NASS: News Annotation Semantic System

Angel L. Garrido*, Oscar Gómez*, Sergio Ilarri† and Eduardo Mena†
*Grupo Heraldo - Grupo La Información
Zaragoza - Pamplona, Spain.
Email: {algarrido, ogomez}@heraldo.es
†IIS Department.
University of Zaragoza. Zaragoza, Spain.
Email: {silarri, emena}@unizar.es

Abstract—**Today in media companies there is a serious problem for cataloging news due to the large number of articles received by the documentation departments. That manual labor is subject to many errors and omissions because of the different points of view and expertise level of each staff member. There is also an additional difficulty due to the large size of the list of words in a thesaurus.**

In this paper, we present a new method for solving the problem of text categorization over a corpus of newspaper articles where the annotation must be composed of thesaurus elements. The method consists of applying lemmatization, obtaining keywords and named entities, and finally using a combination of Support Vector Machines (SVM), ontologies and heuristics to infer appropriate tags for the annotation. We carried out a detailed evaluation of our method with real newspaper articles, and we compared out tagging with the annotation performed by a real documentation department, obtaining really promising results.

Keywords-**Knowledge Discovery; Text Mining; Information Extraction; Natural Language Processing; SVM; Ontologies; Heuristics; Media.**

I. Introduction

In Media, the goal of a documentation department is to help journalists to find information in archived news in order to re-use it in a new article. To this end, these departments have to tag news every day, and the typical way to do that is by using a thesaurus: a set of items (word or phrases) used to classify things. It usually has the structure of a hierarchical list of unique terms. We have worked with real publications belonging to major Spanish media companies, whose news are tagged every day by the documentation department of these companies using a hierarchical thesaurus with near 20,000 terms.

In this paper, we present an overview of a system called NASS *(News Annotation Semantic System)*, which provides a new method to obtain thesaurus tags using semantic tools and information extraction technologies. As we had the chance to compare our results with the real tagging, we have benefited from this real-world experience to evaluate our method.

This paper is structured as follows. Section II explains our method. Section III discusses the results of our experiments. Section IV cites some related works about news categorization. Finally, Section V provides our conclusions.

II. NASS Methodology

According to [1], works like ours could be classified in the field of Ontology-Based Information Extraction (OBIE), an emergent subfield of Information Extraction (IE). The main elements in an OBIE system are a preprocessor that works over the incoming text, an information extraction module usually guided by a semantic lexicon and by a human-made ontology, and finally a Knowledge Database used for storing the system's response. The architecture of our method is represented in Figure 1.

We propose first to obtain the main keywords from the article by using text mining techniques. At the same time, using Natural Language Processing (NLP) [2], the system retrieves another type of keywords called *named entities* [3]. Second, NASS applies Support Vector Machines (SVM) [4] text classification in order to filter only the most relevant articles belonging to a particular topic. We have chosen SVM text categorization because it is a powerful and reliable tool for text categorization. Anyway, we have discovered limitations as soon as we have applied SVM over real newspaper articles: it has strong dependence on the data used in training. While it works very well with texts dealing with highly general themes, it is not the case when the main keywords change in the texts over time (e.g., when we talk about sports).

We have improved the SVM results by using techniques from Ontological Engineering: NASS uses the keywords and the named entities of the filtered texts to query an ontology about the topic these texts are talking about. That ontology has been developed by us using entities and relations existing within the context of the Spanish soccer league. Then, NASS uses the answers of these queries to increase the probability of obtaining a correct thesaurus element in each text, and it updates that score in a table. Finally, the system looks at this table, selects the terms with a score higher than a given threshold, and then labels the text with these tags.

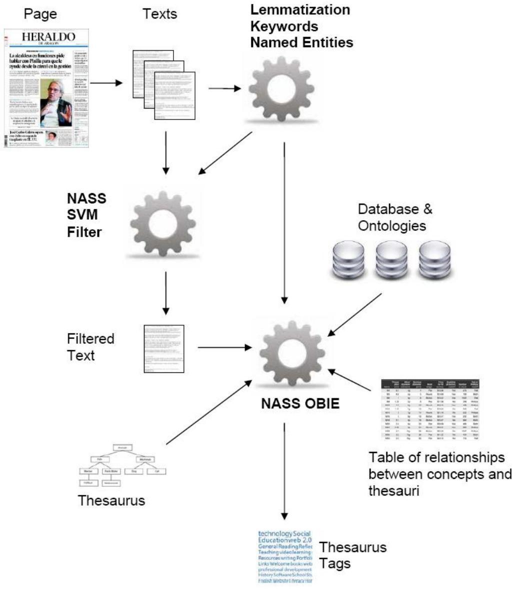

Figure 1. Architecture of the NASS system

III. Experimental Evaluation

In our experiments we have used a corpus of 1755 articles tagged with thesaurus terms manually assigned by the documentation department of the Spanish company *Grupo Heraldo*. The number of articles is limited by the fact that we have used a Spanish soccer ontology which is valid only during one season, as every year the teams, players and coaches could change. Applying NASS, we have reached not only more than 95% of recall and precision, but we had *extended and corrected* the human labeling. The operation is based on performing a proper ontology population. When we removed elements from the ontology, the precision was still very good but the recall dramatically fell.

IV. Related Work

We have combined SVM with OBIE systems. Although OBIE is a relatively new field of study, most researchers believe that it could contribute very much to IE. Many researchers work with this kind of systems obtaining good results. With respect to its use for classifying news, we would like to highlight [5], a recent work in which their authors categorize economic articles using multi-label categorization, [6] based on text mining technics and an SVM categorization engine, and [7], a combination between the use of WordNet and a particular type of artificial neural network, using unsupervised training. NASS adapts and combines some ideas of these works in a completely different way. The main differences between these works and ours is the fact that they have used only a few categories for categorizing, or they have not even been established. In addition, our experimental results are better in terms of precision and recall.

V. Conclusions

We have introduced a text categorization system adapted to Media, and we have performed experiments on real newspaper articles previously tagged manually to evaluate the accuracy of the system. We have contributed to make a good adjustment of the general methods of SVM and OBIE systems over a real set of news. The experimental results show that we are able to get a reasonable number of correct tags using IE methods like SVM, but it can be improved with the use of NLP and semantic tools in scenarios where the training set of the SVM should be updated too often. We propose that documentalists fill the instances of classes in a predefined ontology. Then, our system would have enough information to label the news automatically by using semantic tools. We found this method simpler and more intuitive for end users, and it helps to get better results. Besides, the accuracy is very good, obtaining almost 99% of correct labels. In fact, the current version of NASS is already being successfully used in the Spanish media company *Grupo Heraldo*. We will continue working on improving the system adding new and more elaborated ontologies.

Acknowledgment

This research work has been supported by the CICYT project TIN2010-21387-C02-02.

References

[1] D. C. Wimalasuriya and D. D., “Ontology-based information extraction: An introduction and a survey of current approaches,” *Journal of Information Science*, vol. 36, no. 3, pp. 306–323, 2010.

[2] A. F. Smeaton, *Using NLP or NLP Resources for Information Retrieval Tasks. Natural Language Information Retrieval*. Kluwer Academic Publishers, 1999.

[3] S. Sekine and E. Ranchhod, *Named Entities: Recognition, Classification and Use*. John Benjamins, 2009.

[4] T. Joachims, “Text categorization with support vector machines: learning with many relevant features,” in *Proceedings of ECML 98, 10th European Conference on Machine Learning*. Springer, 1998, pp. 137–142.

[5] S. Vogrincic and Z. Bosnic, “Ontology-based multi-label classification of economic articles,” *Computer Science and Information Systems*, vol. 8, no. 1, pp. 101–119, 2011.

[6] M. Mittermayer and F. K. Gerhard, “Newscats: A news categorization and trading system,” *IEEE International Conference on Data Mining*, pp. 1002–1007, 2006.

[7] S. Wermter and C. Hung, “Selforganizing classification on the reuters news corpus,” in *Proceedings of COLING 02, 19th international conference on Computational linguistics*. Association for Computational Linguistics, 2002, pp. 1–7.