A Methodology for Semantically Annotating a Corpus Using a Domain Ontology and Machine Learning

Alexandros Valarakos*‡, Georgios Sigletos*, Vangelis Karkaletsis*, Georgios Paliouras*

* Software and Knowledge Engineering Laboratory
Institute of Informatics and Telecommunications,
National Centre for Scientific Research "Demokritos"
153 10 Ag. Paraskevi, Athens, Greece
{alexv, sigletos, vangelis, paliourg}@iit.demokritos.gr

‡ Department of Information and Telecommunication Systems
Engineering, School of Sciences,University of the Aegean
83200, Karlovassi, Samos, Greece

Abstract

In this paper we present a methodology for the semantic annotation of domain-specific corpora. This method relies on a domain ontology used initially for identifying and annotating domain-specific instances within the corpus. A machine learning-based information extraction system is then trained on the annotated corpus. The final result of this process is a model which is used to annotate new corpora in the specific domain. We applied the proposed methodology to a Web corpus examining different ontology size using hidden Markov models. The paper presents the proposed methodology together with some first experimental results.

1 Introduction

Annotating Web pages with semantic information is fundamental for accomplishment of the Semantic Web[1] vision, as semantic annotations can be exploited by various Web services (e.g. search engines, information extraction applications). However, annotating a corpus semantically is an expensive and error-prone process. Moreover, the task becomes even more difficult when a variety of knowledge resources needs to be taken into consideration. Thus, it would be very useful to annotate corpora semantically, using existing knowledge resources. In order to achieve this, a promising approach seems to be the combination of natural language processing and machine learning methods.

In this paper we propose a methodology following this approach and evaluate it on the task of semantically annotating with named entities a domain-specific corpus of Web pages. Named Entity Recognition (NER) deals with the identification and categorization of specific names, numerical and temporal expressions, etc., within a corpus and forms an important subtask of the information extraction process.

The proposed methodology exploits a domain ontology and a NER system automatically trained in the specific domain using hidden Markov models (HMMs). The ontology knowledge is used for the initial annotation of the corpus with ontology instances. This corpus is then used to train a NER system using a machine learning method (HMMs is used in the case study presented here). The resulting NER system will be then able to identify new named entities that are not included in the ontology.

At runtime, the named entities identified from both the ontology and the HMM-based system are annotated. The contribution of the HMM-based NER system is important, as it can identify named entities not included in the ontology. However, the contribution of the ontology is also important, as it corrects some entities erroneously identified by the NER system. Experimental results highlight the effective collaboration of the two knowledge sources.

In the following section of this paper we present the proposed methodology. We describe the use of the ontology, its structure and the knowledge it incorporates, as well as the way in which we use the HMMs. In section 3 we describe the conducted experiments and discuss the results. Section 4 presents

[1] http://www.w3c.org/2001/sw

some related work and section 5 concludes discussing potential improvements of our method.

2 Methodology

Our methodology is depicted in Figure 1. A domain ontology is used first to identify ontology instances within a domain-specific corpus. This initial ontology-based annotation stage is very precise, but may suffer from low recall, as the ontology is not expected to contain all of the domain-specific instances. At the 2nd stage, a machine learning-based NER system is trained from the ontology-annotated corpus. It exploits the content and context of the ontology's instances that appear in the corpus. The union of the ontology-based added annotations and the annotations from the trained machine learning-based NER system form the final annotated corpus.

One of the main advantages of a domain ontology is that it centralizes and organizes valuable knowledge in a structured form ready to be exploited. We take advantage of this in order to semantically annotate a corpus with the appropriate knowledge contained in the ontology associated with a specific task. However, we cannot rely entirely on the ontology-based annotation because an ontology may be incomplete or out-of-date. For example, an ontology that has been constructed for the domain of laptops last year, it is possible to lack the latest processor types. For this reason we train a machine learning-based NER system that learns to recognize new instances.

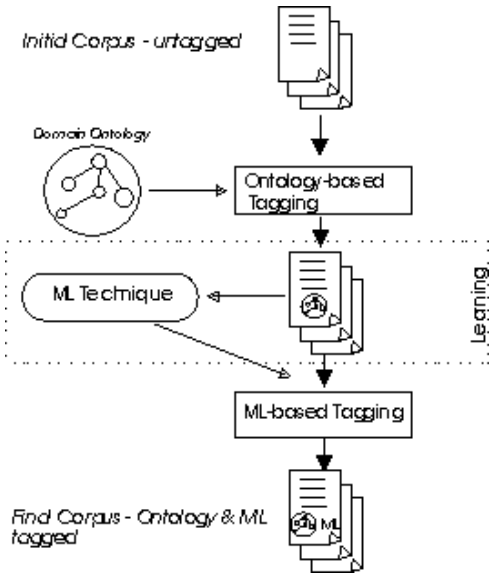

Figure 1: The proposed methodology for semantic annotation of Web corpora.

2.1 Ontology-based Tagging

2.1.1 The CROSSMARC Ontology

The ontology we used in our case study describes laptop products and has been manually constructed in the context of the CROSSMARC[2] project using a version of the Protégé (Noy et al. 2000) ontology editor adapted to the project needs. This ontology consists of the main concept, named laptop, the concepts that correspond to the laptop features (e.g. processor, memory, screen, brand etc.) and the features' attributes (e.g. processor name and speed, screen type, size and resolution, etc.) (Pazienza et al. 2003). Lexicons for the languages of the CROSSMARC project are linked to the ontology concepts, features and attributes. The English lexicon consists of 176 instances. There are also lexicons for the other 3 languages of the project (Greek, Italian, and French).

In our case study, we have chosen a part of the whole ontology to experiment with. It includes the following concepts: ***processor*** with attribute *name*, ***manufacturer*** with attribute *name*, ***screen*** with attributes *resolution* and *type*, ***preinstalled*** with attribute *software* and ***battery*** with attribute *type*. Moreover, we derived three subsets from the initial ontology size of size 75%, 50% and 25% , by appropriately decreasing the number of instances. These subsets were constructed based on the evolution of the instances in time. For instance, in the laptops domain, "*pentium*" is predecessor of "*pentium 3*", thus "*pentium*" was selected to participate into the 25% of the initial ontology. Our aim was to study the effect of the ontology size on the semantic annotation task.

2.1.2 Exploiting the Domain Ontology

The ontology provides us with knowledge related to the named entity type an instance belongs to. Therefore, the ontology is used at a first stage as a NER system. In CROSSMARC the monolingual NER systems are using a common DTD, which specifies the named entity types for a specific domain. Therefore, we had to map the instances of the attributes of the CROSSMARC ontology to the corresponding named entity types of the domain-specific DTD. This mapping is shown in Table 1 for the 1st domain of CROSSMARC (laptops offers).

A simple string matching mechanism is used to annotate the corpus with the ontological knowledge. This instance-matching is biased to select the longest

[2] http://www.iit.demokritos.gr/skel/crossmarc

match lexical expression, i.e. among the expression "Intel Pentium III" and its part "Intel Pentium" expression, the former one will be chosen.

ONTOLOGY (Concept.Attribute)	Named Entities (TYPE)
Processor.Name	Processor
Manufacturer.Name	Manuf
Screen.Resolution	Resolution
Preinstalled.OS	SOFT_OS
Screen.Type	Term
Battery.Type	Term

Table 1: Association of named entities types

Although, we have exploited a particular portion of knowledge related to named entities from the ontology, other types of ontology's knowledge can also be used for other information extraction tasks. For example, fact types of named entities can be extracted as this type of knowledge is also available in the ontology.

At the end of this phase, the ontology-based annotated corpus can be used as a training dataset for the learning algorithm.

2.2 Learning using Hidden Markov Models

Hidden Markov modeling is a powerful statistical learning technique, suited for the modeling of sequential data, such as spoken or written language. The main advantage of HMMs in language modeling is their strong statistical foundations, which provide a sound theoretical basis for the constructed models. HMMs have been successfully used in many language related tasks, including part-of-speech tagging (Kupiec 92), named entity recognition (Bikel et al. 99) and text segmentation (Yamron et al. 1998).

In our task, we use word tokens to train a single HMM for each named entity type in the second column of Table 1, as proposed in (Freitag & McCallum 99) and (Seymore et al. 99). The structure of each HMM is carefully set by hand. The model parameters are estimated in a single pass over the training data by calculating ratios of counts (maximum likelihood estimation). At runtime, each HMM is applied to a Web page, using the Viterbi procedure to identify matches.

3 Experimental Results

The initial corpus consists of 100 English Web pages describing laptops, of which 50 are used for training and 50 for testing. The corpus processing was done using the text engineering platform Ellogon (Petasis et al. 2002)[3]. The application of the proposed methodology requires the preprocessing of the corpus using a tokeniser. The tokenizer identifies text tokens (i.e., words, symbols, etc.) in the web pages and characterizes them according to a token-type tag set which encodes graphological information (e.g. the token is an English upper case word).

According to the proposed methodology, at the first phase we performed three separate annotations of the pre-processed training corpus using 75%, 50% and 25% of the initial ontology, respectively (see Table 2). The difference between instances and examples is that the latter describe the number of instances not uniquely appear in the corpus.

% ontology	# Instances	# examples
75%	146	759
50%	117	555
25%	76	214

Table 2: Number of examples in the training dataset**.**

In the next phase we used the ontology-annotated corpus to train the HMM-based NER system. Finally, we annotated the testing corpus using the ontology-based tagger and the HMM-based tagger. In case of overlapping annotations, the ontology-tagged annotation was preferred from the corresponding HMM-based annotation.

We evaluated the results of each of the three tagging methods (ontology-based, HMM-based, combination) over the testing corpus of 50 Web pages using the corpus comparison tools provided by the text engineering platform Ellogon. The performance of each method is evaluated using the precision and recall metrics. Tables 3, 4 and 5 show the results of the above annotation methods for five types of named entities and for an ontology size of 75%, 50% and 25%, respectively.

The precision of ontology-based tagging was high, as expected. On the other hand, recall was affected by the size of the ontology. The precision of the HMM-based tagging varied between 64% and 77%, while its recall was low strongly depending on the size of the ontology that we were using to create the training data. However, the combination of the two methods performs better recall as the HMM-based tagging provided new annotations not included in the ontology. Furthermore, the precision of the combined approach is higher than that of the HMM-based

[3] http://www.iit.demokritos.gr/skel/Ellogon

tagging, although not perfect as the ontology-based tagging.

Annotation Method	Precision (%)	Recall (%)
Union	74,0	76,0
Ontology	100,0	66,1
HMM	69,2	65,5

Table 3: Using the 75% of the ontology

Annotation Method	Precision (%)	Recall (%)
Union	67,0	57,4
Ontology	100,0	50,0
HMM	62,3	47,7

Table 4: Using the 50% of the ontology

Annotation Method	Precision (%)	Recall (%)
Union	71,3	33,1
Ontology	100,0	24,5
HMM	68,0	26,3

Table 5: Using the 25% of the ontology

Figures 2 and 3 graphically illustrate the results of Tables 3, 4 and 5.

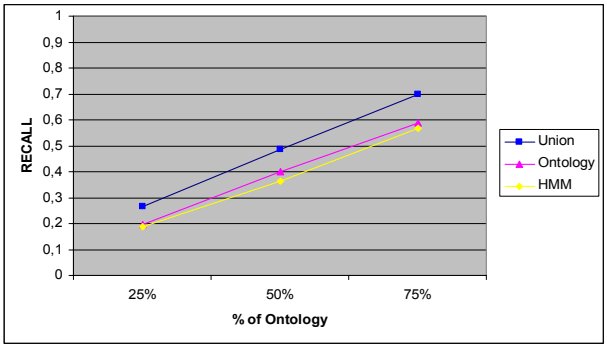

Figure 2: Recall for all tagging methods

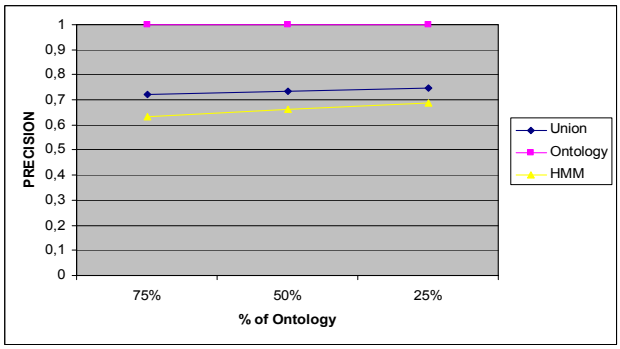

Figure 3: Precision for all tagging methods

4 Related Work

Semantic annotation of a corpus can be performed semi-automatically by various annotation tools (Handschuh et al. 2002; Vargas-Vera et al. 2002), which speed up the whole procedure by providing a friendly interface to a domain expert. A manually annotated corpus can be used to train an information extraction system, which will then annotate it further. (Ciravegna et al. 2002) present a methodology for the interaction between the user and the automatically added annotations, which are derived by an information extraction system. This system is trained initially using representative annotations added manually by the user. The trained system is then used to add automatically semantic annotation to the corpus, which can be modified by the user in order to retrain the system. Our methodology presents similarities with this approach, but instead of using a manually annotated corpus it exploits an ontology-based one. Other relevant approaches are those of (Thelen & Riloff 2002) and (Petasis et al. 2000). The aim of these approaches is the exploitation of an initial small-sized lexicon and a machine learning-based IE system for the lexicon enrichment through an iterative approach.

5 Concluding Remarks

We presented a methodology for the semantic annotation of a corpus exploiting a domain ontology according to the task at hand and machine learning. We applied the methodology to an information extraction subtask: recognition of named entities.

The proposed methodology is based on the combination of an ontology-based matching process and an HMM-based NER system trained on a corpus annotated by the ontology. The only requirement posed by this approach is the existence of a domain-specific ontology with a satisfactory level of coverage for the particular domain. In order to overcome this limitation we plan to examine an iterative application of the methodology through which an initial ontology of limited size can be enriched using the results of the ML-based method.

These initial results encourage us to investigate further the use of the methodology to bootstrap the annotation of a corpus by providing to the human annotator some precise annotations (ontology-based annotations) and some additional ones for consideration (HMM-based annotation).

Acknowledgements

CROSSMARC is an R&D project under the IST Programme of the European Union (IST 2000-25366)[4]. We would like to thank the annotation team

[4] http://www.iit.demokritos.gr/skel/crossmarc

of the University of Edinburgh for providing the annotated corpora and the ontology development team of the University of Roma Tor Vergata for providing the initial ontology.

References

(Bikel et al. 1997) Bikel D.M., Miller S., Schwartz R., Weishedel R., "Nymble: a high performance learning name finder". In *Proceedings of ANLP-97*, 194-201.

(Ciravegna et al. 2002) Fabio Ciravegna, Alexiei Dingli, Daniela Petrelli and Yorick Wilks, "User-System Cooperation in Document Annotation based on Information Extraction". In *Proceedings of the EKAW02*, October 2002, Sigüenza Spain.

(Freitag, McCallum 1999) Freitag, D., McCallum, A.K., "Information Extraction using HMMs and shrinkage". In Proceedings of *AAAI-99 Workshop on Machine Learning for Information Extraction*, pp. 31-36 (1999).

(Handschuh et al. 2002) Siegfried Handschuh, Steffen Staab and Fabio Ciravegna, "S-CREAM - Semi-automatic CREAtion of Metadata". In Proceedings of *EKAW02*, 2002.

(Kupiec 1992) Kupiec, J., "Robust part-of-speech tagging using a hidden Markov model". *Computer Speech and Language*, 6, 225-242.

(Noy et al. 2000) N. F. Noy, R. W. Fergerson, & M. A. Musen. "The knowledge model of Protege-2000: Combining interoperability and flexibility". In Proceedings of *EKAW'2000,* Juan-les-Pins, France, 2000.

(Ohta et al. 2001) Ohta, T., Tateisi, Y., Kim, J.D., Mima, H. and Tsujii, J**.,** "Ontology Based Corpus Annotation and Tools". In Proceedings of the *12th Workshop on Genome Informatics*, pp. 469-470, Dec. 2001.

(Pazienza et al. 2003) M. T. Pazienza, A. Stellato, M. Vindigni, A.Valarakos, V. Karkaletsis, "Ontology Integration in a Multilingual e-Retail System". In Proceedings of the *2nd International Conference on Universal Access in Human-Computer Interaction,* Crete, Greece, June 22-23 2003.

(Petasis et al. 2000) G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C.D. Spyropoulos, "Automatic adaptation of proper noun dictionaries through co-operation of machine learning and probabilistic methods". In *Proceedings of the 23rd ACM SIGIR Conference,* pp. 128-135, Athens, Greece, 2000.

(Petasis et al. 2002) G. Petasis, V. Karkaletsis, G. Paliouras, I. Androutsopoulos and C. D. Spyropoulos, "Ellogon: A New Text Engineering Platform". In Proceedings of *LREC 2002*, Las Palmas, Spain, pp. 72-78, May 2002.

(Poibeau, Dutoit 2002) T. Pibeau and D. Dutoit, "Generating extraction patterns from a large semantic network and an untagged corpus". In Proceedings of *COLING, 2002.*

(Seymore et al. 1999) Seymore, K., McCallum A.K., Rosenfeld, R., "Learning hidden Markov model structure for Information Extraction". *Journal of Intelligent Information Systems* 8(1): 5-28, (1999).

(Thelen, Riloff 2002) Thelen, M. and Riloff, E., "A Bootstrapping Method for Learning Semantic Lexicons using Extraction Pattern Contexts". In Proceedings of *EMNLP 2002*.

(Vargas-Vera et al. 2002) M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt and F. Ciravegna, "MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup". In Proceedings of *EKAW 2002*.

(Yanngarber et al. 2000) R. Yangarber, R. Grishman, P. Tapanainen, S. Huttunen, "Unsupervised Discovery of Scenario-Level Patterns for Information Extraction". In Proceedings of *ANLP-NAACL, 2000*.

(Yamron et al. 1998) Yamron J., Carp I., Gillick L., Lowe S., Van Mulbregt P., "A hidden Markov model approach to text segmentation and event tracking". In Proceedings of *IEEE Conference on Acoustics, Speech and Signal Processing*, vol I., Seattle, WA, pp. 333-336, 1998.