Creating a Knowledge Base From a Collaboratively Generated Encyclopedia

Simone Paolo Ponzetto
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/~ponzetto

Abstract

We present our work on using Wikipedia as a knowledge source for Natural Language Processing. We first describe our previous work on computing semantic relatedness from Wikipedia, and its application to a machine learning based coreference resolution system. Our results suggest that Wikipedia represents a semantic resource to be treasured for NLP applications, and accordingly present the work directions to be explored in the future.

1 Introduction

The last decade has seen statistical techniques for Natural Language Processing (NLP) gaining the status of standard approaches to most NLP tasks. While advances towards robust statistical inference methods (cf. e.g. Domingos et al. (2006) and Punyakanok et al. (2006)) will certainly improve the computational modelling of natural language, we believe that crucial advances will also come from rediscovering the use of symbolic knowledge, i.e. the deployment of large scale knowledge bases.

Arguments for the necessity of symbolically encoded knowledge for AI and NLP date back at least to McCarthy (1959). Symbolic approaches using knowledge bases, however, are expensive and time-consuming to maintain. They also have a limited and arbitrary coverage. In our work we try to overcome such problems by relying on a wide coverage on-line encyclopedia developed by a large amount of users, namely Wikipedia. That is, we are interested in whether and how Wikipedia can be integrated into NLP applications as a knowledge base. The motivation comes from the necessity to overcome the brittleness and knowledge acquisition bottlenecks that NLP applications suffer.

2 Previous Work: WikiRelate! and Semantic Knowledge Sources for Coreference Resolution

Ponzetto & Strube (2006) and Strube & Ponzetto (2006) aimed at showing that 'the encyclopedia that anyone can edit' can be indeed used as a semantic resource for research in NLP. In particular, we assumed its category tree to represent a semantic network modelling relations between concepts, and we computed measures of semantic relatedness from it. We did not show only that Wikipedia-based measures of semantic relatedness are competitive with the ones computed from a widely used standard resource such as WordNet (Fellbaum, 1998), but also that including semantic knowledge mined from Wikipedia into an NLP system dealing with coreference resolution is in fact beneficial.

2.1 WikiRelate! Computing Semantic Relatedness Using Wikipedia

Semantic relatedness measures have been proven to be useful in many NLP applications such as word sense disambiguation (Kohomban & Lee, 2005; Patwardhan et al., 2005), information retrieval (Finkelstein et al., 2002), information extraction pattern induction (Stevenson & Greenwood, 2005), interpretation of noun compounds (Kim & Baldwin, 2005), paraphrase detection (Mihalcea et al., 2006) and spelling correction (Budanitsky & Hirst, 2006). Approaches to measuring semantic relatedness that

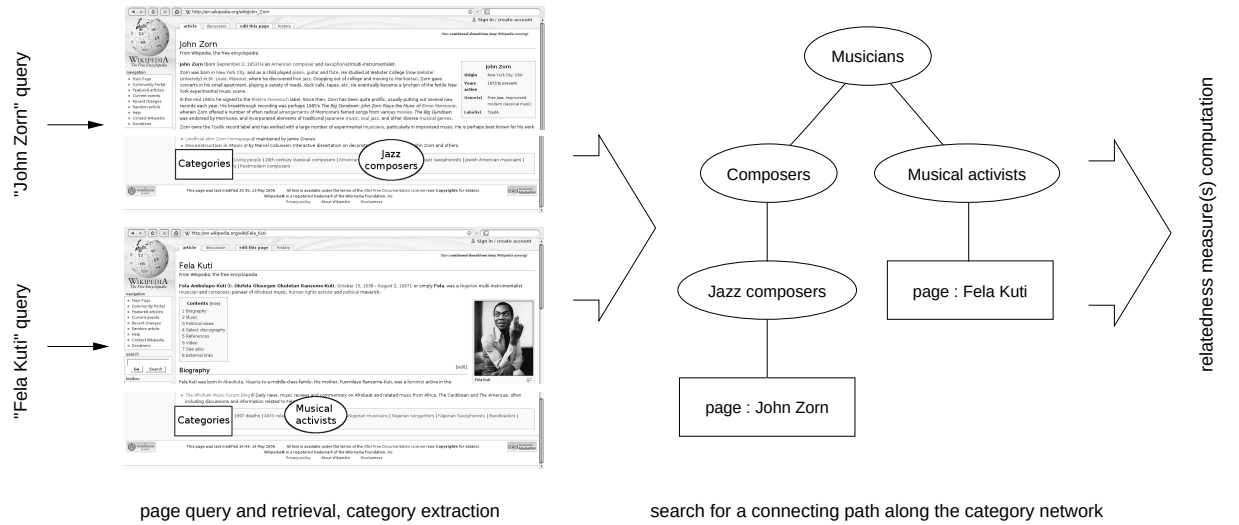


Figure 1: Wikipedia-based semantic relatedness computation. First, target pages for the given queries are retrieved, possibly via disambiguation. Next, categories are extracted to provide an entry point to the category network. Connecting paths are then searched along the category network using a depth-limited search. The paths found are scored and the ones satisfying the measure definitions (i.e. the shortest one for path-length measures, and the most informative one for information-content measures) are returned.

use lexical resources transform that resource into a network or graph and compute relatedness using paths in it[1]. For instance, Rada et al. (1989) traverse MeSH, a term hierarchy for indexing articles in Medline, and compute semantic relatedness as the edge distance between terms in the hierarchy. Jarmasz & Szpakowicz (2003) use the same approach with *Roget's Thesaurus* while Hirst & St-Onge (1998) apply a similar strategy to WordNet.

The novel idea presented in Strube & Ponzetto (2006) was to induce a semantic network from the Wikipedia categorization graph to compute measures of semantic relatedness. Wikipedia, a multilingual Web-based free-content encyclopedia, allows for structured access by means of *categories*: the encyclopedia articles can be assigned one or more categories, which are further categorized to provide a so-called "category tree". Though not designed as a strict hierarchy or tree, the categories form a graph which can be used as a taxonomy to compute semantic relatedness. We showed (1) how to retrieve Wikipedia articles from textual queries and resolve ambiguous queries based on the articles' link structure; (2) compute semantic relatedness as a function of the articles found and the paths between them along the categorization graph (Figure 1). We evaluated the Wikipedia-based measures against the ones computed from WordNet on benchmarking datasets from the literature (e.g. Miller and Charles' (1991) list of 30 noun pairs) and found Wikipedia to be competitive with WordNet.

2.2 Semantic Knowledge Sources for Coreference Resolution

Evaluating measures of semantic relatedness on word pair datasets poses non-trivial problems, i.e. all available datasets are small in size, and it is not always clear which linguistic notion (i.e. similarity vs. relatedness) underlies them. Accordingly, in Ponzetto & Strube (2006) we used a machine learning based coreference resolution system to provide an *extrinsic* evaluation of the utility of WordNet and Wikipedia relatedness measures for NLP applications. We started with the machine learning based

[1] An overview of lexical resource-based approaches to measuring semantic relatedness is presented in Budanitsky & Hirst (2006). Note that here we do not distinguish between *semantic similarity* (computed using hyponymy/hyperonymy, i.e. *is-a*, relations only) and *semantic relatedness* (using all relations in the taxonomy, including antonymic, meronymic, functional relations such as *is-made-of*, etc.), since the relations between categories in Wikipedia are neither semantically typed nor show a uniform semantics (see Section 3).

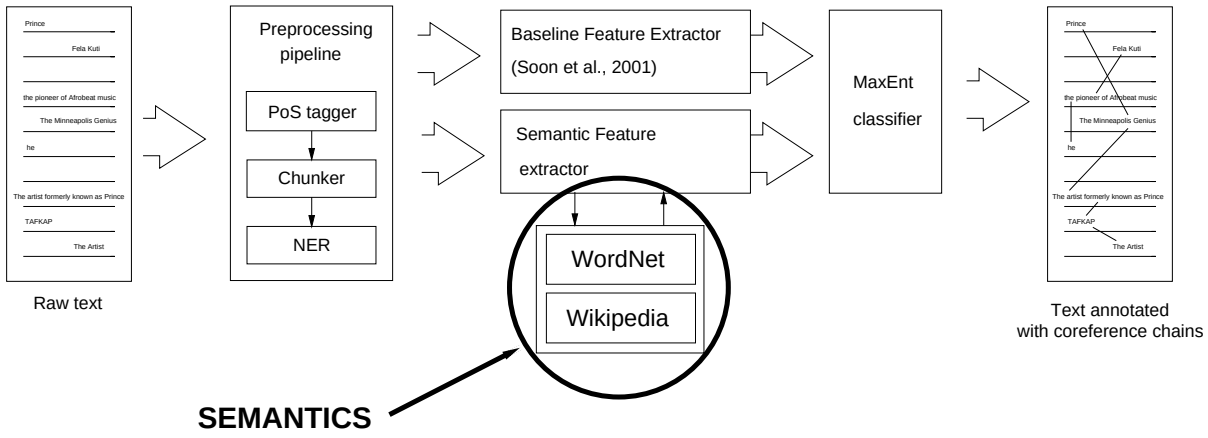


Figure 2: Overview of the coreference system for extrinsic evaluation of WordNet and Wikipedia relatedness measures. We start with a baseline system from Soon et al. (2001). We then include at different times features from WordNet and Wikipedia and register performance variations.

baseline system from Soon et al. (2001), and analyzed the performance variations given by including the relatedness measures in the feature set (Figure 2). The results showed that coreference resolution benefits from information mined from semantic knowledge sources and also, that using features induced from Wikipedia gives a performance only slightly worse than when using WordNet.

3 Future Work: Inducing an Ontology from a Collaboratively Generated Encyclopedia

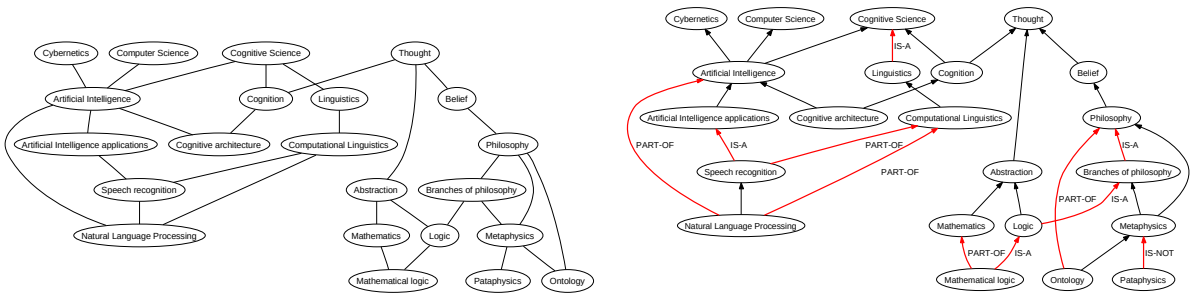
Our results so far suggest that Wikipedia can be considered a semantic resource in its own right. Unfortunately, the Wikipedia categorization still suffers from some limitations: it cannot be considered an ontology, as the relations between categories are not semantically-typed, i.e. the links between categories do not have an explicit semantics such as *is-a*, *part-of*, etc. Work in the near future will accordingly concentrate on automatically inducing the semantics of the relations between Wikipedia categories. This aims at transforming the unlabeled graph in Figure 3(a) into the semantic network in Figure 3(b), where the links between categories are augmented with a clearly defined semantics.

The availability of explicit semantic relations would allow to compute *semantic similarity* rather than *semantic relatedness* (Budanitsky & Hirst, 2006), which is more suitable for coreference resolution. That is, we assume that the availability of hyponymic/hyperonymic relations will allow us to compute lexical semantic measures which will further increase the performance of our coreference resolution system, as well as further bringing forward Wikipedia as a direct competitor of manually-designed resources such as WordNet.

In order to make the task feasible, we are currently concentrating on inducing *is-a* vs. *not-is-a* semantic relations. This simplifies the task, but still allows us to compute measures of semantic similarity. As we made limited use of the large amount of text in Wikipedia, we are now trying to integrate text and categorization. This includes extracting semantic relations expressed in the encyclopedic definitions by means of *Hearst patterns* (Hearst, 1992), detection of *semantic variations* (Morin & Jacquemin, 1999) between category labels, as well as using the categorized pages as bag-of-words to compute scores of *idf-based semantic overlap* (Monz & de Rijke, 2001) between categories. Further work will then concentrate on making this information available to our coreference resolution system, e.g. via semantic similarity computation.

Finally, since Wikipedia is available in many languages, we believe it is worth performing experiments in a multilingual setting. Accordingly, we are currently testing a website[2] that will allow us to collect word relatedness judgements from native speak-

[2] Available at `http://www.eml-research.de/nlp/353-TC`.



(a) current category graph

(b) category graph augmented with semantic relations

Figure 3: Inducing explicit semantic relations between categories in Wikipedia

ers of German, French and Italian, in order to translate the semantic relatedness dataset from Finkelstein et al. (2002) and test our methodology with languages other than English.

4 Conclusions

In this paper we presented our previous efforts on using Wikipedia as a semantic knowledge source. We aim in the future to induce an ontology from its collaboratively generated categorization graph. We believe that our work opens up exciting new challenges for the AI and NLP research community, e.g. how to handle the noise included in such knowledge bases and how to fully structure the information given in the form of only partially structured text and relations between knowledge base entries.

Acknowledgements: This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The author has been supported by a KTF grant (09.003.2004).

References

Budanitsky, A. & G. Hirst (2006). Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1).

Domingos, P., S. Kok, H. Poon, M. Richardson & P. Singla (2006). Unifying logical and statistical AI. In *Proc. of AAAI-06*, pp. 2–7.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database.* Cambridge, Mass.: MIT Press.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman & E. Ruppin (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING-92*, pp. 539–545.

Hirst, G. & D. St-Onge (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 305–332. Cambridge, Mass.: MIT Press.

Jarmasz, M. & S. Szpakowicz (2003). Roget's Thesaurus and semantic similarity. In *Proc. of RANLP-03*, pp. 212–219.

Kim, S. N. & T. Baldwin (2005). Automatic interpretation of noun compounds using WordNet similarity. In *Proc. of IJCNLP-05*, pp. 945–956.

Kohomban, U. S. & W. S. Lee (2005). Learning semantic classes for word sense disambiguation. In *Proc. of ACL-05*, pp. 34–41.

McCarthy, J. (1959). Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pp. 75–91.

Mihalcea, R., C. Corley & C. Strapparava (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proc. of AAAI-06*, pp. 775–780.

Miller, G. A. & W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Monz, C. & M. de Rijke (2001). Light-weight entailment checking for computational semantics. In *Proc. of ICoS-3*, pp. 59–72.

Morin, E. & C. Jacquemin (1999). Projecting corpus-based semantic links on a thesaurus. In *Proc. of ACL-99*, pp. 389–396.

Patwardhan, S., S. Banerjee & T. Pedersen (2005). SenseRelate::TargetWord – A generalized framework for word sense disambiguation. In *Proc. of AAAI-05*.

Ponzetto, S. P. & M. Strube (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL-06*, pp. 192–199.

Punyakanok, V., D. Roth, W. Yih & D. Zimak (2006). Learning and inference over constrained output. In *Proc. of IJCAI-05*, pp. 1117–1123.

Rada, R., H. Mili, E. Bicknell & M. Blettner (1989). Development and application of a metric to semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.

Soon, W. M., H. T. Ng & D. C. Y. Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stevenson, M. & M. Greenwood (2005). A semantic approach to IE pattern induction. In *Proc. of ACL-05*, pp. 379–386.

Strube, M. & S. P. Ponzetto (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of AAAI-06*, pp. 1419–1424.