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, in order 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 indicating a promising path for scalable semantic annotation of documents.

1 Introduction

Content-based reasoning about text documents produced by human experts constitutes a key challenge to every semantics-aware document management system. By and large, the automated reasoning directly from text, allows for the automatic 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 (*Computer-Aided Semantic Annotation of Multimedia*) introduces the concept of computer-aided semantic annotation to accelerate the adoption of semi-automated multimedia annotation in the industry. In this work, we present part of the *KDTA* (*Knowledge-driven Text Analysis*) module of the 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 residing in a given domain ontology. The method is based on the pre-processing of the input text with techniques that extract semantic information from text (e.g., word senses). Its main processing utilizes knowledge bases, like the WordNet thesaurus, and the Wikipedia electronic encyclopedia[2], and combines 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) thorough experimental evaluation of the proposed method, by measuring the precision and recall of the performed annotations in two different data sets, pertaining to the environmental and the bioinformatics domain respectively, and (c) study of the effect that several categories of used techniques have on the performed task (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., 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 that assist in extending the annotation in more fragments of text [6]. In our case, we automatically annotate text with existing ontology concepts without using any type of learning, and without using any information extraction techniques.

In this direction, Cimiano et al. [3] have proposed a method that annotates named entities in a document by first mapping them into several linguistic patterns, which convey competing semantic meanings. The top scoring patterns are the ones selected to indicate the meaning of the named entity. Though this procedure may offer high accuracy, the system only annotates named entities which exist in very specific instances found in the examined Web page, thus limiting the recall of the method. In [4], the authors propose a method of automated semantic annotation of Web pages based on the existence of data-extraction ontologies, which specify for a specific domain its formalized semantics. These ontologies are used to avoid the heuristics of standard information extraction techniques. However, a domain 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.

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

In other approaches, the authors in [5] the idea of mapping text headings to one or more entries in the ontology is utilized. The mapping is performed with exact match 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 able to perform in free text is the need for predefined domain specific patterns, which constitute the basis for the Web documents annotation.

2.2 Measures of Semantic Relatedness and Similarity

Semantic relatedness measures estimate the degree of relatedness or similarity[4] 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] take into account 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, is is assumed that all edges in the path are equally important. Resnik's [13] measure for pairs of concepts is based on the Information Content (IC) of the deepest concept that can subsume both. The measure combines both the hierarchy of the used thesaurus, and statistical information for concept occurrences measured in large corpora. Recent works include 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 [15,16], 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 embedding WSD before 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, since word-to-word measures of semantic relatedness do not take into account the context of the terms. The WSD method we are employing is an unsupervised

[4] In the case of relatedness, not only the hierarchical relations from a thesaurus are used, compared to similarity.

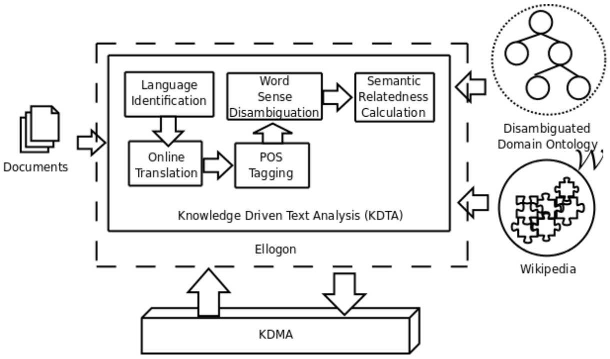
Language Identification
Word Sense Disambiguation
Semantic Relatedness Calculation
Online Translation
POS Tagging
Documents
Knowledge Driven Text Analysis (KDTA)
Ellogon
Disambiguated Domain Ontology
Wikipedia
KDMA

Fig. 1. Overall architecture of the KDTA CASAM Module.

one. 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 in the construction of semantic graphs from text. The graphs are consequently processed in order to select the most appropriate meaning[5] 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 used is described in [14]. Any other WSD approach could have been implemented instead in *CASAM*. However, the method we selected has shown to produce very high accuracy with full coverage for all parts of speech 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, Part of Speech (POS) tagging and the application of Word Sense Disambiguation (WSD) techniques.

The second step performs the semantic annotation of the text with ontology concepts. This step applies methods that are based on the calculation of semantic relatedness between candidate keywords of the text and the concepts of the given ontology, and selects those keywords to be annotated, that are more

[5] 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.

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 in case they are 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 an automated annotation of text with ontology concepts. The overall solution can scale for even large number of documents, 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 made through a large infrastructure [15] that has indexed all pairwise WordNet synsets relatedness values.

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, *CASAM* 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 POS tags. The application 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, a noun or a noun phrase of the input text will probably be annotated. 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). The latter step provides much flexibility on the choice of the calculation of relatedness between words and ontology concepts. Depending on that choice, the third step may rely on all the information obtained by the pre-processing module (POS tags, word senses, etc.), or on a subpart of it.

When the method initiates, the first step searches for lexicalizations of concepts inside the input text. In case of success, it annotates the corresponding word with the lexicalization of the concept, and assigns a relatedness value to that annotation equal to 1. If on the other hand, none concept of the given ontology appears in the text in the form that it appears in the ontology, the second step searches for appearances of its stemmed form. If such a case occurs, the corresponding word is annotated with that ontology concept and a relatedness value equal to 0.9 is assigned to that annotation.

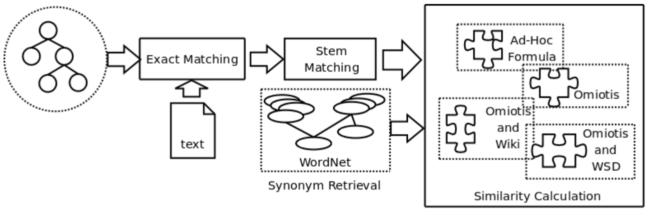

Fig. 2. The proposed annotation method.

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

(a) Baseline Ad-Hoc method - When this method is used, KDTA consults the WordNet, in order to retrieve a list of synonyms for the lexicalization of each concept of the given ontology. The calculation of the relatedness in this method depends to a large extent on the set of the retrieved synonyms. In particular, it tends to assign 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 the relatedness score.

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

Assuming a possible annotation of a keyword with a synonym of an ontology concept, SD would be the relatedness score for that annotation, NS would be the total number of synonyms of the concept in question and CS would be 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 way the annotation was performed, 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 that concept that gives the highest relatedness score. Regarding the calculation 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 enhancement of the one mentioned in the previous paragraph. 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 straightforward. Besides the disambiguation part, the main idea is the same with (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 - the pairs of words that do not appear in WordNet. The measure in [11] is employed, 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: MESH Concepts and LUSA Documents

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, provided by the technical university of Hamburg[7] (TUHH), 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 by the *CASAM* project participants, in order to serve as a gold standard and to be used to derive quantitative results in terms of 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 used 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. As Table 1 shows, the best experimental results were achieved with the use of the baseline

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

[7] http://www.tu-harburg.de

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

method. This behavior was expected to some extent, since 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. Since the rest of the methods rely on the calculation of the relatedness in order to perform the annotation, those cases will not produce very high relatedness values, and thus cannot be tracked by the relatedness-based methods.

From the experimental analysis, we conclude that beyond the baseline method, Omiotis, and Omiotis in conjunction with Wikipedia perform rather similarly. On the other hand, the disambiguation of words before the calculation of relatedness seems to help increase the precision of the method by 3% (p.p.), but decreases the recall. However, the overall F-Measure using WSD is 2% (p.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 range of the values of the experimental results is that the given 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 also aided by a Web service client that requests from Open Calais[8] to perform a named entity recognition to the input text, improving this way the performance of the method by increasing the values of the experimental results about 20%.

4.2 Molecular Biology Domain: GENIA Concepts and MEDLINE Documents

In order to stress out the applicability of our proposed architecture for performing automated text annotation with ontology concepts in a different domain that the environmental, we also experimented on a dataset used in the molecular biology domain. More specifically, we have used the GENIA ontology [9] 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 are 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-Wikipedia approach. From the reported results, we can conclude that the baseline, though achieving an impressive precision of almost 72%, it has very

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

[9] `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

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 13% and 16% (p.p.) respectively, compared to the baseline. Analyzing the results shown in Table 2, the reason of the low recall in the case 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 that exist (directly or through the use of stemming), are very successfully captured by the baseline, and this explains its very high precision. Regarding the performance of the two relatedness approaches, it is in general lower than the one experienced in the first data set, and this stems from the fact that all the produced relatedness values were very low in this case, in contrast to the relatedness values produced in the first experiment.

More specifically, in this second dataset, there were many proposals for each annotation, all having very low relatedness values, close to zero (i.e., between 10^{-5} and $3 \cdot 10^{-1}$). However, the recall of the relatedness methods increased very much the overall performance, and 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 instead for the computation of Omiotis (i.e., 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% in that case. The tested cut-offs tried were 10^{-3}, 10^{-2}, and 10^{-1}, with the latter producing the best precision. In all, a 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

We have presented a method for automated semantic annotation of documents with ontology concepts, based on generic lexicons and Web resources. 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 is used in the context of the *CASAM* project, and we have validated its performance in two case studies, concluding in promising results.

References

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