In the Development of a Spanish Metamap

Francisco Carrero, José Carlos Cortizo
Universidad Europea de Madrid
C/Tajo s/n, 28670, Madrid, Spain

{francisco.carrero, josecarlos.cortizo}@uem.es

José María Gómez
Departamento de I+D, Optenet
C/José Echegaray 8 edf. 3
Las Rozas, Madrid, Spain
+34 902154604

jgomez@optenet.com

Manuel de Buenaga
Universidad Europea de Madrid
C/Tajo s/n, 28670, Madrid, Spain
+34912115611

buenaga@uem.es

ABSTRACT

MetaMap is an online application that allows mapping text to UMLS Metathesaurus concepts, which is very useful interoperability among different languages and systems within the biomedical domain. MetaMap Transfer (MMTx) is a Java program that makes MetaMap available to biomedical researchers. Currently there is no Spanish version of MetaMap, which difficult the use of UMLS Metathesaurus to extract concepts from Spanish biomedical texts.

Our ongoing research is mainly focused on using biomedical concepts for cross-lingual text classification and retrieval. In this context the use of concepts instead of bag of words representation allows us to face text classification tasks abstracting from the language [4]. 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.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *retrieval models.* H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.* H.4.3 [**Information Systems Applications**]: Communication Applications – *Information Browsers.* J.3 [**Life and Medical Sciences**]: Medical Information Systems. I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Machine Translation*

General Terms

Performance, **Experimentation**, **Languages**.

Keywords

Semantic techniques, data pre-processing, information filtering.

1. INTRODUCTION

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]); 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 [1], and MMTx. The development of equivalent tools 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 including similar concepts than in Spanish.

2. SPANISH MMTx

A first simple approach to Spanish MetaMap uses Google Translator to obtain an English version of the text and then applies English MMTx to extract concepts. This approach presents some important mistakes when translating some technical biomedical terms, keeping them in Spanish.

The second approach delegates on Google Translator in order 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.

- 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.

3. EXPERIMENTS

As we needed to evaluate the suitability of develop a Spanish MetaMap, we designed a set of experiments with this orientation. To test the validity of our 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.

For testing our hypothesis, we needed a corpus of biomedical documents in both languages: Spanish and English. MedLine Plus stores health-related news articles both in English and Spanish. All these news articles are tagged with a set of related MedLine Plus pages, which can be considered as topics or categories.

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).

The main goal of these experiments is to compare the translated documents (ENG_TRANS and ENG_UNMKD) to the baseline (ENG) document collection. MMTx representation to a representation for a given text contains a list of concepts. We have developed 4 possible data representations derived from the MMTS output: A1 (uses compound concepts and ambiguity), A2 (uses compound concepts and no ambiguity), B1 (does not use compound concepts and ambiguity), B2 (does not use compound concepts and no ambiguity). We have also applied Zipf's Law to reduce the number of attributes used for describing each document. The global number of concepts after this filtering process is shown in Table 1.

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

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

3.1 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 similarity between document representations generated from translated texts and the representations generated from the English documents

Doc. Repr.	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

4. DISCUSSION AND FUTURE WORK

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. The use of Zipf's law improves the results within the A representations, while makes the values obtained for B get worse.

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 evaluate other general translation systems, such as Systran or PAHOMTS, automatic translation software maintained by the Pan American Health Organization (PAHO).

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] Aronson, AR., 2001, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the American Medical Informatics Association Symposium, 2001, pp. 17-21.

[2] Bodenreider O, 2004, The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 2004, 32:D267-D270

[3] Carrero García, F., Gómez Hidalgo, J.M., Puertas Sanz, E., Maña López, M., Mata, J., 2007, Attribute Analysis in Biomedical Text Classification. Second BioCreAtIvE Challenge Workshop: Critical Assessment of Information Extraction in Molecular Biology, 2007.

[4] Gómez Hidalgo, J.M., Cortizo Pérez, J.C., Puertas Sanz, E., Ruíz Leyva, M. Concept Indexing for Automated Text Categorization. In NLDB 2004, Lecture Notes in Computer Science, Vol. 3136, Springer, pp. 195-206, 2004.