brüninghaus and Kevin D. Ashley. 2005. Generating legal arguments and predictions from case texts. In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 65–74.

David Campos, Srgio Matos, and Jos Lus Oliveir. 2014. Current methodologies for biomedical named entity recognition. In Mourad Elloumi and Albert Y. Zomaya, editors, *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data*. Wiley.

Sandra Carberry. 1990. *Plan Recognition in Natural Language Dialogue*. MIT Press, Cambridge, MA.

Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors. 1990. *Intentions in Communication*. MIT Press, Cambridge, MA.

A.M. Cohen, W.R. Hersh, K. Peterson, and Po-Yin Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219.

Robin Cohen. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics(COLING-84)*, pages 251–255.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence*, 77:321–357.

Richrd Farkas, Veronika Vincze, Gyrgy Mra, Jnos Csirik, and Gyrgy Szarvas. 2010. The conll-2010 shared task: learning to detect hedges and their scope in natural language text. In *Proceedings of CoNLL '10: Shared Task Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task*.

Mark Garzone and Robert E. Mercer. 2000. Towards an automated citation classifier. In *Proceedings of the 13th Biennial Conference of the CSCI/SCEIO (AI-2000)*, pages 337–346.

Nancy L. Green. 2014. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proc. of the First Workshop on Argumentation Mining, ACL 2014*.

Yufan Guo, Roi Reichart, and Anna Korhonen. 2013. Improved information structure analysis of scientific documents through discourse and lexical constraints. In *Proceedings of NAACL-2013*, Atlanta, US.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 381–388, Hyderabad, India. ACL Anthology Ref. I08-1050.

Ken Hyland. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics*, 30(4):437–455.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for conceptualisation and zoning of scientific papers. In *In: Proceedings of LREC-10*, Valetta, Malta.

Greg Myers. 1992. In this paper we report...—speech acts and scientific facts. *Journal of Pragmatics*, 17(4):295–313.

Peter Norvig. 1989. Marker passing as a weak method for text inferencing. *Cognitive Science*, 13(4):569–620.

Diarmuid O'Seaghdha and Simone Teufel. 2014. Unsupervised learning of rhetorical structure with untopic models. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, Dublin, Ireland.

Martha E. Pollack. 1986. A model of plan inference that distinguishes between the beliefs of actors and observers. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics (ACL-86)*, pages 207–214, New York, US.

Martha E. Pollack. 1990. Plans as complex mental attitudes. In P.R. Cohen, J. Morgan, and M.E. Pollack, editors, *Intentions in Communication*, pages 77–103. MIT Press, Cambridge, MA.

Dietrich Rebholz-Schuhmann, H Kirsch, and F Couto. 2005. Facts from textis text mining ready to deliver? *PLoS Biol*, 3(2). doi:10.1371/journal.pbio.0030065.

Francoise Salager-Meyer. 1992. A text-type and move analysis study of verb tense and modality distributions in medical English abstracts. *English for Specific Purposes*, 11:93–113.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Goals, Plans and Understanding*. Lawrence Erlbaum, Hillsdale, NJ.

John Swales, 1990. *Genre Analysis: English in Academic and Research Settings. Chapter 7: Research articles in English*, pages 110–176. Cambridge University Press, Cambridge, UK.

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of EMNLP-09*, Singapore.

Simone Teufel. 1998. Meta-discourse markers and problem-structuring in scientific articles. In *Proceedings of the ACL-98 Workshop on Discourse Structure and Discourse Markers*, pages 43–49, Montreal, Canada.

Simone Teufel. 2000. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK.

Simone Teufel. 2010. *The Structure of Scientific Articles: Applications to Citation Indexing and Summarization*. CSLI Publications.

Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Howard D. White. 2004. Citation analysis and discourse analysis revisited. *Applied Linguistics*, 25(1):89–116.