State of the Art in Knowledge Extraction from Online Polls

A Survey of Current Technologies

Martin Stabauer
Department of Data Processing in Social Sciences, Economics and Business
Johannes Kepler University
Linz, Austria
stabauer@idv.edu

Georg Grossmann
Knowledge and Software Engineering Laboratory
School of Information Technology & Mathematical Sciences
University of South Australia
georg@cs.unisa.edu.au

Markus Stumptner
Advanced Computing Research Centre
School of Information Technology & Mathematical Sciences
University of South Australia
mst@cs.unisa.edu.au

ABSTRACT

The ongoing research and development in the field of Natural Language Processing has lead to a great number of technologies in its context. There have been major benefits when it comes to bringing together the worlds of natural language and semantic technologies, so more and more potential areas of application emerge. One of these is the subject of this paper, in particular the possible ways of knowledge extraction from single-question online polls.

With concepts of the Social Web, internet users want to contribute and express their opinion. As a consequence, the popularity of online polls is rapidly increasing; they can be found in news articles of media sites, on blogs etc. It would be desirable to bring intelligence to the application of polls by using technologies of the Semantic Web and Natural Language Processing as this would allow to build a great knowledge base and to draw conclusions from it.

This paper surveys the current landscape of tools and state-of-the-art technologies and analyses them with regard to pre-defined requirements that need to be accomplished, in order to be useful for extracting knowledge from the results generated by online polls.

CCS Concepts

•**Computing methodologies** → **Information extraction; Lexical semantics;** *Ontology engineering;*

Keywords

Named Entity Recognition; Information Extraction; Online Polls; Semantic Technologies

1. INTRODUCTION

Natural Language Processing (often abbreviated as NLP) has achieved immense progress in the last years. Supported by increased computational power, there have evolved lots of different languages, modelling techniques and extraction, disambiguation or annotation approaches. Whilst some of these approaches focus on specific fields of knowledge like business specifications [28], others try to stay as generic as possible in terms of their knowledge domain. The latter usually focus on very specific purposes.

The great progress in NLP make numerous new potential areas of application possible. One of these are online polls, which will be described in the following sections.

1.1 Motivation and Background

Within the scope of an ongoing and very promising industry project we were confronted with the task to bring some intelligence to online polls. By default, such online polls consist of nothing more than "dumb" texts shaping the question as well as the possible answers. One can find such polls on various media sites like online newspapers, TV stations or blogs. Quite often, selected articles with good traction are being supplemented with polls in order to increase user participation and interaction.

Our research is based on a real-world data pool of more than 3,000 questions with more than 8,000 answers and more than 400,000 user votes given in the years 2014 and 2015. This data pool enabled us to conduct an analysis of a realistic scenario and see what polls created by the average user look like. The results give a good representation of the requirements for a system coping with these polls, which are discussed in detail in Section 2.

The objective of this project is to make these polls more intelligent by analysing their meaning and connect them with semantic concepts bundled in a knowledge tree. One wants to know what can be said about a person who chose a specific answer to a specific question. We could find no signs of research on this particular task; the closest research being done is specifically on question/answer pairs (see e.g., [4, 17]). However, as our focus is very specific, we decided to give an overview on a selection of related technologies and see what can be done by utilising them.

1.2 Online Polls

As mentioned above, the focus of this paper is single-question online polls with their very specific characteristics. Typically, these polls are created and published by media websites or blog owners to match the topic of specific articles and to provide input for statistical analyses; others are

ACSW '16 Multiconference, February 02 - 05, 2016, Canberra, Australia

DOI: http://dx.doi.org/10.1145/2843043.2843378

posed by web shops or companies who want to get to know their target group or need answers to specific questions with regard to market research. Figure 1 shows such a poll.

Figure 1: Online Poll

The main benefits of bringing more intelligence to the application of polls are:

- Clustering polls by their respective topic. This makes various applications possible like automatically showing a poll that fits the content of the article that it will be complementing, or showing users a related poll after they have answered a first one.
- Analysing the website visitors who answer one or more polls. By learning about their users, publishers can verify current assumptions about their target groups and get to know entirely new groups.
- These target groups can then be displayed graphically and/or be used as input for retargeting advertisements across platforms.
- Building up a knowledge base of relevant information that can be connected to other linked data available online.
- Combining the answers of several independent polls for creating more detailed user profiles. The website content can then be adapted to better fit the user profiles.

The online polls of this survey always consist of one question and two or more answer options. As it can be regular users who draft these questions, there are several possible question/answer combinations regarding the knowledge that needs to be extracted that have emerged from our data pool and have to be considered:

- Polls, where the information is embedded mainly in the question. This is the case with Yes/No polls like *Have you ever been to Russia?*
- Polls, where the knowledge to be extracted is embedded mainly in the answers, like *What applies to you? – Vegan / Vegetarian / Omnivore.*
- Polls, where the information is divided, e.g. *How often do you play video games? – Daily / Weekly / Monthly / Never.*
- Polls, where the answers are dependent on each other, like sequences as *Which do you like better? – Cat / Dog.* A person who answers *Cat* to this question should only be seen as somebody who likes cats better than dogs, not as somebody who likes cats in general.

1.3 Related Work

Several fields of research related to our task have emerged in the last years. Whilst they have set their center of attention not really at our specific issue, they still can deliver a valuable contribution and are definitely worth mentioning here. There is, for example, document categorisation, which algorithmically tries to assign a document or a text to one or more classes, see e.g., [2] or [27]. Another major area of research is sentiment analysis, which aims to identify subjective information such as the authors attitude in the input text (e.g., [10, 18, 23]).

The approach of concept mining already comes quite close to one of the main requirements in this paper as it tries to extract concepts taken from some kind of thesaurus out of documents. These thesauri are either created specially for the respective problem, or one of the well-known pre-defined concept models is used, e.g., WordNet [31], or the Wikipedia centred DBpedia [21] and Freebase [5], or those trying to connect these two worlds like BabelNet [22].

The ongoing research on information extraction brings together some of the aforementioned fields as well as numerous other aspects like coreference resolution or relationship extraction (e.g., [9, 11, 20]). The extraction of information leads to the task of storing and managing the obtained knowledge. Knowledge representation and reasoning in ontologies is the area of research that focuses on these issues. Knowledge representation has been well known to research for decades (e.g., [3]). However, enormous progress has been made especially in the last years (e.g., [1, 15]).

1.4 Contribution

The contribution of this paper is three fold: (1) Based on our use case we identified a set of requirements which can be used to select appropriate analysis and knowledge extraction tools, (2) a list of existing state-of-the-art tools have been identified and a preliminary evaluation with our data set has been conducted, and (3) the tools have been compared based on their underlying functionality and applicability for online polls.

2. REQUIREMENTS

A system that copes with knowledge extraction from online polls has to deal with several specific issues. Our analysis of the data pool exposed several particularly important requirements. Some of them are very characteristic for online polls, like the ability to combine questions and answers or identifying and sorting out unsuitable polls.

It seems quite obvious that not all the requirements can be covered by one single system as they are fairly wide spread. Nonetheless, by combining several technologies one could create a suitable solution for appropriately solving the task.

2.1 Knowledge Extraction

This point can be seen as the main requirement and the motivation for our research. The knowledge about people who answer questions online that can be drawn from the respective polls should be modelled in a structured way. Consequently, this knowledge serves several purposes like drawing conclusions and combining several pieces of information to create new knowledge. As aforementioned, NLP tools typically divide this task into two separate techniques: Named Entity Recognition and Entity Linking. Both will be

necessary and are depicted in Figure 2. The example shows the detection of the entity "Bob Marley" as a person and two possible matches, each with its own confidence value.

2.2 Scope: Everything

Whilst many NLP technologies are specialised on a specific topic, or at least the topic is known, this condition does not apply to our online polls. They have very diverse topics, which can't really be narrowed down by any means. This implies that many specific knowledge bases for certain domains can only be used if they are connected and made compatible.

Sometimes, the scope of a poll is limited to a certain region or time, e.g. *Who have you voted for in South Africa's General Elections 2014?* or *Who will win the football world cup?* It would be desirable to determine the specific scope and only show polls to users who are interested in that scope, while not discommoding anybody else.

2.3 Paraphrasing

As the knowledge that needs to be extracted from a poll quite often lies within both the question and the answers, somehow question and answer need to be combined. This can be seen as an exceptional requirement for this kind of input text. Simply analysing the question *How often do you play video games?* will not bring the desired knowledge about the answering user that can be modelled later on. The respective answer also needs to be considered by rephrasing the poll to *I play video games daily.*

In addition, even for simple Yes/No questions there is a need for a positive and a negative version of the question asked. We need to have both *Yes, I have been to Russia* and *No, I have never been to Russia* to get the desired knowledge.

Finally, one must always bear in mind that it is always a specific user these phrases apply to. Semantically correct, the latter two phrases should read *Person X has been to Russia* and *Person Y has never been to Russia* with proper links to the respective users.

2.4 Differing Semantics

Almost certainly, the different types of polls described in Section 1.2 need to be handled in a different way. In order to do that we have to be able to distinguish between them. In addition, not every question is posed following certain logical rules. Some questions are asked using bad grammar or with problematic answer combinations. For example, the poll *Would you ever go skydiving? – Yes, absolutely / Yes / Maybe / I don't think so* provides answers that are for one thing too similar, for another thing very difficult to combine with the question.

Sometimes this even leads to manipulative or suggestive question/answer combinations, where the asking person desires to achieve a certain outcome. In extreme cases, a poll needs to be omitted and not be taken into consideration for further analysis.

2.5 Multiple Entities/Relations

There may be more than one entity of knowledge and/or more than one relation to them in one poll. The system should be able to discover as many of them as possible. For example, somebody who answers *Daily* to the question *How often do you play video games?* should not only be marked as *plays video games daily* but also as *likes video games.* This reasoning can be carried out in the knowledge base or when annotating the poll. There are tools available to discover the relatedness of two sentences, for one of these tools see section 3.7.

2.6 Multiple Languages

Research on NLP mainly focuses on texts in English, other languages have widely been ignored or found way more difficult to implement. Some advantages of the English language in comparison to most others are the rather simple morphology and strict word order. For recent research on German language see e.g., [16] or [26]. Significant progress has also been made in multilingual entity extraction [8].

Whilst English is also the main language in our data pool, there exists an increasing number of polls in other languages, e.g., German, Russian, Spanish, Italian etc.

2.7 Building a Knowledge Base

The knowledge extracted from the polls needs to be structured in some way. As the knowledge domain cannot be narrowed down it could make sense to make use of existing ontologies like the DBpedia as a starting point. Semantic technologies allow for connecting entities in very refined ways, this capability should be put to use as future applications involving polls and the knowledge of their users could benefit from it.

2.8 Manual Intervention

If the system gets something wrong like missing an entity or linking it to a wrong resource, the team of administrators needs to be able to intervene if necessary. Consequently, the system should learn from these interventions and take them into consideration when annotating future polls. It would also be favourable to have recursive self-learning implemented, which also is the approach of modern machine learning systems like discussed e.g., in [30].

2.9 Performance

As the number of polls being asked every day is predicted to increase exponentially, the performance of any algorithm extracting knowledge from these new polls cannot be ignored. There is no specific benchmark the tools need to achieve, but they should be able to handle larger numbers of polls in reasonable time on standard server hardware.

3. SURVEY OF TECHNOLOGIES

As there is a lot of research being done on all kinds of NLP technologies, there are more and more tools available. We are aware of the variety and tried to narrow down the selection for this paper to technologies that are available today and can solve at least one of the requirements described in the last section. This survey does not intend to be complete; in addition to the tools we discuss in this section, there are many more very close to our research topic but that don't fit in completely.

In addition there are rapid changes going on in this field of research. Development of earlier tools is ceased, others like the Ontotext KIM Platform [24] have gone commercial. In this context our selection has to be seen as a snapshot in time. For an overview of alternative technologies like FRED, OpenCalais and others see generic comparison papers like [13] or [14], respectively.

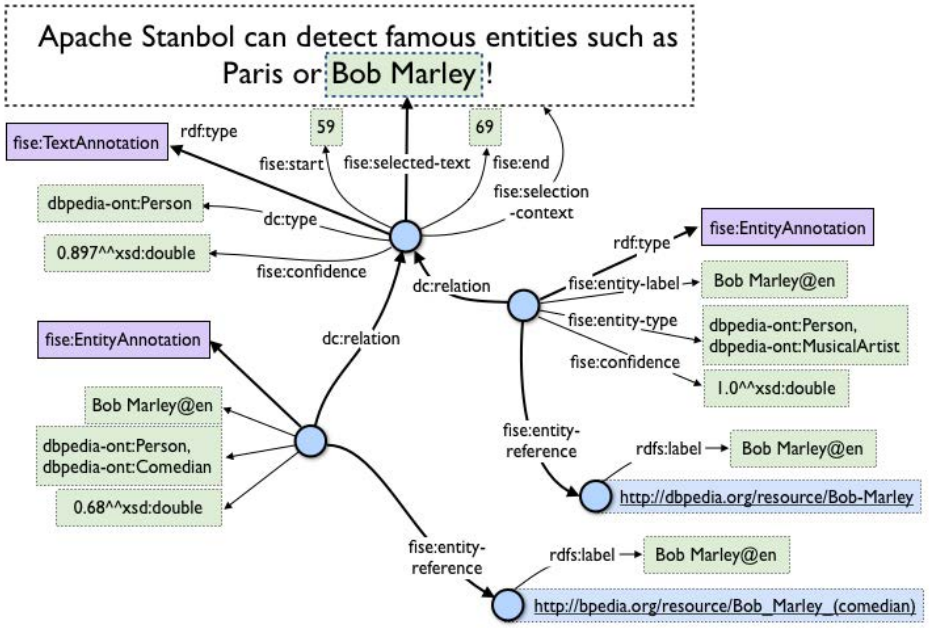

Figure 2: Entity Disambiguation with Apache Stanbol[1]

As the tools surveyed follow very diverse approaches and serve very different purposes there is no scale or performance indicators that would help to assess the candidates quantitatively. We followed a qualitative approach and tried to find the specialties of each technology with the focus on our specific task and requirements.

The following abbreviations are used in this section:

- NLP
  Natural Language Processing can be seen as the superclass to the various technologies explained below.
- NER
  Named Entity Recognition aims at classifying atomic elements in the text into semantic categories like e.g. Person or Location.
- POS
  Part-of-Speech tagging tries to annotate each word with a unique tag indicating its syntactic role. These roles are e.g. noun, adverb, plural, etc. This is not directly applicable to our research task, however, many technologies combine POS with more relevant functions.
- SRL
  Semantic Role Labelling detects shallow semantic arguments of an input sentence like different types of subject (this could be agents, forces and others), object (themes, results, etc.) and relation (verbs) as well as several adjuncts.
- CHK
  Chunking up is the task of finding more general categories for entities like for example generalising from *Dog* to *Animal*.
- NED
  Named Entity Disambiguation determines the identity of found entities in a text and typically links them to a popular knowledge base like DBpedia or Wikidata.

The following subsections discuss the selected and tested tools. Table 1 shows a summary of the overall results.

3.1 ANNIE [NER, POS]

ANNIE is an information extraction plugin in the open source text processing environment GATE [12]. It is multilingual and able to identify entities such as locations, persons, organisations, dates and identifiers from the online poll questions. Although not 100% accurate, e.g., "Victorian" and "Absinthe" were not identified, the majority of entities were tagged correctly. ANNIE is a plugin in GATE, which comes with a user interface as shown in Figure 3. ANNIE also includes a POS Tagger but in some cases the questions were too short and the tagger could not be applied.

ANNIE supports a majority of requirements as shown in Table 1. Support for *Differing Semantics* could not be found though.

3.2 CoreNLP [NER, POS]

CoreNLP [19] is a set of open source natural language tools. Similar to ANNIE, it is multilingual and is able to tag entities with categories like persons, locations and organisations. Like in the previous, "Absinthe" was not recognised but "Victorian" was identified as an adjective and assigned to group *miscellaneous*. The processing is executed on the command line and results are created in an XML file (partly shown in Figure 4) which can be processed further. In addition CoreNLP also provides sentiment values to each token.

Similar to ANNIE, support for *Differing Semantics* and a Knowledge Base could not be identified. Also the support for a domain independent scope and paraphrasing was limited.

[1]https://stanbol.apache.org

Table 1: Summary of Tool Comparison Tests

Requirement	ANNIE	CoreNLP	Senna	Alchemy	Spotlight	Stanbol	MetaMind
Knowledge Extraction	+	+	−	+	+	+	−
Scope: Everything	+	−	−	+	+	+	+
Paraphrasing	+	−	−	−	−	−	+
Differing Semantics	−	−	−	−	−	−	−
Multiple Entities	+	+	+	+	+	+	−
Multiple Languages	+	+	−	−	+	+	−
Knowledge Base	−	−	−	−	−	+	+
Manual Intervention	+	+	+	−	−	+	−
Performance	+	+	+	+	−	−	+

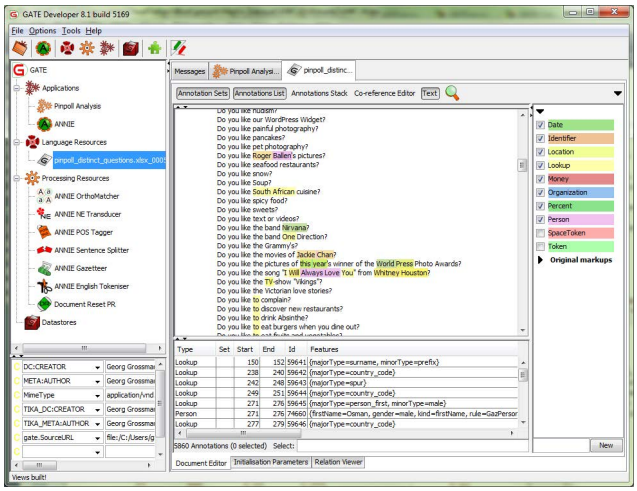

Figure 3: ANNIE information extraction in GATE[3]

```
...
<token id="14">
<word>Whitney</word>
<lemma>Whitney</lemma>
<CharacterOffsetBegin>26300</CharacterOffsetBegin>
    <CharacterOffsetEnd>26307</CharacterOffsetEnd>
<POS>NNP</POS>
    <NER>PERSON</NER>
    <sentiment>Neutral</sentiment>
</token>
<token id="15">
<word>Houston</word>
<lemma>Houston</lemma>
<CharacterOffsetBegin>26308</CharacterOffsetBegin>
    <CharacterOffsetEnd>26315</CharacterOffsetEnd>
<POS>NNP</POS>
    <NER>PERSON</NER>
    <sentiment>Neutral</sentiment>
</token>
...
```

Figure 4: Part of the CoreNLP result

3.3 Senna [POS, CHK, NER, SRL]

Senna is a neural network architecture and learning algorithm that combines various NLP tasks [6, 7]. Table 2 shows the results for an example sentence. The first column contains the "tokens" that were found. These are followed by the POS, CHK, NER and SRL results. For each SRL verb found there is one column after that. For example, column 6 shows A0 (the "Thinker"), S-V (think as a verb) and A1 (the "Thought"), whereby in our example A0 matches the user who has given that answer to the question mainly contained in A1. A0 is therefore correctly identified as "Person" (see column 4).

Whilst Senna can be of great use for analysing the underlying structure of questions and answers, it does not allow for many other tasks emphasised in our requirements. E.g., there is no support for paraphrasing, named entity disambiguation or building a knowledge base. In regard to our requirements, it turned out that Senna can be used particularly well for preliminary tasks like semantic role labelling or named entity recognition.

3.4 AlchemyAPI [CHK, NER, SRL, NED]

AlchemyAPI[4] is a toolset that has crossed the borders of the academic world and is being used commercially. It can be seen as one the most popular web-based text classification APIs [25]. In 2015, IBM bought AlchemyAPI to support its own Watson system. There are numerous APIs for NLP tasks available, many of them being useful for our problem. Detecting the language of a text is one of them, finding entities and concepts and discovering the texts sentiment are others.

The results of concept extraction are of varying usefulness. For example, one of the highest ranked concepts in the sentence *Person XY prefers iOS over Android* is *Google*. This is misleading, as for this specific answer iOS related concepts are way more relevant. Entities are mostly detected correctly, although linking to DBpedia and Freebase URLs did not always work in our tests.

Apparently, some features of AlchemyAPI like language detection have problems with the relatively short texts of online polls. These features work more reliably and precisely with longer texts including more content.

AlchemyAPI fulfils a number of our requirements defined in Section 2, e.g. is there a good support for knowledge extraction and it enables the users to include polls with diverse scopes and multiple entities. The downsides lie in tasks like paraphrasing or allowing for manual intervention. In addition, there is only rudimentary support for text inputs in other languages than English.

[3]https://gate.ac.uk/gate/doc/plugins.html#ANNIE
[4]http://www.alchemyapi.com

Table 2: Senna results

Person	NNP	B-NP	O	-	B-A0	O	(S1(S(NP*
XY	NNP	E-NP	S-PER	-	E-A0	O	*)
thinks	VBZ	S-VP	O	thinks	S-V	O	(VP*
that	IN	S-SBAR	O	-	B-A1	O	(SBAR*
solar	JJ	B-NP	O	-	I-A1	B-A0	(S(NP*
energy	NN	E-NP	O	-	I-A1	E-A0	*)
can	MD	B-VP	O	-	I-A1	S-AM-MOD	(VP*
save	VB	E-VP	O	save	I-A1	S-V	(VP*
the	DT	B-NP	O	-	I-A1	B-A1	(NP*
world.	NN	E-NP	O	-	E-A1	E-A1	*))))))))

3.5 DBpedia Spotlight [NED]

DBpedia Spotlight is an open-source system for annotating texts specifically with DBpedia resources [21]. It provides a REST-based web service for recognising phrases and entity linking. Like most tools it was originally designed for the English language, but today it supports several other major languages as well and with steadily improving performance [8].

The main task of finding DBpedia entities in the input text is done with great confidence and performance, polls in other languages than English can also be annotated with DBpedia URIs. The mobile operating systems in the German poll *Bevorzugst du iOS oder Android?* were found and linked correctly.

Regarding the derived requirements it turned out that DBpedia Spotlight can be very helpful for knowledge extraction from polls of very diverse scopes and languages. On the other hand, there is no support for rephrasing questions and answers or for automatically/manually sorting out unusable polls.

3.6 Apache Stanbol [NER, SRL, POS, CHK, NED, etc.]

Apache Stanbol is an integrated NLP framework for a variety of tasks to our research challenge. It works with so called Enhancement Chains that define which engines in what order are used to process the input content. There are several engines preconfigured with the possibility to include one's own extensions. The available engines include a large number of OpenNLP-based services, OpenCalais integration, DBpedia Spotlight annotation and many more. The results of the Enhancement Chain can then be processed by other Stanbol components like OWL Reasoners or inference rule engines. Figure 5 shows the big picture.

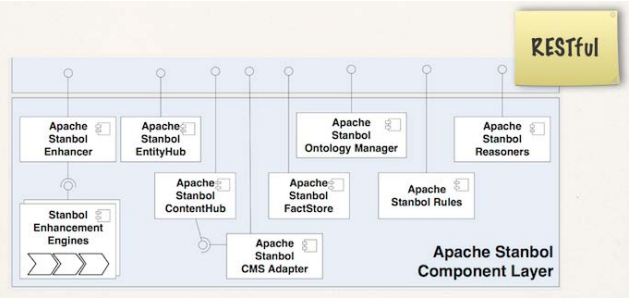

Figure 5: Apache Stanbol Components[5]

Interestingly enough, in our tests some entities that were correctly identified and linked by DBpedia Spotlight were not found by the Stanbol engines. Not only was the food in *Do you like sushi?* missed, the sentence was also marked as Swahili language. However, as Apache Stanbol allows for developing and integrating one's own extensions, it might emerge as a useful platform for many of the discovered requirements for online polls.

Despite the discovered downsides, Apache Stanbol fulfils many of the presented requirements. The most outstanding feature is its plug-in architecture, which may help to overcome the missing features in the future. At the moment, these are especially paraphrasing and categorising polls. In addition, some performance issues during our tests reduced the usability.

3.7 MetaMind

While MetaMind[6] mainly focuses on image recognition and sentiment analysis, there also is a tool based on dependency tree recursive neural networks (first introduced in [29]) that allows to calculate the relatedness of two sentences. E.g. *Person XY has been to Russia* is related to the sentence *Person XY likes travelling* with a factor of 4 where 1 would be not related at all and 5 an almost perfect paraphrase. These results could help drawing conclusions about a user.

In regard to the requirements in Section 2, MetaMind can be used for a limited number of use cases. Its very specialised set of features allows for checking the relatedness of two sentences and consequently the quality of paraphrased polls. This also helps with building up a knowledge base of related concepts.

4. FUTURE RESEARCH DIRECTIONS

Regarding the requirements specified in Section 2 it turned out that with tools currently available they can be met with very differing success. Whilst the usage of well-known knowledge bases like DBpedia or Wikidata lead to good results in entity recognition and disambiguation, there is still only very little research on paraphrasing for combining question/answer pairs.

As the requirements are very diverse, a single solution for all requirements is not yet in sight. Therefore, this study offers a first step towards finding a combination of technologies that can satisfy as many requirements as possible. Especially tool support for multiple languages and paraphras-

[5]https://stanbol.apache.org
[6]https://www.metamind.io

ing have proven unsatisfactory and further research in this domain is absolutely essential.

The relevance of online polls is consistently increasing and bringing more intelligence seems as much a desirable objective as bringing the power and possibilities of the Semantic Web to further applications. All it needs is a reliable and effective solution for knowledge extracting from question/answer combinations as described, but at least today some building blocks are still missing.

5. CONCLUSION

We have presented the results of a survey in the area of knowledge extraction from online polls. We have analysed the requirements for a system coping with the specific characteristics of online polls and have investigated current state of the art technologies in this field. One conclusion of this study is that some technologies deliver reliable and effective results while others lack consistency or support for our specific input.

It turned out that no single system can fulfil all requirements and it seems unrealistic to expect one in the foreseeable future. Therefore, best chances lay in systems that are designed as platforms to host and combine multiple technologies. These platforms like Apache Stanbol allow for combining the best of various other research outcomes.

6. ACKNOWLEDGMENTS

Georg Grossmann was partly funded by the Data to Decisions Cooperative Research Centre (D2D CRC).

7. REFERENCES

[1] S. Amir and H. Ait-Kaci. Cedar: Efficient reasoning for the semantic web. In *Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems*, pages 157–163. IEEE, 2014.

[2] R. Bekkerman and M. Gavish. High-precision phrase-based document classification on a modern scale. In *Proceedings of the KDD*. ACM, 2011.

[3] T. J. M. Bench-Capon. *Knowledge Representation: An Approach to artificial intelligence*. Academic Press Ltd, 1990.

[4] J. Berant and P. Liang. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1415–1425, 2014.

[5] K. Bollacker, R. Cook, and P. Tufts. Freebase: A shared database of structured general human knowledge. In *22nd National Conference on Artificial Intelligence*, 2007.

[6] R. Collobert. Deep learning for efficient discriminative parsing. In *International Conference on Artificial Intelligence and Statistics*, 2011.

[7] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, August 2011.

[8] J. Daiber, M. Jakob, c. Hokamp, and P. N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems*, 2013.

[9] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. ACL, 2011.

[10] R. Feldman. Techniques and applications for sentiment analysis. In *Communications of the ACM*, volume 56, pages 82–89, 2013.

[11] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the ACL*, pages 363–370, 2005.

[12] R. Gaizauskas, H. Cunningham, Y. Wilks, P. Rodgers, and K. Humphreys. GATE - an environment to support research and development in natural language engineering. In *Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence*, 1996.

[13] A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In *The Semantic Web: Semantics and Big Data*, pages 351–366. Springer Berlin Heidelberg, 2013.

[14] S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating NLP using linked data. In *The Semantic Web-ISWC*, pages 98–113, 2013.

[15] L. Hotz, A. Felfernig, M. Stumptner, A. Ryabokon, C. Bagley, and K. Wolter. *Knowledge-Based Configuration*, chapter Configuration Knowledge Representation and Reasoning. Elsevier Science Inc, 2014.

[16] L. Kallmeyer and W. Maier. Data-driven parsing using probabilistic linear context-free rewriting systems. *Computational Linguistics*, 39(1):87–119, March 2013.

[17] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, M. Iyyer, I. Gulrajani, and R. Socher. Ask me anything: Dynamic memory netorks for natural language processing. *CoRR*, abs/1506.07285, 2015.

[18] B. Liu. *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers, May 2012.

[19] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[20] S. Martschat and M. Strube. Recall error analysis for coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2070–2081. ACL, 2014.

[21] P. N. Mendes, M. Jakob, A. Garcia-Silva, and C. Bizer. DBpedia Spotlight: Shedding light on the web of documents. *I-Semantics*, 2011.

[22] R. Navigli and S. P. Ponzetto. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, 2010.

[23] B. Pang and L. Lee. Optinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[24] B. Popov, A. Kiryakov, A. Kirilov, D. Manov,

D. Ognyanoff, and M. Goranov. Kim - semantic annotation platform. In *The Semantic Web-ISWC*, pages 834–849, 2003.

[25] D. Quercia, H. Askham, and J. Crowcroft. TweetLDA: Supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 247–250, 2012.

[26] A. Rafferty and C. D. Manning. Parsing three german treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German at ACL*, pages 40–46, 2008.

[27] T. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

[28] M. Selway, G. Grossmann, W. Mayer, and M. Stumptner. Formalising natural language specifications using a cognitive linguistic/configuration based approach. *Information Systems*, 2015.

[29] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[30] R. Socher, C. C.-Y. Lin, A. Y. Ng, and C. D. Manning. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[31] P. Vossen, E. Agirre, N. Calzolari, C. Fellbaum, S.-K. Hsieh, C.-R. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon, and J. VanGent. KYOTO: A system for mining, structuring, and distributing knowledge across languages and cultures. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, 2008.