Converting Opinion into Knowledge
Improving User Experience and Analytics of Online Polls

Martin Stabauer(✉), Christian Mayrhauser, and Michael Karlinger

Johannes Kepler University, Linz, Austria
martin.stabauer@jku.at

Abstract. A vast majority of internet users has adopted new ways and possibilities of interaction and information exchange on the social web. Individuals are becoming accustomed to contribute and express their opinion on various platforms and websites. Commercial online polls allow operators of online newspapers, blogs and other forms of media sites to provide such services to their users. Consequently, their popularity is rapidly increasing and more and more potential areas of application emerge. However, in most cases the expressed opinions are stored and displayed without any further actions and the knowledge that lies in the answers is discarded.

This research paper explores the possibilities, advantages and limits of applying semantic technologies to these online polls. For this purpose, a list of requirements was assembled and possible system architectures for semantic knowledgebases were investigated with the focus on providing consistent and extensive data for further processing. In a next step, the current state of the art of relevant visualization technologies was analyzed and further research challenges were identified.

Our results discuss possible applications within the scope of a challenging case study. A comprehensive data pool provided by our industry partner allows for testing various improvements to user experience and traction of the polling system.

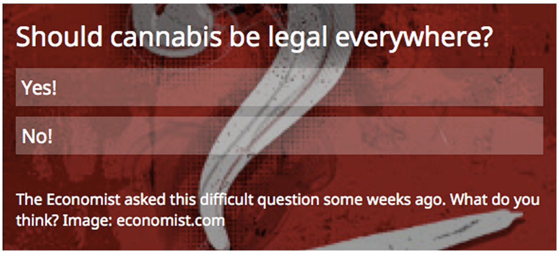
Keywords: Online polls · Named entity recognition · Information extraction · Semantic technologies · Ontology engineering · Dashboards · Graphical user interfaces

1 Introduction

1.1 Online Polls

Online polls are becoming more and more popular on a large variety of websites, e.g., online newspapers, blogs and other forms of media sites. These single-question polls allow users of the respective sites to express their opinion and contribute to the outcome of a question drafted by the operator of the website. This opportunity is appreciated very much by large numbers of internet users. Figure 1 shows such an online poll.

F.F.-H. Nah and C.-H. Tan (Eds.): HCIBGO 2016, Part I, LNCS 9751, pp. 330–340, 2016.
DOI: 10.1007/978-3-319-39396-4_30



Should cannabis be legal everywhere?

The Economist asked this difficult question some weeks ago. What do you think? Image: economist.com

Fig. 1. Online Poll

Many of today's single-question online polling systems are operated by the website owners themselves. Giving their users the possibility to contribute to parts of the site's content is first priority for most systems; it seems that great usability or analytical features are only insufficiently considered. However, gaining knowledge from the answers to polls – in contrast of simply displaying the results and then discarding the information contained therein – could bring enormous benefits:

- Polls can be clustered and categorized 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.
- The website visitors who answer one or more polls can be analyzed in regard to their specific attitudes and preferences. 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, e.g., via Venn or Euler diagrams. This gives the website operators a better overview of their users and allows them to select specific groups of persons for further actions.
- One of these further actions is using the target groups as input for retargeting advertisements across platforms. This type of advertising has emerged as one of the most widely used across the internet and facilitates custom-tailored ads for segmented user groups.
- A semantic knowledgebase of relevant information can be generated and consequently connected to other linked data available online. This database is intended to show interconnections and dependencies in a more detailed and precise way than classic relational data structures. New knowledge can be discovered by techniques of the semantic web like reasoning.
- Combining the answers of several independent polls for creating more detailed user profiles becomes feasible. The website content can be adapted to better match the discovered user profiles.
- Extracting and aggregating private information about claims that can not be verified turn out to be of great value and online polls can contribute in doing that. [11]

1.2 Methodology and Contribution

Natural Language Processing (NLP), the languages of the semantic web and other technologies relevant to this study's field of research have achieved immense progress in the last years. This paper explores the possibilities of tools and techniques for improving user experience for both common internet users and website operators. While the former can profit from a better quality of suggestions for further articles and polls as well as from better categorized question/answer pairs, the latter can benefit from a greatly improved admin dashboard that provides whole new possibilities for analyzing and illustrating the outcomes of their polls and use them as basis for further applications.

To achieve these advances, current state-of-the-art technologies in the field of knowledge extraction (KE) from natural language question/answer pairs are analyzed. Consequently, the implementation of a semantic knowledgebase designed specifically to the requirements of online polling systems is demonstrated within the scope of an extensive case study based on a real-world data pool of more than 10,000 questions with approx. 36,000 answers and 653,000 user votes given worldwide in the years 2014 and 2015. This is followed by discussing UI elements for the administrator's dashboard like displaying suggestions for future target groups by employing Euler diagrams as well as prototypical advertising capabilities.

1.3 Related Work

Great advances in NLP and the Semantic Web have led to various fields of research related to our task. Recent examples are NLP for question/answer pairs as described in [14] or [5], and paraphrasing (e.g., [2]). Knowledge representation has been in the center of attention of research for decades (e.g., [4,13]) and still great progress is being achieved (e.g., [1,3,8]).

Applications of the Semantic Web like DBpedia [15] and other knowledgebases are making it possible to link shared knowledge and build new solutions on top of existing ones. While early research was mainly done for texts in English, globalization of Linked Data brought the necessity to deal with different languages. Significant progress has been made in multilingual entity extraction [6]. This is exemplified by research on German language (e.g., [12,17]).

Modern approaches suggest recursive self-learning methods when it comes to entity detection and extraction [9] and various methods for semi-automatic ontology development [16].

2 Semantic Technologies

2.1 Knowledge Extraction

To be helpful for extracting knowledge from single-question online polls, a semantic system needs to fulfil a number of requirements. Some of these are:

- While most ontologies and NLP technologies specialize on a certain domain, online polls can relate to pretty much everything. In most cases, the topic of a poll is not even known in advance. This means that if there are external knowledgebases for specific domains involved, they need to be made compatible and get interconnected. Finding out the scope of a poll is also very important to determine the relevance for specific user groups. The scope of some questions is limited to a certain time or region, e.g. *Who will win the football world cup?* or *Who have you voted for in South Africa's General Elections?*.
- The knowledge to be extracted in many cases lies within the question combined with the chosen answer. The question *How often do you play video games?* by itself does not contain any knowledge about the user who answers the question, the chosen answer *Daily* needs to be taken into consideration as well. This means that there is the need to combine question and answer and/or to paraphrase the question. Even relatively simple Yes/No polls like *Hillary Clinton: First female U.S. President?* need to be converted to a positive and a negative version for further processing.
- The focus in NLP research traditionally lies on English language. However, the online polls in our case study are being created in many different languages. The findings of NLP for texts in English can not be transferred to other languages, but there is an increasing number of research projects on multilingual entity extraction.
- Another special requirement for a KE system for online polls is that it needs to have the capability to sort out inappropriate, manipulative or suggestive polls. These can occur, because most of the polls in the data pool are created and published by media websites or blog owners who can pose all the questions they like. These "bad" polls should not be taken into consideration for further analysis.

A comprehensive set of requirements for a system coping with knowledge extraction from online polls can be found in the work of Stabauer, Grossmann and Stumptner [20]. For a comparison of knowledge extraction tools, see [7]. At the time of writing there is no known system able to deal with all the aforementioned special requirements. Therefore, there is the need for an alternative way of annotating polls manually. Figure 2 shows a mockup. For each possible answer to a question the administrator can choose from predefined relations and semantic concepts, thus creating new object properties in the knowledgebase.

2.2 Knowledgebase

The semantic knowledgebase in our case study is embedded in a complex system architecture, Fig. 3 gives an overview. The analytical subsystem is responsible for storing and analyzing the knowledge that is extracted from the polls and the users' answers, respectively. Some information in the analytical subsystem (e.g., meta information of polls and votes) is stored in an auxiliary database. This divide is due to size limitations in the semantic knowledgebase and to the structure of the data to be stored. All databases and systems work together seamlessly

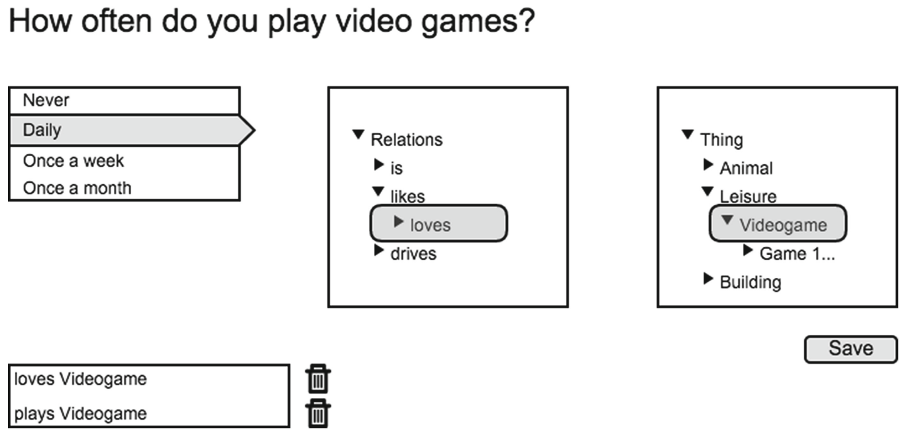


Fig. 2. Semantic Annotation

and enable the administrator to gradually build up a consistent, accurate and extensive knowledgebase.

Following the approach of semantic annotation as depicted in Fig. 2, knowledge about a specific user is stored in RDFS/OWL triples as follows:

```
<http://polling.com/pollees#Pollee123>
        <http://polling.com/ontology#loves>
                <http://polling.com/ontology#Videogame>.
```

To comply with the ideas of Linked Data, the concepts in the knowledgebase (*#Videogame* in the example above) are linked to external knowledgebases. In this case there might be a link to DBpedia as follows:

```
<http://polling.com/ontology#Videogame>
        rdfs:subClassOf
                <http://dbpedia.org/ontology/VideoGame>.
```

2.3 Reasoning and Analysis

The knowledge about users and concepts that is stored in the analytical subsystem, is consequently being analyzed by a series of algorithms, beginning with standard RDFS and OWL reasoners. This enables clustering users by their preferences and characteristics and so very advanced retargeting applications become feasible. Additionally, users can profit as they are given suggestions for further articles and polls that meet their specific interests without breaking the simplicity and anonymity of the polling process. Immediate results of the reasoning process are sets of persons with specific properties, their intersections and the respective set sizes. These will be used for visualization in the administrator dashboard and for further applications in retargeting advertising.

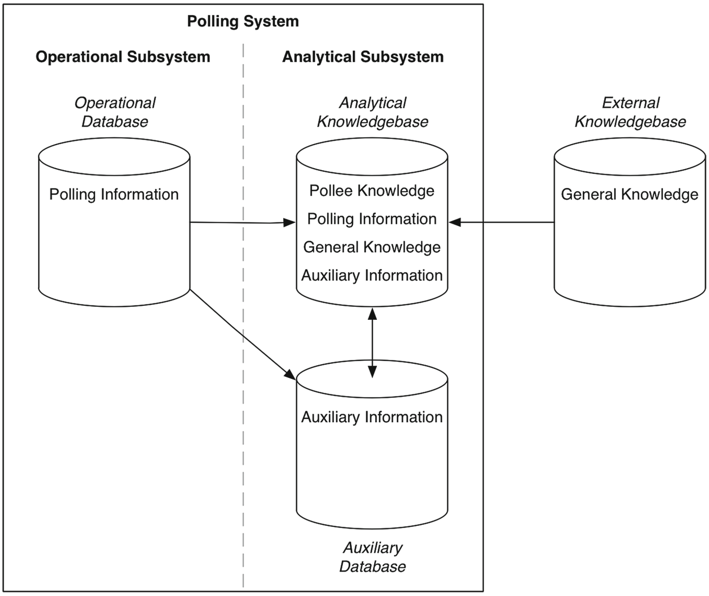


Fig. 3. Polling System Architecture

3 Knowledge Visualization

The consequent step after building up a consistent and extensive knowledge base is to use that data to improve certain aspects of the polling process. In this case we try to enhance the administrator's user experience and provide additional functionality by visualizing the findings of the analysis within the framework of the existing dashboard. This chapter gives an overview of existing means of visualizing bigger data sets and their issues.

Common representation techniques can be used to gain new knowledge about given data. Both Euler and Venn diagrams show the size of data sets that are built from a data pool as well as the correlations between these sets. Euler and Venn diagrams use geometrical shapes to represent data sets and intersecting sets, which contain named data sets. To build an Euler or Venn diagram two sets of data are needed: A set to store all data sets M and a set of intersections U, where every set $e \in M$ is at least part of one intersection set in U.

Following the definition of Venn diagrams, all possible intersections have to be shown, and intersections that are not contained in U must be marked as empty. Figure 4 shows common Euler and Venn diagrams of the sets $M = \{A, B, C\}$ and $U = \{A, B, C, AB, AC\}$ and reveals the problem of Venn diagrams in regard to empty intersection sets. If the amount of sets in M increases, Venn Diagrams get more and more complex and harder to understand. This problem makes it clear that Venn diagrams can not be efficiently used for the visualization of bigger sets such as the ones in our case study.

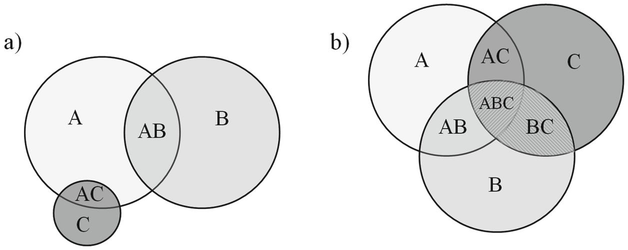


Fig. 4. Examples of an Euler diagram (a) and a Venn diagram (b)

3.1 Visualization Technologies

Calculating Euler diagrams gets more complicated and difficult when the complexity of the diagram grows [19]. The complexity is defined by the amount of sets in the intersection sets in U, which can be as high as 2^n, whereby n equals the amount of sets in M. More issues become apparent when the complexity increases. To verify if an Euler or Euler-like diagram (such as described in e.g., [19]) is correctly drawn, an agreement about the visualization is needed. The following points need to be fulfilled by the diagram in order to be recognized as correctly drawn:

- Every set $e \in M$ needs to be visualized by at least one marked geometrical form and e has to be at least part of one set in U.
- The area of the used geometrical form representing $e \in M$ has to be sized in reference to the amount of elements in the set e.
- For every set $e \in M$ that is part of an intersection set $i \in U$, which contains more than one set, there needs to be an area where every set $s \in i$ intersects with every other set in i.
- If two or more geometrical forms are intersecting, there has to be at least one intersection set in U that contains the sets of the intersecting forms. Furthermore, the intersecting area has to be sized according to the amount of elements in the geometrical form.
- Only sets that are part of M and intersection sets that are part of U are allowed to be visualized.

Stapleton et al. describe 3 base methods to generate Euler diagrams: Dual graphs, inductive and using particular shapes [21]. The first method calculates a dual graph based on the intersection sets in U and draws the geometrical forms in a way, so that every geometrical form representing a set $e \in M$ contains every node that is included in the set e. The inductive method uses a step by step procedure to calculate the position of the geometrical forms. The last method changes the shapes of the geometrical forms to create drawable Euler diagrams.

3.2 Issues and Solutions

On the basis of the afore-mentioned generation methods and the used geometrical forms, several issues become apparent. Euler diagrams using circles as geometrical indicator of sets, such as shown in Fig. 4, can not always be drawn in a way that every requirement defined in Chap. 3.1 is fulfilled. As an example, a diagram containing the set $M = \{A, B, C\}$ and the set $U = \{A, B, AC, BC\}$ is not drawable with one circle per set $e \in M$ and without creating a fictional intersection set with C and $\{\}$ in U.

In order to avoid creating fictional intersection sets, a simple method is to remove intersection sets that violate the requirements defined in Chap. 3.1. Another method is to remove certain sets $e \in M$ and the intersection sets that contains e. As an example: Fig. 4 shows the desired facts by removing the conflict set C or the conflict sets A and B. This creates two potential, rough approximations of the given example. The conflict sets could be calculated through variants of the MinRelax or the QuickXPlain algorithms [10]. For every visualization problem, there are 1 to n conflict sets linked with the problem, which could be removed to solve the problem.

Another way to create drawable Euler diagrams is to split or clone sets [19]. The newly created or cloned sets can then be drawn as several geometrical forms that are not intersecting each other. All sets have to be marked as either the starting set or be linked together to improve the readability of the diagram. The total amount of intersections, which the newly created sets are part of, should equal the amount of sets in U containing the starting set. According to the defined requirements for Euler diagrams in Chap. 3.1, the size of the forms represents the amount of objects in one set. If a set is cloned, the sum of the areas of the cloned sets will represent the wrong amount of objects in the starting set. Figure 5 shows the issues and the solutions of the afore-mentioned example.

Since geometrical forms in our study represent a set of persons with specific properties, generalizing these properties to create drawable forms is another option. With the assumption that a property is composed of a type and a noun, both could be generalized individually. As an example, the property "loves Dog" could be generalized to "likes Dog", "likes Animal", "loves Animal", "loves Thing", and so on. Every set of properties with the option to be generalized will form a combined new set in M, which contains the generalized properties. The starting set of properties has to be removed from M. Properties in U that do not exist any longer in M need to be replaced with the new generalized set. With reference to the shown approach to split or clone sets [19], it is possible

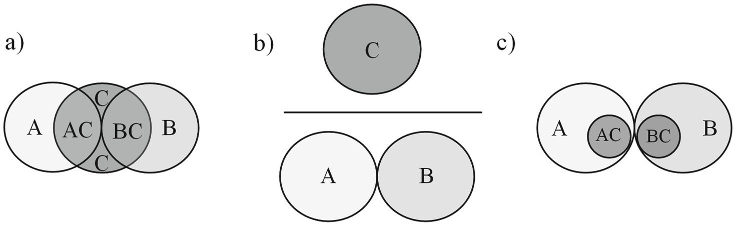


Fig. 5. (a) shows the sets $M = \{A, B, C\}$ and $U = \{A, B, AC, BC\}$, (b) shows rough approximations of the starting diagram and (c) shows a solution for the problem.

to generalize only one part or clone of a set. The generalization is supposed to reduce the amount of intersections between sets in M, and the generalized set has to be one of the sets that were identified with the described method before, creating rough approximations.

Sets in Euler diagrams can also be visualized by using abstract forms [18]. The diagrams can be created and calculated with the use of planar graphs and triangles. Used planar graphs represent the set U and extend the set with a null set for every set in U which only contains one set. The nodes in the planar graph are represented by sets in U, while the edges are represented as lines between two nodes. A line is drawn between two nodes a and b, if the set of a is fully contained in b. The line between a and b is omitted, if there exists a node c that fully contains the set of a and its set is fully contained in b. The drawing area is divided into several triangular sectors that can be used to calculate the area of several abstract forms for the different sets in M [18]. Figure 6 shows an example.

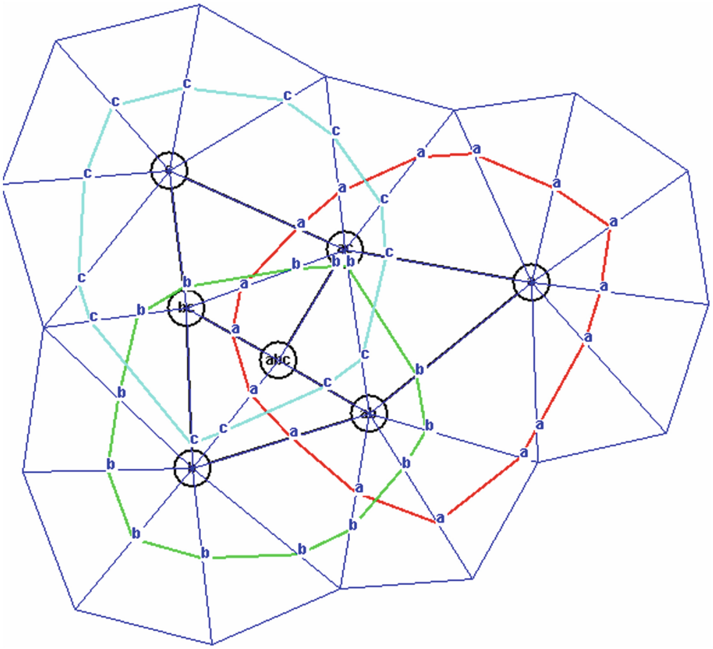


Fig. 6. A diagram for the sets $M = \{a, b, c\}$ and $U = \{a, b, c, ab, bc, abc, ac\}$ [18]

4 Conclusions and Future Research Directions

We have presented the findings of a conducted case study that turned the results of simple textual single question online polls into extensive knowledge about polls, answers and above all, the users. We have shown the structure and functioning of an analytical subsystem that complements the existing operational polling system responsible for the basic functions of displaying polls and collecting votes. It does so without breaking the main strengths of simplicity and anonymity and with maintaining full independency from third-party APIs.

We also have presented possibilities and challenges of visualizing the obtained knowledge in an administration dashboard. The created diagrams not only serve as a source of information but also let the administrator select groups of users for further usage in retargeting advertising. There are still some issues in visualizing semantic models using common Euler diagrams that need to be solved.

The interaction of semantic technologies with visualization strategies turned out to be quite challenging. However, there were promising advances in basic visualization of a limited number of sets, which proved to be very useful and intuitive for administrators of polling systems and could create a fair quantity of new useful knowledge. While not being explicitly designed for further refining of the knowledgebase, the visualizations do contribute to a better understanding of the collected knowledge. This is of particular importance when building extensively large ontologies like the one in our case study, where it is hard to stay on top of things.

Future research will include improvements in analysis of the knowledgebase. Building on standard RDFS and OWL reasoning many useful extensions need to be considered, e.g., calculation of probabilities (*has been to Russia* could have a probability of 80 % for *likes travelling*) in order to extend target groups in case they have been conceived too narrowly. Another research direction shall be alternative visualization techniques for different applications within the administrator dashboard like conveniently navigating the knowledgebase or displaying explanations of inferences being made. As the relevance for polls being elements of the social web keeps growing, investing more efforts in them will certainly be worthwhile.

References

1. Amir, S., Ait-Kaci, H.: Cedar: efficient reasoning for the semantic web. In: Proceedings of the 10th International Conference on Signal-Image Technology and Internet-Based Systems, pp. 157–163. IEEE (2014)
2. Androutsopoulos, I., Malakasiotis, P.: A survey of paraphrasing and textual entailment methods. J. Artif. Intell. Res. **38**, 135–187 (2010)
3. Bekkerman, R., Gavish, M.: High-precision phrase-based document classification on a modern scale. In: Proceedings of the KDD. ACM (2011)
4. Bench-Capon, T.J.M.: Knowledge Representation: An Approach to Artificial Intelligence. Academic Press Ltd, London (1990)
5. Berant, J., Liang, P.: Semantic parsing via paraphrasing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1415–1425 (2014)
6. Daiber, J., Jakob, M., Hokamp, C., Mendes, P.N.: Improving efficiency and accuracy in multilingual entity extraction. In: Proceedings of the 9th International Conference on Semantic Systems (2013)

7. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 351–366. Springer, Heidelberg (2013)
8. Hellmann, S., Lehmann, J., Auer, S., Brümmer, M.: Integrating NLP using linked data. In: Alani, H., Kagal, L., Fokoue, A., Groth, P., Biemann, C., Parreira, J.X., Aroyo, L., Noy, N., Welty, C., Janowicz, K. (eds.) ISWC 2013, Part II. LNCS, vol. 8219, pp. 98–113. Springer, Heidelberg (2013)
9. Huang, E.H., Socher, R., Manning, C.D., Ng, A.Y.: Improving word representations via global context and multiple word prototypes. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 873–882 (2012)
10. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems - An Introduction, Chapter Knowledge-Based Recommendation, pp. 81–123. Cambridge University Press, New York (2010)
11. Jurca, R., Faltings, B.: Incentives for expressing opinions in online polls. In: Proceedings of the 9th ACM Conference on Electronic Commerce, pp. 119–128. ACM (2008)
12. Kallmeyer, L., Maier, W.: Data-driven parsing using probabilistic linear context-free rewriting systems. Comput. Linguist. **39**(1), 87–119 (2013)
13. Kamp, H.: A theory of truth and semantic representation. In: Groenendijk, J., Janssen, T., Stokhof, M. (eds.) Formal Methods in the Study of Language, pp. 277–322. Mathematisch Centrum, University of Amsterdam (1981)
14. Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., Ondruska, P., Iyyer, M., Gulrajani, I., Socher, R.: Ask me anything: dynamic memory netorks for natural language processing. CoRR, abs/1506.07285 (2015)
15. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semant. Web J. **6**(2), 167–195 (2015)
16. Pazienza, M.T., Stellato, A. (eds.): Semi-Automatic Ontology Development - Processes and Resources. IGI Global, Hershey (2012)
17. Rafferty, A., Manning, C.D.: Parsing three german treebanks: lexicalized and unlexicalized baselines. In: Proceedings of the Workshop on Parsing German at ACL, pp. 40–46 (2008)
18. Rodgers, P., Zhang, L., Stapleton, G., Fish, A.: Embedding wellformed euler diagrams. In: Proceedings of the 12th International Conference on Information Visualisation, pp. 585–593. IEEE (2008)
19. Simonetto, P., Auber, D.: Visualise undrawable euler diagrams. In: Proceedings of the 12th International Conference on Information Visualisation, pp. 594–599 (2008)
20. Stabauer, M., Grossmann, G., Stumptner, M.: State of the art in knowledge extraction from online polls: a survey of current technologies. In: Proceedings of the Australasian Computer Science Week Multiconference, vol. 58, pp. 1–8. ACM (2016)
21. Stapleton, G., Zhang, L., Howse, J., Rodgers, P.: Drawing euler diagrams with circles. In: Goel, A.K., Jamnik, M., Narayanan, N.H. (eds.) Diagrams 2010. LNCS, vol. 6170, pp. 23–38. Springer, Heidelberg (2010)