Overview of a Semantic Disambiguation Method for Unstructured Web Contexts

Jorge Gracia
IIS Department
University of Zaragoza
Zaragoza, Spain
jogracia@unizar.es

Eduardo Mena
IIS Department
University of Zaragoza
Zaragoza, Spain
emena@unizar.es

ABSTRACT

In this paper we give an overview of a multiontology disambiguation method, targeted to discover the intended meaning of words in unstructured web contexts. It receives an ambiguous keyword and its context words as input and provides a list of possible senses for the keyword, scored according to the probability of being the intended one. It accesses any pool of online ontologies as source of word senses, in addition to other available resources. This method is targeted to be used in unstructured contexts that lack well-formed sentences, such as user keywords or folksonomy tags.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods

General Terms

Algorithms

Keywords

Semantic Web, knowledge selection, disambiguation

1. INTRODUCTION

The ability of computers to automatically determine the right sense of words, according to the context where they appear, can help bridge the gap between syntax and semantics required for the full development of the Semantic Web. However, the applicability of these techniques is sometimes hampered by the unrestricted way in which humans use keywords or annotate resources on the Web.

In this work we tackle the problem of determining what are the context words that better help in the disambiguation of keywords in unstructured contexts. It is the starting point of a complete disambiguation method that finds possible semantic descriptions for an ambiguous keyword and picks the most suitable one according to the context. It combines different techniques to operate: Web-based relatedness, overlap of semantic descriptions, and frequency of use of senses.

2. OVERVIEW OF THE METHOD

An scheme of our method (that improves the one presented in [4]) is shown in Figure 1.

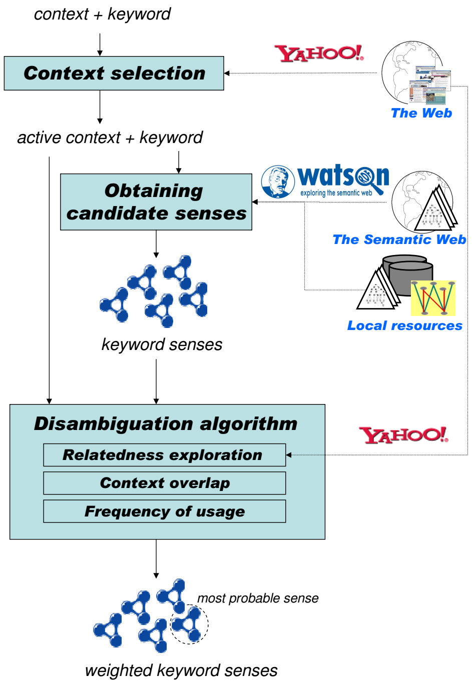

Figure 1: Scheme of the disambiguation method.

K-CAP'09, September 1–4, 2009, Redondo Beach, California, USA.
ACM 978-1-60558-658-8/09/09

It represents how an ambiguous keyword and its context are introduced as input, and a score for each possible keyword sense is given as output. The following process is carried out[1]:

1. *Context selection.* We rely on the hypothesis that the most suitable context words for disambiguation are the ones most highly related to the ambiguous keyword. Based on that, we compute the web-based relatedness [3] between each context word and the keyword to disambiguate, retaining the ones that score above a certain threshold. We call this set of words *active context.*

2. *Obtaining candidate senses.* After that, online and local resources [2, 5] are accessed to provide a set of candidate senses for the keyword. The output of this process is a set of candidate *senses* that describe the possible meanings of the keyword to disambiguate. Each sense corresponds to an ontology term or to the integration of various ontology terms of the same type.

3. *Disambiguation algorithm.* Finally, our disambiguation algorithm is run and the senses are weighted according to their likeliness of being the right one. It is performed in three steps:

   (a) *Relatedness exploration.* We first explore the semantic relatedness among the keyword senses and the words in the context, using the relatedness computation described in [3].

   (b) *Context overlap.* We add a factor that measures the overlap between the words that appear in the context, and the words that appear in the semantic definition of the sense. It adapts the method in [1].

   (c) *Frequency of usage.* If sense scores are too close each other, we consider the frequency of usage of senses (if available) as an additional factor for disambiguation.

For simplicity, we consider only one keyword to disambiguate, although the algorithm can be iteratively repeated if more keywords need disambiguation.

3. EXPERIMENTAL EVALUATION

We have experimented with a corpus of 350 pictures, extracted from searches of ambiguous words in Flickr[2] and annotated by humans with WordNet [5] senses. We have chosen WordNet as source of knowledge in this experiment to compare with the "most frequent sense" baseline that it provides.

For each test case, the input to our disambiguation process was the ambiguous keyword and, as context, the set of tags that annotate each picture. We have compared the results provided by our disambiguation method with respect to the human judgement and computed *accuracy.*

The results shows that our method (58% accuracy) outperforms both the random and the "most frequent sense" (MFS) baselines in this experiment (20% and 43% accuracy respectively). This is a remarkable achievement, because the state of the art indicates that rarely non supervised techniques score above MFS baseline [6].

4. CONCLUSIONS

We have presented a method that disambiguates words in web contexts, by computing relatedness measures, studying semantic overlap, and utilizing frequencies of usage of senses. It is intended to be used in situations where other disambiguation methods have difficulties to operate, for example when: 1) Dealing with unstructured contexts, as folksonomy tags, search query terms, etc. instead of well-formed texts and sentences; 2) knowledge sources are not known in advance, and must be selected dynamically; and 3) maximizing the coverage of possible interpretations of a word is required.

Our experimental results support our ideas and encourage us to further improve our method and its applications in the future.

Acknowledgments. This work is supported by the CICYT project TIN2007-68091-C02-02.

5. REFERENCES

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proc. of the 18th International Joint Conference on Artificial Intelligence*, pages 805–810, August 2003.

[2] M. d'Aquin, C. Baldassarre, L. Gridinoc, S. Angeletou, M. Sabou, and E. Motta. Characterizing knowledge on the semantic web with Watson. In *5th International EON Workshop, at ISWC'07, Busan, Korea*, 2007.

[3] J. Gracia and E. Mena. Web-based measure of semantic relatedness. In *Proc. of 9th International Conference on Web Information Systems Engineering (WISE'08), Auckland, New Zealand*, pages 136–150. LNCS, September 2008.

[4] J. Gracia, R. Trillo, M. Espinoza, and E. Mena. Querying the web: A multiontology disambiguation method. In *Sixth International Conference on Web Engineering (ICWE'06), Palo Alto (California, USA)*. ACM, July 2006.

[5] G. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11), November 1995.

[6] S. Pradhan, E. Loper, D. Dligach, and M. Palmer. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic, June 2007. ACL.

[1] The figure represents Yahoo! (http://yahoo.com) and Watson (http://watson.kmi.open.ac.uk/WatsonWUI/) as sources of web frequencies and online information respectively, but others can be used.

[2] http://www.flickr.com/