IADIS International Journal on Computer Science and Information Systems
Vol. 4, No. 2, pp. 73-85
ISSN: 1646-3692

# A NOVEL SEMANTIC APPROACH TO DOCUMENT COLLECTIONS

Andrea Addis. *Department of Electrical and Electronic Engineering, University of Cagliari. P.za d'Armi, I-09123, Cagliari, Italy.*

*addis@diee.unica.it*

Manuela Angioni. *CRS4, Center of Advanced Studies, Research and Development in Sardinia, Parco Scientifico e Tecnologico, Ed.1 I-09010 Pula (CA), Italy.*

*angioni@cras4.it*

Giuliano Armano. *Department of Electrical and Electronic Engineering, University of Cagliari. P.za d'Armi, I-09123, Cagliari, Italy.*

*armano@diee.unica.it*

Roberto Demontis. *CRS4, Center of Advanced Studies, Research and Development in Sardinia, Parco Scientifico e Tecnologico, Ed.1 I-09010 Pula (CA), Italy.*

*demontis@crs4.it*

Franco Tuveri. *CRS4, Center of Advanced Studies, Research and Development in Sardinia, Parco Scientifico e Tecnologico, Ed.1 I-09010 Pula (CA), Italy.*

*tuveri@crs4.it*

Eloisa Vargiu. *Department of Electrical and Electronic Engineering, University of Cagliari. P.za d'Armi, I-09123, Cagliari, Italy.*

*vargiu@diee.unica.it*

## ABSTRACT

Available document collections are more and more required for supervised text categorization tasks. They are typically collections of documents classified by domain engineers. In this paper, we propose a semantic text categorization approach able to automatically create document collections in which documents are classified according to WordNet Domains taxonomy. Experiments have been performed by training a classifier with an automatic document collection and comparing results with those obtained by training the same classifier with a document collection classified by domain engineers. Experimental results point out that, on average, the performances of the automatic approach are quite similar to those obtained on a document collection classified by hand.

**KEYWORDS**

Text Categorization, Document Collections, Intelligent Software Systems, Machine Learning.

# 1. INTRODUCTION

Text categorization can be defined as the task of determining and assigning topical labels to content. The more the amount of available data (e.g., in digital libraries), the greater the need for high-performance text categorization algorithms. In particular, text categorization is a key technology in several information processing tasks, including controlled vocabulary indexing, routing and packaging of news and other text streams, content filtering, information security, and help desk automation.

In the literature, many machine learning approaches have been proposed, both in the field of supervised [Sebastiani02] and unsupervised [Ghahramani04] learning. Supervised approaches use only labeled data during the training phase (easy to use but difficult to collect). Unsupervised approaches use unlabeled data (easy to collect but difficult to use). Semi-supervised learning, which stands in between, tries to solve this problem by using large amount of unlabeled data, together with labeled data and/or with user feedback, to build better classifiers [Zhu05].

As for unsupervised text categorization, only few works have been proposed in the literature [Sahami96]. In this scenario, available labeled document collections are more and more required. They are typically standard collections to which humans have assigned categories from a predefined set [Lewis96, Yang99, Lewis04], so that researchers are able to test their algorithms in a controlled benchmarking setting. Unfortunately, existing document collections suffer from one or more of the following drawbacks: (i) few documents, (ii) lack of the document full text, (iii) inconsistent or incomplete category assignment, (iv) peculiar textual properties, and (v) limited availability. Moreover, often researchers do not have documentation on how collections were produced, and on the nature of the underlying categories.

So far, only few attempts to automatically create document collections have been proposed [Ko00] [Kohonen00]. In particular, semantic approaches to text categorization are rarely applied to this specific issue. Many works attempted to address this task by incorporating semantic information into document representation. In [Zesch06] a corpus-based system for the automatic creation of test datasets has been proposed, which annotates pairs of similar words in terms of semantic relatedness. Other researchers studied techniques to extract semantic information calculating probabilities between topic signatures and single words, using semantic relatedness [Achananuparp08]. All the corresponding systems must work on organized and predefined corpora of documents.

The approach proposed in this paper differs from those mentioned above in the fact that it is aimed at creating document collections from generic sources of documents by adopting a fully-automated semantic approach. Each text document is suitably labeled according to

WordNet Domains, a predefined taxonomy of classes [Magnini00]. Experimental results point out that the proposed method allows to create reliable document collections.

This work is part of DART, Distributed Agent-Based Retrieval Toolkit, a research project aimed at studying, developing and testing patterns and integrated tools to achieve a semantic, distributed geo-sensible search engine [Angioni07a, Angioni07b].

The remainder of the paper is organized as follows: Section 2 illustrates our semantic approach to text categorization. Section 3 reports and discusses experimental results. Finally, Section 4 draws conclusions.

## 2. A SEMANTIC APPROACH FOR BUILDING DOCUMENT COLLECTIONS

One main goal of the DART project was to categorize a great number of web resources, different in content and type, that are not always mapped on taxonomies or ontologies. In fact, the web is a vast collection of heterogeneous documents: they can differ in type and format (e.g. text, HTML, PDF, audio, images, web services), in language, vocabulary and may even be automatically generated (e.g. log files or output from a database). Hence, it is not always easy to map generic resources to taxonomies or to find a corpus of tagged resources to categorize them by means of traditional categorization techniques. To address all these problems, we adopted a semantic approach to develop a categorizer, i.e., a module able to manage resources and queries exploiting semantic text categorization techniques.

The proposed approach consists of mapping words of WordNet [Miller98] to pages of Wikipedia in order to obtain a classification of the contents. We adopted the classification given by WordNet Domains [Magnini02, Magnini04] to the word-senses of WordNet where each synset is labeled with one or more domain labels selected from a set of 167 labels, hierarchically organized, extracted from the Dewey Decimal Classification system [Dewey09].

Figure 1 gives an overview of our semantic approach to text analysis. The corresponding system, inspired by the one proposed in [Scott98], performs for each phrase a syntactic and a semantic disambiguation of the textual content of the Wikipedia page and extracts its meanings mapping them to categories. The overall system categorizes the whole document, analyzing it syntactically and semantically, extracting the most important and frequent synsets referred to the real sense expressed in the content, and associating to each synset the categories provided by WordNet Domains.

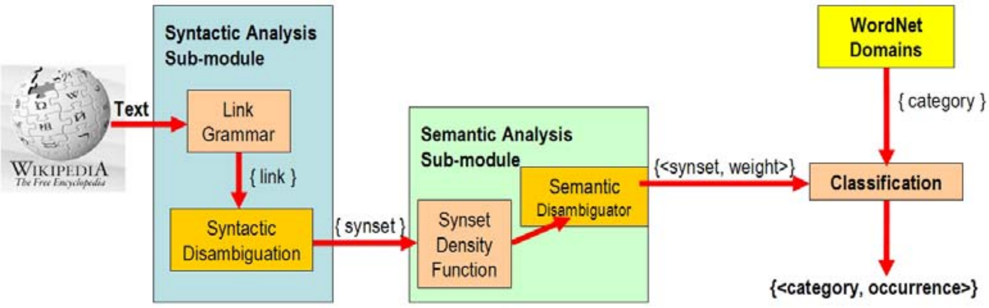
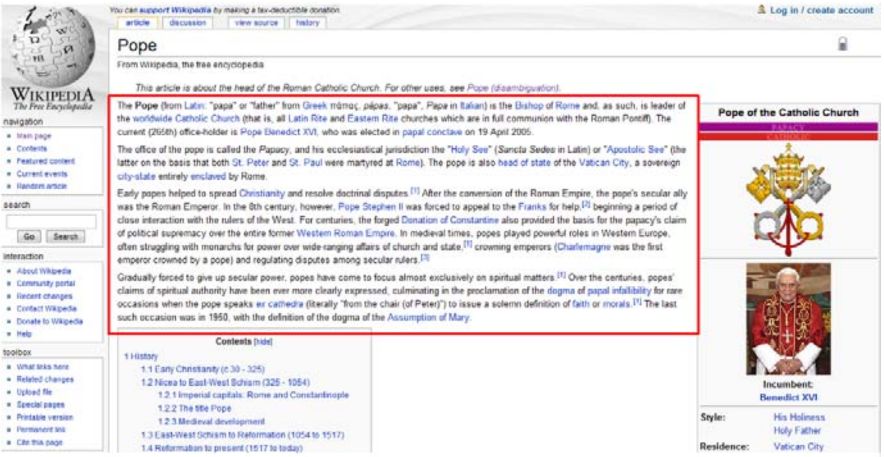


Figure 1. Schema of the semantic classifier

Since we are mainly interested in classifying document belonging to the World Wide Web, to test the capabilities of the proposed approach, we decided to adopt documents extracted by a multidisciplinary, multilingual, web-based, free content document encyclopedia, i.e., Wikipedia. The use of Wikipedia is relevant because pages have a standard structure that permits to isolate the text describing the main concepts of each topic in order to assign few specific categories to each page. An ad hoc parser, based on this characteristic, has been developed. The parser analyzes the text tagged in the Wiki format and extracts the definition of the topic of each page together with a description of the content. Figure 2 shows the selected portion of the Wikipedia page for the term *Pope*. Phrases are then analyzed in order to extract the related links and the plain text that will be passed to the syntactic module as shown in the Figure 1.



Pope

From Wikipedia, the free encyclopedia

*This article is about the head of the Roman Catholic Church. For other uses, see Pope (disambiguation).*

The **Pope** (from Latin: "papa" or "father" from Greek πάπας, *pápas*, "papa", *Papa* in Italian) is the Bishop of Rome and, as such, is leader of the worldwide Catholic Church (that is, all Latin Rite and Eastern Rite churches which are in full communion with the Roman Pontiff). The current (265th) office-holder is Pope Benedict XVI, who was elected in papal conclave on 19 April 2005.

The office of the pope is called the *Papacy*, and his ecclesiastical jurisdiction the "Holy See" (*Sancta Sedes* in Latin) or "Apostolic See" (the latter on the basis that both St. Peter and St. Paul were martyred at Rome). The pope is also head of state of the Vatican City, a sovereign city-state entirely enclaved by Rome.

Early popes helped to spread Christianity and resolve doctrinal disputes.[1] After the conversion of the Roman Empire, the pope's secular ally was the Roman Emperor. In the 8th century, however, Pope Stephen II was forced to appeal to the Franks for help,[2] beginning a period of close interaction with the rulers of the West. For centuries, the forged Donation of Constantine also provided the basis for the papacy's claim of political supremacy over the entire former Western Roman Empire. In medieval times, popes played powerful roles in Western Europe, often struggling with monarchs for power over wide-ranging affairs of church and state,[1] crowning emperors (Charlemagne was the first emperor crowned by a pope) and regulating disputes among secular rulers.[3]

Gradually forced to give up secular power, popes have come to focus almost exclusively on spiritual matters.[1] Over the centuries, popes' claims of spiritual authority have been ever more clearly expressed, culminating in the proclamation of the dogma of papal infallibility for rare occasions when the pope speaks ex cathedra (literally "from the chair (of Peter)") to issue a solemn definition of faith or morals.[1] The last such occasion was in 1950, with the definition of the dogma of the Assumption of Mary.

Contents [hide]

Pope of the Catholic Church

PAPACY

CATHOLIC

Incumbent:
Benedict XVI

| | |
|---|---|
| Style: | His Holiness<br>Holy Father |
| Residence: | Vatican City |

Figure 2. Isolating the text from a Wikipedia page

## 2.1 Syntactic Analysis

The purpose of the syntactic analysis is to determine the structure of sentences, queries or text contained in resources, as well as to perform a syntactic disambiguation in order to overcome problems related to structural ambiguity of words in a text. To address such problems, we use the LinkGrammar [Sleator93] library, a syntactic parser of English that assigns to each sentence a syntactic structure consisting of a set of labeled links connecting words. The approach of LinkGrammar to the analysis of texts differs from other grammatical systems because it analyzes phrases in terms of relationships between pairs of words and not in terms of constituents, like “verb phrases” or syntactic functions (subject or object).

Considering the sentence “The office of the Pope is called the Papacy”, LinkGrammar outputs the following structure:

```
+---------Ss---------+
|      +---Js---+    |            +-----Os----+
+--Ds--+--Mp-+    +--Ds-+     +--Pv--+     +--DG-
+|
the office.n of the Pope.n is.v called.v the
Papacy
```

In particular, it labels *office* and *Pope* as nouns (*.n*) and *is* and *called* as related verbs (*.v*). Furthermore,  in the structure above, *Ss* specifies a link between a subject and a verb, *Ds* connects a determiner to a noun, *Js* connects a preposition to the object, *Os* is a link between a transitive verb and a direct or indirect object, *Mp* is used for prepositional phrases modifying nouns, *Pv* is used to connect forms of  "be” to passive participles,  and *DG* connects a word the with proper nouns.

Like all phrase recognition parsers, LinkGrammar is not able to assign a unique role to a word in a sentence, giving several possible relations among them. Hence, in order to reduce the number of links between words, we apply a weight function that assigns to each link a value proportional to how many times the link appears in possible relations and inversely proportional to the number of these relations. In fact, the goal of the syntactic disambiguation function is to reduce the set of possible results given by the syntactic analysis, understanding the role and the possible meaning of each word within the syntactical context in which it is used. Links between terms are then converted in links between unique identifiers (IDs) of terms by means of a semantic disambiguation that allows to further reduce results given by the syntactic disambiguation.

In the example above, the Syntactic Analysis sub-module selects the relevant terms and passes the nouns *Pope*, *office* and *Papacy* with all their possible synsets to the Semantic sub-module (see Figure 1) in order to perform the semantic disambiguation.

## 2.2 Semantic Analysis

The semantic analysis of a sentence expressed in natural language allows to assign to each term its most likely meaning in the context of the phrase, choosing between all available synsets. The Semantic Disambiguator performs this task by resorting to the Java WordNet Library (JWNL), a Java API for accessing WordNet. Let us recall, here, that after grouping nouns, verbs, adjectives, and adverbs, the dictionary organizes them into synonyms sets, called synsets, each expressing a distinct concept uniquely identified by a synsetID. In WordNet synsets are linked by means of conceptual-semantic and lexical relations, such as synonymy, meronymy/holonymy, hyperonymy/hyponymy. The relation between terms and synsets is not unique, but different senses of the same word (synsets) are possible: WordNet produces about 200,000 couples word-synsetID.

The goal of the semantic disambiguation task is to reduce the number of synsets that have been activated by the syntactic analysis, evaluating the use of words in the context of a sentence and the use of sentences in the context of the document. To solve word sense ambiguity in each sentence, the Semantic Disambiguator receives useful information from the Syntactic Disambiguator about the labeling of words (noun, adverb, verb, and adjective). Going back to the given example, the word Pope is contained in the following synsets:

```
09774028 • Pope, Catholic Pope, Roman Catholic Pope, Pontiff, Holy
Father,
    Vicar of Christ, Bishop of Rome (the head of the Roman Catholic
Church)
10513534 • Pope, Alexander Pope (English poet and satirist (1688-
1744))
```

According to the Semantic Disambiguator, only the first synset is selected, the second being not relevant in this case.

The semantic analysis identifies all synsets related to phrases, which are the keys for indexing and retrieving a document. If a term or a specific sense of a word is not included in the WordNet dictionary, it is considered anyway by adding the prefix “NO_WN_” to the term itself. For instance, if a specific sense of the term *photogrammetry* is not contained in WordNet, we can consider the generic unique *NO_WN_photogrammetry* to account for those senses of the word *photogrammetry* that are not reported in WordNet.

In order to take into account only synsets that really characterize the document, we consider both the results of the syntactic analysis (i.e., the syntactic relations among synsets) and a density function that assigns a weight to each synset in a phrase and in a document. The weight of a synset in a document is given by the number of occurrences of the synset itself in the document normalized by the overall number of synsets in the document. Starting from the weight of each synset in a document, we can categorize it by selecting all categories the synsets belong to. The weight of each category is obtained by summing the weights of the synsets related to it.

More formally, given a phrase $P$, let us denote with $T = \{ t_1,......,t_n \}$ the set of terms contained in $P$, and with $S(t_i) = \{ \lambda ,s_1,..., s_m \}$ the set of the synsets related to the term $t_i$ $(i = 1, 2, ..., n)$, where $\lambda$ is the default synset that accounts for additional senses not found in WordNet. The weight of a generic synset $s_k$, in the set $S(t_i)$ is:

$$W( s_k \mid t_i) = \begin{cases} \dfrac{1}{card(S(t_i))} & s_k \in S(t_i) \\ 0 & otherwise \end{cases}$$

Similarly, the weight of a generic synset $s_k$ in a phrase $P_j$, is defined as:

$$W(s_k \mid P_j) = \frac{\sum_{t \in P_j} W(s_k|t)}{card(P_j)}$$

Let us note that the weight of a synset related to a single term, in the case it is included only in a set *S(t$_i$)* is:

$$\frac{1}{n \cdot card(S(ti))} = \frac{1}{n \cdot (m+1)}$$

Summarizing, the weight of the phrase is given by the weights of all the synsets that are included in the phrase. Moreover, depending on the structure of the document the phrase belong to, a coefficient ωi is assigned to the phrase. For instance, the title of a document is weighted more than a generic phrase that occurs in the body of the document. In symbols, given a document D = { P$_1$,.., P$_n$ }, the weight of a phrase W(P$_j$ | D) is defined as:

$$W(P_j | D_h) = \frac{\omega_j}{\sum_i \omega_i}$$

The total weight of the document, given by the sum of weights calculated for each phrase, is 1.

Thus considering both the weight of the synset s$_k$ in the phrase W(s$_k$ | P$_j$) and the weight of the phrase in a document W(P$_j$ | D), the weight of a generic synset s$_k$ in a document can be defined as:

$$W(s_k | D_h) = \sum_{P \in D_h} W(s_k|P) \cdot W(P | D_h)$$

## 2.3 Classification

The classifier is capable to automatically categorize web documents, using the mapping between synsets of WordNet and a set of categories as proposed in WordNet Domains. The classifier assigns "coarse grain" domains categories to each synset extracted from the terms in phrases and applies to all synsets the density function W(s$_k$ | D) defined in the previous section.

Let us recall that the weight of each category associated to a document is given by the sum of the weights of the synsets related to such category. In such evaluation, a categorization error may occur also influenced by the number of senses λ not included in WordNet and not represented by a synset. The resulting set of categories is further reduced by the application of a function that takes into account only categories characterized by a density value bigger than a given threshold.

Table 1. Semantic text categorization related to the term *Pope* (text extracted from the Wikipedia page: http://en.wikipedia.org/wiki/Pope)

| Categories | Religion | School | Geography | Politics | Theology |
|---|---|---|---|---|---|
| Occurrences | 1688 | 245 | 150 | 159 | 98 |

Table 2. Semantic text categorization related to the term *Jellyfish* (text extracted from the Wikipedia page: http://en.wikipedia.org/wiki/Jellyfish)

| Categories | Animals | Biology | Anatomy | Person | Chemistry |
|---|---|---|---|---|---|
| Occurrences | 3639 | 3556 | 328 | 183 | 171 |

Tables 1 and 2 show some results of our approach in text categorization applied to the content of two different Wikipedia pages (we refer to the Wikipedia version of the 2006-07-02), the first related to the word *Pope* and the second related to the word *jellyfish*. The first table evidences *Religion* as the main category, being the page related to the *Pope*, while the second reveals the categories *Animals* and *Biology* as the most important topic of the document related to *jellyfish*. These peaks are due to the fact that Wikipedia pages are generally referred to a unique topic and each word (*Pope* and *jellyfish*) related to the page expresses a unique sense of the word itself.

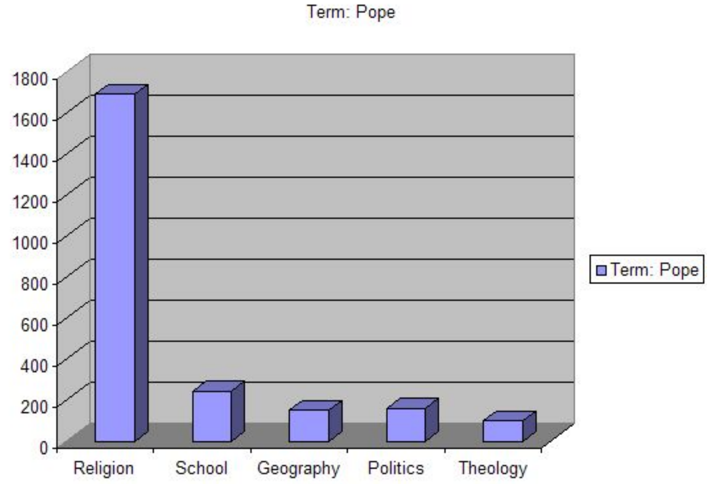


Figure 3. Classification of the page *Pope*

Moreover, in the Figure 3, the resulting sets of categories for the page *Pope* are represented on the X axis, while the Y axis depicts the values associated to the category, calculated as the sum of the weights of the synsets related to the category.

## 3. EXPERIMENTAL RESULTS

To assess the performances of the proposed approach, first a supervised classifier, based on the wk-NN technology [Cost93] (WKNN_A), has been trained starting from an automatic document collection. Then, it has been tested by using a set of documents classified by hand. Finally, its performances have been compared with those obtained by the same supervised classifier trained with a hand-made document collection (WKNN_H) and tested with the same test set adopted to test WKNN_A.

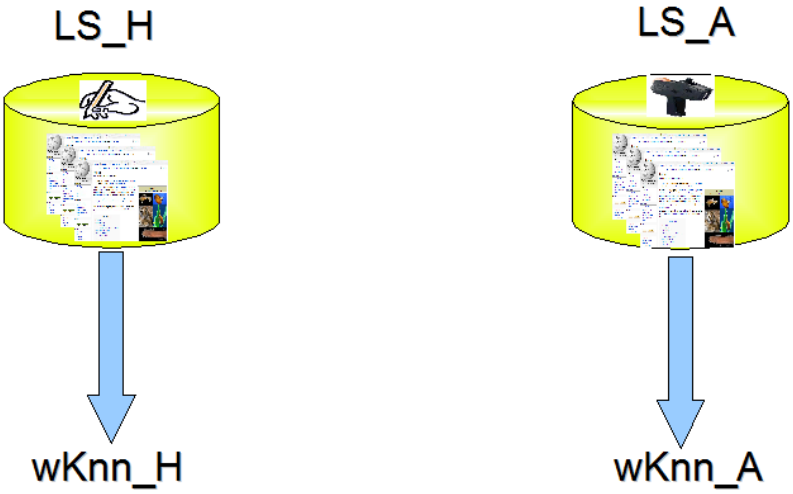


Figure 4. Training phase

### 3.1 Setting up the Experiments

In order to put into evidence that the proposed semantic classifier can be adopted to create document collections, first we selected five WordNet Domains categories: Animals, Chemistry, Geography, Medicine, and Plants. For each category, 600 documents have been classified by some domain engineers whereas 300 documents have been selected from the ones classified by the semantic classifier. Then, for each category, a learning set of 300 documents classified by hand (LS_H), with a balanced set of positive and negative examples, have been set up together with a further learning set of 300 documents classified by the system (LS_A). Each learning set will be adopted to train a classifier based on the wk-NN technology, WKNN_H and WKNN_A, respectively as shown in Figure 4.

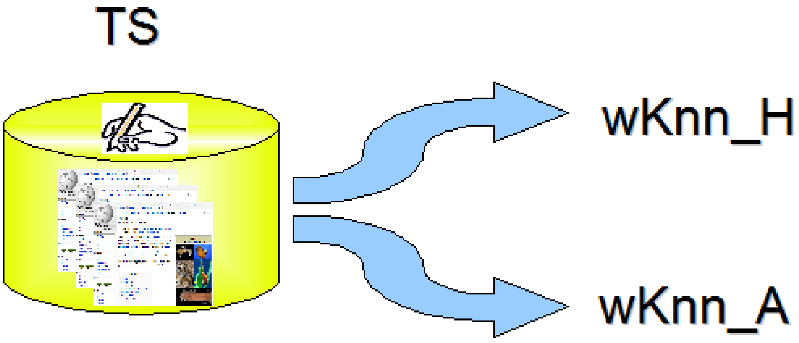


Figure 5. Test phase

As for the test, a training set of 300 documents has been generated using the 300 remaining documents classified by hand (see Figure 5).

## 3.2 Performances

To assess the capabilities of the proposed approach, we resorted to different metrics aimed at evaluating precision (π) and recall (ρ). In particular, we used micro- and macro-averaging, together with the F1 measure obtained by moving the acceptation threshold of each classifier under investigation over the range [0,1]. A schematic recall of the corresponding definitions follows (the interested reader may consult the corresponding literature, e.g. [Sebastiani02]).

As for micro- and macro-averaging, they are aimed at obtaining estimates of π and ρ relative to the whole category set. In particular, micro-averaging evaluates the overall π and ρ by globally summing over all individual decisions. In symbols (the μ superscript stands for micro-averaging):

$$\pi^{\mu} = \frac{\sum_{i=1..m} TP_i}{\sum_{i=1..m} (TP_i + FP_i)} \qquad \rho^{\mu} = \frac{\sum_{i=1..m} TP_i}{\sum_{i=1..m} (TP_i + FN_i)}$$

On the other hand, macro-averaging first evaluates $\pi$ and $\rho$ “locally” for each category, and then “globally” by averaging over the results of the different categories. In symbols (the *M* superscript stands for macro-averaging):

$$\pi^{\mathrm{M}} = \frac{\sum_{i}^{m} \pi_i}{m} \qquad \rho^{\mathrm{M}} = \frac{\sum_{i}^{m} \rho_i}{m}$$

As for F1 [vanRijsbergen79], it is obtained from a more general definition by imposing that π and ρ have the same degree of importance. In symbols:

$$F1 = \frac{2PR}{P + R}$$

Table 3 and 4 summarize our results. In particular, in Table 3 for each category the performances of WKNN_A are compared with the ones obtained by WKNN_H. Results show that the performances of WKNN_A are always worse, but definitely comparable with the ones corresponding to WKNN_H. In Table 4 micro- and macro-averaging are presented and comparing. Also, these results show that, as expected, WKNN_H performs better than WKNN_A, but, on the average, the results are quite similar. Therefore, datasets created adopting the semantic classifier can be used as trained sets.

Table 3. Experimental results

| | ***π*** | | ***ρ*** | | ***F1*** | |
|---|---|---|---|---|---|---|
| | WKNN_H | WKNN_A | WKNN_H | WKNN_A | WKNN_H | WKNN_A |
| ***Animals*** | 0.898 | 0.829 | 0.941 | 0.920 | 0.919 | 0.872 |
| ***Chemistry*** | 0.864 | 0.813 | 0.886 | 0.920 | 0.875 | 0.863 |

| | | | | | | |
|---|---|---|---|---|---|---|
| ***Geography*** | 0.974 | 0.926 | 0.917 | 0.955 | 0.945 | 0.940 |
| ***Medicine*** | 0.915 | 0.877 | 0.934 | 0.962 | 0.924 | 0.917 |
| ***Plants*** | 0.802 | 0.836 | 0.980 | 0.925 | 0.882 | 0.878 |

Table 4. Micro and macro-averaging

| | ***π*** | | ***ρ*** | | ***F1*** | |
|---|---|---|---|---|---|---|
| | WKNN_H | WKNN_A | WKNN_H | WKNN_A | WKNN_H | WKNN_A |
| ***micro-averaging*** | 0.886 | 0.857 | 0.933 | 0.937 | 0.909 | 0.895 |
| ***macro-averaging*** | 0.891 | 0.856 | 0.932 | 0.936 | 0.909 | 0.894 |

### 3.3 Discussion

Results point out that, on average, the performances of the automatic approach are quite similar to those obtained adopting a document collection classified by some domain engineers. This suggest that our approach can be effectively adopted to create document collections, in which documents are classified according to all the 167 classes of WordNet Domains. It turns out that, our approach being quite general, it is very easy to adapt it to a different set of classes in order to create different document collections. The main advantages of adopting our semantic approach to create document collections are: (i) very large document collections can be created; (ii) the full document text is preserved; (iii) multi-label classification is supported; (iv) document properties are known; and (v) no limitation regarding document availability is given.

Document collections created by the proposed system have been adopted in experimenting information retrieval and filtering systems [Addis08, Addis09].

## 4. CONCLUSIONS

In this paper, a semantic text categorization approach able to automatically create document collections has been presented. In these collections, documents are classified according to WordNet Domains taxonomy. The approach applies a classification algorithm that associates a set of categories and a weight to each document. Only relevant categories are taken into account. Experiments have been performed by training a supervised classifier with an automatic document collection and comparing its results with those belonging to the same supervised classifier trained with a hand-made document collection. Results point out that, on the average, the performances of the automatic approach are quite similar to those obtained by adopting a document collection classified by domain engineers. Therefore, the approach can be used to automatically create reliable document collections.

## AKNOWLEDGEMENT

This paper illustrates some results of the DART, Distributed Agent-based Retrieval Tools, project (http://www.dart-project.org). DART has been supported by the Italian "Ministero dell'Istruzione, dell'Università della Ricerca" and developed by the NDA team of the CRS4, Tiscali S.p.A. and the DIEE, Dept. of Electrical and Electronic Engineering, University of Cagliari.

We would like to thank all the researchers who worked with us to the development of the architecture of the project for their precious contribution.

## REFERENCES

Achananuparp, P. et al, 2008, Semantic Representation in Text Classification Using Topic Signature Mapping. IJCNN 2008, Hong Kong, China, pp. 1034-1040.

Addis, A. et al 2008, WIKI.MAS: A MultiAgent Information Retrieval System for Classifying Wikipedia Contents. *Communications of SIWN*, Vol. 3, June 2008, pp. 83-87.

Addis, A. et al 2009, Profiling Users to Perform Contextual Advertising. *Proceedings of the 10th Workshop dagli Oggetti agli Agenti (WOA 2009)*, 9-10 July 2009, Parma (Italy).

Angioni, M. et al, 2007a, DART: The Distributed Agent-Based Retrieval Toolkit. *Proceedings of CEA 07*. Gold Coast, Australia, pp. 425-433.

Angioni, M. et al, 2007b, User Oriented Information Retrieval in a Collaborative and Context Aware Search Engine. *In WSEAS Transactions on Computer Research*, Vol. 2, No. 1, pp. 79-86.

Cost, W. and Salzberg, S. 1993, A weighted nearest neighbor algorithm for learning with symbolic features. In *Machine Learning*, Vol. 10, No. 1, pp 57-78.

'Dewey Services' (2009), [Online], Available: http://www.oclc.org/dewey/

Ghahramani, Z., 2004, Unsupervised Learning. *In Advanced Lectures in Machine Learning. Lecture Notes in Computer Science* ,Vol. 3176, pp. 72-112.

Ko, Y. and Seo, J., 2000, Automatic Text Categorization by Unsupervised Learning. *Proceedings of COLING-00, the 18th International Conference on Computational Linguistics*. Saarbrücken, Germany, pp. 453-460.

Kohonen, T. et al, 2000, Self organization of a massive document collection, *In Neural Networks, IEEE Transactions*, Vol.11, No.3, pp.574-585.

Lewis, D.D. et al, 1996. Training algorithms for linear text classifiers. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 96)*. Zurich, Switzerland, pp. 298–306.

Lewis, D.D. et al, 2004. Rcv1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, Vol. 5, pp. 361–397.

Magnini, B. and Cavaglià, G., 2000, Integrating subject field codes into WordNet. *Proceedings of LREC, 2nd International Conference on Language Resources and Evaluation*. Athens, Greece.

Magnini, B. et al, 2002. The Role of Domain Information in Word Sense Disambiguation. *In Natural Language Engineering, special issue on Word Sense Disambiguation*, Vol. 8, No.4, pp. 359-373.

Magnini, B. and Strapparava C., 2004. User Modelling for News Web Sites with Word Sense Based Techniques. *User Modeling and User-Adapted Interaction* Vol. 14, No. 2, pp. 239-257

Miller et al., 1998, *WordNet: An Electronic Lexical Database*, MIT Press.

Sahami, M. et al, 1996, Applying the Multiple Cause Mixture Model to Text Categorization. *Proceedings of the International Conference on Machine Learning*, pp. 435-443.

Scott S., Matwin S., 1998, Text Classification using WordNet Hypernyms. *COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada.

Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47.

Sleator, D.D. and Temperley D., 1993, Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies,* Tilburg, The Netherlands.

Van Rijsbergen, C., 1979, *Information Retrieval*. Butterworths, London.

Yang, Y., 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, Vol. 1, No. 1/2, pp. 67–88.

Zesch, T. and Gurevych, I., 2006, Automatically Creating Datasets For Measures Of Semantic Relatedness. *Workshop On Linguistic Distances.* Sydney, Australia, pp. 16-25.

Zhu, X., 2005. Semi-Supervised Learning Literature Survey. *Technical Report n° 1530*, Computer Sciences, University of Wisconsin-Madison.