

# Testing Concept Indexing in Crosslingual Medical Text Classification

Francisco Carrero, Jose Carlos Cortizo  
Universidad Europea de Madrid  
{francisco.carrero, josecarlos.cortizo}@uem.es

Jose Maria Gomez  
Departamento de I+D, Optenet  
jgomez@optenet.com

## Abstract

*MetaMap is an online application that allows mapping text to UMLS Metathesaurus concepts, which is very useful for interoperability among different languages and systems within the biomedical domain. MetaMap Transfer (MMTx) is a Java program that makes MetaMap available to biomedical researchers in controlled, configurable environment. Currently there is no Spanish version of MetaMap, which difficult the use of UMLS Metathesaurus to extract concepts from Spanish biomedical texts. Developing a Spanish version of MetaMap would be a huge task, since there has been a lot of work supporting the English version for the last sixteen years.*

*Our ongoing research is mainly focused on using biomedical concepts for cross-lingual text classification. In this context the use of concepts instead of bag of words representation allows us to face text classification tasks abstracting from the language. In this paper we show our experiments on combining automatic translation techniques with the use of biomedical ontologies to produce an English text that can be processed by MMTx in order to extract concepts for text classification.*

## 1. Introduction

Information overload is common nowadays in our society. This is also the case for biomedical information, available from a variety of sources, as scientific papers, databases of summaries, structured or semi-structured databases and web services and clinical records of patients. In this domain, professionals in general need tools oriented to provide facilities for accessing and visualizing the adequate information for their needs. Medline, the most important and consulted bibliographical database in the biomedical domain, constitutes a main example. Medline contains more than 16 million references, with an increment between 2.000 and 4.000 references per day, and over 670,000 total added in 2007 [1].

In order to increase the retrieval and interoperability be-

tween biomedical resources, one of the key solutions may lie in the development of common terminologies acting as a metadata layer allowing link elements from various resources. UMLS (Unified Medical Language System) [4] constitutes a major repository of biomedical standard terminologies including controlled vocabularies and resources, such as MeSH, ICD-10, the Gene Ontology or SNOMED-CT that have served well in their respective domain. This knowledge has proved useful for many applications including decision support systems, management of patient records, information retrieval and data mining.

Several systems have been developed having as main goal the identification of concepts based on the text analysis of documents, ranging applications from genomics, drugs identification, and concrete aspects such as protein-protein interaction [5, 13] The MetaMap system [2] is nowadays the standard application developed at the National Library of Medicine (NLM) that identifies biomedical concepts from free-text documents and maps them to entries in UMLS.

### 1.1. Project Description

In this paper we present MIRCAT (Multilingual Information Retrieval based on Concepts and Automated Translation), a cross-lingual system to retrieve biomedical documents significantly related to medical records. Given a query in Spanish submitted by a person, it firstly retrieves a list of medical records ordered by relevance in two steps: 1) the query is expanded using concepts included in a biomedical ontology (i.e.: UMLS); 2) medical records are ranked using a representation based on biomedical concepts. Then, the user can choose a record and the system will retrieve several lists of ranked documents as follows: 1) Spanish news; 2) English news; 3) Spanish article abstracts; and 4) English article abstracts. This last step is done by using concepts to rank the documents against the selected medical record.

Throughout all the phases we need to obtain a semantic document representation, which makes it definitely crucial to use an accurate system to extract concepts from text. Keeping in mind that we are mainly working with UMLS,

Medical Records

- [Report ID: 97659154 - \[Related\]](#)  
This is a 7-month - old female who had a chest x-ray on 1/2 which was read as normal, but the patient is still coughing and congested
- [Report ID: 97660995 - \[Related\]](#)  
20-month - old male with cough. Evaluate for pneumonia.
- [Report ID: 97662371 - \[Related\]](#)  
Coarse markings with segmental disease in the right middle and left lower lobe. This may represent areas of atelectasis and/or pneumonia. The information was provided to the physician's office
- [Report ID: 97666817 - \[Related\]](#)  
Six year old with cough. The lungs are well - expanded and clear. There is no focal infiltrate or pleural effusion. The cardiac and mediastinal silhouette is within normal limits. No bony abnormalities are seen.
- [Report ID: 97669484 - \[Related\]](#)  
This is a 3-year - old with cough. Findings are suggestive of a viral process vs reactive airway disease without focal infiltrate.
- [Report ID: 97671492 - \[Related\]](#)  
The lungs are well expanded, but not hyperinflated. Lung parenchyma is clear, without peribronchial thickening or focal infiltrate. The cardiac and mediastinal silhouette is normal. No bony abnormalities are seen.
- [Report ID: 97690215 - \[Related\]](#)  
1. Left basilar atelectasis or pneumonia. 2. Very dense nodule right lateral inferior thorax. This could represent a calcified granuloma, but is nonspecific; or it is outside the thorax and in the region of the skin or subcutaneous tissues.

news papers

- [10% of U.S. Kids Using Cough Medicine Every Week](#)  
...While cough and cold medications for children are widely marketed in the United States, how frequently they are used had not been scientifically studied. This new finding, from researchers at Boston University's Slone Epidemiology Center, gives increased weight to recent revelations that cough and cold medication use can lead to serious adverse effects, including death...  
[http://www.nlm.nih.gov/medlineplus/news/fullstory\\_64193.html](http://www.nlm.nih.gov/medlineplus/news/fullstory_64193.html)
- [Health Woes Not Always to Blame for Chronic Cough](#)  
...In the first paper, the authors provided information on chronic cough, which is defined as a cough lasting longer than eight weeks. The condition, which affects 9 percent to 33 percent of the population in many areas of Europe and the United States, is often associated with cigarette smoking. Compared to non-smokers or ex-smokers, smokers are three times more likely to have chronic cough...  
[http://www.nlm.nih.gov/medlineplus/news/fullstory\\_63622.html](http://www.nlm.nih.gov/medlineplus/news/fullstory_63622.html)
- [ACE Inhibitor as Effective as More Expensive Blood Pressure Drug](#)  
...“The main value of ARBs as monotherapy in patients with cardiovascular disease is, in my opinion, as a substitute for an ACE inhibitor in a patient who cannot tolerate the ACE inhibitor because of cough,” said Dr. John J.V. McMurray, a professor of medical cardiology at the University of Glasgow in Scotland, who wrote an editorial accompanying the journal report...  
[http://www.nlm.nih.gov/medlineplus/news/fullstory\\_62832.html](http://www.nlm.nih.gov/medlineplus/news/fullstory_62832.html)
- [Ethnicity Plays Role in Parents' Treatment of Childhood Fever](#)  
...“It's a natural response for a parent to worry when a child has a fever and to want to fix it, so every pediatrician must have the fever talk with parents every time they bring a sick child to the office,” study author Dr. Michael Crocetti said in a prepared statement. “We must remind parents not all fevers are dangerous, that fever is a sign of the body's revved up defenses fighting infection, and that fever-reducing medications carry their own risks.”...  
[http://www.nlm.nih.gov/medlineplus/news/fullstory\\_64242.html](http://www.nlm.nih.gov/medlineplus/news/fullstory_64242.html)

**Figure 1. Prototype of the system’s interface showing the medical records retrieved for a given query and a set of related news to a certain medical record.**

we face the issue that currently there is only an English version of MetaMap, the tool that maps arbitrary text to concepts in UMLS Metathesaurus, and MMTx , a generic, configurable environment to make the MetaMap program available to biomedical researchers. The development of an equivalent tool in Spanish would require a huge amount of work and specific knowledge and, although it would be a very valuable task, we wonder if it is really a must.

The key point for us at current stage is to evaluate the necessity to develop a Spanish version of MMTx, against the possibility of using automatic translation systems (such as Google Translator or Systran) to obtain an English representation for a Spanish text.; and then, to apply MMTx to English text and obtain a semantic representation that should include (almost) the same concepts as in Spanish.

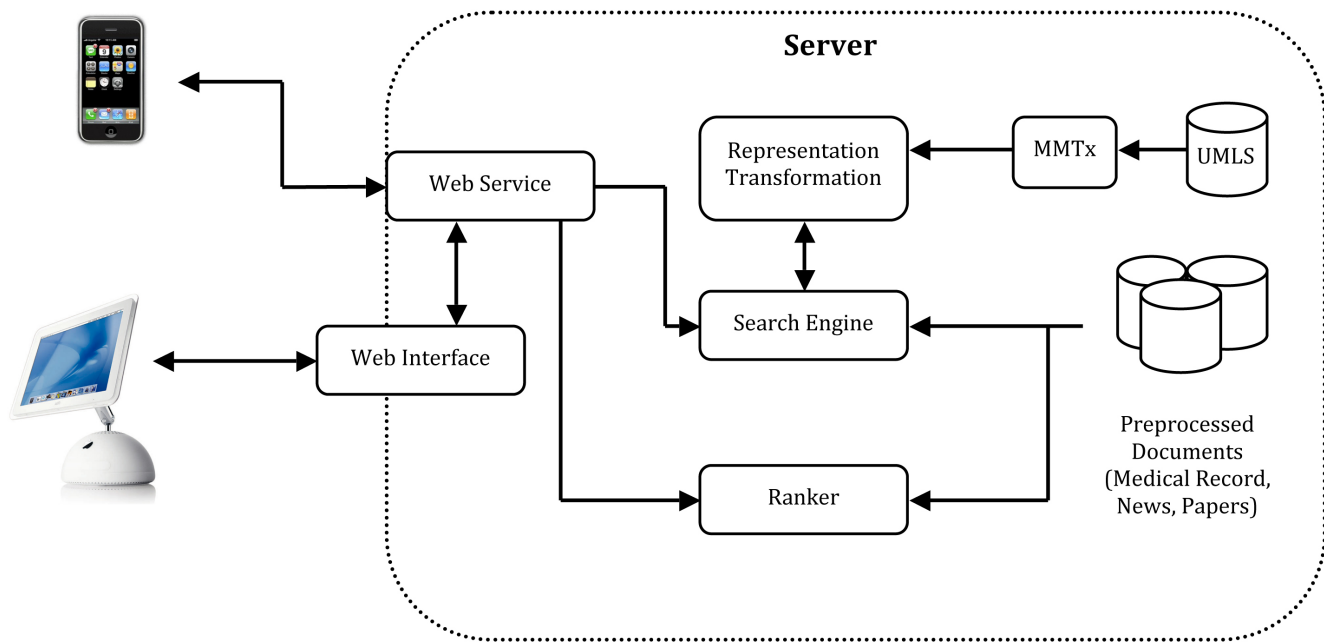
**1.2. Architecture**

MIRCAT is a Web Information System [14] that may be accessed from desktop computers or from mobile devices. This is an important issue when designing the architecture, due to the restrictions of some mobile devices [8].

In MIRCAT's system architecture (see Figure 2), there is a Web Service [19] that standardizes the access to the main functionality of the system. It is responsible of receiving the queries, pre-process the parameters, execute the system

modules according to the needs and send back the response of the system. The Web Service response is in XML format, which is a standard format to interoperate systems but it cannot be used as a direct response to the user. When accessed from a mobile device, an application installed on the clients device communicates with the Web Service, analyzes the response and shows it. When the user is using a desktop computer to access the system, it works as a web application. The user opens its browser, introduces the URL of MIRCAT, and the browser connects to the Web Interface. This Web Interface communicates with the Web Service and produces an interface in HTML according to the response of the system. The Web Service acts like a middleman between the user interfaces and the information retrieval system. It can receive two kind of requests:

- Given a simple query, to retrieve the more relevant medical records. As an example, a medical doctor that is taking care a new patient searches into the system giving one or some of the symptoms as the initial query.
- Given a medical record, to retrieve all the related documents. Following the use case proposed in the previous point, the medical doctor is interested in one of the medical records and clicks on the Related Documents button near the medical record he is interested



**Figure 2. Global system's architecture.**

on. Figure 1 shows an example of the interface showing medical records and related documents.

According to the request, the Web Service will call the Search Engine or the Ranker. The Search Engine module is used to retrieve medical records from simple queries. It receives the query and pre-processes it transforming its representation from bag-of-words to a set of concepts (using MMTx and UMLS). Then, it uses the set of concepts as a query to search in a concepts inverted index previously constructed from pre-processed medical records.

Ranker module is used to retrieve documents related to a given medical record. It doesn't need to transform any representation as the medical records and the other documents (news and papers) are previously pre-processed and indexed using a concept-based representation. From a simple perspective, the Ranker can be seen as a search engine where the query is a complete medical record. In the practice, we are currently developing more sophisticated techniques to retrieve similar documents using probabilistic models, machine learning algorithms [9] and feature selection techniques [7].

## 2. Related Work

The most widely used text representation in text classification like Information Retrieval (IR) or Automated Text

Categorization (ATC) tasks has been, by far, the bag of words model [16, 17]. In this representation, a document is represented as vector of terms and associated weights. Terms are usually stemmed words, and weights are computed as a function of their occurrences in documents and the whole text collection, like TF.IDF weights. This representation does not capture the full meaning of texts, but it is enough to build reasonably effective text classifiers.

However, there have been several attempts to design text representations that better capture the semantics of documents. These approaches rely on the emergence of wide-coverage semantic resources like WordNet. For instance, some authors have demonstrated that using WordNet concepts (synsets) instead of, or added to, words can improve Information Retrieval [11] and Text Categorization [10].

A major point is that concepts can be language-independent (as in EuroWordNet), what allows full cross-language retrieval and categorization [12]. However, concept based representations (generally named concept indexing) are doomed with the limited effectiveness of current free text Word Sense Disambiguation (WSD) approaches. The effectiveness of an average WSD system rarely exceeds 60% on ambiguous words (see e.g. Senseval [18] results) on running text, a level that is hardly reached on short texts like search engine queries. On the other side, the previous works have demonstrated that the effectiveness of text classification can be improved even in the presence of an im-

portant percentage of disambiguation errors. Moreover, our approach takes full medical records as queries, providing a better context for disambiguation.

A promising issue is that there are high quality semantic resources in the domain of biomedicine, like the Unified Medical Language System (UMLS) or SNOMED. These resources have been successfully used in several text classification tasks. For instance, [20] reports good results when using UMLS concepts for concept indexing in the European Project MUCHMORE. Also, [15] presents the MorphoSaurus system, which makes use of UMLS for concept indexing in cross-language retrieval, in comparison with query translation, with results that support concept indexing.

Regarding translation, a full report of the current state of the art is beyond the scope of this paper. Instead, let us remark that the system we employ, Google statistical translator, has top performed in the most recent NIST Open Machine Translation Competition (2006). The strength of this translation tool relies on the huge amount of data it makes use for computing the statistical metrics of its language model.

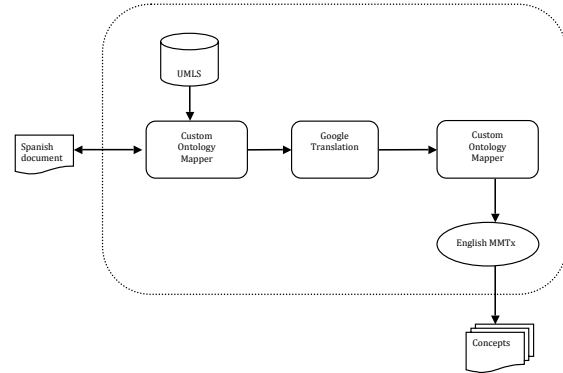
### 3. Spanish MMTx

We have developed two versions of Spanish MMTx: A first simple approach uses Google Translator to obtain an English version of the text and then applies English MMTx to extract concepts. This approach, ignoring the quality of general translation, presents some important mistakes when translating some technical biomedical terms, keeping them in Spanish.

The second approach is illustrated in Figure 3. It delegates to Google Translator to obtain the general translation, but uses a custom UMLS ontology mapper to translate biomedical terms. The first version of the custom UMLS ontology mapper has been created building a sub-ontology of UMLS by using only the isa relation. Then, for each of the concepts included, all Spanish and English string representations have been stored. Considering this mapper, this second approach involves the following steps:

- Search the original Spanish text and substitute each of the found concepts by its concept ID. In case of ambiguity, the chosen concept is the one with higher level in the ontology.
- Send the text from the first step to Google Translator, retrieving an English version with the general translation.
- Search the English version and replace the concept IDs with a string representation. If there are several representations, we chose to use the shortest one.

- Use the English MMTx to extract the concepts.



**Figure 3. Automated translation process using ontologies to translate biomedical specific terms.**

## 4. Experiments

As we needed to evaluate the suitability of develop a Spanish MetaMap, we designed a set of experiments with this orientation. In the previous section we stated our hypothesis: using automated translation combined with the use of domain ontologies and MMTx, we could avoid the need of an specific Spanish MetaMap for text classification tasks. To test the validity of this hypothesis, we need to compare the classification of a certain classifier built from the english texts versus the classification obtained by a classifier built from Spanish texts translated with an automated translation system as explained in previous sections.

### 4.1. Collection

For testing our hypothesis, we needed a corpus of biomedical documents in both languages: Spanish and English. MedLine Plus stores health-related news articles extracted from Reuters Health and HealthDay . All these news articles are tagged with a set of related MedLine Plus pages, which can be considered as topics or categories (there are over 750 different diseases or conditions treated as topics).

One interesting characteristic of MedLine Plus is that it contains medical information in English and also some of the contents in Spanish. Not all the contents are in both languages, and also not all the news. We developed a spider that, once a month, downloaded all the English and Spanish

news articles and checked the correspondence among news. From over 2000 news downloaded since December 2007, we were able to detect 600 news articles available in both languages and we built the collection using those items.

Categories in MedLine Plus are organized in a hierarchy that contains 5 first-level concepts: body locations, disorders and conditions, diagnosis and therapy, demographic groups and health and wellness. There are 42 second-level categories and 755 leaf-categories. As there are much more categories than documents, we have decided to create a binary class that indicates whether a news article belongs a category contained inside the super-category demographic groups or not.

## 4.2. Description and Goals

In order to evaluate the suitability of the proposed hypothesis to avoid the need of a Spanish MetaMap, we have designed a set of experiments. These experiments makes a general comparison among different concept based document representation and then compares the performance of a classifier built using concepts extracted from original English text to classifiers built using concepts extracted from the same text automatically translated from Spanish.

From our original bilingual collection of news articles, we have generated 3 different collections:

- ENG: Containing the original English documents.
- ENG\_TRANS: Containing the Spanish documents automatically translated to English using Google Translator.
- ENG\_UNMKD: Containing the Spanish documents translated to English by means of Google Translator and domain ontologies (UMLS), as described in section 3 (Figure 3).

The main goal of these experiments is to compare the translated documents (ENG\_TRANS and ENG\_UNMKD) to the baseline (ENG) document collection. For each document, MMTx produces a list of pairs containing the phrase number and the concepts in that phrase. As an example, we copy a subset of the list produced by MMTx:

```
1: C0331964|C0585027
3: C0585027
4: C0015967
5: C0282425
7: C0004339
7: C0018767
8: C0026867
9: C0205388|C0439227|C0439228
9: C0205388|C0439227|C1561539
12: C1511253
13: C0038454|C0030705
15: C1279919|C1555688
15: C1279919|C0237820
16: C0035173
```

```
17: C0008972
17: C0947630
17: C0557651
```

This representation is not a standard approach to document representation for any text mining task. Usually, documents are represented as a bag-of-word or a bag-of-concepts. So we should transform the data representation that MMTx outputs to a more standard one.

## 4.3. Possible Documents Representations

There are two important considerations from the MMTx representation. A string like C0331964 represents each concept. Some phrases are represented by a conjunction of strings, which is represented by several strings connected by —, for example C0205388—C0439227—C0439228. There are some phrases that appear several times, that means there are ambiguities or different possible concepts or combination of concepts that represents that phrase.

We should translate the MMTx representation to a representation containing a list of concepts. Paying attention to the previous considerations of the MMTx representation, we should deal with compound concepts and with ambiguities. We have developed 4 possible data representations according to this: A1, A2, B1, B2:

- Document representations starting with an A (A1 and A2) uses compound concepts. That means that a compound concept like C0205388—C0439227—C0439228 would be treated as a simple one like C0331964.
- Document representations starting with a B (B1 and B2) do not use compound concepts. Instead, they use the simple concepts that they are compound of as indexing units. That means that a concept like C0205388—C0439227—C0439228 is transformed into 3 different concepts (C0205388, C0439227 and C0439228).
- Document representations ending with a 1 (A1 and B1) resolves the ambiguity by adding all the concepts contained in all the possible interpretations of the phrase. Following the previous example, the phrase 7 that presents 2 possible interpretations (concepts C0004339 or C0018767) is represented by the two concepts.
- Document representations ending with a 2 (A2 and B2) ignores the ambiguities by choosing the first possibility for each phrase.

A1 document representation is more complex and nearer to the human understanding and B2 document representation is the simplest one and nearer to the standard machine

**Table 1. Number of different concepts for each document representation and number of concepts after filtering.**

Doc. Rep.	Total	Filtered
A1	45.280	2.368
A2	21.257	1.415
B1	9.990	2.293
B2	8.148	1.653

representation for text mining tasks. More complex document representation generates more different concepts. Table 1 shows the number of global concepts for each document representation.

Data representations containing a lot of features do not usually perform very well in text tasks, especially in text classification, as many classifiers degrade in prediction accuracy when faced with many irrelevant features or redundant/correlated ones [6]. The explanation to this phenomenon may be found in the curse of dimensionality, which refers to the exponential growth of the number of instances needed to describe the data as a function of dimensionality (number of attributes) [3]. Zipfs Law [21] can be used to solve this problem without facing any concrete task, by filtering the features appearing in more than M% of the documents and the ones appearing in less than N% of the documents. We have filtered the concepts according to this, with M=10% and N=1%. The global number of concepts after this filtering process is shown in Table 1.

#### 4.4. Results

We have computed the similarity between the original ENG documents and the translated ones (ENG\_TRANS and ENG\_UNMKD) for each possible representation. Then, we have calculated the average value and standard deviation for the 600 news items contained in the global collection. Table 2 resumes the results of these experiments.

We have also compared the classification performance (accuracy and AUC) of a baseline classifier built upon the different document representations. As baseline classifier we have chosen the Nave Bayes (NB) because it has a privilege position [7] due to its simplicity, its resilience to noise, its time and space efficiency, its understandability and its results both in performance and speed in the area of information retrieval and automated text categorization.

For each possible document representation (words, A1, A2, B1 and B2), we have computed the performance of the NB classifier built upon the original English documents and the performance obtained by building the NB classifier upon the Spanish documents automatically translated by Google translation service (TRANS) and the documents

**Table 2. Average similarity between document representations generated from translated texts and the representations generated from the English documents.**

Doc. Rep.	TRANS	UNMKD
A1	58.86 ± 8.37	54.31 ± 7.90
A1+Zipf	65.87 ± 11.11	63.23 ± 10.99
A2	60.79 ± 6.78	58.07 ± 6.40
A2+Zipf	65.80 ± 9.56	62.94 ± 9.51
B1	79.42 ± 6.43	76.55 ± 5.54
B1+Zipf	77.63 ± 8.85	75.00 ± 8.56
B2	78.38 ± 6.21	74.76 ± 5.38
B2+Zipf	76.38 ± 8.53	73.59 ± 8.18
Words	75.11 ± 6.13	72.69 ± 8.09
Words+Zipf	73.45 ± 5.21	70.30 ± 7.55

translated as explained in Section 3 (UNMKD). Table 3 shows the performance increments (+) or decrements (-) of the classifiers compared with the baseline NB built from the original documents.

Table 4 and Table 5 show the results of the comparisons after selecting the 10% or 1% of best attributes by means of a Information Gain filtering process.

## 5. Results Discussion

### 5.1. Translation

Considering the four representations described above, the worst results in terms of similarity are achieved with the most complex and near-to-humans representation (A1). On the other side, B1 is a less complex and near-to-humans representation, and produces the best results of the series. This proves that our model seems to be more suitable for tasks that manage the concepts on a plain bag-of-concepts way.

The use of Zipfs law improves the results within the A representations, while makes the values obtained for B get worse. The reason for A may possibly be that this representation produces too many different concepts, because some of them are made up of combinations of simpler ones and many of them appear few times on the text. Since we keep only the most relevant concepts, it seems to eliminate some of the concepts that make the difference for each pair of documents. The loss of precision obtained with representation B may come from the fact that the initial number of concepts is already low.

Relating to the difference between the results when applying simple or complex custom UMLS concepts mapper, it is clear that the complex one currently does not improve

**Table 3. Comparison of the performance, measured in accuracy and AUC, of the NB classifier built upon the different document representations.**

Doc. Rep.	P.M.	TRANS	UNMKD
Words+Zipf	Acc.	+4.66	+4.83
	AUC	+7.1	+7.8
A1+Zipf	Acc.	-0.33	+2.0
	AUC	-1.3	+1.0
A2+Zipf	Acc.	-0.33	+1.33
	AUC	-1.5	+0.3
B1+Zipf	Acc.	+4.66	+6.16
	AUC	+6.2	+6.7
Words+Zipf	Acc.	-0.66	+2.1
	AUC	+2.66	+2.9

**Table 4. Comparison of the performance, measured in accuracy and AUC, of the NB classifier built upon the different document representations after selecting the 10% of best attributes using Information Gain filter.**

Doc. Rep.	P.M.	TRANS	UNMKD
Words+Zipf	Acc.	-1.33	+1.99
	AUC	-1.1	+1.7
A1+Zipf	Acc.	+2.33	+5.16
	AUC	+3.6	+7.0
A2+Zipf	Acc.	+0.5	+5.5
	AUC	0.0	+5.9
B1+Zipf	Acc.	-0.5	-1.33
	AUC	-1.2	-0.7
Words+Zipf	Acc.	-0.33	+2.66
	AUC	-0.1	+1.2

the translation over the simple one, although the difference isn't too high. It may be to some extent due to several limitations on the translator that are described below but, however, there are enough things to improve on the mapper.

Regarding the best results obtained for all experiments, we consider that values of 79,42% for the simple mapper and 76,55% for the complex one are promising enough to continue with our research on improving the models. Specially, we find that there is a broad field to improve the complex UMLS ontology mapper.

## 5.2. Text Classification

Regardless of translation quality discussed above, our approach of combining translation with English MMTx throws good results for text classification. In tables 3, 4

**Table 5. Comparison of the performance, measured in accuracy and AUC, of the NB classifier built upon the different document representations after selecting the 1% of best attributes using Information Gain filter.**

Doc. Rep.	P.M.	TRANS	UNMKD
Words+Zipf	Acc.	-3.0	+0.66
	AUC	-0.6	+2.2
A1+Zipf	Acc.	+0.33	+4.33
	AUC	+0.7	+6.4
A2+Zipf	Acc.	+0.83	+5.5
	AUC	+2.1	+7.8
B1+Zipf	Acc.	-1.66	-1.0
	AUC	-1.7	-0.4
Words+Zipf	Acc.	-3.16	-2.0
	AUC	-3.9	-3.3

and 5 we can see that all results are always comparable to classification on original English texts, and in some cases are even better. For instance, the best result is achieved when using ENG\_UNMKD with A2+Zipf, which gives a difference of 7.8% on AUC. Worst value comes from ENG\_TRANS with B2+Zipf, with an AUC difference of -3.9%. We hope further experiments will reveal whether these good results are only a question of the dataset or they can be related to differences on original English and Spanish documents.

Furthermore, the use of specific ontologies during translation process improves over the approach of simple translation to Spanish using Google Translate. This is a point to be taken into account, since results on tables 1 and 2 show that translation quality is worse when using ontologies.

## 6. Limitations

Current version of Google Translator has imposed some limitations at different levels. First, we wanted to send a Spanish document including some kind of code to mark out the concept strings found using our custom UMLS ontology mapper. However, although we found some marks that seemed to work, sometimes they were eliminated after the translation, or translation lost part of the meaning. For instance, we tried to delimitate using quotes, but they were occasionally removed; and we tried to mark out with (: and :), but translation considered marked text as an explanation without any grammatical connection.

The second limitation seems to be related to the API version. Current version is 0.40, and it does not allow sending texts with more than 512 bytes length. When our texts were longer, we had to split them into several parts, thus using

syntactic tools that otherwise would not be needed.

## 7. Conclusions and Future Work

We have presented MIRCAT, a system to improve the access to cross-lingual information related to medical records. It makes use of semantic information to represent all the documents. To date, there isn't an effective tool to extract UMLS concepts from Spanish texts. Our experiments on creating a Spanish MMTx combining existing English MMTx and automatic translators have shown to be promising for tasks such as Text Categorization and Information Retrieval as the concept based representation of translated text does not vary much from the concept based representation of English documents. However, it is out of the scope to evaluate the correctness and quality of translation. Of course, a specific Spanish MMTx will always be more accurate than this model, but the key point is to consider if such a huge task would improve further results in TC and IR.

First experiments on Text Classification suggest that our approach can be considered to be viable. However, our next steps will be focused on applying the Spanish MMTx to diverse Information Retrieval tasks, as well as working on different Text Classification data sets to confirm and even improve our first results. Testing documents representations evaluated in this paper on different text tasks will allow us to conclude if there is any need to build a Spanish MMTx from scratch.

We will evaluate other general translation systems, such as Systran, and some more domain-specific like PAHOMTS, an automatic translation software maintained by the Pan American Health Organization (PAHO).

We will also modify our custom UMLS ontology mapper, using more semantic relations and keeping only those concepts that can be considered to belong to the biomedical domain.

## 8. Acknowledgments

This work has been partially funded by the Spanish Ministry of Education and Science and the European Union from the ERDF (TIN2005-08988-C02), and the Spanish Ministry of Industry as part of the PROFIT program (FIT-350300-2007-75).

## References

- [1] Medline factsheet.
- [2] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the American Medical Informatics Association Symposium*, pages 17–21, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA. alan@nlm.nih.gov, 2001.
- [3] R. Bellman. *Adaptive Control Processes: a Guided Tour*. Princeton. University Press, 1961.
- [4] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue), January 2004.
- [5] F. Carrero, J. M. Gomez, E. Puertas, M. Maña, and J. Mata. Attribute analysis in biomedical text classification. In *Second BioCreAtIvE Challenge Workshop: Critical Assessment of Information Extraction in Molecular Biology*, 2007.
- [6] J. C. Cortizo and J. I. Giráldez. Discovering data dependencies in web content mining. In J. M. Gutierrez, J. J. Martinez, and P. Isaías, editors, *IADIS International Conference WWW/Internet*, 2004.
- [7] J. C. Cortizo, J. I. Giráldez, and M. C. Gaya. Wrapping the naive bayes classifier to relax the effect of dependences. In *IDEAL 2007*, volume 4881 of *Lecture Notes in Computer Science*, pages 229–239. Springer Verlag, 2007.
- [8] D. Gachet, M. de Buenaga, and E. Puertas. Mobile access to patient clinical records and related medical documentation. In *Proceedings of the 1 International Conference on Ubiquitous Computing: Applications, Technology and Social Issues*, 2006.
- [9] M. C. Gaya, I. Giraldez, and J. C. Cortizo. Uso de algoritmos evolutivos para la fusion de teorías en minería de datos distribuida. In *Actas de la XII Conferencia de la Asociación Española para la Inteligencia Artificial – CAEPIA/TTIA*, 2007.
- [10] J. M. Gomez Hidalgo, M. d. Buenaga, and J. C. Cortizo. The role of word sense disambiguation in automated text categorization. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems*, Lecture Notes in Computer Science, pages 298–309. Springer Verlag, 2005.
- [11] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarrán. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL 1998 Workshop on Usage of WordNet for Natural Language Processing*, 1998.
- [12] J. Gonzalo, F. Verdejo, C. Peters, and N. Calzolari. Applying eurowordnet to cross-language text retrieval. *Computers and the Humanities*, 32(2-3):185–207, 1998.
- [13] L. Hunter. Opendmap: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. *BMC Bioinformatics*, 9(78), 2008.
- [14] T. Isakowitz, M. Bieber, and F. Vitali. Web information systems. *Communications of the ACM*, 41(7):78–80, 1998.
- [15] K. Marko, S. Schulz, and U. Hahn. Morphosaurus—design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4):537–545, 2005.
- [16] G. Salton. *Automatic text processing: the transformation, analysis and retrieval of information by computer*. Addison-Wesley, 1989.
- [17] F. Sebastiani. Machine learning in automated text categorization. *ACM Computer Surveys*, 34(1):1–47, 2002.

- [18] B. Snyder and M. Palmer. The english all words task. In *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 2004.
- [19] K. Umapathy and S. Purao. A theoretical investigation of the emerging standards for web services. *Information Systems Frontiers*, 9(1):119–134, 2007.
- [20] M. Volk, B. Ripplinger, S. Vintar, and P. Buitelaar. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1-3):97–112, 2002.
- [21] G. K. Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley, 1949.