

NLP & DBpedia: An Upward Knowledge Acquisition Spiral

Sebastian Hellmann¹, Agata Filipowska^{2,3}, Caroline Barrière⁴, Pablo N. Mendes⁵, and Dimitris Kontokostas¹

¹ University of Leipzig, Institute of Computer Science, AKSW Group,
Augustusplatz 10, D-04009 Leipzig, Germany
{lastname}@informatik.uni-leipzig.de,
<http://aksw.org>

² Poznan University of Economics, Faculty of Informatics and Electronic Economy,
Department of Information Systems,
Al. Niepodleglosci 10, 61-875 Poznan, Poland
{firstname.lastname}@kie.ue.poznan.pl,
<http://www.kie.ue.poznan.pl>

³ Instytut Informatyki Gospodarczej Sp. z o.o.,
ul. Rubiez 12G/6, 61-612 Poznan, Poland
{firstname.lastname}@i2g.pl,
<http://www.i2g.pl/>

⁴ Centre de Recherche Informatique de Montréal,
Montréal, Canada
{firstname.lastname}@crim.ca,
<http://crim.ca>

⁵ Kno.e.sis Center, Wright State University, USA
{firstname}@knoesis.org,
<http://knoesis.org>

Abstract. Recently, the DBpedia community has experienced an immense increase in activity and we believe, that the time has come to explore the connection between DBpedia & Natural Language Processing (NLP) in a yet unprecedented depth. DBpedia has a long-standing tradition to provide useful data as well as a commitment to reliable Semantic Web technologies and living best practices.

As the extraction of the Wikipedia's infoboxes by DBpedia matures, we can shift our focus to new challenges such as extracting information from an unstructured article text as well as becoming a testing ground for multilingual NLP methods. DBpedia has the potential to create an upward knowledge acquisition spiral as it provides a small amount of general knowledge allowing to process text, derive more knowledge, validate this knowledge and improve text processing methods.

The goal of this workshop was to present existing research, systems and resources, but also to allow discussion about different points of convergence and divergence of the NLP and DBpedia community with a special focus on challenges that lie ahead. We would like to take part in the debate on how to use DBpedia for NLP and NLP for DBpedia.

Keywords: DBpedia, Natural Language Processing, RDF

1 Introduction

Communities interested in Natural Language Processing (NLP) and in the Semantic Web, in particular DBpedia, come together to explore different ways of collaborating, and helping each other, towards a common goal of understanding and representing information.

Resources such as DBpedia are a step towards a solution to the knowledge acquisition bottleneck, so often mentioned in earlier days of NLP [10]. A prerequisite of text processing and understanding is the availability of knowledge about words, concepts and ways of expressing information. But then, to acquire such knowledge, we are required to automatically process text or immerse in costly and error-prone manual knowledge engineering.

Where formerly, there was a chicken and egg problem with a serious bootstrapping issue, we now have structured data in DBpedia, which is readily available to turn the bottleneck into an *upward knowledge acquisition spiral* – a small amount of general knowledge allowing to process text, create more knowledge, validate this knowledge and improve text processing for more acquisition (and so on).

The recent years have seen a major change, mostly through crowd-sourcing for the construction of the largest encyclopaedic resource, Wikipedia. Although first, mainly made of unstructured data (paragraphs), the addition of infoboxes, and the expansion of interest towards the Semantic Web, have led to DBpedia – one of the largest openly shared structured resource available today.

However, any resource not curated nor scrutinized by experts will be prone to noise, and that becomes a new and different challenge for NLP. Also, any resource, even as large as DBpedia, is not complete. So far, mainly the infoboxes, which are already semi-structured, are used to build the RDF repository. But even then, Aprosio et al. [1] (this volume) mention that more than 50% of Wikipedia articles do not include an infobox. So if the article text is analysed, the spiral can turn further, using DBpedia as input for the NLP process and then create more RDF triples to add and integrate into DBpedia [12].

This workshop’s aim is right in the knowledge acquisition spiral, bringing together researchers in both areas to see how NLP can benefit DBpedia and how DBpedia can benefit NLP. The contributions in the workshop allow to highlight multiple facets of this duality. In the remainder of this article, we discuss the contributions to the NLP&DBpedia workshop. Our main interest, however, are the challenges that the readers can expect to stay unresolved, that is the many interesting underlying issues brought forward by these articles. Another goal of this workshop was to present existing research, systems and resources to allow discussion about different points of convergence and divergence of the NLP and DBpedia community. It is also interesting to illustrate when both communities actually tackle very similar problems, with different approaches.

2 Knowledge acquisition and structuring

To some extent [7] (this volume) explore the problem of the above-mentioned knowledge acquisition bottleneck, by comparing information extraction systems, in particular NELL [4], which spirals on the large corpus ClueWeb09⁶ to acquire more and more knowledge, with database extraction approaches based on crowd-sourcing resources such as DBpedia.

While the main focus of [7] is more about how to structure the acquired knowledge than on the acquisition method itself, their work raises an important question: to what extent can we (or should we) use Wikipedia and DBpedia to structure and organize data extracted from text? This relates to an issue known in NLP, computational terminology and even more in library science – the debate between classifying (finding which terms in a thesaurus to associate to a document) and free-characterisation (extracting any terms from the text for its representation). The former obliges a thesaurus-like structure to be built before the text is analysed. But then many questions of how such structure was made arise. The latter allows the structure (or none) to emerge from the analysed text, but makes it difficult to compare information extracted from different texts, as there is no agreed-upon schema and synonyms stay unresolved.

The proposal of [19], is clearly on the acquisition of knowledge to be "fitted" into a known schema, that of the DBpedia ontology. Their proposal suggests the extension of DBpedia through Wikipedia list pages. The main problem is the actual matching between the extracted knowledge and the ontology. Knowledge sharing and matching is always problematic because of two main issues in semantics, that of polysemy (multiple concepts for a word) and synonymy (multiple words for a concept). Furthermore, there are also two main issues in ontology design and knowledge structuring, that of purpose-based versus non-purposed based ontologies, and that of the granularity of the information represented. All those issues combined make it quite difficult to attempt any kind of ontology expansion.

3 Representation of knowledge

As we look at NLP and DBpedia, we see that NLP requires knowledge about words, not only about concepts. Obviously the notion of labels exists in DBpedia, but there is more to language than labels. Should this lexical information be represented the same ways as conceptual information is?

The separation between lexical, conceptual, terminological, encyclopaedic, and other kinds of knowledge has been a debate for years. Can a single schema allow all types of knowledge? Lexical approaches usually start from words, going from a word to all its senses, and sometimes terminological approaches will start from concepts, and defining all the words that illustrate such concept. If DBpedia is more concept-based, we can then wonder how lexical information would be

⁶ <http://lemurproject.org/clueweb09/>

attached to it, or a more general question of how lexical knowledge has its place within the Semantic Web?

[26] (this volume) present a *lemon* lexicon for DBpedia and discuss different issues of lexicalization of conceptual structures.

The BabelNet[15] resource, resulting from a merge of WordNet [9] (a widely-used lexical resource in NLP) and Wikipedia, is an example of a mixed-level representation in which lexical, conceptual and encyclopaedic knowledge is combined. BabelNet is used in the work of [8] (this volume) for the task of QALD (Question Answering over Linked Data) as we will see in the next section. Also [27] (this volume) talk of developing their own representation, SAR-Graphs (Semantically Associated Relations Graphs) to express not only lexical knowledge, but sentence-based knowledge, that is useful for verbalizing simple predicates but also combined predicates (child of child, for example). These three contributions stimulate a debate on the granularity of the representation of any language resource. Such debate is present in corpus studies, where experts study the value of not only terms, but also phrases (phraseology) in the understanding of language use [24].

4 NLP tasks and applications

Although different tasks are mentioned in our workshop’s contributions, three of them are more prominent, that of NER (Named Entity Recognition), Relation Extraction, and Question Answering over Linked Data (QALD).

4.1 Named Entity Recognition

Named Entity Recognition is defined as the task of assigning a class to entities found in a text, such as person, location, organization, date, etc. NER is a well-recognized task in the NLP community since the beginning of the Message Understanding Conferences (MUC) in 1987 (see [11] for a good overview of information extraction and the early MUC conferences). Although not called as such at the time, early work on information extraction looked at text to find *Who did What When How* discovering entities such as places, people and dates. Extracted entities were not necessarily typed, or classified, but as information extraction templates were used, such types were implicitly given by the roles the entities filled (Agent, Place, Date).

Later on, researchers, such as Sekine ([21]) defined a hierarchical schema of classes for the NER task. Although, the more fine-grained the classes are, however, the more difficult it is to obtain (or even measure) classification results. Obviously, integration and comparison of these hierarchies can have high complexity, if no reference hierarchy is agreed upon. One such reference hierarchy is the recently created NERD ontology [20], however, containing only 84

types⁷ which is coarse grained when compared to the over 500 DBpedia Ontology classes⁸, which are used in [6] (this volume).

As mentioned in [23] (this volume) Named Entity Disambiguation is a further step towards identifying not only that an entity is a Person, but who this person actually is by establishing a link to a more specific reference id or URI in a knowledge base. New names are given to the NED or NERD task, that of Entity linking and "wikifiers" [6] (this volume) and the list of emerging tools, which belong to this class of wikifiers is quite huge and growing steadily: Zemanta, OpenCalais, Ontos, Evri, Extractiv, Alchemy API and many more⁹.

Wikipedia (and therefore DBpedia) is limited to encyclopaedic knowledge, but often terminological knowledge (how different terms describe different domain specific concepts) as well as lexical knowledge (common words) are available for interlinking with text, thus resembling Word-Sense Disambiguation (WSD), i.e. taking any word in a text and being able to connect the appropriate URI. In [8] (this volume), both tasks (NED and WSD) are tackled using BabelNet.

4.2 Relation extraction

The task of relation extraction is sometimes seen as a step following that of NER. After entities are extracted, it would be interesting to see how they are related. But sometimes a more "template-like" strategy, as was suggested in early information extraction is done. For example, a system would look for "merger" relations between companies, to find out which companies merged. In such case, the relation is known in advance, and we look in text for both the relation and the participants in such relation.

Different types of relations have been investigated over the years, and as NLP and DBpedia come closer, relations found in DBpedia tend to be used. [16] (this volume) focus on ten different relations found in DBpedia. They identify such relations in text through developed lexical extraction rules. The work of [1] (this volume) focuses on seven different properties found in DBpedia. By properties, they mean relations for which the subject is most likely a named entity, but the object could be a literal, such as the property *populationTotal*. The line is fuzzy between properties and relations (for example, both contributions mentioned above use the *birthDate* as a relation to extract in text), and could bring an interesting discussion and debate about this topic. The work of [27] (this volume) does not target any specific relation and is mostly about the development of a representational schema (as mentioned before) for the English expression of relations.

The explicit expression of relations in text is a topic of interest in the NLP community for a while. Different methods, either statistical [25] or pattern-based

⁷ accessed Oct. 10th, 2013 <http://nerd.eurecom.fr/ontology>

⁸ An up to date version can be downloaded from <http://mappings.dbpedia.org/server/ontology/dbpedia.owl>

⁹ http://en.wikipedia.org/wiki/Knowledge_extraction#Tools contains an up-to-date overview

are developed and experimented on [2]. This is an interesting place for NLP and the Semantic Web to meet as both communities are interested in finding links between concepts and extract facts.

4.3 Question Answering over Linked Data

The tasks of Information Retrieval and Question Answering, within the NLP community, provided some of the early attempts towards a more systematized approach to making the field of NLP grow. Those tasks encouraged the development of challenges and competitions with common data (TREC, [22]) which we discuss in the next section. The more recent task of Question Answering over Linked Data¹⁰ is a very interesting task, certainly promoting a communication and shared interest between the NLP and the Semantic Web community, and also providing some early attempts within the Semantic Web community at sharing data and evaluation standards.

Three contributions look into QALD. The work of [8] (this volume), addresses the task of QALD, with a particular strategy which involves NED and word sense disambiguation, as we mentioned above. In [3] (this volume), the QALD task is not just tackled, but they go further into the study of inconsistency detection when gathering knowledge to answer questions. They look into English, German, French and Italian chapters of DBpedia, and try to detect inconsistencies and supporting evidence among the different answers. In [26] (this volume) the task of QALD is not performed in itself, but it is mentioned as an extrinsic evaluation of the coverage of the *lemon* lexicon, saying that the verbalizations found in the lexicon cover many of the questions.

5 Resources

As most workshop contributions combine some techniques from NLP with the Semantic Web, they talk about different resources that would be useful to the community. We don't want to reinvent the wheel. Obviously, even if alternative Semantic Web resources, such as Yago (<http://www.mpi-inf.mpg.de/yago-naga/yago/>) and Freebase (<http://www.freebase.com>) exist, this workshop focuses on DBpedia, which therefore is the Semantic Web resource most referred to in the different contributions.

On the NLP side, many frameworks and typical resources exist as well. Wordnet (<http://wordnet.princeton.edu/>) for example, has been a resource much used in the community for English. More recently, Babelnet (<http://babelnet.org>), mentioned earlier, has been developed to merge Wikipedia and Wordnet. Also GATE, an open source development framework (<http://gate.ac.uk>), is used in [6] (this volume).

We can think that the primary resource for NLP is text, but which text? There has been work in NLP on different types of texts, from news articles to

¹⁰ The first challenge started in 2011, and information can be found at <http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/>

scientific articles, to blogs, to web data. In the present day, textual content is abundant, and the appropriateness of which text should be analysed for which purpose is a pertinent question. In fact, if we see NLP for DBpedia, at the service of expanding DBpedia, then the chosen text should be informative, factual, accurate. As we saw above, mining Wikipedia for more information is an interesting direction, it is not the only one. We also saw (with NELL) that a large crawled Web corpus is a possibility, as it brings large coverage, but it can also bring noise.

Different ways of filtering noise exists, either by trying to evaluate the source of information (trust), or by looking at how consistent or inconsistent different information is, looking at redundancy and conflicts. In [3] (this volume), the general problem of inconsistent information is tackled.

If we reverse our point of view and see DBpedia at the service of NLP, then the text on which NLP techniques are used is quite arbitrary and depends on further purposes and applications. For example, in [6] (this volume), both news articles and tweets are explored, which are two very different types of texts.

The question of language is valid whether we are looking at "NLP for DBpedia" or "DBpedia for NLP". In [16] (this volume), French text is analysed, and in [3] (this volume), four different language chapters of DBpedia are used. This is a minority of contributions exploring other languages than English. As always, work on English is more prominent than that on other language, and it brings awareness that it would be interesting for both communities to work on different languages.

5.1 Gold and silver standards

The topic of evaluation is both an important one, and a much debated one. In NLP, there has been a tendency in the past 15 years to perform experiments for which there are well defined gold standards and datasets. There has been an increase in the number of competitions and challenges in many sub-fields of NLP, such as automatic summarization [17], word-sense disambiguation [14], textual entailment [5], etc.

In the Semantic Web community, there is less of such rigid evaluation, as the field is younger than NLP, and is still looking at pushing the field with different ideas and concepts without imposing rigid evaluations. Certainly, one of the purposes of this workshop was to start discussion towards bringing more of gold standards and evaluation datasets into the community. Although there are some competitions in other areas, such as the OAEI (Ontology Alignment Evaluation Initiative¹¹) which has been happening for a few years now, as well as the QALD (see above) and the plethora of benchmarks for triplestores such as the DBPSB (DBpedia SPARQL Benchmark [13]). In the field of NER/NED, there are not many datasets or gold standards and only few challenges. The work of [6] (this volume) paves the way towards the standardization of NER and NED benchmarking in an implemented benchmarking system.

¹¹ <http://oaei.ontologymatching.org/>

As a first important step to develop such a gold standard, it is also good to review and question existing work. The work of [23] (this volume) is an extensive comparison of NED benchmarks and characterizes them to see, if they could be biased for particular types of algorithms, or types of test data. The contribution therefore opens the debate as to how we should develop such benchmarks and provides a solid foundation to built upon.

When gold standards are hard (costly, time-consuming) to develop, it can be interesting to develop silver standards that are the results of well-known methods, or the combined results of different methods. Such standards do not replace gold standards, but they at least give an indication of the direction of progress for particular algorithms. One possibility when two communities come together is to take the results of one to become the "silver standard" of the other. [18] (this volume) describes such a silver standard and discusses its benefits as well as its limitations.

In some work, such as [16] (this volume) and [1] (this volume), DBpedia's network of relations is used as a gold standard in relation extraction. Also Wikipedia/DBpedia entities have become the most predominant link targets in NED. [20] reports of 7 out of 10 tools that attach Wikipedia/DBpedia URLs as annotations (3 out of 10 for the DBpedia Ontology). Although this is an interesting way to proceed, we can debate whether we are using gold or silver standards and how to unify benchmarks for comparison.

6 Summary

We conclude by highlighting a few issues brought forward by the contributions in this workshop. First, the selected papers discuss many problems that have been recognized within the NLP community for a long time, but have only recently been introduced to Semantic Web researchers. The main challenges here concern:

- consensus upon annotation guidelines,
- development of extraction rules and agreed upon hierarchies that may be used to unify semantic enrichment and benchmarks,
- identification of well-defined tasks and problem classes,
- transferability of NLP tasks, resources and tools to other research communities (e.g. library and life sciences) as well as other languages and application areas,
- building practical resources and infrastructures, which do not target one single research question, but can be exploited in a more universal manner by NLP tools,
- unlock higher layers of semantic annotation to enable state-of-the art OWL-based reasoning on a combination of noisy NLP data and LOD and DBpedia based knowledge structures.

Second, and perhaps more importantly, new possibilities emerge from the combination of the communities, and we hope to further push such possibilities to have more NLP for DBpedia and more DBpedia for NLP, continuing the

knowledge spiral, and fighting together to open the knowledge acquisition bottleneck. We hope that the readers of this volume will find all papers interesting. We invite you to join our community and attend future workshop editions.

Acknowledgments.

We especially thank all contributors to DBpedia and the DBpedia Internationalisation committee¹². This work was supported by grants from the European Union's 7th Framework Programme provided for the projects LOD2 (GA no. 257943) and GeoKnow (GA no. 318159).

Programme Committee

We would like to thank all reviewers that have helped us and especially the authors with their comments and feedback.

- Guadalupe Aguado, Universidad Politécnica de Madrid, Spain
- Chris Bizer, Universität Mannheim, Germany
- Volha Bryl, Universität Mannheim, Germany
- Paul Buitelaar, DERI, National University of Ireland, Galway
- Charalampos Bratsas, OKFN, Aristotle University of Thessaloniki, Greece
- Philipp Cimiano, CITEC, Universität Bielefeld, Germany
- Samhaa R. El-Beltagy, Nile University, Egypt
- Daniel Gerber, AKSW, Universität Leipzig, Germany
- Jorge Gracia, Universidad Politécnica de Madrid, Spain
- Max Jakob, Neofonie GmbH, Germany
- Anja Jentzsch, Hasso-Plattner-Institut, Potsdam, Germany
- Ali Khalili, AKSW, Universität Leipzig, Germany
- Daniel Kinzler, Wikidata, Germany
- David Lewis, Trinity College Dublin, Ireland
- John McCrae, Universität Bielefeld, Germany
- Uroš Milošević, Institut Mihajlo Pupin, Serbia
- Roberto Navigli, Sapienza, Università di Roma, Italy
- Axel Ngonga, AKSW, Universität Leipzig, Germany
- Asunción Gómez Pérez, Universidad Politécnica de Madrid, Spain
- Lydia Pintscher, Wikidata, Germany
- Elena Montiel Ponsoda, Universidad Politécnica de Madrid, Spain
- Giuseppe Rizzo, Eurecom, France
- Harald Sack, Hasso-Plattner-Institut, Potsdam, Germany
- Felix Sasaki, Deutsches Forschungszentrum für künstliche Intelligenz, Germany
- Mladen Stanojević, Institut Mihajlo Pupin, Serbia
- Ricardo Usbeck, AKSW, Universität Leipzig, Germany
- Hans Uszkoreit, Deutsches Forschungszentrum für künstliche Intelligenz, Germany
- Rupert Westenthaler, Salzburg Research, Austria
- Feiyu Xu, Deutsches Forschungszentrum für künstliche Intelligenz, Germany

¹² <http://wiki.dbpedia.org/Internationalization>

References

1. A. P. Aprosio, C. Giuliano, and L. Alberto Lavelli. Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
2. A. Auger and C. Barrière. *Probing Semantic Relations: Exploration and identification in specialized texts*. John Benjamins, benjamins edition, 2010.
3. E. Cabrio, J. Cojan, S. Villata, and F. Gandon. Argumentation-based Inconsistencies Detection for Question-Answering over DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
4. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an architecture for never-ending language learning. In M. Fox and D. Poole, editors, *AAAI*. AAAI Press, 2010.
5. D. Cristea. Textual entailment. *Computational Linguistics*, (June):1140–1143, 2009.
6. M. Dojchinovski and T. Kliegr. Datasets and GATE Evaluation Framework for Benchmarking Wikipedia-Based NER Systems. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
7. A. Dutta, C. Meilicke, M. Niepert, and S. Ponzetto. Integrating Open and Closed Information Extraction: Challenges and First Steps. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
8. K. Elbedweihy, S. Wrigley, and F. Ciravegna. Using BabelNet in Bridging the Gap Between Natural Language Queries and Linked Data Concepts. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
9. C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.
10. W. A. Gale, K. W. Church, and D. Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5-6):415–439, 1992.
11. R. Grishman. Information Extraction: Techniques and Challenges. *New York*, i(4):10–27, 1997.
12. M. Héder and P. N. Mendes. Round-trip semantics with sztakipedia and dbpedia spotlight. In A. Mille, F. L. Gandon, J. Misselis, M. Rabinovich, and S. Staab, editors, *WWW (Companion Volume)*, pages 357–360. ACM, 2012.
13. M. Morsey, J. Lehmann, S. Auer, and A.-C. Ngonga Ngomo. DBpedia SPARQL Benchmark – Performance Assessment with Real Queries on Real Data. In *ISWC 2011*, 2011.
14. R. Navigli, D. Jurgens, and D. Vannella. SemEval-2013 Task 12 : Multilingual Word Sense Disambiguation. In *Proceedings of the 7th International Workshop on Semantic Evaluation SemEval 2013 in conjunction with the Second Joint Conference on Lexical and Computational Semantics SEM 2013*, 2013.

15. R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.
16. K. Nebhi. A Rule-Based Relation Extraction System using DBpedia and Syntactic Parsing. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
17. T. Okumura, Manabu Fukusima and H. Nanba. Text Summarization Challenge 2 - Text summarization evaluation at NTCIR Workshop 3. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 49–56, 2003.
18. H. Paulheim. DBpediaNYD, A Silver Standard Benchmark Dataset for Semantic Relatedness in DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
19. H. Paulheim and S. P. Ponzetto. Extending DBpedia with Wikipedia List Pages. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
20. G. Rizzo, R. Troncy, S. Hellmann, and M. Brummer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In *LDOW*, 2012.
21. S. Sekine and C. Nobata. Definition, dictionaries and tagger for extended named entity hierarchy. In A. Zampolli and M. T. Lino, editors, *Proceedings of the Language Resources and Evaluation Conference LREC*, pages 1977–1980. European Language Resources Association, 2004.
22. K. Sparck Jones. Further reflections on TREC. *Information Processing & Management*, 36(1):37–85, 2000.
23. N. Steinmetz, M. Knuth, and H. Sack. Statistical Analyses of Named Entity Disambiguation Benchmarks. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
24. M. Stubbs. An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti, editor, *Corpus linguistics 25 years on*, number 62, pages 89–105 385. Rodopi, 2007.
25. P. D. Turney and M. L. Littman. Corpus-based Learning of Analogies and Semantic Relations. *Machine Learning*, 60(1-3):1–3, 2005.
26. C. Unger, J. Mccrae, S. Walter, S. Winter, and P. Cimiano. A lemon lexicon for DBpedia. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.
27. H. Uszkoreit and F. Xu. From Strings to Things SAR-Graphs: A New Type of Resource for Connecting Knowledge and Language. In *Proceedings of 1st International Workshop on NLP and DBpedia, October 21-25, Sydney, Australia*, volume 1064 of *NLP & DBpedia 2013*, Sydney, Australia, October 2013. CEUR Workshop Proceedings.