Scalable Semantic Annotation of Text Using Lexical and Web Resources

Elias Zavitsanos[1], George Tsatsaronis[2], Iraklis Varlamis[3], and Georgios Paliouras[1]

[1] Institute of Informatics & Telecommunications, NCSR "Demokritos"
[2] Department of Computer and Information Science, Norwegian University of Science and Technology
[3] Department of Informatics and Telematics, Harokopio University
izavits@iit.demokritos.gr, gbt@idi.ntnu.no, varlamis@hua.gr, paliourg@iit.demokritos.gr

Abstract. In this paper we are dealing with the task of adding domain-specific semantic tags to a document, based solely on the domain ontology and generic lexical and Web resources. In this manner, we avoid the need for trained domain-specific lexical resources, which hinder the scalability of semantic annotation. More specifically, the proposed method maps the content of the document to concepts of the ontology, using the WordNet lexicon and Wikipedia. The method comprises a novel combination of measures of semantic relatedness and word sense disambiguation techniques to identify the most related ontology concepts for the document. We test the method on two case studies: (a) a set of summaries, accompanying environmental news videos, (b) a set of medical abstracts. The results in both cases show that the proposed method achieves reasonable performance, thus pointing to a promising path for scalable semantic annotation of documents.

1 Introduction

Reasoning about the contents of text documents, as achieved by human readers constitutes a key challenge to every semantics-aware document management system. Automated reasoning directly from text aims at the automated inference of new knowledge. 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 need, the *CASAM* research project[1] introduces the concept of computer-aided semantic annotation to accelerate the adoption of semi-automated multimedia annotation in the industry. In the context of this work, we present part of the *KDTA* (*Knowledge-driven Text Analysis*) module of the

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

S. Konstantopoulos et al. (Eds.): SETN 2010, LNAI 6040, pp. 287–296, 2010.

overall project architecture, that is responsible for the automated annotation of text documents. In particular, this work presents a new method for the automated annotation of plain text with ontology concepts from a given domain ontology. The method is based on the pre-processing of the input text and the extraction of semantic information (e.g. word senses) from text. The text processing techniques utilize knowledge bases, like the WordNet thesaurus, and the Wikipedia electronic encyclopedia[2], and combine measures of semantic relatedness and word sense disambiguation (WSD) algorithms to annotate text words with ontology concepts.

The contributions of this work lie in the following: (a) a novel method for semantic annotation of plain texts with ontology concepts, (b) experimental evaluation of the proposed method, by measuring the precision and recall of the annotations in two different data sets, pertaining to the environmental and the bioinformatics domain respectively, and (c) a study on the effects of the various techniques involved on the performance of semantic annotation (e.g. the effect of WSD techniques).

In what follows, we discuss the related work on automated or semi-automated text annotation with ontology concepts, as well as on measures of semantic relatedness and WSD techniques (Section 2). Section 3 introduces the proposed method. Section 4 presents our experimental evaluation, and Section 5 concludes the paper.

2 Related Work

2.1 Automated or Semi-automated Text Annotation with Ontology Concepts

Text annotation with ontology concepts constitutes a fundamental technology for intelligent Web applications, e.g. the Semantic Web. Usually the task is performed in a semi-automated manner, starting from an initial set of manual annotations. An automated system is then suggesting new annotations to the user and assists in extending the annotation to more fragments of text [6]. In our case, we automatically annotate text with existing ontology concepts without using any type of learning or information extraction.

In this direction, Cimiano et al. [3] propose a method for annotating named entities in a document. The method first maps entities into several linguistic patterns, which convey competing semantics, and then selects the top scoring patterns to indicate the meaning of the named entity. Though this procedure may offer high accuracy, it has limited recall, since it annotates only certain kinds of named entity.

In [4], the authors propose a method for automated semantic annotation of Web pages, which is based on the existence of data-extraction ontologies that specify formalized semantics for each domain. These ontologies are used to avoid the heuristics of standard information extraction techniques. However, a domain

[2] `http://www.wikipedia.org/`

expert is required to import the formalized semantics of the domain, in order for the system to detect candidate instances to annotate with concepts of the original domain ontology.

In the approach presented in [5], the idea of mapping text headings to one or more entries in the ontology is introduced. The mapping is performed with exact matching of the segment titles and the used ontology concepts. N-grams and simple transformations, such as stemming, are employed in order to improve the method's performance. Finally, in [8] the authors present the *Ontea* system, which is based on the application of regular expression patterns and methods of lemmatization. In this case the caveat, which prohibits this approach from being applicable to free text, is the need for predefined domain specific patterns that constitute the basis for the Web document annotation.

2.2 Measures of Semantic Relatedness and Similarity

Semantic relatedness measures estimate the degree of relatedness or similarity[3] between two concepts in a thesaurus. Such measures can be classified to dictionary-based, corpus-based and hybrid. Among dictionary-based measures, the measures in [1] and [9] consider factors such as the density and depth of concepts in the set, or the length of the shortest path that connects them, or even the maximum depth of the taxonomy. However, in most such measures, it is assumed that all edges in the path are equally important. Resnik's [13] measure for a pair of concepts A,B is based on the Information Content (IC) of the deepest concept C that can subsume both A and B. The measure combines both the hierarchy of the used thesaurus, and statistical information for concept occurrences measured in large corpora. Recent work includes the measure in [12], which utilizes the gloss words found in the word's definitions to create WordNet-based context vectors, and several Wikipedia-based measures [7, 11]. We encourage the reader to consult the analysis in [2] for a detailed discussion on relatedness measures. Although any of the aforementioned measures of semantic similarity or relatedness could fit our method, in this work, we use the *Omiotis* measure of semantic relatedness between two words [16, 15], which was shown to provide the highest correlation with human judgments among the dictionary-based measures of semantic relatedness. For the cases where one of the words does not exist in WordNet, we use the Wikipedia-based measure of Milne and Witten [11], since among the offered Wikipedia-based alternatives, this is the fastest, and provides very high correlation with human judgements.

2.3 Word Sense Disambiguation

In the proposed method, we also explore the merits of sense disambiguation prior to computing the semantic relatedness between words. Thus, before computing semantic relatedness between text terms and ontology concepts, we first disambiguate the text terms, so as to compute even more precise relatedness values,

[3] Similarity measures use only the hierarchical relations from a thesaurus, whereas relatedness measures employ all the available relations.

since word-to-word measures of semantic relatedness do not take into account the context of the terms. The WSD method that we are employing is unsupervised. Though supervised methods outperform their unsupervised rivals, they require extensive training in large data sets. Unsupervised approaches comprise corpus-based [17], knowledge-based [10] and graph-based [14] methods. However, the graph-based methods demonstrate high performance and seem to be a promising solution for unsupervised WSD. Such methods rely on the construction of semantic graphs from text. The graphs are consequently processed in order to select the most appropriate meaning[4] of each examined word, in its given context. In this work, we use a graph-based approach, that constructs semantic networks and processes them with an altered PageRank formula that takes into account edge weights. The PageRank-based method is described in [14]. Any other WSD approach could have been implemented instead in *CASAM*. However, the method that we selected has demonstrated high accuracy with full coverage for all parts of speech when tested in benchmark WSD data sets [14].

3 Semantic-Based Automated Annotation of Text Documents with Ontology Concepts

This section presents the proposed automated semantic annotation method that is followed in *CASAM*. The overall architecture of the KDTA module is depicted in Figure 1. Given a text document written in natural language, the preprocessing phase starts with the identification of the text language, its translation to English, if necessary, and the application of Part of Speech (POS) tagging and Word Sense Disambiguation (WSD) techniques.

Then, the text is semantically annotated with ontology concepts. For this purpose, we calculate the semantic relatedness between candidate keywords of the text and the concepts of the domain ontology, and select for annotation the keywords that are more closely related to ontology concepts than others, in the sense of having higher relatedness values. In addition, KDTA exploits the senses of the ontology concepts, where available, as well as other external resources, such as WordNet and Wikipedia, for the calculation of semantic relatedness. Thus, given a text document, the proposed solution depicted in Figure 1, produces a ranking of proposed annotations of text segments with ontology concepts. The highest ranked proposals can be used for the annotation of the text with ontology concepts. The overall solution can scale up to large document collections, since the language identification, online translation, POS tagging, and WSD modules do not require any type of training or learning, and the computation of semantic relatedness values is supported by a powerful infrastructure [15] that has indexed all pairwise WordNet synsets relatedness values in order to accelerate computations.

[4] In the remaining of the paper, the words *concept*, *sense*, and *synset* may be used interchangeably to describe the meaning of a word, among the several offered by a dictionary or a word thesaurus.

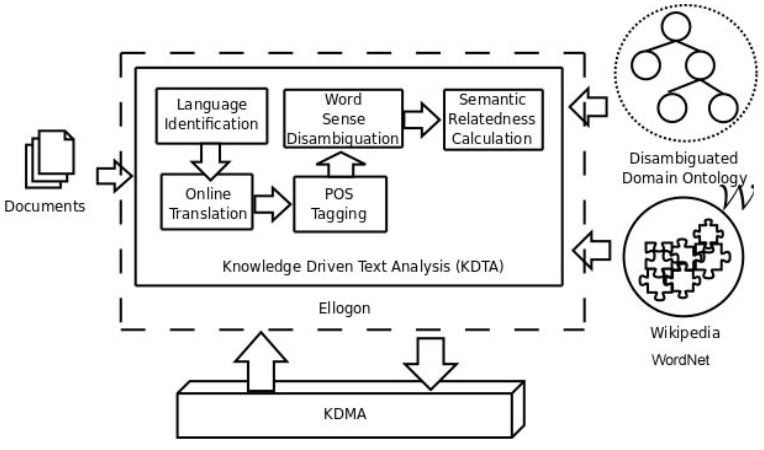

Fig. 1. Overall architecture of the KDTA CASAM Module

KDTA is implemented as a system on the general-purpose text engineering platform Ellogon[5]. Apart from providing basic pre-processing modules, Ellogon facilitates an open and flexible architecture for KDTA and provides efficient handling of text document information.

3.1 Pre-processing Phase

Given an input text, a language identifier is called in order to detect the language of the text. At this stage, KDTA operates on English documents, and thus, in case the text appears in another language, online translation services are exploited to translate the input into English. The next step is the annotation of the text with part-of-speech (POS) tags. The use of such a tagger is important, since the POS tag provides useful information to the disambiguation process and it is also helpful in the identification of candidate keywords to be annotated with ontology concepts. Particularly in *CASAM*, the domain ontology comprises mainly nouns, and thus, nouns or a noun phrases in the input text are more likely to be linked to concepts of the ontology. The last step of the pre-processing phase is the disambiguation of the input text. This process results in finding the correct sense of each word, by consulting WordNet. In particular, we use the PageRank-based method in [14] to find the sense that corresponds to each word.

3.2 Annotating Text Words with Ontology Concepts

The annotation procedure, as shown in Figure 2, comprises three consecutive steps: exact matching, stem matching and semantic matching (similarity calculation).

At the first step of exact matching, the method searches for lexicalizations of concepts inside the input text. In case of success, the document is annotated

[5] `http://www.ellogon.org/`

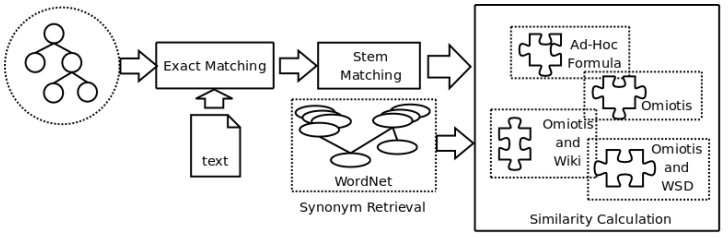

Fig. 2. The proposed annotation method

with the corresponding concept, and a relatedness value equal to 1 is assigned to that annotation. If, on the other hand, none of the concepts of the given ontology appears in the text, in its original form (i.e. as it appears in the ontology), the second step searches for appearances of its stemmed form. If such a case occurs, the document is annotated with the corresponding concept and a relatedness value equal to 0.9 is assigned to the annotation.

The third step is responsible for a more advanced annotation procedure. Four different methods are implemented:

(a) Baseline Ad-Hoc method - When this method is used, KDTA consults WordNet to retrieve a list of synonyms for the lexicalization of each concept of the given ontology. The calculation of relatedness in this method depends to a large extent on the set of the retrieved synonyms. In particular, it assigns high relatedness scores in cases where the semantic distance between a concept and its synonym is small, and lower relatedness scores otherwise. The semantic distance is actually the length of the path in WordNet between the concept and its synonym. Equation (1) incorporates the above constraints in the calculation of SD, the relatedness score between a text keyword and a synonym of an ontology concept:

$$SD = \frac{1}{NS * \frac{\log CS}{\log NS}} \tag{1}$$

NS is the total number of synonyms of the concept in question and CS is the semantic distance, expressed as the length of the path in WordNet, between the concept and its synonym.

(b) Relatedness-based Annotation with Omiotis - In contrast to the Baseline method, where the relatedness is calculated according to the distance of the synonym from the domain concept, this method relies on the relatedness between two words (in this case between a word of the text and an ontology concept), in order to perform the annotation. Specifically, after exploiting a list of standard English common words to reduce the term space of the input text, the underlying idea is to measure the relatedness between each of the resulting words and each ontology concept. Only words that are related to concepts, in the sense of having relatedness score greater than zero, are annotated, and in particular, we annotate a specific word with the concept that gives the highest

relatedness score. Regarding the computation of relatedness between two terms, i.e. a candidate word and the lexicalization of a concept, we use the measure of *Omiotis* [16], which was shown to provide the highest correlation with human judgments among the dictionary-based measures of semantic relatedness.

(c) Relatedness-based Annotation with Omiotis and WSD - This method is an extension of the previous method. It exploits additional information, derived from the pre-processing phase, in order to construct a specific structure for each word, comprising its POS tag and its sense. This structure is further exploited by *Omiotis*, in order to calculate the semantic relatedness between the word and an ontology concept, and provide a more accurate score. However, this method requires the ontology concepts to be disambiguated as well, and thus its direct application, using any ontology is not always straightforward. Besides the disambiguation part, the main idea is the same as in (b).

(d) Relatedness-based Annotation with Omiotis and Wikipedia - The last annotation method employs an additional Wikipedia-based measure, in order to handle those cases not supported by Omiotis, i.e., the words that do not appear in WordNet. The method employs the measure of Milne and Witten [11], which is the fastest among several alternatives and provides very high correlation with human judgements.

4 Experimental Evaluation

This section presents the empirical evaluation of our semantic annotation method in two datasets. Subsection 4.1 presents evaluation results in the LUSA dataset, regarding the environmental domain, while 4.2 presents the performance of the method in the Genia dataset from the molecular biology domain.

4.1 Environmental Domain: LUSA Corpus

The first dataset that was used for the empirical evaluation of the proposed annotation method comprises 51 documents provided by the LUSA Agency[6], regarding the environmental domain. The corresponding ontology, developed in the *CASAM* project, comprises 230 concepts, covering environmental concepts, such as "Wind", "Water", "Solar Energy", "Alternative Energy", etc., entities, such as "Person", "Profession Name", etc., and technological concepts, such as "Media Equipment", "Car", "Building", etc.

For the evaluation of the proposed method in the given documents, a ground truth dataset was created in *CASAM*, in order to serve as a gold standard and assist in deriving quantitative results using Macro Average Precision and Recall. The ground truth dataset contains manual annotations of terms residing in the 51 documents, with ontology concepts from the used ontology. Furthermore, the ontology concepts were manually disambiguated with WordNet senses.

Table 1 presents the performance of the proposed method for the four alternative approaches of the advanced annotation step, discussed in 3.2. The best

[6] `http://www.lusa.pt/`

Table 1. Evaluation results for the LUSA dataset

	Baseline	Omiotis	Omiotis&WSD	Omiotis&Wiki
Macro Avg. Precision	0.73	0.51	0.54	0.51
Macro Avg. Recall	0.76	0.57	0.55	0.58
Macro Avg. Fmeasure	0.73	0.49	0.51	0.50

results were achieved with the use of the baseline method. This behavior is explained by the fact that in many cases the manually annotated data set contained cases as simple as the annotation of a term, with its stem, which exists in the ontology. Those cases do not produce high relatedness values, and thus cannot be tracked by the relatedness-based methods.

Beyond the baseline method, Omiotis, and its enhancement with Wikipedia perform rather similarly. On the other hand, the disambiguation of words seems to help increase the precision of the method by 3p.p., but decreases recall. The overall F-Measure using WSD is 2p.p. higher than the simple case, which shows that WSD can help in the computation of more accurate relatedness values.

A final point regarding the interpretation of the experimental results is that the domain ontology comprises many concepts regarding entities, such as "Person", "Person Name", "Profession Name", "Organization", "Date", etc. In the context of the *CASAM* project, the proposed method is extended by the recognition of entities, using the Open Calais[7] service. In this manner, the performance of the method can be improved further by about 20p.p., achieving nearly perfect results.

4.2 Molecular Biology Domain: GENIA Corpus

In order to test the applicability of our proposed architecture in a different domain, we also experimented on a dataset used in the molecular biology domain. More specifically, we have used the GENIA ontology[8] comprising 49 concepts and a set of 2000 MEDLINE abstracts, which have been annotated with GENIA concepts. Since we know the correct annotations per document, we were able to measure the macro-average precision, recall and F-Measure, as previously. Table 2 shows the results for the baseline, the Omiotis, and the Omiotis+Wiki approach. From the reported results, we can observe that the baseline achieves a very high precision of almost 72%, but also a very low recall, and a total F-Measure of 15%. In contrast, the Omiotis and Omiotis+Wiki approaches, increase the recall and the total F-Measure by 13p.p. and 16p.p. respectively, compared to the baseline. The reason for the low recall of the baseline method stems from the fact that there are rarely exact matchings between terms and ontology concepts in the ground truth answers. On the other hand, the few

[7] http://www.opencalais.com/

[8] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html

Table 2. Evaluation results for the GENIA dataset

	Baseline	Omiotis	Omiotis&Wiki
Macro Avg. Precision	0.72	0.30	0.36
Macro Avg. Recall	0.08	0.26	0.27
Macro Avg. Fmeasure	0.15	0.28	0.31

that exist (directly or in stemmed form), are very successfully captured by the baseline, and this explains its very high precision.

Regarding the performance of the two relatedness approaches, it is overall lower than in the first data set, due to the very relatedness values that were calculated. More specifically, in this second dataset, there were many proposals for each annotation, all having close to zero (i.e., between 10^{-5} and $3 \cdot 10^{-1}$) relatedness values. Compared to the baseline, the recall of the relatedness methods improved their overall performance. This is due to the fact that the relatedness methods can capture annotations even between a text segment and an ontology concept that contain different parts of speech, or are connected through a really long path in WordNet or Wikipedia, which is often the case in this dataset. A possible improvement in this case could occur from the use of an additional knowledge base, that would be more specific to the domain, i.e., a molecular biology lexicon. This would solve the problem of low relatedness values, since for each term candidate, the lemmas from the lexicon could be used for the computation of Omiotis. Omiotis can also compute the relatedness between two sentences, or even between a term - like an ontology concept - and a sentence.

Since the relatedness approaches seem to improve the overall performance in this dataset, but mostly due to increased recall, we have also experimented for various thresholds of the Omiotis values, i.e., below which values, we do not consider the proposals at all. Our results showed that the macro-averaged precision can reach up to almost 95% for the Omiotis and the combined Omiotis-Wikipedia approaches, but the respective recall drops to almost 3%. The cut-offs that we tested were 10^{-3}, 10^{-2}, and 10^{-1}, with the latter producing the best precision. Further investigation of how to tune automatically the relatedness variants of our approach, seems promising and may lead to even more interesting results in the future.

5 Conclusions

This work presented a method for automated semantic annotation of documents with ontology concepts, based on generic lexicons and Web resources. The use of generic lexical and Web resources removes the need for trained semantic classifiers, thus constituting the method scalable. The proposed method consists of a novel combination of measures of semantic relatedness and word sense disambiguation techniques, in order to identify the most related ontology concepts for a given document. The proposed method forms the basis for the Knowledge-driven Text Analysis (KDTA) module, in the context of the *CASAM*

project, and we have validated its performance in two case studies, obtaining promising results.

Acknowledgments

This work has been partially funded by the *CASAM* Project, under the EU FP7 programme (contract number FP7-217061). We would like to thank our partners in *CASAM* for providing us with the ontology and the data that we used in the first experiment.

References

1. Agirre, E., Rigau, G.: A proposal for word sense disambiguation using conceptual distance. In: International Conference on Recent Advances in NLP (1995)
2. Budanitsky, A., Hirst, G.: Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics 32(1), 13–47 (2006)
3. Cimiano, P., Ladwig, G., Staab, S.: Gimme' the context: context-driven automatic semantic annotation with c-pankow. In: WWW, pp. 332–341 (2005)
4. Ding, Y., Embley, D.W.: Using data-extraction ontologies to foster automating semantic annotation. In: ICDE Workshops (2006)
5. El-Beltagy, S.R., Hazman, M., Rafea, A.A.: Ontology based annotation of text segments. In: SAC (2007)
6. Erdmann, M., Maedche, A., Schnurr, H.P., Staab, S.: From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. ETAI Journal - Section on Semantic Web 6(2) (2001)
7. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using wikipedia-based explicit semantic analysis. In: IJCAI, pp. 1606–1611 (2007)
8. Laclavik, M., Seleng, M., Gatial, E., Balogh, Z., Hluchý, L.: Ontology based text annotation - ontea. In: EJC (2006)
9. Leacock, C., Miller, G., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. Computational Linguistics 24(1), 147–165 (1998)
10. Lesk, M.: Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. In: SIGDOC (1986)
11. Milne, D., Witten, I.H.: An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In: AAAI Workshop on Wikipedia and Artificial Intelligence (2008)
12. Patwardhan, S., Pedersen, T.: Using wordnet based context vectors to estimate the semantic relatedness of concepts. In: EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together (2006)
13. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
14. Tsatsaronis, G., Varlamis, I., Nørvåg, K.: An experimental study on unsupervised graph-based word sense disambiguation. In: CICLing (2010)
15. Tsatsaronis, G., Varlamis, I., Nørvåg, K., Vazirgiannis, M.: Omiotis: A thesaurus-based measure of text relatedness. In: ECML-PKDD (2009)
16. Tsatsaronis, G., Varlamis, I., Vazirgiannis, M.: Text relatedness based on a word thesaurus. Journal of Artificial Intelligence Research 37, 1–39 (2010)
17. Yarowsky, D.: Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In: Int. Conf. on Compuitational Linguistics (1992)