Demo: Enriching Text with RDF/OWL Encoded Senses

Delia Rusu, Tadej Štajner, Lorand Dali, Blaž Fortuna, Dunja Mladenić,

Jožef Stefan Institute, Ljubljana, Slovenia
{delia.rusu, tadej.stajner, lorand.dali, blaz.fortuna, dunja.mladenic}@ijs.si

Abstract. This demo paper describes an extension of the Enrycher text enhancement system, which annotates words in context, from a text fragment, with RDF/OWL encoded senses from WordNet and OpenCyc. The extension is based on a general purpose disambiguation algorithm which takes advantage of the structure and/or content of knowledge resources, reaching state-of-the-art performance when compared to other knowledge-lean word sense disambiguation algorithms.

Keywords: RDF/OWL word sense representation.

1 Introduction

A variety of Semantic Web resources in the Linked Open Data (LOD) cloud can serve as knowledge bases for identifying word senses; more general, like DBpedia, W3C WordNet, OpenCyc, or more domain specific like the Gene Ontology, just to name a few of them. Moreover, these resources complement each other, as they span across several domains from music to chemistry and biology.

Enrycher [8] is a service-oriented natural language processing and information extraction framework. It annotates text at various levels, listing: subject – predicate – object triplets (interesting statements) visually interconnected in a semantic graph representation, co-referenced named entities linked to DBpedia, Yago and OpenCyc, keywords and DMOZ categories. In this Demo paper we present an extension of Enrycher[1], relying on a general purpose algorithm which can take advantage of several Semantic Web resources to disambiguate text. This extension annotates words in context with RDF/OWL encoded senses from WordNet [1] and OpenCyc[2]. Given an input text fragment, every word or collocation (word sequence) will be annotated with the appropriate sense in context, and linked to the associated RDF resources defining the sense, in both WordNet and OpenCyc. The motivation behind adding this extension is to provide richer disambiguated annotations of words that are not named entities, and to improve the semantic graph quality, by merging nodes that refer to the same disambiguated concept.

Word sense disambiguation (WSD) is defined as identifying the meaning of words in a given context, and has become a prerequisite for several Semantic Web specific

[1] Demo video: http://marquis.ijs.si/delia/
[2] http://sw.opencyc.org/

tasks like ontology mapping and reasoning. WSD techniques have been previously introduced to validate ontology mappings, by analyzing the semantics of the ontological terms; they exploit ontological context, as well as information provided by WordNet. Aside from WordNet, another knowledge resource, namely Wikipedia has been used for building sense tagged corpora, which have further been employed to train a classifier, obtaining promising results [4]. Wikipedia was also used to automatically extend WordNet with semantic relations (such as synonymy, antonymy, hyponymy, etc.) [6]. However, the existing disambiguation systems mainly retrieve WordNet senses that are not readily usable for Semantic Web applications. Our extension can be easily integrated in other applications that require WSD as a preprocessing step, as word senses are labeled with the corresponding disambiguated RDF/OWL resource. Moreover, we take advantage of ontologies to find word senses, and in future work we plan to add some domain ontologies that can better disambiguate domain specific terminology.

The paper is structured as follows: we start by describing the Enrycher extension integration in Section 2, continue with presenting the disambiguation algorithm in Section 3 and conclude with a section on future work and the demo presentation.

2 Enrycher Extension Integration

Our RDF/OWL word sense annotation extension of Enrycher relies on the Text Preprocessing component which performs sentence splitting, tokenization, part-of-speech tagging and keyword extraction based on a bag-of-words model (see Fig. 1). Both WordNet 3.0 and OpenCyc are processed offline, in order to extract structure and content information. By structure we refer to the semantic relations: synonymy, hypernymy, etc. specific to WordNet, as well as the generalization, specialization, etc. relations encoded in OpenCyc. The content is given by the WordNet *glosses* and the OpenCyc *comments*, and provides descriptions of the word sense.

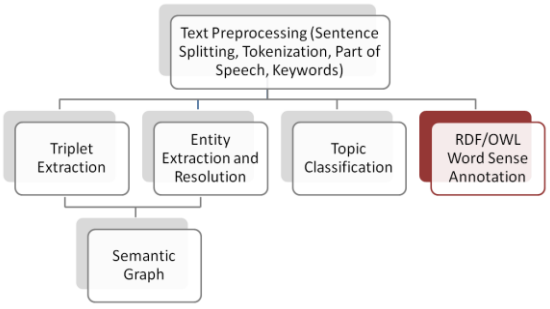

Fig. 1. Enrycher components and their dependencies.

Given an input text fragment, every word or collocation will be annotated with the appropriate sense in context from the aforementioned knowledge resources. If existent, both RDF resources corresponding to WordNet and OpenCyc will be linked. The following section elaborates on the proposed general purpose disambiguation algorithm.

3 Word/Collocation Annotation

We have implemented an unsupervised semantic knowledge based word sense disambiguation algorithm. It relies on the Viterbi algorithm for Hidden Markov Model (HMM) part-of-speech tagging [2], and an initial version was described in [7]. The Viterbi algorithm is a common decoding algorithm for HMM, which was first applied to speech and language processing in the context of speech recognition. We have adapted the algorithm in order to determine, given the senses of words in a sentence, the best sequence of senses that disambiguates the sentence. We start by looking for the senses of nouns, verbs, adjectives and adverbs in one of the two aforementioned knowledge resources. The sequence of observations $O = o_1 o_2 ... o_T$ will represent the T words we disambiguate, while the set of states $Q = q_1 q_2 ... q_N$ define the N senses for a given observation. The sequence of *observation likelihoods* $B=b_i(o_t)$ expresses the probability of an observation o_t being generated from a state i. They are obtained by computing the *cosine similarity* between an ambiguous word description, as defined by the knowledge resource, and information provided by the context (at the level of the sentence, paragraph, etc.). The transition probability matrix $A = a_{11} a_{12} ... a_{n1} ... a_{nm}$ is determined by computing the semantic relatedness between the two senses in state i and j respectively. There have been several relatedness measures proposed in the literature, some of them relying on the knowledge resource structure, others on its content. We have implemented four such relatedness measures, one of which exploiting the resource structure – *Lexical Chains*, while the others take the resource content into account – *Adapted Lesk*, *Vector* and *Vector Pairwise* [5].

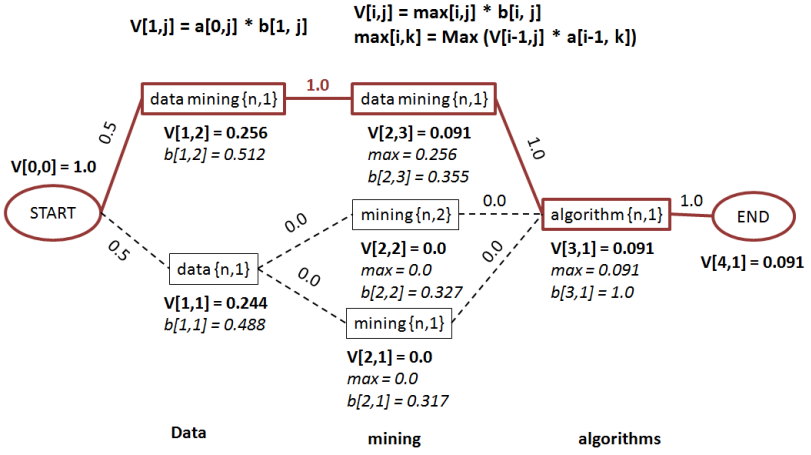

Fig. 2. Disambiguating the phrase *Data mining algorithms* using the proposed algorithm.

We explain the algorithm with the aid of the following example in Fig. 2. To disambiguate the phrase "Data mining algorithms" using WordNet 3.0 as a sense repository, we consider the senses of all words (the word "mining" having the sense of "excavating" or "minelaying"), and in addition the sense of the collocation "data mining" (data processing). We denote the sense part of speech and number in curly brackets. The edges are labeled by state transitions. The collocation is modeled by copying the corresponding sense state, and setting the transition between these two

states to 1.0. There is equal probability to reach any of the sense states of the first word from the start state. Once the final state is reach, we back trace to find the states with the highest associated scores.

We compared our system with others participating in the SemEval 2007 coarse grained all words English disambiguation task based on WordNet senses, obtaining precision/recall/F1 measures of 77.3, lower than the most frequent sense baseline of 78.9, but higher than the best unsupervised disambiguation algorithm participating in the task (SUSSX-FR, based on parsing text and identifying the k nearest neighbors of each word [3]) – 77.0. We also evaluated OpenCyc using a labor-on-demand platform, asking people to determine the correct sense for a given word in context, from a subset of OpenCyc sense definitions, obtaining an average F1 score of 37.55.

4 The Demo and Future Work

The demo will show how the implemented system's web interface annotates words/collocations in a given text fragment with RDF/OWL encoded senses from WordNet and OpenCyc. We are also going to show how to make usage of the system output programmatically, using the LarKC (the Large Knowledge Collider) platform, in order to build Semantic Web applications that rely on WSD.

As for the future work, we plan to integrate other Semantic Web resources from LOD datasets, such as DBpedia, and investigate differences in disambiguation results when using distinct resources and the potential for combining different resources in the same task. Additionally, we aim to apply our WSD algorithm to improve the Enrycher generated semantic graphs.

References

1. Fellbaum, Ch., WordNet: An Electronic Lexical Database. MIT Press (1998)
2. Jurafsky, D., Martin, J. H. Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall Series in Artificial Intelligence. (2008).
3. Koeling, R. and D. McCarthy. Sussx: WSD using Automatically Acquired Predominant Senses. In Proceedings of the 4th SemEval. pp 314--317. Prague (2007).
4. Mihalcea, R., Using Wikipedia for Automatic Word Sense Disambiguation. In Proceedings of the North American Chapter of the ACL (NAACL), Rochester, NY (2007)
5. Pedersen, T., Patwardhan, S. and Michelizzi, J. WordNet::Similarity - Measuring the Relatedness of Concepts. In Proceedings of NAACL, pp 38--41, Boston, MA (2004).
6. Ponzetto, S.P., Navigli, R., Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems. In Proceedings of the 48th ACL. pp 1522--1531. Uppsala, (2010).
7. Rusu, D., Fortuna, B. Mladenic, D. Improved Semantic Graphs with Word Sense Disambiguation. Poster. 8th ISWC. Washington, DC (2009).
8. Stajner, T., Rusu, D., Dali, L., Fortuna, B., Mladenic, D., and Grobelnik, M. Enrycher: Service Oriented Text Enrichment. In Proceedings of the 12th Int. Multiconference Information Society. pp. 203--206. Ljubljana, (2009).