# Identifying Disputed Topics in the News

Orphee De Clercq<sup>1,3</sup>, Sven Hertling<sup>2</sup>, Veronique Hoste<sup>1</sup>, Simone Paolo Ponzetto<sup>3</sup>, and Heiko Paulheim<sup>3</sup>

 LT3, Language and Translation Technology Team, Ghent University {orphee.declercq,veronique.hoste}@ugent.be
 Knowledge Engineering Group, Technische Universität Darmstadt hertling@ke.tu-darmstadt.de
 Research Group Data and Web Science, University of Mannheim

{simone,heiko}@informatik.uni-mannheim.de

**Abstract.** News articles often reflect an opinion or point of view, with certain topics evoking more diverse opinions than others. For analyzing and better understanding public discourses, identifying such contested topics constitutes an interesting research question. In this paper, we describe an approach that combines NLP techniques and background knowledge from DBpedia for finding disputed topics in news sites. To identify these topics, we annotate each article with DBpedia concepts, extract their categories revealing significant deviations in polarity across different media. We illustrate our approach in a qualitative evaluation on a sample of six popular British and American news sites.

Keywords: Linked Open Data, DBpedia, Sentiment Analysis, Online News

## 1 Introduction

The internet has changed the landscape of journalism, as well as the way readers consume news. With many newspapers providing a website available offering news for free, many people are no longer local readers who are subscribed to one particular newspaper, but receive news from many sources, covering a wide range of opinions. At the same time, the availability of online news sites allows for indepth analysis of topics, their coverage, and the opinions about them. In this paper, we explore the possibilities of current basic Semantic Web and Natural Language Processing (NLP) technologies to identify topics carrying *disputed* opinions.

There are different scenarios in which identifying those disputed opinions is interesting. For example, media studies are concerned with analyzing the political polarity of media. Here, means for automatically identifying conflicting topics can help understanding the political bias of those sources. Furthermore, campaigns of paid journalism may be uncovered, e.g. if certain media have significant positive or negative deviations in articles mentioning certain politicians. In this paper, we start with the assumption that DBpedia categories help us identify specific topics. Next, we look at how the semantic orientation of news articles, based on a lexicon-based sentiment analysis, helps us find disputed news. Finally, we apply our methodology to a web crawl of six popular news sites, which were analyzed for *both* topics and sentiment. To this end, we first annotate articles with DBpedia concepts, and then use the concepts' categories to assign topics to the articles. Disputed topics are located by first identifying significant deviations of a topics' average sentiment per news site from the news site's overall average sentiment, and selecting those topics which have both significant positive and negative deviations.

This work contributes an interesting application of combining Semantic Web and NLP techniques for a high-end task. The remainder of this paper is structured as follows: in the next section we describe related work (Section 2). Next, we present how we collected and processed the data used for our system (Section 3). We continue by describing some interesting findings of our approach together with some of its limitations (Section 4). We finish with some concluding remarks and prospects for future research (Section 5).

# 2 Background and Related Work

Text and data mining approaches are increasingly used in the social science field of media or content analysis. Using statistical learning algorithms, Fortuna et al. [6] focused on finding differences in American and Arab news reporting and revealed a bias in the choice of topics different newspapers report on or a different choice of terms when reporting on a given topic. Also the work by Segev and Miesch [17], which envisaged to detect biases when reporting on Israel, found that news reports are largely critical and negative towards Israel. More qualitative studies were performed, such as the discourse analysis by Pollak et al.[14] which revealed contrast patterns that provide evidence for ideological differences between local and international press coverage. These studies either focus on a particular event or topic [14,17] or use text classification in order to define topics [6], and most often require an upfront definition of topics and/or manually annotated training data. In this work, instead, we use semantic web technologies to semantically annotate newswire text, and develop a fully automatic pipeline to find disputed topics by employing sentiment analysis techniques.

Semantic annotation deals with enriching texts with pointers to knowledge bases and ontologies [16]. Previous work mostly focused on linking mentions of concepts and instances to either semantic lexicons like WordNet [5], or Wikipediabased knowledge bases [7] like DBpedia [9]. DBpedia was for example used by [8] to automatically extract topic labels by linking the inherent topics of a text to concepts found in DBpedia and mining the resulting semantic topic graphs. They found that this is a better approach than using text-based methods. Sentiment analysis, on the other hand, deals with finding opinions in text. Most research has been performed on clearly opinionated texts such as product or movie reviews [15], instead of newspaper texts which are believed to be less opinionated.

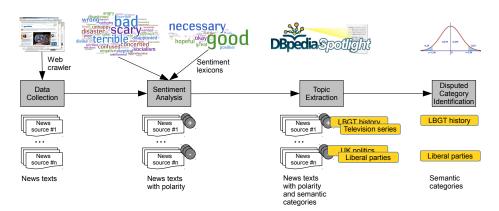


Fig. 1. An illustrative overview of our approach for identifying disputed topics.

An exception is the work performed by [2] in the framework of the European Media Monitor project [18].

While the combination of sentiment analysis and semantic annotation for the purpose discussed in this paper is relatively new, some applications have been produced in the past. The DiversiNews tool [20], for example, enables the analysis of text in a web-based environment for diversified topic extraction. Closely related are DisputeFinder [4] and OpinioNetIt [1]. The former is a browser extension which highlights known disputed claims and presents the user with a list of articles supporting a different point of view, the latter should allow to automatically derive a map of the opinions-people network from news and other web documents.

## 3 Approach

Our process comprises four steps, as depicted in Fig. 1. First, data is collected from online news sites. Next, the collected texts are augmented with sentiment scores and semantic categories, which are then used to identify disputed categories.

#### 3.1 Data Collection

We have collected data from six online news sites. First, we looked at those having a high circulation and online presence. Another criterion for selection was the ability to crawl the website, since, e.g., dynamically loaded content is hard to crawl.

The six selected news sites fulfilling these requirements are shown in Table 1. We work with three UK and three US news sites. As far as the British news sites are concerned, we selected one rather conservative news site, the Daily Telegraph which is traditional right-wing; one news site, the Guardian, which

**Table 1.** Data Sources used in this paper. We report the news sites we used, the number of articles crawled, and the average article length in words.

Newspaper	Country	Website	# Art	Avg. artlen
The Daily Mirror	UK	http://www.mirror.co.uk/	1,024	422
The Daily Telegraph	UK	http://www.telegraph.co.uk	1,055	599
The Guardian	UK	http://www.guardian.co.uk/	1,138	638
The Huffington Post	US	http://www.huffingtonpost.com/	1,016	446
Las Vegas Review-Journal	US	http://www.reviewjournal.com/	1,016	618
NY Daily News	US	http://www.nydailynews.com/	1,016	338

can be situated more in the middle of the political spectrum though its main points of view are quite liberal; and finally also one tabloid news site, the Mirror, which can be regarded as a very populist, left-wing news site.<sup>1</sup> For the American news sites, both the Las Vegas Review–Journal and the Huffington Post can be perceived as more libertarian news sites<sup>2</sup>, with the latter one being the most progressive [3], whereas the NY Daily News, which is also a tabloid, is still liberal but can be situated more in the center and is even conservative when it comes to matters such as immigration and crime.

The news site articles were collected with the python web crawling framework  $Scrapy^3$ . This open-source software focuses on extracting items, in our case, news site articles. Each item has a title, an abstract, a full article text, a date, and an URL. We only crawled articles published in the period September 2013 – March 2014. Duplicates are detected and removed based on the article headlines.<sup>4</sup>

#### 3.2 Sentiment Analysis

We consider the full article text as the context to determine the document's semantic orientation. The basis of our approach to define sentiment relies on word lists which are used to determine positive and negative words or phrases.

We employ three well-known sentiment lexicons. The first one is the Harvard General Inquirer lexicon – GenInq [19] – which contains 4,206 words with either a positive or negative polarity. The second one is the Multi-Perspective Question Answering Subjectivity lexicon – MPQA [22] – which contains 8,222 words rated between strong and weak positive or negative subjectivity and where morphosyntactic categories (PoS) are also represented. The last one is the AFINN lexicon [12], which includes 2,477 words rated between -5 to 5 for polarity.

Before defining a news article's polarity, all texts were sentence-split, tokenized and part-of-speech tagged using the LeTs preprocessing toolkit [21]. In

<sup>&</sup>lt;sup>1</sup> Cf. results of 2005 MORI research: http://www.theguardian.com/news/datablog/ 2009/oct/05/sun-labour-newspapers-support-elections.

 $<sup>^2 \ \</sup>texttt{http://articles.latimes.com/2006/mar/08/entertainment/et-vegas8}$ 

<sup>&</sup>lt;sup>3</sup> http://scrapy.org/

<sup>&</sup>lt;sup>4</sup> The dataset and all other resources (e.g. RapidMiner processes) are made freely available to the research community at http://dws.informatik.uni-mannheim.de/en/research/identifying-disputed-topics-in-the-news.

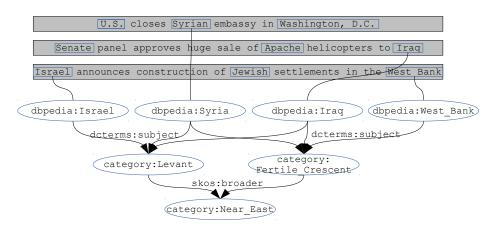


Fig. 2. Example of news texts annotated with DBpedia concepts, and categories extracted. The graph only shows a portion of the categories and their super category relationships.

a next step, various sentiment scores were calculated on the document level by performing a list look-up. For each document, we calculated the fraction of positive and negative words by normalizing over text length, using each lexicon separately. Then, in a final step we calculated the sum of the values of identified sentiment words, which resulted in an overall value for each document. That is, for each document d, our approach takes into consideration an overall lexicon score defined as:

$$lexscore(d) = \sum_{i=1}^{n} v_{w_i} .$$
 (1)

where  $w_i$  is the *i*-th word from *d* matched in the lexicon at hand, and  $v_{w_i}$  its positive or negative sentiment value.

#### 3.3 Topic Extraction

We automatically identify the topics of our news articles on the basis of a twostep process. First, we identify concepts in *DBpedia* [9]. To that end, each article's headline and abstract are processed with *DBpedia Spotlight* [10]. Next, categories for each concept are created, corresponding to the categories in Wikipedia: we extract all direct categories for each concept, and add the more general categories two levels up in the hierarchy.

These two phases comprise a number of generalizations to assign topics to a text. First, processing with DBpedia Spotlight generalizes different surface forms of a concept to a general representation of that concept, e.g. Lebanon, Liban, etc., as well as their inflected forms, are generalized to the concept dbpedia:Lebanon. Second, different DBpedia concepts (such as dbpedia:Lebanon, dbpedia:Syria) are generalized to a common category (e.g. category:Levant). Third, categories (e.g. category:Levant, category:Fertile\_Crescent) are generalized to super

Newspaper	# Concepts	# Categories
The Daily Mirror	971	10,017
The Daily Telegraph	784	8,556
The Guardian	605	6,919
The Huffington Post	400	6,592
Las Vegas Review-Journal	227	2,761
NY Daily News	942	10,540
Total	2,825	22,821

**Table 2.** Number of concepts identified by DBpedia Spotlight, and DBpedia categories

 extracted by source. The totals represent the number of unique concepts and categories.

categories (e.g. category: Near\_East). We provide an illustration of this generalization process in Fig. 2.

The whole process of topic extraction, comprising the annotation with DBpedia Spotlight and the extraction of categories, is performed in the *RapidMiner Linked Open Data Extension* [13]. Table 2 depicts the number of concepts and categories extracted per source. It can be observed that the number of categories is about a factor of 10 larger than the number of concepts found by DBpedia Spotlight alone. This shows that it is more likely that two related articles are found by a common category, rather than a common concept.

#### 3.4 Disputed Categories Extraction

We identify disputed categories (and hence, topics) as follows:

- 1. First, for each news site we produce a global sentiment-orientation profile based on the overall sentiment scores (Equation 1): this is meant to model the coarse sentiment bias of a specific news source and avoid effects occurring due to the typical vocabulary of a news source.
- 2. Next, we identify those DBpedia categories for which the sentiment score deviates significantly from the global sentiment score. Since these follow a Gaussian distribution (Cf., Fig. 3, we can apply a z-test. From the overall number of texts n collected from a news site, the mean  $\mu$  and standard deviation  $\sigma$  of the sentiment scores are computed, together with the average sentiment M(c) of each category c. The latter is computed as

$$M(c) = \frac{1}{|C|} \sum_{d \in C} lexscore(d),$$
(2)

where C is the set of all articles annotated with category c, and lexscore(d) is one of our three lexicon-based scoring functions. We can then compute the category's z score as:

$$z(c) = \frac{M(c) - \mu}{\frac{\sigma}{n}} \tag{3}$$

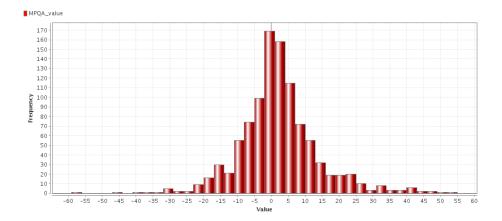


Fig. 3. Example distribution of the sentiment scores. The histogram shows the distribution of the MPQA sentiment score for texts from The Guardian.

If the z score is positive, articles in the category c are more positive than the average of the news source and the other way around. By looking up that z score in a Gaussian distribution table, we can discard those deviations that are statistically insignificant. For instance, the Mirror contains three articles annotated with the category **Church of Scotland**, with an average AFINN sentiment score of 20.667, which is significant at a z-value of 2.270.

3. In the last step, we select those categories for which there is at least one significant positive and one significant negative deviation. If two disputed categories share the same extension of articles (i.e. the same set of articles is annotated with both categories), we merge them into a cluster of disputed categories.

## 4 Analysis

The output of our system is presented in Table 3, showing that up to 19 disputed topics can be identified in our sample. In what follows we present some interesting findings based on a manual analysis of the output and we also draw attention to some limitations of our current approach. In general, we opt in this work for a validation study of the system output – as opposed, for instance, to a gold-standard based evaluation. This is because, due to the very specific nature of our problem domain, any ground truth would be temporally bound to a set of disputed topics for a specific time span.

#### 4.1 Findings

If we look at the different percentages indicating the amount of articles found with a significant positive or negative sentiment, we see that these numbers differ

	GenInq		MPQA			AFINN			
Newspaper	$\mathbf{pos}$	neg	%	pos	$\operatorname{neg}$	%	$\operatorname{pos}$	neg	%
The Daily Mirror	197	362	54.59	0	0	0	103	120	21.78
The Daily Telegraph	268	99	34.79	0	0	0	185	138	30.62
The Guardian	152	455	53.34	389	154	47.72	176	266	38.84
The Huffington Post	165	159	31.89	285	48	32.78	140	95	23.13
Las Vegas Review-Journal	70	75	14.27	54	68	12.01	92	68	15.75
NY Daily News	329	192	51.28	150	270	41.34	305	223	51.97
Disputed	19		11			17			

**Table 3.** Number of categories with a significant positive and/or negative sentiment per source, percentage of number of topics covered using the different lexicons, and the total number of disputed topics.

among the lexicons. The Daily Mirror seems to contain most subjective articles when using the GenInq lexicon, a role played by The Guardian and The Daily News NY when using the MPQA lexicon and the AFINN lexicon, respectively. The largest proportions are found within the Daily Mirror and the NY Daily News, which is not surprising since these are the two tabloid news sites in our dataset. Though the Daily Telegraph and the Daily Mirror seem to have no significant deviations using the MPQA lexicon<sup>5</sup>, we nevertheless find disputed topics among the other four news sites. Consequently, the MPQA has the fewest (11), followed by AFINN (17) and GenInq (19).

Initially, we manually went through the output list of disputed topics and selected two topics per lexicon that intuitively represent interesting news articles (Table 5). What draws the attention when looking at these categories is that these are all rather broad. However, if we have a closer look at the disputed articles we clearly notice that these actually do represent contested news items. Within the category Alternative\_medicine, for example, we find that three articles focus on *medical marijuana legalization*. To illustrate, we present these articles with their headlines, the number of subjective words with some examples, and the overall GI lexicon value<sup>6</sup>.

- NY Daily News. "Gov. Cuomo to allow limited use of medical marijuana in New York"  $\rightarrow$  7 positive (e.g. great, tremendous) and 5 negative (e.g. difficult, stark) words; GI value of 2.00.
- NY Daily News: "Gov. Cuomo says he won't legalize marijuana Coloradostyle in New York",  $\rightarrow$  5 positive (e.g. allow, comfortable) and 8 negative (e.g. violation, controversial) words; GI value of -3.
- Las Vegas Review: "Unincorporated Clark County could house Southern Nevada medical marijuana dispensaries",  $\rightarrow 26$  positive (e.g. ensure, accommodate) and 10 negative (e.g. pessimism, prohibit) words; GI value of 16.

<sup>&</sup>lt;sup>5</sup> This might be due to MPQA's specific nature, it has different gradations of sentiment and also PoS tags need to be assigned in order to use it

<sup>&</sup>lt;sup>6</sup> However, as previously mentioned in Section 3, for the actual sentiment analysis we only considered the actual news article and not its headline or abstract.

Table 4. Article and sentiment statistics of two categories for each lexic
--

Category/topic	# articles $#$	≠ UK 7	# US	# pos #	neg #	neut
GenInq Alternative_medicine	5	3	2	3	2	0
Death	13	11	2	5	7	1
MPQA Government_agencies	13	7	6	7	5	1
LGBT_history	11	5	6	7	4	0
AFINN Democratic_rights	7	4	3	4	3	0
Liberal_parties	34	5	29	26	7	1

Though the last article is clearly about a difficult issue within this whole discussion, we see that the Las Vegas Review-Journal reports mostly positive about this subject which could be explained by its libertarian background. Whereas the NY Daily News, which is more conservative regarding such topics, reports on this positive evolution by using less outspoken positive and even negative language. A similar trend is reflected in the same two news sites when reporting on another contested topic, i.e. *gay marriage*, which turns up using the MPQA lexicon in the category LGBT\_history. We again present some examples.

- Las Vegas Review: "Nevada AG candidates split on gay marriage"  $\rightarrow 25$  positive: 16 weak (allow, defense) and 9 (clearly, opportunity) are strong subjective and 13 negative: 10 weak (against, absence) and 3 (heavily, violate) strong subjective. MPQA value of 19.
- NY Daily News: "Michigan gov. says state won't recognize same-sex marriages",  $\rightarrow$  7 positive: 5 weak (reasonable, successfully) and 2 strong (extraordinary, hopeful) subjective and 9 negative: 4 weak (little, least) and 5 strong subjective (naive, furious). MPQA value of -5.

Another interesting finding we discover is that for four out of six categories, the articles are quite evenly distributed between UK and US news sites and that two categories stand out: Death seems to be more British and Liberal\_parties more American. If we have a closer look at the actual articles representing these categories we see 9 out of the 11 Death articles actually deal with murder and were written for the Daily Mirror which is a tabloid news site focusing more on sensation. As far as the 34 American articles regarding liberal parties are concerned, we notice that all but six were published by the Las Vegas Review-Journal which is known for its libertarian editorial stance.

These findings reveal that using a basic approach based on DBpedia category linking and lexicon-based sentiment analysis already allows us to find some interesting, contested news articles. Of course, we are aware that our samples are too small to make generalizing assumptions which brings us to a discussion of some of the limitations of our current approach.

#### 4.2 Limitations

In order to critically evaluate the limitations of our approach, we first had a look at the actual "topic representation". Since we use the lexicons as a basis to find disputed topics, we randomly select 20 news articles that show up under a specific category per lexicon and assess its representativeness. We found that, because of errors in the semantic annotation process, out of these 60 examples, only 34 were actually representative of the topic or category in which they were represented. If we look at the exact numbers per lexicons, this amounts to an accuracy of 55% in the GenInq, one of 70% in MPQA and one of 40% in AFINN. Examples of mismatches, i.e. where a DBpedia Spotlight concept was misleadingly or erroneously tagged, are presented next:

- AFINN, category:Television\_series\_by\_studio, tagged concepts: United\_States\_Department\_of\_Veterans\_Affairs, Nevada, ER\_TV\_series  $\rightarrow$  article is about a poor emergency room, not about the TV series *ER*.
- GenInq, category:Film\_actresses\_by\_award, tagged concepts: Prince\_Harry\_of\_Wales, Angelina\_Jolie → article is about charity fraud, Angelina Jolie is just a patron of the organization.

We performed the same analysis on our manually selected interesting topics (cf. Table 4) and found that actually 74 out of the 83 articles were representative.

When trying to evaluate the sentiment analysis we found that this is a difficult task when no gold standard annotations or clear guidelines are available. Various questions immediately come to mind: does the sentiment actually represent a journalist's or newspaper's belief or does it just tell something more about the topic at hand? For example, considering the news articles in the Guardian dealing with murder it might be that words such as "murder", "kill",... are actually included as subjective words within the lexicon. However, at the moment this latter question is overruled by our disputed topic filtering step, which discards topics that are negative across all news sites.

#### 5 Conclusions and Future Work

In this paper, we have discussed an approach which finds disputed topics in news media. By assigning sentiment scores and semantic categories to a number of news articles, we can isolate those semantic categories whose sentiment scores deviate significantly across different news media. Our approach is entirely unsupervised, requiring neither an upfront definition of possible topics nor annotated training data. An experiment with articles from six UK and US news sites has shown that such deviations can be found for different topics, ranging from political parties to issues such as drug legislation and gay marriage.

There is room for improvement and further investigation in quite a few directions. Crucially, we have observed that the assignment of topics is not always perfect. There are different reasons for that. First, we annotate the whole abstract of an article and extract categories. Apart from the annotation tool (DBpedia Spotlight) not working 100% accurately, this means that categories extracted for minor entities have the same weight as those extracted for major ones. Performing keyphrase extraction in a preprocessing step (e.g. as proposed by Mihalcea and Csomai [11]) might help overcoming this problem. In our approach, we only assign a global sentiment score to each article. A more fine-grained approach would assign different scores to individual entities found in the article. This would help, e.g. handling cases such as articles which mention politicians from different political parties. In that case, having a polarity value per entity would be more helpful than a global sentiment score. Furthermore, more sophisticated sentiment analysis combining the lexicon approach with machine learning techniques may improve the accuracy.

Our approach identifies many topics, some of which overlap and refer to a similar set of articles. To condense these sets of topics, we use categories' extensions, i.e. the sets of articles annotated with a category. Here, an approach exploiting both the extension as well as the subsumption hierarchy of categories might deliver better results. Another helpful clue for identifying media polarity is analyzing the *coverage* of certain topics. For example, campaigns of paid journalism can be detected by a news site having a few articles on products from a brand, which are not covered by other sites.

Although many issues remain open, we believe this provides a first seminal contribution that shows the substantial benefits of bringing together NLP and Semantic Web techniques for high-level, real-world applications focused on a better, semantically-driven understanding of Web resources such as online media.

## Acknowledgements

The work presented in this paper has been partly funded by the PARIS project (IWT-SBO-Nr. 110067) and the German Science Foundation (DFG) project Mine@LOD (grant number PA 2373/1-1). Furthermore, Orphee De Clercq is supported by an exchange grant from the German Academic Exchange Service (DAAD STIBET scholarship program).

## References

- Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. Opinionetit: Understanding the opinions-people network for politically controversial topics. In *Pro*ceedings of CIKM '11, pages 2481–2484, 2011.
- Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. In *Proc. of LREC'10*, 2010.
- Jon Bekken. Advocacy newspapers. In Christopher H. Sterling, editor, *Encyclopedia of Journalism*. SAGE Publications, 2009.
- Rob Ennals, Beth Trushkowsky, John Mark Agosta, Tye Rattenbury, and Tad Hirsch. Highlighting disputed claims on the web. In ACM International WWW Conference, 2010.
- Christiane Fellbaum, editor. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass., 1998.
- Blaž Fortuna, Carolina Galleguillos, and Nello Cristianini. Detecting the bias in media with statistical learning methods. In *Text Mining: Theory and Applications*. Taylor and Francis Publisher, 2009.

- Eduard Hovy, Roberto Navigli, and Simone Paolo Ponzetto. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. Artificial Intelligence, 194:2–27, 2013.
- Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. Unsupervised graph-based topic labelling using dbpedia. In *Proc. of WSDM '13*, pages 465–474, 2013.
- Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sren Auer, and Christian Bizer. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2013.
- Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In Proc. of the 7th International Conference on Semantic Systems (I-Semantics), 2011.
- Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In Proc. of CIKM '07, pages 233–242, 2007.
- 12. Finn Nielsen. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In Proc. of the ESWC2011 Workshop on Making Sense of Microposts: Big things come in small packages, 2011.
- 13. Heiko Paulheim, Petar Ristoski, Evgeny Mitichkin, and Christian Bizer. Data mining with background knowledge from the web. In *RapidMiner World*, 2014. To appear.
- Senja Pollak, Roel Coesemans, Walter Daelemans, and Nada Lavrač. Detecting contrast patterns in newspaper articles by combining discourse analysis and text mining. *Pragmatics*, 5:1947–1966, 2011.
- Ana-Maria Popescu and Orena Etzioni. Extracting product features and opinions from reviews. In Anne Kao and Stephen R. Poteet, editors, *Natural Language Processing and Text Mining*, pages 9–28. Springer London, 2007.
- Lawrence Reeve and Hyoil Han. Survey of semantic annotation platforms. In Proc. of the 2005 ACM symposium on Applied computing, pages 1634–1638. ACM, 2005.
- 17. Elad Segev and Regula Miesch. A systematic procedure for detecting news biases: The case of israel in european news sites. *International Journal of Communication*, 5:1947–1966, 2011.
- 18. Ralf Steinberger, Bruno Pouliquen, and Erik Van der Goot. An introduction to the europe media monitor family of applications. *CoRR*, abs/1309.5290, 2013.
- Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis.* MIT Press, Cambridge, MA, 1966.
- Mitja Trampus, Flavio Fuart, Jan Bercic, Delia Rusu, Luka Stopar, and Tadej Stajner. Diversinews a stream-based, on-line service for diversified news. In SiKDD 2013, pages 184–188, 2013.
- Marjan Van de Kauter, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Veronique Hoste. Lets preprocess: The multilingual lt3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120, 2013.
- Theresa Wilson, Janyce Wiebe, and Paul Hoffman. Recognizing contextual polarity in phrase-level sentiment analysis. In Proc. of HLT05, pages 347–354, 2005.

12