The GENIE System - Classifying Documents by combining Mixed-Techniques

Angel L. Garrido, Maria G. Buey, Sandra Escudero, Alvaro Peiro, Sergio Ilarri, and Eduardo Mena

IIS Department, University of Zaragoza, Zaragoza, Spain
http://sid.cps.unizar.es/SEMANTICWEB/GENIE/

Abstract. Today, the automatic text classification is still an open problem and its implementation in companies and organizations with large volumes of data in text format is not a trivial matter. To achieve optimum results many parameters come into play, such as the language, the context, the level of knowledge of the issues discussed, the format of the documents, or the type of language that has been used in the documents to be classified. In this paper we describe a multi-language rule-based pipeline system, called GENIE, used for automatic document categorisation. We have used several business corpora in order to test the real capabilities of our proposal, and we have studied the results of applying different stages of the pipeline over the same data to test the influence of each step in the categorization process. The results obtained by this system are very promising, and in fact, the GENIE system is already being used on real production environments with very good results.

Keywords: documents categorization; text mining; ontologies; NLP.

1 Introduction

In almost any public or private organization that manages a considerable amount of information, activities related to text categorisation and document tagging can be found. To do this job, large organizations have documentation departments. However, the big amount of information in text format that organizations usually accumulate cannot be properly processed and documented by these departments. Besides, the manual labour of labeling carried out by these people is subject to errors due to the subjectivity of the individuals. That is why a tool that automates categorisation tasks would be very useful, and would help to improve the quality of searches that are performed later over the data.

To perform these tasks, software based on statistics and the frequency of use of words can be used, and it is also very common to use machine learning systems. However, we think that other kinds of tools capable of dealing with aspects related to Natural Language Processing (NLP) are also necessary to complement and enhance the results provided by these techniques. Moreover, to perform any task related to the processing of text documents, it is highly recommended to own the know-how of the organization, so it is highly advisable

to manage ontologies and semantic tools such as reasoners to make knowledge explicit and reason over it, respectively. Furthermore, it is very common for organizations to have their own catalog of labels, known as thesaurus, so it is basic that the system is able to obtain not only keywords from the text, but also know how to relate them to the thesaurus descriptors.

Our purpose is to bring together these techniques into an architecture that enables the automatic classification of texts, with the particular feature that it exploits different semantic methods. Although there are some researches in text categorization that takes into account Spanish texts as examples, there are no tools especially focused on the Spanish language. Moreover, the proposed system has been implemented to be open to allow the possibility to add the analysis of other languages, like English, French, or Portuguese.

Other important characteristics of the architecture is that it has been proposed as a pipeline system and it has been implemented with different modules. We consider these as important features because a pipeline system gives us the chance to control the results at each phase of the process and also the structure with different modules allows us to easily upgrade its individual components. For example, geographic or lexical databases change over time, and our modular architecture easily accommodates these changes. The fact that the system is implemented in different modules is also interesting because it is ideal when performing the analysis of a text. Sometimes, we may want not to have to use all the modules that make up the architecture to achieve a desired result. For example, we may want to extract only statistical information from the words present in a text, but nothing related about their semantics. Also, it is possible that we need to change the order of the modules a text passes through depending on the type of analysis of the text we want to perform. For these reasons it is important to consider a modular architecture: it makes the system easy to use and it facilitates improving it over time.

This paper provides two main contributions: Firstly, we present a tool called GENIE, whose general architecture is valid for text categorisation tasks in any language. This system has been installed and tested in several real environments using different datasets. The set-up of our algorithm is rule-based and we use for inference the document's features as well as the linguistic content of the text and its meaning. Secondly, we experimentally quantify the influence of using linguistic and semantic tools when performing the automatic classification, working on a real case with Spanish texts previously classified by a professional documentation department.

The system has been also used for classifying and labeling documents in different scenarios such as supporting query expansion systems [1] or recommendation algorithms [2]. In both papers we used the GENIE system with very satisfactory results.

The rest of this paper is structured as follows. Section 2 explains the general architecture of the proposed categorisation system. Section 3 discusses the results of our experiments with real data. Section 4 analyzes other related works. Finally, Section 5 provides our conclusions and future work.

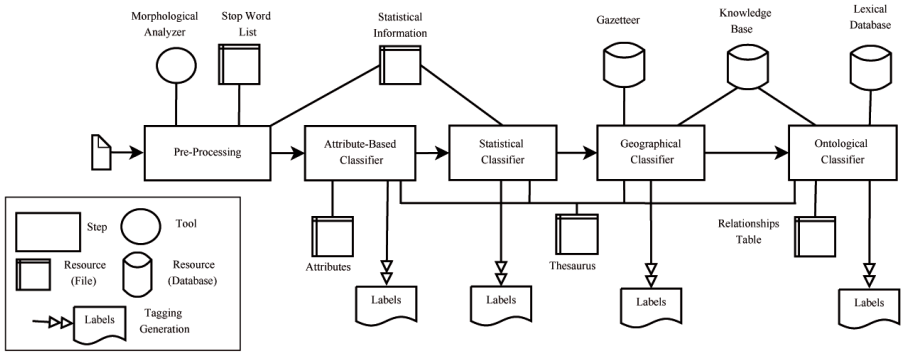

Fig. 1. General pipeline of GENIE, the proposed text categorisation system.

2 PROPOSED ARCHITECTURE

In this section, we explain the general architecture of the proposed system as well as and the corresponding working methodology. The system relies on the existence of several resources. First, we will describe these resources, and then we will explain in detail the classification process (see Figure 1).

2.1 Resources

Regarding resources, we have to consider both static data repositories and software tools:

- *Thesaurus.* A thesaurus is a list of words and a set of relations among them, used to classify items. We use its elements as the set of tags that must be used to categorize the set of documents. Examples of thesaurus entries are words like HEALTH, ACCIDENT, FOOTBALL, BASKETBALL, REAL-MADRID, CINEMA, JACK_NICHOLSON, THEATER, etc. The terms can be related. For example, FOOTBALL and BASKETBALL could depend hierarchically on SPORTS. Each document may take a variable number of terms in the thesaurus during the categorisation process.
- *Gazetteer.* It is a geographic directory containing information about places and place names [3]. In GENIE, it is used to identify geographic features.
- *Morphological Analyzer.* It is an NLP tool whose mission is the identification, analysis and description of the structure of a set of given linguistic units. This analyzer consists of a set of different analysis libraries, which can be configured and used depending on the working language, and a custom middle-ware architecture which aims to store all the different analysis results in structures that represent the desired linguistic units, such as words, sentences and texts. With this approach we can provide the same entities

to the other modules that work with NLP, resulting in an architecture that can work with multiple analysis tools and languages.
- *Lexical Database.* A lexical database is a lexical resource which groups words into sets of synonyms called *synsets*, including semantic relations among them. Examples could be *WordNet* [4] and *EurowordNet* [5].
- *Stop Word List.* This is a list of frequent words that do not contain relevant semantic information. In this set we may include the following types of words: articles, conjunctions, numbers, etc.
- *Knowledge Base.* This refers to the explicit representation of knowledge related to the topics covered in the documents that have to be catalogued. As a tool for knowledge representation in a software system, we use ontologies. The idea is to represent in these ontologies the concepts that could help to label a document in a given context, and to populate the ontologies with as many instances as possible.
- *Statistical Information.* This consists of a set of files with information about the use frequency of each word, related to the attributes of the text and to the set of elements in the thesaurus. For example: the word "ONU" appears more frequently in documents of type "International" and it is related with the descriptor INTERNAT in a thesaurus used in the documentation department of a newspaper we have worked with. These frequencies allow us to estimate if a given text can be categorized with a particular element of the thesaurus.
- *Relationships Table.* This table relates items in the Gazetteer and knowledge base concepts with thesaurus elements. It may be necessary in an organization because the concepts stored in the semantic resources available may not match the labels in the thesaurus that must be considered for classification. The construction of this table could be manual or automatic, using any machine learning method.

As we will show in the experimental evaluation, the use of some resources is optional, leading to different results in terms of the expected performance of the system. This system could be used with different languages by changing the language-dependent resources, i.e. the Gazetteer, the NLP tool, the lexical database, and the stop word list.

2.2 Process Pipeline

We have used a pipeline scheme with separated stages. Each of the stages is associated with only one type of process and they communicate between themselves through different files. Although it is a pipeline system, the process can be configured so that each of the tasks can be activated or deactivated depending on whether we want the text document to go through certain phases or not. This choice has three purposes: Firstly, the early stages perform a more general classification, and later phases make more specific labeling that requires more precise resources. We have verified, through experimental evaluation, that taking advantage of a filter to select the most appropriate resources for the later stages improves the results. Secondly, separating each stage simplifies control for

evaluation. We know that there are certain tasks that could be parallelized, but the aim is to analyze the results in the best possible way, rather than to provide an optimized algorithm. Finally, we have more freedom to add, delete or modify any of the stages of the pipeline if they are independent. If we would like to use a different tool in any of the stages, changing it is very easy when there is a minimum coupling between phases.

Our system works over a set of text documents, but we have to note that each of them could have a variable number of attributes (author, title, subtitle, domain, date, section, type, extension, etc.), that we will use during the categorisation process. These attributes vary according to the origin of the document: a digital library, a database, a website, etc. Numeric fields, dates, strings, or even HTML tags may be perfectly valid attributes to the system. As a very first stage, the system includes specific interfaces to convert the original documents into XML files with a specific field for plain text and others for attributes.

The tasks for the proposed automatic text categorisation system are:

1. *Preprocessing* of the text of the document, which consists of three steps:
   (a) *Lemmatization.* Through this process we obtain a new text consisting of a set of words corresponding to the lemmas (canonical forms) of the words in the initial text. This process eliminates prepositions, articles, conjunctions and other words included in the *Stop Words List.* All the word information (Part of Speech, gender, number) is stored in the corresponding structure, so it can be recovered later for future uses.
   (b) *Named Entities Recognition (NER).* Named entities are atomic elements in a text representing, for example, names of persons, organizations or locations [6]. By using a named entity extractor, this procedure gets a list of items identified as named entities. This extractor can be paired with a statistical Named Entity Classification (NEC) in a first attempt to classify the named entity into a pre-defined group (person, place, organization) or leave it undefined so the following tasks (Geographical Classifier) can disambiguate it.
   (c) *Keywords Extraction.* Keywords are words selected from the text that are in fact key elements to consider to categorize the document. We use the lemmatized form of such words and the TF/IDF algorithm [7].

   These processes produce several results that are used in subsequent stages. The resources used in this stage are the morphological analyzer, the Stop Word List and the statistical data.
2. *Attributes-Based Classifier.* Taking advantage of the attributes of each of the documents, this ruled-based process makes a first basic and general tagging. For example, if we find the words "film review" in the "title" field the system will infer that the thesaurus descriptor CINEMA could be assigned to this document. At the same time, it establishes the values of the attributes to be used for the selection of appropriate resources in the following steps, choosing for instance an ontology about cinema for the Ontological Classifier stage.
3. *Statistical Classifier.* Using machine learning techniques [8], the document text is analyzed to try to find patterns that correspond to data in the files

storing statistical information. This step is mainly useful to try to obtain labels that correspond to the general themes of the document. Trying to deduce if a document is talking about football or basketball could be a good example.

4. *Geographical Classifier*. By using the gazetteer, named entities (NE) corresponding to geographical locations are detected. This stage is managed by a ruled-based system. Besides, it can deal with typical disambiguation problems among locations of the same name and locations whose names match other NE (e.g., people), by using the well-known techniques described in [9]: usually there is only single sense per discourse (so, an ambiguous term is likely to mean only one of its senses when it its used multiple times), and place names appearing in the same context tend to show close locations. Other important considerations that GENIE takes into account are to look at the population of the location candidates as an important aspect to disambiguate places [9] and consider the context where the text is framed to establish a list of bonuses for certain regions [10]. Other used techniques are to construct an N-term window on both sides of the entity considered to be a geographic term, as some words can contribute with a positive or negative modifier [11], or to try to find syntactic structures like "city, country" (e.g. "Madrid, Spain") [12]. Finally, using techniques explained in [13], the system uses ontologies in order to capture information about important aspects related to certain locations. For example: most important streets, monuments and outstanding buildings, neighborhoods, etc. This is useful when a text has not explicit location identified. Besides, it takes advantage too of the results of previous stages. For example, if in the previous stages we got the descriptor EUROPE we can assign higher scores to the results related to European countries and major European cities than to results related to locations in other continents. The geographical tagging unit is very useful because, empirically, near 30% of tags in our experimental context are related to locations.
5. *Ontological Classifier*. To perform a detailed labeling, the knowledge base is queried about the named entities and keywords found in the text. If a positive response is obtained, it means that the main related concepts can be used to label the text. A great advantage is that these concepts need not appear explicitly in the text, as they may be inferred from existing ontological relations. If there is an ambiguous word, it can be disambiguated [14] by using the Lexical Database resource (for a survey on word sense disambiguation, see [15]). As soon as a concept related to the text is found, the relations stored in the *Relationships Table* are considered to obtain appropriate tags from the thesaurus. As explained before, the fact that at this phase we have a partially classified document allows us to choose the most appropriate ontologies for classification using configurable rules. For example, if we have already realised with the statistical classifier that the text speaks of the American Basketball League, we will use a specific ontology to classify the document more accurately finding out for instance the teams and the players, and we will not try to use any other resource. This particular

ontology could be obtained and re-used from the Web. This ontology would probably be hand-made, or it would be adapted from other similar ontology, because this kind of resources are difficult or impossible to find for free on the Web. So, our system is generic enough to accommodate the required and more appropriate ontologies (existing or hand-made) for the different topics covered in the texts.

The way to obtain the tags is asking about keywords and NE to the ontology by using SPARQL[1], a set of rules, and the relationship table to deduce the most suitable tags. The behaviour of the ontology is not only to be a simple *bag-of-words*, because it can contain concepts, relations and axioms, all of them very useful to inquire the implicit topics in the text.

In summary, the text categorization process that GENIE performs consists of following each of the proposed tasks that constitute the system's pipeline. This process begins with the preprocessing of the input text, which implies labours of lemmatization of the text and extraction of named entities and keywords from the text. Then it analyzes a set of attributes that are given with the text that is being analyzed in order to extract the first basic and general labels. Afterwards, it applies a statistical classification method based on machine learning techniques to obtain labels that correspond to the general themes of the document. Then it applies a geographic classifier for the purpose of identifying possible geographical references included in the text. Finally, it applies an ontological classifier in order to carry out a more detailed classification of the text, which performs an analysis of named entities and keywords obtained from the text, consults the appropriate ontology, and uses a lexical database to remove possible ambiguities.

3 EXPERIMENTAL EVALUATION

We have performed a set of experiments to test and compare the performance of our architecture with others tools. For this purpose, we have tested in a real environment using three corpus of news previously labeled by a professional documentation department of several major Spanish Media: *Heraldo de Aragón*[2], *Diario de Navarra*[3] and *Heraldo de Soria*[4]. Each corpus had respectively 11,275, 10,200, and 4,500 news. These corpora are divided in several categories: local, national, international, sports, and culture. Every media has a different professional thesaurus used to classify documents, with more than 10,000 entries each. For classification, each document can receive any number of descriptors belonging to the thesaurus. The ideal situation would be that the automatic text categorization system could perform a work identical to the one performed by the real documentation departments.

[1] http://www.w3.org/TR/2006/WD-rdf-sparql-query-20061004/
[2] http://www.heraldo.es/
[3] http://www.diariodenavarra.es/
[4] http://www.heraldodesoria.es/

These news are stored in several databases, in tables where different fields are used to store the different attributes explained in Section 2 (title, author, date, section, type, extension, etc.). For experimental evaluation, we have extracted them from the databases and we have put each text and the data of its fields in XML files. We have used this corpus of XML files as the input of the system, and the output is the same set of files but with an additional field: classification information. This new XML node contains the set of words (descriptors) belonging to the thesaurus used to categorize the document, i.e., this node contains the different tags that describe the XML file. As the news in the dataset considered had been previously manually annotated by the professionals working in the documentation department, we can compare the automatic categorization with that performed by humans. So, we can evaluate the number of hits, omissions and misses.

3.1 Experimental Settings

In the experiments, we have examined the following measures, commonly used in the Information Retrieval context: the *precision*, the *recall*, and the *F-Measure*. The dataset used initially in the experiments has been the Heraldo de Aragón corpus. We have used the information from this dataset to define most of the rules of the various processes associated with each of the stages of the classification system. These rules are integrated in a configuration file which contains all the information necessary to lead the process and obtain the correct result. The other two datasets (Diario de Navarra and Heraldo de Soria) have been used just to double-check if the application of those rules also produced the desired result; for comparison purposes, at the end of this section we will also present some experimental results based on them.

We have performed four experiments with the GENIE system. Each stage of the pipeline can be enabled or disabled separately. Regarding the resources and tools considered, we have used Freeling[5], as the Morphological Analyzer and Support Vector Machines (SVM) [16] to automatically classify topics in the Statistical Classifier. To obtain the frequencies we have used a different corpus of 100,000 news, in order to get a realistic frequency information. Finally, we have chosen Eurowordnet as the Lexical Database and *Geonames*[6] as the Gazetteer.

To train this Statistical Classifier we have used sets of 5,000 news for each general theme associated to one descriptor (FOOTBALL, BASKET, CINEMA, HANDBALL, and MUSIC). These sets of news are different from the datasets used in the experiments (as is obviously expected in a training phase). For each possible descriptor, we have an ontology, in this case we have designed five ontologies using OWL [17] with near a hundred concepts each one.
Next, there is an example of a piece of news:

[5] http://nlp.lsi.upc.edu/freeling/
[6] http://www.geonames.org/

"This weekend is the best film debut for the movie "In the Valley of Elah". The story revolves around the murder of a young man who has just returned from the Iraq war, whose parents try to clarify with the help of the police. As interpreters we have: Tommy Lee Jones, Susan Sarandon and Charlize Theron. Writer-director Paul Haggis is the author of "Crash" and writer of "Million Dollar Baby", among others."

In this case, the system analyzes and classifies the text with the descriptor CINEMA. Moreover, the news can be tagged with tags such as C_THERON, IRAQ, TL_JONES, etc.

3.2 EXPERIMENTAL RESULTS

We have compared our classification of the 11,275 news in the first dataset with the original classification made by professionals. The results can be seen in Figure 2. Below we analyze the experiments:

1. In the first experiment (*Basic*) we have used the process presented in Section 2 without the Pre-Processing step and without the Ontological Classifier. We have trained the system with SVM to classify 100 themes. In this case, as we do not use the steps of Pre-Processing and the Ontological Classifier, the system has not performed the lemmatization, the named entities recognition, the keywords extraction, and the detailed labeling of the text. For this reason, the precision and the recall are not good, as it is essential to embed semantic information and conceptual patterns in order to enhance the prediction capabilities of classification algorithms.
2. In the second one (*Semantic*), we have introduced the complete Pre-Processing stage and its associated resources, we have used the Lexical Database EuroWordNet [5] to disambiguate keywords, and we have introduced the Ontological Classifier, with five ontologies with about ten concepts and about 20 instances each. In this experiment the precision and the recall slightly improved because, as explained before, the step of Pre-Processing is important to obtain a better classification.
3. In the third one (*Sem + Geo*) we have included the Geographical Classifier but we have used only the Gazzetteer resource. Here we have improved the recall of the labeling but in exchange of a decrease in the precision. By analyzing the errors in detail, we observe that the main cause is the presence of misclassifications performed by the Geographical Classifier.
4. Finally, in the fourth experiment (*Full Mode*), we have executed all the pipeline, exploited all the resources and populated the ontologies with about one hundred instances, leading to an increase in both the precision and the recall. Ontology instances added in this experiment have been inferred from the observation of the errors obtained in previous experiments. The motivation to add them is that otherwise the text includes certain entities unknown to the system, and when they were incorporated this helped to improve the classification.

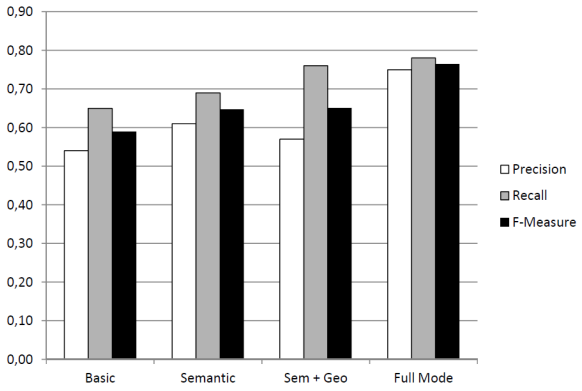

Fig. 2. Results of the four document categorisation experiments with news in the dataset 1.

If we look at the overall results obtained in the experiment 1 and the experiment 2 in the Figure 2, we could say that the influence of using semantic and NLP tools is apparently not so significant (about 20%). However, it seems clear that these tools significantly improve the quality of labeling in terms of precision, recall and F-measure, reaching up to about the 80%. Therefore, the use of semantic techniques can make a difference when deciding about the possibility to perform an automatic labeling.

After evaluating the results obtained in the reference dataset (Heraldo de Aragón), we repeated the same experiments with the two other datasets. These dataset were not considered while designing the ontologies, in order to maintain the independence of the tests. The results can be seen in Figure 3. The results obtained with datasets different from the one used for Heraldo de Aragón, which was used to configure the rule-based system, are only slightly different (differences smaller than 10%). In Figure 3, it can also be seen that the trends of the results are very similar regardless of the data. This shows the generality of our approach, since the behaviour of the classification system has been reproduced with several different corpora. Experimental results have shown that with our approach, in all the experiments, the system has improved the results achieved by basic machine learning based systems.

4 RELATED WORK

Text categorisation represents a challenging problem for the data mining and machine learning communities, due to the growing demand for automatic information retrieval systems. Systems that automatically classify text documents into predefined thematic classes, and thereby contextualize information, offer a promising approach to tackle this complexity [8].

Document classification presents difficult challenges due to the sparsity and the high dimensionality of text data, and to the complex semantics of natural language. The traditional document representation is a word-based vector where each dimension is associated with a term of a dictionary containing all the words that appear in the corpus. The value associated to a given term reflects its frequency of occurrence within the corresponding document and within the entire corpus (the *tf-idf* metric). Although this is a representation that is simple and commonly used, it has several limitations. Specifically, this technique has three main drawbacks: (1) it breaks multi-word expressions into independent features; (2) it maps synonymous words into different components; and (3) it considers polysemous words as one single component. While a traditional preprocessing of documents, such as eliminating stop words, pruning rare words, stemming, and normalization, can improve the representation, its effect is also still limited. So, it is essential to embed semantic information and conceptual patterns in order to enhance the prediction capabilities of classification algorithms.

Research has been done to exploit ontologies for content-based categorisation of large corpora of documents. WordNet has been widely used, but their approaches only use synonyms and hyponyms, fail to handle polysemy, and break multi-word concepts into single terms. Our approach overcomes these limitations by incorporating background knowledge derived from ontologies. This methodology is able to keep multi-word concepts unbroken, it captures the semantic closeness to synonyms, and performs word sense disambiguation for polysemous terms.

For disambiguation tasks we have taken into account an approximation described in [18], that is based on a semantic relatedness computation to detect the set of words that could induce an effective disambiguation. That technique receives an ambiguous keyword and its context words as input and provides a list of possible senses. Other studies show how background knowledge in form of simple ontologies can improve text classification results by directly addressing these problems [19], and others make use of this intelligence to automatically generate tag suggestions based on the semantic content of texts. For example [20], which extracts keywords and their frequencies, uses WordNet as semantics and an artificial neural network for learning.

Among other related studies that quantify the quality of an automatic labeling performed by using ontologies, we could mention [21], but it was focused on a purely semantic labeling. More related to our study, it is interesting to mention the work presented in [22], although it does not include much information about the use of ontologies. Examples of hybrid systems using both types of tools include the web service classifier explained in [23], the system *NASS (News Annotation Semantic System)* described in [24, 25], which is an automatic annotation tool for the Media, or *GoNTogle* [26], which is a framework for general document annotation and retrieval.

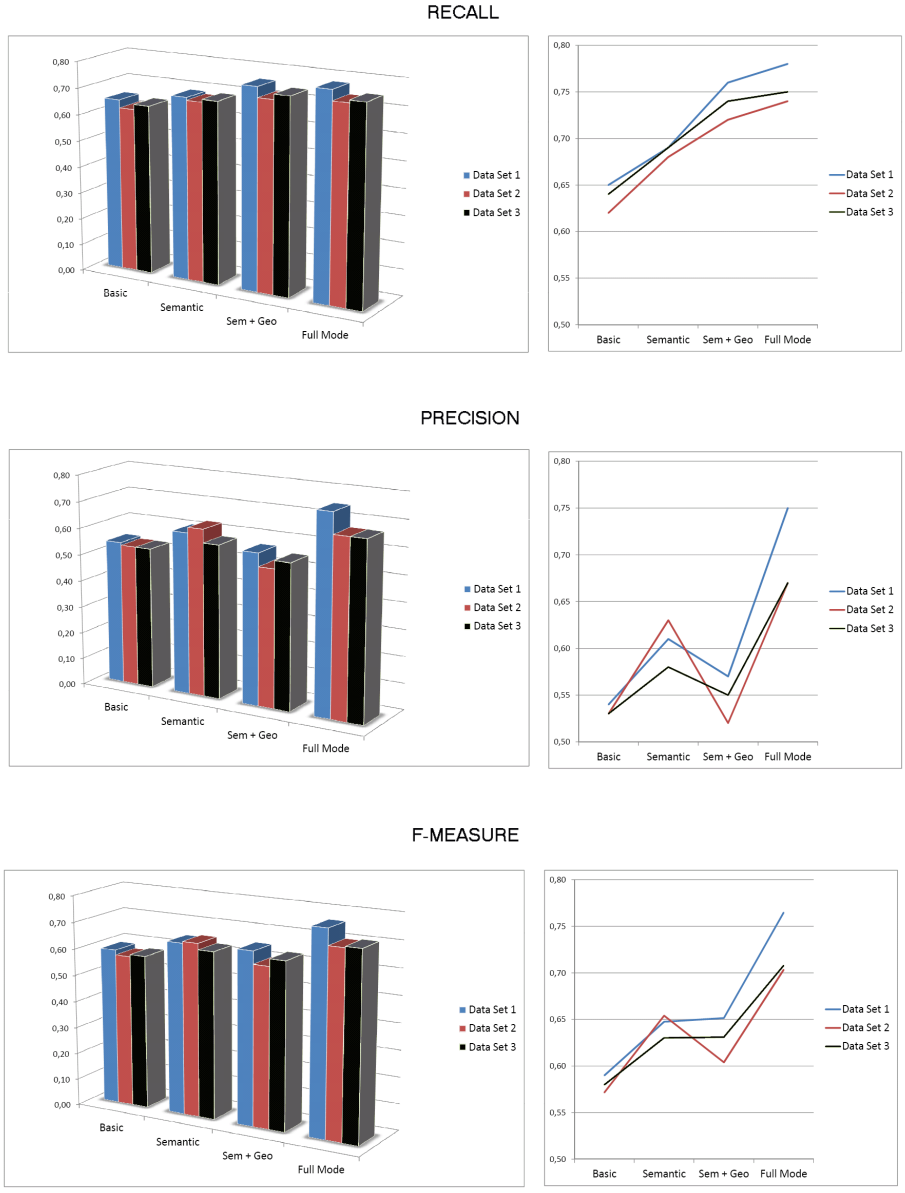

Fig. 3. Comparative results of the automatic categorisation experiments.

5 CONCLUSIONS AND FUTURE WORK

A tool for automating categorisation tasks is very useful nowadays, as it helps to improve the quality of searches that are performed later over textual repositories like digital libraries, databases or web pages. For this reason, in this paper we have presented a pipeline architecture to help in the study of the problem of automatic text categorisation using specific vocabulary contained in a thesaurus. Our main contribution is the design of a system that combines statistics, lexical databases, NLP tools, ontologies, and geographical databases. Its stage-based architecture easily allows the use and exchange of different resources and tools. We have also performed a deep study of the impact of the semantics in a text categorisation process.

Our pipeline architecture is based on five stages: preprocessing, attribute-based classification, statistical classification, geographical classification, and ontological classification. Although the experimental part has been developed in Spanish, the tool is ready to work with any other language. Changing linguistic resources suitable for the intended language is enough to make the system work, since the process is sufficiently general to be applicable regardless of the language used. The main contribution of our work is, apart from the useful and modular pipeline architecture, the experimental study with real data of the problem of categorization of natural language documents written in Spanish. There are many studies related to such problems in English, but it is difficult to find them in Spanish. Besides, we have compared the impact of applying techniques that rely on statistics and supervised learning with the results obtained when semantic techniques are also used. There are two remarkable aspects. Firstly, enhancing the amount of knowledge available by increasing the number of instances in the ontologies leads to a substantial improvement in the results. Secondly, the use of knowledge bases helps to correct many errors from a Geographical Classifier.

Spanish vs. English language. Our research on this topic focuses on transfer projects related to the extraction of information, so for us it is very important to work with real cases. Therefore, the comparison of our work with typical benchmark data sets in English is not fundamental to us, since they are not useful to improve the performance of our system in Spanish, and we have seen that the ambient conditions (language, regional context, thematic news, etc.) have a great influence on the outcome of experiments. Many researchers have already analyzed the differences between working in NLP topics in English and in Spanish, and they have made it clear the additional difficulties of the Spanish Language [27, 28], which could explain the poor performance of some software applications that work reasonably well in English. Just to mention some of these differences: in Spanish words contain much more grammatical and semantic information than the English words, the subject can be omitted in many cases, and verbs forms carry implicit conjugation, without additional words. That, coupled with the high number of meanings that the same word can have, increases the computational complexity for syntactic, semantic and morphological analyzers, which so behave differently in Spanish and English. Spanish is the third language

in the world according to the number of speakers, after Mandarin and English, but in terms of studies related to NLP we have not found many scientific papers.

Impact of NLP and semantics. Our experimental evaluation suggests that the influence of NLP and semantic tools is not quantitatively as important as the classic statistical approaches, although their contribution can tip the scales when evaluating the quality of a labeling technique, since the difference in terms of precision and recall is sufficiently influential (near 20%). So, our conclusion is that a statistical approach can be successfully complemented with semantic techniques to obtain an acceptable automatic categorisation. Our experience also proves that facing this issue in a real environment when professional results are needed, the typical machine learning approach is the best option but is not always enough. We have seen that it should be complemented with other techniques, in our case semantic and linguistic ones. Anyway, the main drawback of the semantic techniques is that the work of searching or constructing the ontologies for each set of tags of every topic, populating them, and building the relationship tables, is harder than the typical training of the machine learning approaches. So, although the results are better, the scalability could be problematic. Sometimes it can be quite costly, especially if detailed knowledge of the topic to tag is required in order to appropriately configure the system.

NLP future tasks In some categorisation scenarios, like bigger analysis (novels, reports, etc.) or groups of documents of the same field, it can be interesting to obtain a summary of the given inputs in order to categorise them with their general terms before entering a more detailed analysis which requires the entire texts. These summaries, alongside with the previous defined tasks, can led to a more suitable detailed labeling, providing hints of which knowledge bases might be interesting to work with. In order to achieve this, we can perform syntactic analysis to simplify the sentences of the summaries, as we have seen in works like [29], and then we will use the obtained results to filter unnecessary information and select the most relevant sentences without compromising the text integrity. Although the required structures have been implemented and some approaches as [30] are being designed and tested, they are into an early stage and they require more work before trying to use it inside the categorisation pipeline.

Open tasks. As future work, we plan to increase the number of methods used in the pipeline, and to test this methodology in new contexts and languages. It is noteworthy that a piece of news is a very specific type of text, characterized by objectivity, clarity, and the use of synonyms and acronyms, the high presence of specific and descriptive adjectives, the tendency to use impersonal or passive constructions, and the use of connectors. Therefore it is not sufficient to test only with this kind of text, and to make a more complete study it is necessary to work with other types. In fact, some tests have been made with GENIE with other types of documents very different from news, such as book reviews, business reports, lyrics, blogs, etc., and the results are very promising, but it is early to assert the generality of the solution in different contexts because the studies are still in progress.

ACKNOWLEDGEMENTS

This research work has been supported by the CICYT project TIN2013-46238-C4-4-R and DGA-FSE. Thank you to Heraldo Group and Diario de Navarra.

References

1. M. G. Buey, A. L. Garrido, S. Escudero, R. Trillo, S. Ilarri, and E. Mena, "SQX-Lib: Developing a semantic query expansion system in a media group," in *European Conference on Information Retrieval*, pp. 780–784, 2014.
2. A. L. Garrido, M. S. Pera, and S. Ilarri, "SOLE-R, a Semantic and Linguistic Aproach for Book Recommendations," in *14th IEEE International Conference on Advanced Learning Technologies - ICALT*, pp. 524–528, IEEE Computer Society, 2014.
3. M. F. Goodchild and L. Hill, "Introduction to digital gazetteer research," *International Journal of Geographical Information Science*, vol. 22, no. 10, pp. 1039–1044, 2008.
4. G. A. Miller, "WordNet: a lexical database for english," *Communications of ACM*, vol. 38, no. 11, pp. 39–41, 1995.
5. P. Vossen, *EuroWordNet: A multilingual database with lexical semantic networks*. Kluwer Academic Boston, 1998.
6. S. Sekine and E. Ranchhod, *Named Entities: Recognition, Classification and Use*. John Benjamins, 2009.
7. G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management*, vol. 24, no. 5, pp. 513–523, 1988.
8. F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.
9. E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 273–280, ACM, 2004.
10. G. Quercini, H. Samet, J. Sankaranarayanan, and M. D. Lieberman, "Determining the spatial reader scopes of news sources using local lexicons," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 43–52, ACM, 2010.
11. E. Rauch, M. Bukatin, and K. Baker, "A confidence-based framework for disambiguating geographic terms," in *HLT-NAACL 2003 Workshop on Analysis of Geographic References, vol. 1*, pp. 50–54, Association for Computational Linguistics, 2003.
12. H. Li, R. K. Srihari, C. Niu, and W. Li, "Location normalization for information extraction," in *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pp. 1–7, Association for Computational Linguistics, 2002.
13. A. L. Garrido, M. G. Buey, S. Ilarri, and E. Mena, "GEO-NASS: A semantic tagging experience from geographical data on the media," in *17th East-European Conference on Advances in Databases and Information Systems (ADBIS 2013), Genoa (Italy)*, vol. 8133, pp. 56–69, Springer, September 2013.
14. P. Resnik, "Disambiguating noun groupings with respect to WordNet senses," in *Natural Language Processing Using Very Large Corpora*, pp. 77–98, Springer, 1999.

15. R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, no. 2, pp. 10:1–10:69, 2009.
16. T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Tenth European Conference on Machine Learning (ECML'98)*, pp. 137–142, Springer, 1998.
17. D. L. McGuinness, F. Van Harmelen, *et al.*, "OWL web ontology language overview," *W3C recommendation 10 February 2004*, 2004.
18. R. Trillo, J. Gracia, M. Espinoza, and E. Mena, "Discovering the semantics of user keywords," *Journal of Universal Computer Science*, vol. 13, pp. 1908–1935, dec 2007.
19. S. Bloehdorn and A. Hotho, "Boosting for text classification with semantic features," in *Advances in Web mining and Web Usage Analysis*, pp. 149–166, Springer, 2006.
20. S. O. K. Lee and A. H. W. Chun, "Automatic tag recommendation for the web 2.0 blogosphere using collaborative tagging and hybrid and semantic structures," *Sixth Conference on WSEAS International Conference on Applied Computer Science (ACOS'07), World Scientific and Engineering Academy and Society (WSEAS)*, vol. 7, pp. 88–93, 2007.
21. D. Maynard, W. Peters, and Y. Li, "Metrics for evaluation of ontology-based information extraction," in *Workshop on Evaluation of Ontologies for the Web (EON) at the International World Wide Web Conference (WWW'06)*, 2006.
22. M. Scharkow, "Thematic content analysis using supervised machine learning: An empirical evaluation using German online news," *Quality and Quantity*, vol. 47, no. 2, pp. 761–773, 2013.
23. M. Bruno, G. Canfora, M. Di Penta, and R. Scognamiglio, "An approach to support web service classification and annotation," in *2005 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05)*, pp. 138–143, IEEE, 2005.
24. A. L. Garrido, O. Gomez, S. Ilarri, and E. Mena, "NASS: News Annotation Semantic System," in *23rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2011), Boca Raton, Florida (USA)*, pp. 904–905, IEEE Computer Society, November 2011.
25. A. L. Garrido, O. Gomez, S. Ilarri, and E. Mena, "An experience developing a semantic annotation system in a media group," in *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, pp. 333–338, Springer, 2012.
26. N. Bikakis, G. Giannopoulos, T. Dalamagas, and T. Sellis, "Integrating keywords and semantics on document annotation and search," in *On the Move to Meaningful Internet Systems (OTM 2010)*, pp. 921–938, Springer, 2010.
27. R. Carrasco and A. Gelbukh, "Evaluation of TnT Tagger for Spanish," in *Proceedings of ENC, Fourth Mexican International Conference on Computer Science*, pp. 18–25, IEEE, 2003.
28. G. Aguado de Cea, J. Puch, and J. Ramos, "Tagging Spanish texts: The problem of 'se'," in *Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pp. 2321–2324, 2008.
29. S. B. Silveira and A. Branco, "Extracting multi-document summaries with a double clustering approach," in *Natural Language Processing and Information Systems*, pp. 70–81, Springer, 2012.
30. A. L. Garrido, M. G. Buey, S. Escudero, S. Ilarri, E. Mena, and S. B. Silveira, "TM-gen: A topic map generator from text documents," in *25th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2013), Washington DC (USA)*, pp. 735–740, IEEE Computer Society, 2013.