Using Semantic Web Resources for Solving Winograd Schemas: Sculptures, Shelves, Envy, and Success*

Peter Schüller
Computer Engineering Department
Faculty of Engineering, Marmara University
Istanbul, Turkey
peter.schuller@marmara.edu.tr

Mishal Kazmi
Faculty of Engineering and Natural Science
Sabanci University
Istanbul, Turkey
mishalkazmi@sabanciuniv.edu

ABSTRACT

Winograd Schemas are sentences where a pronoun must be linked to one of two possible entities in the same sentence. Deciding correctly which entity should be linked was proposed as an alternative to the Turing test. Knowledge is a critical component of solving this challenge and Linked Data resources promise to be useful to that end. We discuss two example Winograd Schemas and related knowledge that can be discovered by manual search in WikiData, DBPedia, BabelNet, freebase, WordNet, VerbNet, and the Component Library. We find that these resources are difficult to leverage because (i) they mix named entities with expert jargon and generic ontological knowledge, (ii) annotation tools are lacking, and (ii) commonsense knowledge is kept implicit.

1. INTRODUCTION

The Winograd Schema Challenge (WSC) [20, 12] was proposed as a more practical alternative for the Turing Test. An example is the following Winograd Schema (WS):

[The sculpture]$_a$ rolled off [the shelf]$_b$ because [it]$_X$ wasn't anchored. (ScAnchor)

[The sculpture]$_a$ rolled off [the shelf]$_b$ because [it]$_X$ wasn't level. (ScLevel)

Each sentence in such a schema poses a coreference ambiguity problem between the phrases marked in square brackets. This example has two candidate solutions: $X = a$ and $X = b$. An important property of WSs is, that the correct solution is different in both sentences, and that the sentences differ only by a single word ('level' vs. 'anchored'): the correct solution of (ScAnchor) is $X = a$ while for (ScLevel) it is $X = b$. Because of this property it has been argued, that purely statistical approaches will be insufficient for beating the WSC and that methods of (symbolic) knowledge representation and reasoning will be necessary [12].

*This work has been supported by Scientific and Technological Research Council of Turkey (TUBITAK) Grant 114E777.

Reasoning requires knowledge, the biggest repository of knowledge is arguably the Internet, however it is mostly unstructured information. The Linked Data effort [4] structures data in a way that it becomes *machine readable*, hence it can be used for automated reasoning. Therefore using the Semantic Web as a knowledge resource for tackling the WSC is highly suggestive. But knowledge is more than data, so how far can we get with existing resources?

In this work we discuss two examples of WSs and attempt to resolve them using data repositories typically considered part of the Semantic Web and other resources. We show that repositories like WikiData, DBPedia, BabelNet, and freebase are necessary but not sufficient by themselves: they contain mostly taxonomic knowledge and (historical) facts about named entities.

On the contrary, most existing Winograd Schemas [7] do not refer to historical events or entities, they can be understood *out of the blue* (i.e., without additional context) using *Commonsense knowledge* [14] that is shared by humans because they live in a similar world as other humans.[1] In the above WS such knowledge is that anchoring/fixing an object (usually) prevents its movement.

But is such Commonsense knowledge represented in Semantic Web resources?

In this work we first outline how to perform reasoning, following the idea that many schemas can be resolved using correlation [2].[2] Then we give an — in our opinion — representative part of background knowledge obtained from existing resources by manual search

Our contribution is to point out potential use of Semantic Web resources towards tackling Winograd Schemas and to show problems that become apparent while doing so. The main issues we point out are as follows.

- Misinterpreting the topic of a sentence causes annotation of many wrong entities, in particular if knowledge about named entities or expert jargon is preferred over generic concepts. Therefore tools that annotate text with the correct links (concepts or entities) are crucial.
- A high level of detail in textual descriptions, or a varying detail of such descriptions, can mislead reasoning.
- Missing Commonsense knowledge is a limiting factor, but annotating Web content with links to common vo-

[1] We will disregard the question where commonsense knowledge ends and where culture-dependent knowledge starts.
[2] Note that such reasoning need not be based on symbolic logic, we can similarly envision to realize it statistically.

cabularies has the potential to enable future work that mines such knowledge from the (annotated) Web.

2. REASONING ABOUT CORRELATION

Why is correlation [2] a possibility for resolving coreferences in the WSC?

If we split (ScAnchor) and (ScLevel) into two sentences, including all possible resolutions of the pronoun, we obtain the following simple statements.

The sculpture rolled off the shelf. (1)

The sculpture wasn't anchored. (2)

The shelf wasn't anchored. (3)

The sculpture wasn't level. (4)

The shelf wasn't level. (5)

In the original schema, the word 'because' raises an expectation in the reader, namely that the second sentence is a *plausible reason* for the first sentence. One way to handle this plausibility is to reduce it to correlation: For example by checking whether (1) has a higher correlation with (2) than with (3) we can find the correct solution.

But what kind of correlation do we use in this case?

The topic of all three sentences is related with movement or its impossibility. Hence the correlation can be about properties within that topic. (1) and (2) pertain to movement of the sculpture, while (3) pertains to movement of the shelf. In absence of knowledge about the meaning of anchoring this can be enough to infer the correct solution, namely that (1) and (2) are correlated more than (1) and (3). If we additionally know that anchoring prevents rolling, and not anchoring allows rolling (this can be seen as a positive correlation) then we can also infer the correct solution: (2) better fulfills the expectation raised by 'because', yielding the solution $X = a$.

Note that we deal with (ScLevel) in the next section.

Our second example is the following schema.

[Pete]$_a$ envies [Martin]$_b$ because [he]$_X$ is very successful. (EnvyBecause)

[Pete]$_a$ envies Martin$_b$ although [he]$_X$ is very successful. (EnvyAlthough)

Extracted parts of this schema are

Pete envies Martin. (6)

Pete is very successful. (7)

Martin is very successful. (8)

Note that this time the same three sentences are used for resolving (EnvyBecause) and (EnvyAlthough). The only difference between (EnvyBecause) and (EnvyAlthough) is the connective between the sentences: 'because' (again) raises an expectation of positive correlation, however 'although' raises an expectation of an exception: although we would normally assume that Pete is not envious (because he is successful), he actually is.

Intuitively, the object of 'to envy' is correlated with being successful, and the subject of 'to envy' is correlated with not being successful. Therefore the solution for (EnvyBecause) can be found by maximizing correlation, while we need to minimize correlation for (EnvyAlthough).

Theoretical Justification. The expectation of correlation is explained by linguistic theories about discourse structure and discourse coherence (e.g., [13, 1]). For simplicity, our examples show only *explicit* discourse structure indicated by 'because' or 'although'. However in a coherent text all sentences are related in a structure, often a tree structure, and often not explicitly marked. Two examples of frequent implicit discourse relations are temporal order (time usually progresses forward from one sentence to the subsequent one); and elaboration (a topic is explained in more detail in a subsequent sentence). Examples for further explicit discourse connectives are 'but', 'hence', 'in order to', ...

3. SEMANTIC WEB KNOWLEDGE

We now investigate how to obtain the required knowledge from resources integrated into the Semantic Web and from similar resources built by other communities. We consider WikiData [19], which is a language- and presentation-independent annotated data backend for Wikipedia, DBPedia [10], which contains RDF triples extracted from Wikipedia infoboxes, freebase [5], which is a community-built knowledge graph repository, and BabelNet [17], which connects several Wikipedia projects with the linguistic resources WordNet [15]. Furthermore we use the linguistic resources WordNet, VerbNet [9], and the Component Library Clib [3], which is a Commonsense knowledge resource.

For WikiData, freebase, WordNet, VerbNet, and Clib, we lookup single words. We also search in the Falcons Semantic Web search engine [18] which performs search and ranking results in most of the above resources.

Additionally we perform annotation of the whole schemas using the annotation engines DBPedia Spotlight[3] [6] and the Babelfy [16] annotation engine for linking to BabelNet.[4]

3.1 Disambiguating the Anchored Schema

(ScAnchor) can be disambiguated if we have knowledge that (i) sentences (1), (2), and (3) are about the same topic (movement of a physical object); and that (ii) anchoring prevents movement and rolling is movement.

We will take for granted that we have *linguistic knowledge and mechanisms* that allow us to identify subject and object of 'rolled' and how to handle the predicate 'wasn't': these are research areas on their own.

Useful Knowledge. WikiData has definitions for 'sculpture' as well as 'shelf', classifying them as 'three-dimensional work of art' and as 'furniture', respectively, and both are a subclass of 'artificial physical object'.

VerbNet contains an entry classifying 'roll off'as 'move'.

Babelfy produces several correct annotations: 'rolled' is linked to 'roll' which is a kind of 'move'; 'anchored' is linked to 'anchor' which is a kind of 'fix' which is a kind of 'attach'.

Falcons finds WordNet and WikiData concepts for 'sculpture', 'to roll', and 'shelf', ranking correct data high but not in first place. Results for 'anchored' are not helpful, but searching for 'anchor' reveals the correct WordNet entry.

While attaching intuitively prevents moving, this knowledge is cannot be found easily.

Clib contains nearly enough information to infer this causal relation: 'Move' has a precondition that the object to be moved does not have the property 'Be-Restrained' which inherits from 'Be-Obstructed'. The problem is that obstruction or restraint is caused by 'Move' and the separate

[3] `http://spotlight.dbpedia.org/rest/annotate`

[4] `http://babelfy.org/index`

event 'Attach' is not causally related to restraining ('Be-Restrained' contains the linguistic annotation 'fixed' but fixing and attaching are modeled as different concepts).

While it seems that enough knowledge is available, automatically linking that knowledge in the right way is not trivial. Next we show that there is also additional knowledge that could be linked and that is *counterproductive* towards the goal of reasoning about the intuition of our example.

Misleading knowledge. freebase provides several entries for 'sculpture', mostly about the art form of sculpturing, and few about (very specific) physical objects. The closest helpful entry is 'statue' which, according to its definition, '... is a sculpture representing one or more people or animals ....' For 'shelf' the first hit is the correct entry, however in its definition we find that 'A shelf is a flat horizontal plane ... to hold items ... It is raised off the ground and usually anchored/supported on its shorter length sides by brackets.'

The definition of sculpture (or statue) does not contain information about anchoring, hence a system that heuristically evaluates correlation will conclude that shelves are more likely to be anchored than sculptures. Therefore (ScAnchor) will be disambiguated in the wrong way, even though we have only truthful evidence. The problem is that the heuristic fails because (by chance) one definition contains misleading information. (Actually searching the web yields many do-it-yourself forums with information about proper ways to anchor shelves in the wall, and not nearly as much for anchoring statues, hence correlation of anchoring seems to be higher for shelves than for statues, however we need to consider correlation between rolling an object and not anchoring that object.)

Babelfy provides the following annotations: 'sculpture' is a 'three-dimensional figure' which is a 'shape' which is a 'mathematical object', moreover 'sculpture' is subclass of several art forms; 'shelf' is a 'furniture' which is a 'decorative art' which is a 'perceptible object', moreover 'shelf' is a 'support' which is a 'machine' and a 'tool'. In summary 'sculpture' is classified as an intangible abstract concept, and 'shelf' is classified also as a tool which can be misleading.

DBPedia Spotlight annotates 'sculpture' with a particular species of sea snail, 'shelf' with 'Shelf life' (shelf here means shallow coastal area of the sea). While 'rolled' and 'anchored' are not associated with any entity, a search on the web reveals more potentially misleading information: there are 'roll anchors' for anchoring ships in the shelf zone, moreover rolling is a specific movement of ships induced by wind.

While the presence of this (expert jargon) knowledge in DBPedia is no problem, its usage is a problem: it should be linked only when significant evidence suggests that the text *is about* anchoring ships near the coast. It seems that Babelfy performs better than DBPedia Spotlight, although this can be a random effect due to the limited number of examples we are looking at. While Falcons contains a possibility to choose between 'object' and 'concept' in the search, this does not seem to provide the required distinction: expert jargon is always contained in search results.

Note that we mainly discussed 'sculpture', because for other content words, misleading knowledge cannot be found to such an extent. Due to the amount of available knowledge, separating useful from irrelevant knowledge is crucial.

3.2 Disambiguating the Level Schema

For (ScLevel) we do not need correlation: if we can show that among the two candidates (4) and (5), the second one is a property that is reasonable while the first is an unreasonable one, then we find the correct result.

To show this we require the following knowledge: (i) a shelf is usually a flat entity, (ii) a sculpture is usually not flat, and (iii) 'level' is a potential property of flat entities.[5]

Useful knowledge. freebase's entry for 'shelf' states 'A shelf is a flat horizontal plane [...]' and most of its entries for 'level' refer to 'horizontal' or 'plane' in their definitions. This allows us to infer that 'level' is a more likely property of 'shelf' than of 'sculpture', yielding the solution $X = b$.

Misleading knowledge. DBPedia Spotlight wrongly links 'level' to 'deck of a ship', again using the wrong topic area.

Babelfy wrongly links 'level' to 'level of a game'.

Falcons provides many pages of search results, but the order of results is misleading: the first five entries are related with 'level of visibility'.

3.3 Disambiguating both Envy Schemas

(EnvyBecause) and (EnvyAlthough) differ only in the rhetorical relation, therefore the same knowledge is relevant.

Useful knowledge. DBPedia Spotlight correctly links 'envies' to an emotion which 'occurs when a person lacks another's superior quality, achievement, or possession and either desires it or wishes that the other lacked it'.

VerbNet does not contain an entry for 'to envy' but for 'success' which is a potential property of humans according to several of its free-text definitions.

WikiData provides as first results for 'envy' the same entry as DBPedia Spotlight. Additionally for 'success' it contains an entry for 'achievement of a goal' and one for 'victory'. (For 'successful' there are only entries related to arts pieces.)

Babelfy links 'envies' to 'to envy' which is a subclass of 'to admire' which is a subclass of 'to think', moreover 'successful' is linked to the entry of the same name but this entry does not contain any classification.

Falcons provides useful results, linking 'successful' to the corresponding WordNet entry, and 'to envy' to 'jealousy'.

These pieces of knowledge can be sufficient for our purpose, in particular the definition of 'to envy' in connection with linking success to 'achievement'.

Misleading knowledge. However there is also misleading and missing knowledge.

freebase provides many results for 'envy', 'envies', 'success', and 'successful', most of them names of arts pieces.

DBPedia Spotlight links 'Pete' and 'Martin' to TV programs and characters, respectively. Note that interpreting these names is not useful for disambiguating this schema.

Clib does not contain any information about envy or success, it does not even contain the concepts of feeling, emotion, attitude, or thinking.

In summary, (EnvyBecause) and (EnvyAlthough) can be disambiguated automatically with existing resources, if we manage to ignore irrelevant search results.

4. CONCLUSION

Authors of content in the Web rarely describe how the world works, mostly they give an efficient account of what happened, why, when, and how it happened. Such an efficient

[5]The words 'usually' and 'potential' point out that this knowledge is *default knowledge* and can be defeated by more specific knowledge (to account for atypical cases).

use of natural language omits certain content that can be inferred by the reader, therefore computers must *interpret* natural language to reason with it. Similarly, if data is published in non-annotated unstructured form, humans can often guess which part of that data is a name or a location. Computers cannot do that, therefore the Linked Data initiative aims to *annotate* data with type information in common ontologies and information about its relation to other data.

Linked Data, as the name indicates, is about data, annotated with its type and further meta information. However, various Semantic Web resources do not only contain data about named entities and events, they also contain a bit of (mostly taxonomic) commonsense knowledge. This knowledge is used to organize the meta information and is often mixed with the other knowledge.

Using Linked Data for reasoning requires to connect it with additional commonsense knowledge that is currently not contained in existing resources. Moreover, connecting Linked Data to natural language texts (i.e., with unstructured data), requires *annotation tools* like Babelfy or DBPedia Spotlight that annotate words and phrases in a text with appropriate URIs to resources in the Semantic Web. Such tools are often based on (or supported by) machine learning.

In this work we saw that only Babelfy provides reasonable automatic annotations, so its ranking scheme seems to be superior to DBPedia Spotlight. Babelfy distinguishes between knowledge about concepts and named entities, and internally uses coherence measures. Our examples of misleading knowledge consider only the correct POS, however NER detection could help to separate between common nouns and names. Falcons distinguishes between 'concepts' and 'objects', however it returns expert jargon in both result types and the tfidf ranking [18] produces misleading search results.

About the issue of expert jargon (e.g., 'sculpture' as a certain type of mollusc) we note that already in the CYC project [11] there were 'microtheories' for separating more specific from more generic knowledge. However, in none of the resources discussed in this work we found a method of distinguishing between these types of knowledge.

RDFa allows Web authors to annotate parts of their website content (i.e., words or phrases) with type information and links to common vocabularies such as WikiData. This eliminates the need for disambiguation and can make *machine reading* more feasible on these websites. Therefore we think that widespread usage of RDFa could be a crucial enabler for *mining commonsense knowledge* from the web, in efforts similar to [8].

We conclude that Linked Data can be used for reasoning, but we need better tools that automatically annotate a given text with the most suitable Semantic Web URIs. Additionally, only if we manage to integrate Linked Data with commonsense knowledge, we can *use this data as knowledge*.

5. REFERENCES

[1] N. Asher and A. Lascarides. *Logics of Conversation*. Cambridge University Press, 2003.

[2] D. Bailey, A. Harrison, Y. Lierler, V. Lifschitz, and J. Michael. The Winograd Schema Challenge and Reasoning about Correlation. In *Symposium on Logical Formalizations of Commonsense Reasoning*, 2015.

[3] K. Barker, B. Porter, and P. Clark. A Library of Generic Concepts for Composing Knowledge Bases. In *International Conference on Knowledge Capture (K-CAP)*, New York, USA, 2001. ACM Press.

[4] T. Berners-Lee, C. Bizer, and T. Heath. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (Special Issue on Linked Data)*, 5(3):1–22, 2009.

[5] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database For Structuring Human Knowledge. In *ACM SIGMOD International Conference on Management of Data*, pages 1247–1249, 2008.

[6] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction Categories and Subject Descriptors. In *Semantic Systems*, pages 3–6, 2012.

[7] E. Davis. A collection of Winograd Schemas. `http://www.cs.nyu.edu/faculty/davise/papers/WS.html`.

[8] A. S. Gordon. Mining Commonsense Knowledge From Personal Stories in Internet Weblogs. In *Workshop on Automated Knowledge Base Construction (AKBC)*, pages 8–15, 2010.

[9] K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, 2007.

[10] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia - A Large-scale , Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 1(1-5):1–29, 2012.

[11] D. B. Lenat. CYC: A Largs-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[12] H. J. Levesque, E. Davis, and L. Morgenstern. The Winograd Schema Challenge. In *Principles of Knowledge Representation and Reasoning (KR)*, pages 552–561. AAAI Press, 2012.

[13] W. C. Mann and S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[14] J. McCarthy. Programs with Common Sense. `www-formal.stanford.edu/jmc/mcc59.ps`, 1959.

[15] G. A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[16] A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014.

[17] R. Navigli and S. P. Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[18] Y. Qu and G. Cheng. Falcons concept search: A practical search engine for web ontologies. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 41(4):810–816, 2011.

[19] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[20] T. Winograd. *Understanding Natural Language*. Academic Press, 1972.