KDTA: Automated Knowledge-Driven Text Annotation

Katerina Papantoniou[1], George Tsatsaronis[2,⋆], and Georgios Paliouras[1]

[1] Institute of Informatics Telecommunications, NCSR "Demokritos", Greece
[2] Dept. of Computer and Information Science,
Norwegian University of Science and Technology
kpapantoniou@iit.demokritos.gr, gbt@idi.ntnu.no,
paliourg@iit.demokritos.gr

Abstract. In this paper we demonstrate a system that automatically annotates text documents with a given domain ontology's concepts. The annotation process utilizes lexical and Web resources to analyze the semantic similarity of text components with any of the ontology concepts, and outputs a list with the proposed annotations, accompanied with appropriate confidence values. The demonstrated system is available online and free to use, and it constitutes one of the main components of the *KDTA* (*Knowledge-Driven Text Analysis*) module of the *CASAM* European research project[1].

1 Introduction

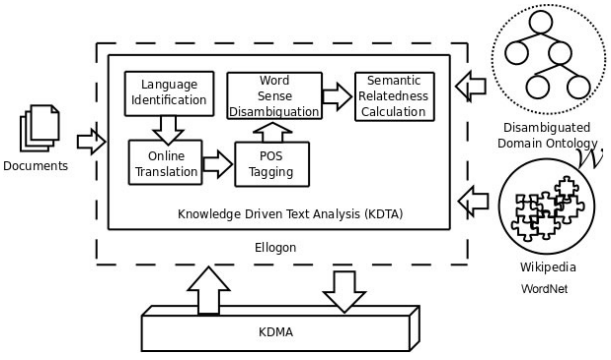
Reasoning about the content of text documents constitutes a key challenge to every semantics-aware document management system. One step towards this direction is the design and development of new methods that enable the automated annotation of plain text with ontology concepts. Such techniques enable the transfer of useful information from text documents to ontology structures, and vice versa. Motivated by this effort, the *CASAM* research project introduces the concept of computer-aided semantic annotation to accelerate the adoption of semi-automated multimedia annotation by the industry. In previous work, we presented part of the *KDTA* (*Knowledge-driven Text Analysis*) module of the overall project architecture [5], that is responsible for the automated annotation of text documents. In this work we demonstrate the component that automatically annotates text documents with ontology concepts.

The contributions of this work lie in the following: (a) a prototype implementation of a system that automatically annotates plain text with domain ontology concepts, using thesauri (e.g., *WordNet*) and Web resources (e.g., *Wikipedia*), and (b) a Web demo that is, publicly accessible. The rest of the paper is organized as follows: Section 2 summarizes the methodology of the text annotation with ontology concepts, while Section 3 presents the on-line demo that is publicly available. Finally, Section 4 concludes and provides pointers to future work.

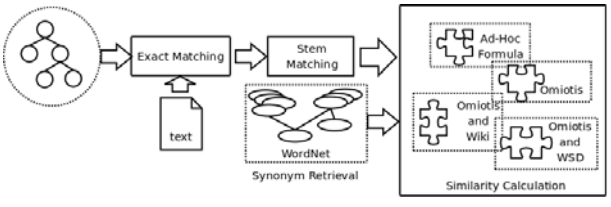
⋆ The author conducted part of this work while in the Bioinformatics Group, Biotechnological Center, Technische Universität Dresden.

[1] *Computer-Aided Semantic Annotation of Multimedia* - http://www.casam-project.eu

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 611–614, 2010.



(a) KDTA Architecture



(b) Annotation Process

Fig. 1. The KDTA overall architecture (a), and the details of the annotation process (b)

2 Automated Annotation of Text with Domain Ontology Concepts

Text annotation with ontology concepts constitutes a fundamental technology for intelligent Web applications, e.g., Semantic Web, and for this reason, there has been a large focus in this research direction over the past few years, e.g., [1,2]. However, most of the previous approaches required a lot of human intervention, or were able to annotate only specific parts of text, like named entities. In our recent work [5], we presented a new method for the automated annotation of plain text with concepts residing in a given domain ontology. The method combines a pre-processing and a semantic annotation phases (Fig. 1). In the pre-processing phase the text is processed syntactically and semantically by generic tools. The semantic annotation phase, which is the core of the method, utilizes the WordNet thesaurus[2] and the Wikipedia electronic encyclopedia[3]. The proposed method combines measures of semantic relatedness and word sense disambiguation (WSD) algorithms to annotate text words with ontology concepts.

In Fig. 1 the architecture of the *KDTA* component of *CASAM*, as well as the details of the semantic annotation process are shown (Fig. 1(a) and 1(b) respectively). The latter is the core of the proposed method and utilizes *WordNet*, and *Wikipedia* as knowledge bases, as well as two respective measures of semantic relatedness: a dictionary-based measure, namely *Omiotis* [4], which is based on *WordNet*, and a *Wikipedia*-based

[2] `http://wordnet.princeton.edu/`
[3] `http://www.wikipedia.org/`

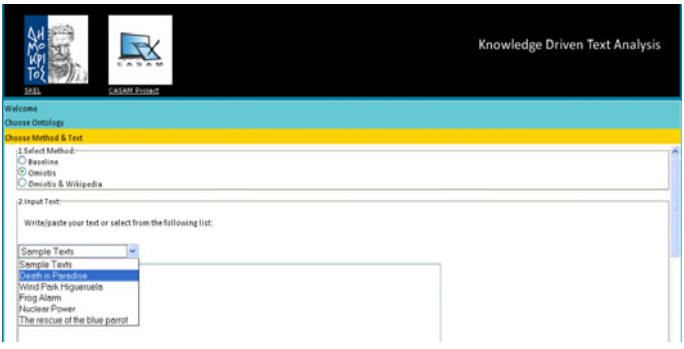


Fig. 2. Loading text in the KDTA annotation demo

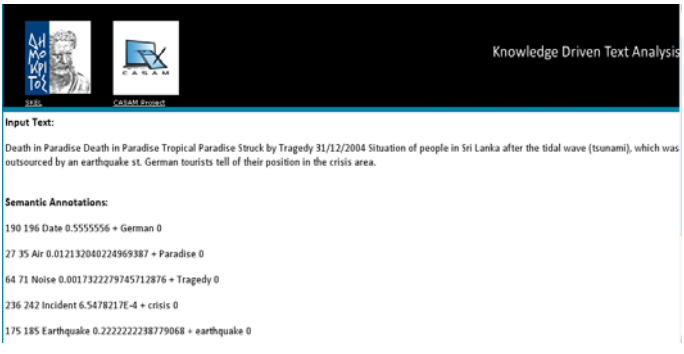


Fig. 3. Ontology-based annotation suggestions in the provided text document

measure [3]. Both measures have been shown to achieve state-of-the-art performance in measuring word-to-word semantic relatedness [4]. Both methods are also very efficient.

3 System Demonstration

A Web application that implements KDTA is publicly available online[4]. Firstly, the user may upload an ontology in *OWL* format or may select an already existing ontology by leaving the corresponding browsing path empty. This will load the default ontology of the environmental domain. Next, the user may select the text to be annotated. As Fig. 2 shows, there is already a list of text documents in the server that can be used, pertaining to environmental news. Alternatively, the user may write the text to be annotated in the provided text area. Finally, in Fig. 3 the results of the suggested annotations are shown. Column 1 is the id of the annotation, columns 2 and 3 are the start and end offsets of the annotation measured in characters, column 4 is the ontology concept that matches the respective part of the text, column 5 is a confidence value from 0 to 1, and column

[4] `http://phoebe.iit.demokritos.gr:8480/KDTA/`

6 is the source term from the given text, that led to the choice of the annotation. The method was experimentally evaluated in two domains, environmental and biomedical. For the environmental domain two datasets were examined provided by LUSA[5] and Deutche Welle[6], while for the biomedical domain MEDLINE abstracts were used. The performance was measured in terms of Macro Average Precision, Recall and F-measure with the aid of a manually created gold standards for each dataset. The results show that the method can achieve high F-measures, up to 73% in some cases. Analytical results and a detailed description of the evaluation process have been presented in [5].

4 Conclusions and Future Work

In this paper we have presented briefly an on-line web application of a system that annotates automatically text documents with concepts from a given domain ontology. The annotation system has already been embedded in the *CASAM* prototype successfully, and evaluation results have shown that provides accurate text annotation with ontology concepts. In the future, we plan to investigate the usage of more measures of semantic relatedness, and attempt to ensemble their confidence.

Acknowledgements

This work has been supported by the EU, in the context of the CASAM project (Contract number FP7-217061, Web site: www.casam-project.eu). We would like to thank our partners in the project for providing us with the environmental ontology and the corresponding data.

References

1. Cimiano, P., Ladwig, G., Staab, S.: Gimme' the context: context-driven automatic semantic annotation with c-pankow. In: WWW, pp. 332–341 (2005)
2. El-Beltagy, S., Hazman, M., Rafea, A.: Ontology based annotation of text segments. In: SAC (2007)
3. Milne, D., Witten, I.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: AAAI Workshop on Wikipedia and Artificial Intelligence (2008)
4. Tsatsaronis, G., Varlamis, I., Vazirgiannis, M.: Text relatedness based on a word thesaurus. Journal of Artificial Intelligence Research 37, 1–39 (2010)
5. Zavitsanos, E., Tsatsaronis, G., Varlamis, E., Paliouras, G.: Scalable semantic annotation of text using lexical and web resources. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds.) SETN 2010. LNCS, vol. 6040, pp. 287–296. Springer, Heidelberg (2010)

[5] http://www.lusa.pt/lusaweb/
[6] http://www.dw-world.de/