# An Experience Developing a Semantic Annotation System in a Media Group*

Angel L. Garrido[1], Oscar Gómez[1], Sergio Ilarri[2] and Eduardo Mena[2]

[1] Grupo Heraldo - Grupo La Información. Zaragoza - Pamplona, Spain.
{algarrido, ogomez}@heraldo.es
[2] IIS Department, University of Zaragoza, Zaragoza, Spain.
{silarri, emena}@unizar.es

**Abstract.** Nowadays media companies have difficulties for managing large amounts of news from agencies and self-made articles. Journalists and documentalists must face categorization tasks every day. There is also an additional trouble due to the usual large size of the list of words in a thesaurus, the typical tool used to tag news in the media.

In this paper, we present a new method to tackle the problem of information extraction over a set of texts where the annotation must be composed by thesaurus elements. The method consists of applying lemmatization, obtaining keywords, and finally using a combination of Support Vector Machines (SVM), ontologies and heuristics to deduce appropriate tags for the annotation. We have evaluated it with a real set of changing news and we compared our tagging with the annotation performed by a real documentation department, obtaining very good results.

**Keywords:** Semantic tagging and classification; Information Extraction; NLP; SVM; Ontologies; Text classification; Media; News.

## 1 Introduction

In almost every company in the media industry, activities related to categorization can be found: news production systems must filter and sort the news at their entry points, documentalists must classify all the news, and even journalists themselves need to organize the vast amount of information they receive. With the appearance of the Internet, mechanisms for automatic news classification often become indispensable in order to enable their inclusion in web pages and their distribution to mobile devices like phones and tablets.

To do this job, medium and big media companies have documentation departments. They label the published news, and the typical way to do that is by using thesauri. A thesaurus [1] is a set of items (words or phrases) used to classify things. These items may be interrelated, and it has usually the structure of a hierarchical list of unique terms.

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

In this paper, we focus on the inner workings of the NASS system (*News Annotation Semantic System*), which provides a new method to obtain thesaurus tags using semantic tools and information extraction technologies. The seminal ideas of this project have been recently presented in [2] and with this paper we want to delve deeper into the operation of the system.

In our system we propose to obtain the main keywords from the article by using text mining techniques. Then, using Natural Language Processing (NLP) [3], the system retrieves other type of keywords called *named entities*. After, NASS applies Support Vector Machines (SVM) [4] text classification in order to filter articles. The system uses the keywords and the named entities of the filtered texts to query an ontology about the topic these texts are talking about, using the answers to those queries to increase the probability of obtaining correct thesaurus elements for each text, and it updates that matching score in a table. Finally, NASS looks at this table and selects the terms with a score higher than a given threshold, and then it labels the text with the corresponding tags. We have tested this method over a set of thousands of real news published by a leading Spanish media. As we had the chance to compare our results with the real tagging performed by the documentation departments, we have benefited from this real-world experience to evaluate our method.

This paper is structured in the following way. Section 2 explains our method. Section 3 discusses the results of our experiments. Section 4 cites some related work about news categorization. Finally, Section 5 provides our conclusions.

## 2 NASS Methodology

The outline of our method is as follow. First, NASS obtains from the text of every piece of news the names of the main characters, places and institutions, and then it guesses which the key ideas and major themes are. Second, the system uses this information in order to find related thesaurus terms. Finally, NASS assigns corresponding thesaurus terms to the text. The architecture of our solution which is shown in Figure 1, fits the general schema of an Ontology-Based Information Extraction (OBIE), as presented by Wimalasuriya and Dou [5].

Before obtaining keywords, NASS lemmatizes all the words in the text. Lemmatization can be defined as the process of obtaining the canonical form representing one word, called *lemma*. This procedure simplifies the task of obtaining keywords and reduces the number of words the system has to consider later. It can also help to obviate *stop words*: prepositions, conjunctions, articles, numbers and other meaningless words. Moreover, it also provides us clues about the kinds of words appearing in the text: nouns, adjectives, verbs, and so on. For this purpose, we have used Freeling [6]. Freeling is an open source suite of language analyzers developed at TALP Research Center, at BarcelonaTech (Polytechnic University of Catalunya). After this process, the system has a list of significant words, which is the input to the next process: obtaining keywords.

We propose merging two methods to obtain a set of significant keywords from a given text. The first method that NASS applies is a simple term frequency

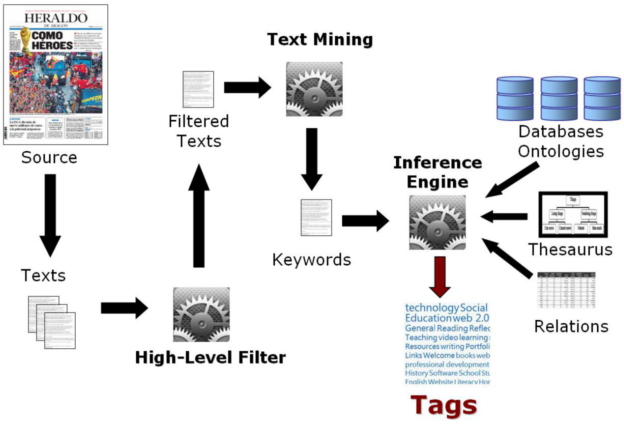

**Fig. 1.** General architecture of the NASS System.

algorithm, but with some improvements. We call it *TF-WP* (*Term Frequency – Word Position*), which is obtained by multiplying the frequency of a term with a position score that decreases as the term appears for the first time towards the end of the document. This heuristic is very useful for long documents, as more informative terms tend to appear towards the beginning of the document. The TF-WP keyword extraction formula is as follows:

$$TF - WP = (\frac{1}{2} + \frac{1}{2} * \frac{nrWords - pos}{nrWords}) * TF$$

$$TF = \frac{nrRepetitions}{nrWords}$$

where $nrWords$ is the total number of terms in the document, $pos$ is the position of the first appearance of the term, $TF$ is the frequency of each term in the document, and $nrRepetitions$ is the number of occurrences of that term.

The second method is the well-known *TF-IDF* (*Term Frequency – Inverse Document Frequency*), based on the word frequency in the text but also taking into account the whole set of documents, not only the text considered. We have merged both methods by adding the two values TF-WP and TF-IDF after applying them a weight $\alpha$ and $\beta$, respectively. At the end of this task NASS obtains a list of keywords with their number of repetitions and weights.

In news, it is very common to find names of people, places or companies. These names are usually called *named entities* [7]. For example, if the text contains the words "Dalai Lama" both words could be considered as a single named entity to capture its actual meaning. This is a better option than considering the words "Dalai" and "Lama" independently. To do this task we have used Freeling too, which uses this identification method and provides a confidence threshold to decide whether to accept a named entity or not. In our system, this threshold is set at 75% in order to ensure a good result. NASS retrieves the more relevant named entities identified, replaces whitespaces by underscores (for example, "Dalai_Lama"), and finally adds them to the same collection of keywords obtained before.

At this point, NASS has a list of the most important keywords and named entities in the input article. It could tag the text with this information, but we want to take a step forward by using the thesaurus for tagging. The question now is: how could NASS obtain thesaurus terms from a list of keywords, when the words that best summarize the text are not present in it? There are a lot of interesting ways to *infer and deduce* terms according to their meaning and their relationships with other terms. We have chosen SVM text categorization because it is a powerful and reliable tool for text categorization [4]. Regarding the type of SVM used, we have used a modified version of the Cornell SVM-Light implementation with a Gaussian radial basis function kernel and the term frequency of the keywords as features [8].

However, we have discovered some limitations as soon as we apply SVM over real news sets. SVM has a strong dependence on the data used for training. While it works very well with texts dealing with highly general topics, it is not the case when we need to classify texts on very specific topics not included in the training stage, or when the main keywords change over time. So, we have improved the SVM results by using techniques from Ontological Engineering [9]. We advocate the use of knowledge management tools (ontologies, semantic data models, inference engines) due to the benefits they can provide in this context.

The first step is to design an ontology that describes the items we want to tag, but this is not an easy task in media business. The reason is that media cover many different themes, and therefore it is a disproportionate task to try to develop an ontology about all the publishable topics. We think that a better approximation is to design an ontology for every interesting subject we want to tag, whenever SVM results are not able to get good results. Our ontologies are not only composed by named entities, as we have also introduced relationships and actions that join concepts providing semantic information, which is a substantial difference that brings advantages not contemplated by SVM.

NASS tries to match each keyword with one of the words in the ontology. For this, we have prepared in advance a table with the help of the documentation department and based on experimental and statistical analysis of the tags introduced manually. This table has two columns. In the first column we put ontology concepts. In the second column we put the probability of talking about a topic usually labeled with a term of the thesaurus when the system detects that concept in a text. Then, NASS submits SPARQL queries against the ontology and it uses Jena as a framework and Pellet as an inference engine. As soon as it finds a keyword that matches a term of the ontology, it looks at its associated concept and then the system uses the previous table to retrieve the corresponding probability. At this point, it is important to mention that some keywords could be related with one or more thesaurus tags, and also a thesaurus tag could be related to one or more keywords. NASS increases the probability of tagging the article with a term each time it accesses a row of this table. A high number of accesses to the same thesaurus term guarantees that it can be used as a tag on the article. Through an extensive experimental evaluation we found useful to use 60% as a threshold to accept a term. Finally, NASS returns the thesaurus tags obtained by this method in order to label the text.

## 3 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*. We found that the highest number of well-labeled articles occurs when we suitably populate the ontology, and if NASS fails it is due to a lack of semantic information. When the ontology has fewer elements, then the recall drops significantly (73% vs. 95%) but the precision is the same (98%). Summing up, the results obtained were really good. Furthermore, our system was able to detect additional labels that are relevant (even though they were not selected in the manual annotation) and avoid labels that were wrongly chosen by the documentation department.

## 4 Related Work

As commented along the paper, among other techniques, we have used a method commonly used for text categorization: the Support Vector Machine, or SVM [10]. This method has some appropriate features to face text classification problems, for example its capability to manage a huge number of attributes and its ability to discover which of them are important to predict the category of the text after a training stage. It is based on solid mathematical principles related to statistical learning theory and there are plenty of articles and books based on such mechanisms, not only for text classification but for any work related to cataloging all kinds of elements and entities.

As an example of project combining SVM and ontologies we can reference [11], a recent work where SVM is used to categorize economic articles using multi-label categorization. The big difference is that they use ontologies to create labels prior to the categorization process, and then use different types of SVM only in that process, which does not let them obtain neither a high degree of accuracy nor a high number of categories, whereas our proposal avoids these problems. We would also like to mention the work performed by Wu et al. [12], which faced this kind of problems using a quite interesting unsupervised method based on a Naive-Bayes classifier and natural language processing techniques.

## 5 Conclusions and Further Work

In this paper, we have presented a tagging system that helps media companies in their daily labeling labor when they use a thesaurus as the annotation tool.

We have performed experiments on real news previously tagged manually by the staff of the documentation department and our experimental results show that we are able to get a reasonable number of correct tags using methods like SVM, but the accuracy improves with the combined use of NLP and semantic tools when the training set of the SVM must be updated frequently. Instead of having to label news each year to create a reliable set for training, we propose that documentalists fill the instances of classes in a predefined ontology. Then, our system has enough information to label the news automatically by using semantic tools. We found this simpler and more intuitive for end users, and it helps to get better results. Besides, the accuracy of the automatic assignment of tags with our system is very good, obtaining 99% of correct labels. In fact, the current version of NASS is already being successfully used in several companies. In the future, we plan to introduce new and more powerful methods to enrich the system providing it with greater speed, wider scope and better accuracy.

## References

1. A. Gilchrist. 2003. Thesauri, taxonomies and ontologies: an etymological note. Journal of Documentation vol. 59(1): pp. 7-18.
2. A. L. Garrido, O. Gomez, S. Ilarri and E. Mena. 2011. NASS: news annotation semantic system. Proceedings of ICTAI 11, International Conference on Tools with Artificial Intelligence: pp. 904-905. IEEE.
3. A. F. Smeaton. 1997. Using NLP or NLP resources for information retrieval tasks. Natural Language Information Retrieval. Kluwer Academic Publishers.
4. T. Joachims. 1998. Text categorization with support vector machines: learning with many Relevant Features. Proceedings of ECML 98, European Conference on Machine Learning: pp. 137-142. Springer.
5. D.C. Wimalasuriya and D. Dou. 2010. Ontology-Based Information Extraction: an introduction and a survey of current approaches. Journal of Information Science vol. 36(3): pp. 306-323. Sage Publications.
6. X. Carreras, I. Chao, L. Padró, and M. Padró. 2004. Freeling: an open-source suite of language analyzers. Proceedings of the 4th International Conference on Language Resources and Evaluation: pp. 239-242. European Language Resources Association.
7. S. Sekine and E. Ranchhod. 2009. Named entities: recognition, classification and use. John Benjamins.
8. E. Leopold and J. Kindermann. 2002. Text categorization with support vector machines. How to Represent Texts in Input Space? Machine Learning vol. 46: pp. 423-444. Kluwer Academic Publishers.
9. M. Fernandez-Lopez and O. Corcho. 2004. Ontological engineering. Springer.
10. C. Cortes, V. N. Vapnik. 1995. Support-vector networks. Machine Learning vol. 20(3):273-297. Kluwer Academic Publishers.
11. S. Vogrincic and Z. Bosnic. 2011. Ontology-based multi-label classification of economic articles. Computer Science and Information Systems, vol. 8(1): pp. 101-119. ComSIS Consortium.
12. X. Wu, F. Xie, G. Wu, W. Ding. 2011. Personalized news filtering and summarization on the web. Proceedings of ICTAI 11, International Conference on Tools with Artificial Intelligence: pp. 414-421. IEEE.