

Building a Spanish MMTx by using Automatic Translation and Biomedical Ontologies

Francisco Carrero¹, José Carlos Cortizo^{1,2}, José María Gómez³

¹ Universidad Europea de Madrid, C/Tajo s/n, Villaviciosa de Odón, 28670, Madrid, Spain
{francisco.carrero, josecarlos.cortizo}@uem.es

² Artificial Intelligence & Network Solutions S.L., <http://www.ainetsolutions.com/jccp>
jccp@ainetsolutions.com

³ Departamento de I+D, Optenet, Parque Empresarial Alvia, 28230, Las Rozas, Madrid, Spain
jgomez@optenet.com

Abstract. The use of domain ontologies is becoming increasingly popular in Medical Natural Language Processing Systems. A wide variety of knowledge bases in multiple languages has been integrated into the Unified Medical Language System (UMLS) to create a huge knowledge source that can be accessed with diverse lexical tools. MetaMap (and its java version MMTx) is a tool that allows extracting medical concepts from free text, but currently there not exists a Spanish version. Our ongoing research is centered on the application of biomedical concepts to cross-lingual text classification, what makes it necessary to have a Spanish MMTx available. We have combined automatic translation techniques with biomedical ontologies and the existing English MMTx to produce a Spanish version of MMTx. We have evaluated different approaches and applied several types of evaluation according to different concept representations for text classification. Our results prove that the use of existing translation tools such as Google Translate produce translations with a high similarity to original texts in terms of extracted concepts.

Keywords: Semantic techniques, data pre and post processing, information filtering, recommender systems.

1 Introduction

The volume of published biomedical information is increasing every year. The proliferation of online sources such as scientific repositories, clinical records databases, knowledge databases, etc., has produced an information overload that surpass the amount of information that researchers can cope with. This scenario makes it necessary to develop tools that help access and visualization of specific information useful for biomedical researchers. Pubmed¹, a service of the U.S. National Library of Medicine, constitutes an example of a huge source of biomedical

¹ <http://www.ncbi.nlm.nih.gov/pubmed/>.

information, since it includes over 17 million citations from multiple life science journals, among which stand out Medline [1].

Several attempts to develop common biomedical terminologies that help improving interoperability between medical resources have appeared within the last few years. However, it has not been until the appearance of UMLS (Unified Medical Language System) [3] that there is a unified way to access multiple and complementary resources. UMLS includes more than 60 families of controlled vocabularies and resources such as SNOMED-CT, MeSH, ICD-10 or Gene Ontology. This knowledge has proved useful for many applications including decision support systems, management of patient records, information retrieval and data mining.

The MetaMap system [2] is an online application developed at the National Library of Medicine (NLM) 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 evaluate the possibility of combining automatic translation techniques with the use of biomedical ontologies to produce an English text that can be processed by MMTx.

1.1 Project Description

In this paper we present GALEN, 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, 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.

2 Related Work

The most widely used text representation in text classification like Information Retrieval (IR) or Text Categorization (TC) tasks has been, by far, the bag of words model [14,15]. 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 [10] and Text Categorization [9].

A major point is that concepts can be language-independent (as in EuroWordNet), what allows full cross-language retrieval and categorization [11]. 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 [16] 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 important 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, [17] reports good results when using UMLS concepts for concept indexing in the European Project MUCHMORE. Also, [13] 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.

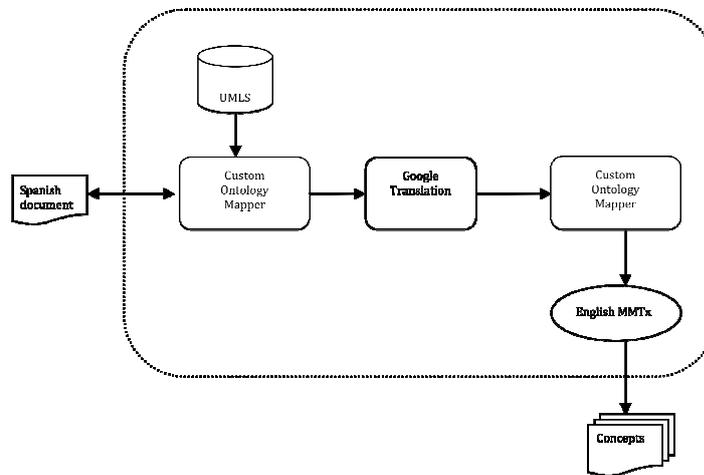


Fig. 1. Automated translation process using Ontologies to translate biomedical specific terms.

The second approach is illustrated in Figure 1. 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.

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 a specific Spanish MetaMap. To test the validity of this hypothesis, we need to compare the concepts extracted by MMTx from English texts to the concepts extracted by MMTx from Spanish texts previously translated to English.

4.1 Description an Goals

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). MedLine Plus contains medical information in English and also some of the contents in Spanish. 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.

For any text task (classification, retrieval or filtering), the possible document representations are similar [8]. Our approach in this paper evaluates the concept based document representations not taking care about a particular task. Our main goal is to establish whether our approach could produce benefits to any text task or if it should not be considered.

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 1).

4.2 Possible Documents Representation

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.
- We have also tested a word based representation as baseline.

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 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.

Table 1 Different concepts for each document representation and number of concepts after filtering

Document representation	Total	Filtered
A1	45.280	2.368
A2	21.257	1.415
B1	9.990	2.293
B2	8.148	1.653
Word	15.966	2.665

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 [5]. 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). Zipf’s Law 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.

4.3 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.

Table 2 Average similarities between document representations generated from translated texts and the representations generated from the original English texts. Best results are bold-faced.

Document representation	TRANS	UNMKD
A1	56.86±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
Word	75.11±6.13	72.69±8.09
Word+Zipf	73.45±5.21	70.30±7.55

5 Discussion of the Results

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 Zipf’s 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 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.

It is interesting to see how simple conceptual representations (B1 and B2) obtains better similarity values than baseline word-based representations. Also, 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.

6 Conclusions and Future Work

We have presented GALEN, 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.

Our next step will be to apply the Spanish MMTx to diverse text mining tasks, like Text Categorization or Information Retrieval. Testing the documents representations evaluated in this paper on real text tasks, will allow us to conclude if there is any need to build a Spanish MMTx from scratch.

We will try modifying our custom UMLS ontology mapper, using more semantic relations and keeping only those concepts that can be considered to belong to the biomedical domain. From a more practical point of view, we are currently developing more sophisticated techniques to retrieve similar documents based on conceptual representations using probabilistic models, machine learning algorithms [7] and feature selection techniques [6].

7 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. <http://www.nlm.nih.gov/pubs/factsheets/medline.html>
2. Aronson, AR., Effective mapping of biomedical text to the UMLS Metathesaurus. Proceedings of the American Medical Informatics Association Symp., pp. 17-21. 2001.
3. Bodenreider O, The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 2004, 32:D267-D270, 2004.
4. Carrero García, F., et al., Attribute Analysis in Biomedical Text Classification. Second BioCreAtIvE Challenge Workshop: Critical Assessment of Information Extraction in Molecular Biology, Spanish National Cancer Research Centre (CNIO), Madrid, SPAIN, 2007.
5. Cortizo, J.C., Giraldez, I. 2004. Discovering Data Dependencies in Web Content Mining. In Proceedings of the IADIS International Conference WWW/Internet 2004 (Madrid, Spain, October 6-9, 2004), 881-884.
6. Cortizo, J.C., Giraldez, I., Gaya, M.C. Wrapping the Naïve Bayes Classifier to Relax the Effect of Dependences. In Proceedings of the Intelligent Data Engineering and Automated Learning – IDEAL 2007 LNCS Vol. 4881, 229-239. 2007.
7. Gaya, M.C., Giraldez, I., Cortizo, J.C. 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, Vol. 2, 121-130. 2007.
8. Gómez Hidalgo, J.M., Tutorial on Text Mining and Internet Concept Filtering. 13th European Conference on Machine Learning (ECML'02) and 6th European Conference on Principles and Practice of Knowledge Discovery in DataBases (PKDD'02), 2002.
9. Gómez Hidalgo, J.M., et al. Concept Indexing for Automated Text Categorization. In 9th International Conference on Applications of Natural Language to Information Systems, Lecture Notes in Computer Science, Vol. 3136, Springer, pp. 195-206, 2004.
10. Gonzalo, J., et al., Indexing with WordNet synsets can improve Text Retrieval. Proceedings of the COLING/ACL 1998 Workshop on Usage of WordNet for Natural Language Processing, Montreal, 1998.
11. Gonzalo, J., et al., Applying EuroWordNet to Cross-Language Text Retrieval. *Computers and the Humanities*, 32, 2-3, 185-207. 1998.
12. Hunter L, et al., 2008, 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* 2008, 9:78.
13. Marko, K., Schulz, S., Hahn, U., MorphoSaurus--design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4), pp. 537-45. 2005.
14. Salton, G. Automatic text processing: the transformation, analysis and retrieval of information by computer. Addison-Wesley, Reading, US. 1989.
15. Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1):1-47. 2002.
16. Snyder, B., Palmer, M. The English all words task. SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, 2004.
17. Volk, M., et al. Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67 (1-3), pp. 97-112. 2002.