Bilingual Terminology Acquisition from Unrelated Corpora

Rogelio Nazar
Universitat Pompeu Fabra

This paper presents a simple yet effective technique for the extraction of term equivalents in different languages. In general, techniques for bilingual lexicon extraction have been related to the elaboration of parallel corpora and have yielded accurate results. However, parallel corpora of different domains and languages are not easy to compile. Because of this, some authors have explored techniques to extract a bilingual lexicon from nonparallel but comparable corpora, which are pairs of texts that are not exactly translations of each other but that roughly "talk about the same things". This paper describes an algorithm that performs bilingual terminology extraction without the need of large amounts of data; dealing with infrequent units; needing not the corpora to be comparable nor other resources like an initial bilingual lexicon to use as seed words. In spite of its simplicity, the results of this algorithm are comparable to those of the state of the art techniques, however it supersedes them considering that it offers a domain and language independent method specially suitable for the extraction of specialized terminology, which is the most dynamic part of the lexicon and the most difficult to acquire.

1. Introduction and Related Work

The acquisition of a bilingual lexicon has traditionally been related to the compilation of parallel corpora, which are pairs of texts that are translations of each other. Usually the texts are first aligned at a sentence level and then in a second step a word level alignment is performed. The most common strategy exploited is the cooccurrence of pairs of words in the aligned sentences. If a pair of words cooccurrs more often than expected by chance in the aligned pair of sentences of both languages, then it is to expect that they are translations of each other. This can be calculated using measures of statistical association like Mutual Information or chi-square. This has been the traditional approach since the works of Brown et al. (1990), Gale and Church, (1991) and Gale and Church (1993). These techniques have started an industry of the elaboration of parallel corpora, including mechanisms for the extraction of parallel corpora from the web, like the one proposed by Resnik (1999).

The main limitation of the parallel corpora approach is that to compile them is expensive and time consuming, and moreover it is not always possible, particularly in the case of the extraction of a specialized bilingual lexicon. An alternative for clean parallel corpora is comparable corpora, which are texts of different languages that are not exactly translations of each other but that roughly talk about the same thing. Comparable corpora may be a collection of newspapers in different languages from a same period of time, because they are supposed to share the same international news and events. Unrelated corpora, on the contrary, are texts of different languages that do not hold any kind of relation and are, obviously, the most abundant resource.

The extraction of bilingual equivalents from non-parallel corpora is a much more difficult task, and the techniques from parallel corpora are useless in this scenario. Fung (1995, 1998) and Fung and McKeown (1997), which are among the most cited articles in this field, propose the compilation of vectors for each of the unknown words in the source language and the possible translation candidates in the target language, in order to compute then a similarity score to find the best translation. The vectors are not compiled using simply term frequency (TF) information, because then they would be populated with words that are frequent but not necessarily related, like words of general use. The use of some association measure is a way to cope with this problem. Fung uses Inverse Document Frequency (IDF), thus the term weighting for each value in the vector is TF x IDF. Using an initial general bilingual dictionary as "seed words", every known word in the vector of the source language is translated to the target language, and then the most similar vector from the target language is selected, using cosine as similarity measure. She reports results of a 30% accuracy using only the top candidate, 76% when top 20 are counted and 88% when top 40 are counted.

Fung also tried a measure that she calls "context heterogeneity" as similarity measure between words and their translations in the attempt to compile bilingual lexicon entries from a non-parallel English-Chinese corpus. She explains that there is a limited number of word bigrams (x, W) and (W, y) where W is, for instance, the word "air". The fact that the number of these bigrams is limited indicates what she calls the degree of heterogeneity of that word. Once the vectors are populated, a similarity measure between two context heterogeneity vectors is computed, using this time simple Euclidean distance instead of cosine. In this case she reported results of over 50% accuracy considering the top 10 candidates.

Rapp (1999) uses a somewhat similar approach with a German and English corpus, since the primary resource is word cooccurrence, but he reports better results. His proposal begins by selecting a word of the source language whose translation is to be determined. He computes a cooccurrence vector for that word and, like Fung, translates as much as possible words in that vector to the target language using an initial bilingual dictionary. Naturally, not all the words in the vector find their translation in the dictionary, but they seem to be enough to compute a similarity with all the vectors in a cooccurrence matrix of the target language. He removed multi-word entries from the initial bilingual lexicon, function words from the text and lemmatized the rest of the corpora. As similarity measure, he used the city-block metric, and he used it to compare each of the German vectors to all vectors of the English matrix, selecting the most similar vector as translation candidates. He reports an accuracy of 72% out of a list of 100 test words, considering only the cases where an acceptable translation was ranked first and 89% considering the top 10 candidates.

2. The algorithm

Supposing that English is the source language and Spanish the target language[1], this algorithm consists basically in three simple steps:

Accept an English term as input.

Download 100 documents in Spanish where that English term occurs[2].

See which is the most frequent word in that collection of documents.

3. Further details

Some simple operations can be undertaken to offer a cleaner output, thus easier to read. If no cleaning of the results was made, then one would find many frequent non-content words along with the correct translation, typically function words or n-grams, such as *has been* in English, or *puede* in Spanish. These undesirable lexical units can be easily removed and this section explains how to do it.

3.1. *Filtering non-informative candidates*

One way to filter out non-informative units is using a stoplist and, in addition, a model for the expectation of the frequencies of a word in a certain language. To construct a model for the expected frequency of a word it suffices to take a reference corpus of general language. In this experiment, a corpus of two million words for each of the treated languages proved to be enough. As a result, instead of ranking the words of the downloaded document collection by

[1] This does not mean, of course, that this is the only pair of languages that can be used.

[2] Automatically recognizing the language in which a text is written is a trivial task. It is a particular instance of text categorization. However, even this operation is not necessary, since most search engines allow the selection of the language of the results.

their absolute frequency, terms are weighted dividing the term frequency (TF) by the frequency that the word shows in the model (MF), in the case it exists. In case it does not exists, the MF receives an arbitrary value of one, which means that the TF remains unaffected.

$$W = \log (TF / MF)$$

Equation 1. The term weighting.

In the case of multi-word expressions that are not contained in the model, the computation of its score is just the average of the computation of each of its components. The next subsection explains how to extract these multi-word units.

3.2. *N-gram extraction*

One of the characteristics of a large part of the specialized terminology is that it has many words. As a consequence, one cannot simply split the text in orthographic words because in that way it is likely that the desired term will be destroyed. One way to cope with this difficulty is to perform n-grams extraction sorting them by decreasing frequency. An n-gram is defined here simply as a sequence of words. A bigram is a sequence of two, a trigram of three, etc. The listing considered here included an *N* from 1 to 5. One of the properties of specialized multi-word terminology is that of being consistent, that is, its components tend to remain together. Thus, the sorting of n-grams by frequency helps to highlight these terms, because the rest of the vocabulary, which is in free combination, does not have this binding property.

N-grams of lower *N* that appear more or less the same times than n-grams of higher *N* can be safely removed, since it is not necessary to keep something like “Tunnel Compression Neuropathy” when an equal number of occurrences of “Carpal Tunnel Compression Neuropathy” has already been registered.

3.3. *Elimination of outliers*

Candidates for translation should be as fewer as possible. In section 2 it is said that the most frequent word among the downloaded collection of documents is the desired translation. However, this is taken to be the absolute frequency and does not take into account the possible differences in extension between documents inside that collection.

Terms that occur only in one or two documents are discarded since it is likely that the desired translation will have a larger dispersion in the collection. Here, dispersion (D) is calculated multiplying the term weight (W) by the document frequency (DF, the number of documents where that term occurs).

$$D = W \, . \, DF$$

Equation 2. The term dispersion.

3.4. *Recognition of morphological similarity*

A productive strategy to boost a translation candidate, particularly in the case of specialized terminology, is the use of some kind of similarity measure for the recognition of cognates, which are very frequent in some scientific and technical domains. The algorithm for the recognition of morphological similarity is the following: given to strings of text, it computes a Dice similarity coefficient by transforming the two strings into vectors that have bigrams of characters as components. This device is, though, not going to be useful all the times, since not always the cognate is the best translation. However it can be an extra clue in case the decision could not be made on the basis of the previous steps.

4. Evaluation

The evaluation consisted in a random sample of 100 English medical terms from a database of diseases (Karolinska Institute, online) that contains around 4000 records. The algorithm looped through that list offering the 30 most probable candidates for translation in Spanish. Results where then checked using the Mosby (2001) bilingual dictionary and the MedlinePlus database (online).

Accuracy	*Top N candidates*
57%	top 5
61%	top 10
66%	top 20
69%	top 30

Table 1. Progression of the accuracy as more candidates are considered

English term	*Spanish Equivalent*
Tachycardia	Taquicardia
Constipation	Estreñimiento
Telangiectasis	Telangiectasias
Vaginal Prolapse	Prolapso Vaginal
Pseudotumor Cerebri	Pseudotumor Cerebral
Moebius Syndrome	Síndrome de Moebius
Christmas Disease	Hemofilia
Scapuloperoneal Form of Spinal Muscular Atrophy	NO DATA
Cryptogenic Chronic Hepatitis	NO DATA
Takatsukis Syndrome	NO DATA

Table 2. Examples of correct and incorrect trials.

In 69 cases (69%) the correct translations were among the first 30 candidates. Table 1 shows the progression of the accuracy as more candidates are considered. For illustration, table 2 shows a selection of 7 correct and 3 incorrect cases. In the vast majority of the failure cases the reason was lack of data, meaning that there are still not enough documents on the web where those terms occur. Considering, though, the rate of grow of the web, we can expect that in the near future this information gap will be filled.

5. Conclusions

It is certainly surprising that this simple method can yield so good results without having been previously reported. We may pose it as the following question: why is it that the English term appears in the Spanish text? The explanation for this is probably of a socio-linguistic nature. It is a fact that the English language pervades most modern science. Therefore, we expect that the Spanish speaking scientific community (as well as those of other languages) will often include the English version of the terms they are using. Even when this is not the case, it is highly probable that many titles of the works that appear in the reference section of their Spanish written papers will be in English. That is, it is very likely that a paper devoted for instance to the study of *Breast Cancer*, even if it is written in Spanish, will have a title in its bibliographical reference section with that English term. Another factor is the incidence of the inclusion of an English version of the abstracts in the papers, as well as the keywords after the abstract, both in English and in the language of the paper.

In comparison with other authors' results, even when this performance might not at first seem good enough, there are reasons to believe it is. In first place, the technique is much more simple than those proposed by other authors, it does not require resources and steps like comparable corpora acquisition, lemmatization, availability of initial bilingual lexicons, etc. Another difference is that this technique is not limited to the study of frequent lexical units, and this is good because the real challenge is not on the frequent words but in the most rare ones. In addition, here we deal with specialized and multi-word lexical units that are precisely the most valuable material for translation and terminology and not with units of the general language, that are already contained in most dictionaries.

These results could have been undoubtedly better if a basic bilingual lexicon had been used, because in that case we could have translated common terms such as *syndrome* for *síndrome; disease* for *enfermedad*, etc., and in that way we could have easily recognized the relation that holds between, for example, *Reiter Syndrome* and *Síndrome de Reiter.* However, the purpose of this paper was to show a language independent and domain independent methodology. Language and domain specific strategies can be valuable from an engineering perspective, but a broader generalization is preferable from a scientific point of view.

As a final comment, even if this performance is still not useful from a practical point of view, the paper would still be a valuable contribution from a theoretical perspective, since it shows how equivalent terms in different languages are statistically related, a fact about language that had gone unnoticed until now. There is a possible explanation for this too, and that is the already mentioned size of the Internet. Surely, this same technique could not have worked well ten years ago, when there was still not so much information available on the web, and probably the same method will work better a few years from now. In the same way, results may differ in other languages, like French or German. Replication of the experiment with those languages should be one of the future lines of research.

6. Future work

There is one obvious drawback with the methodology exposed so far, and it is that it will only work with specialized terminology, terms that have a precise denotation or that refer to a specific concept, such as *Rheumatoid Arthritis.* It will probably not work for terms that are not specialized or that have a predicative function instead of a denotation, like, for example, the term *arrange*. The problem with this type of terms, single word predicates with high frequency and a considerable amount of polysemy, is that in certain technical domains they may have a specific equivalent. For example, in certain domains a term like this one may usually be translated as *situado* in Spanish and not *colocado* or *dispuesto*, which are synonyms when considered out of context.

There is already a line of research and another paper (whose title should be something like "Two Step Flow in Bilingual Lexicon Extraction from Unrelated Corpora") that explains a method to tackle with this problem. This new methodology is slightly more complex than the present one because, as the title suggests, it involves two different kinds of actions. The first step would be to disambiguate the term in the source language and the second would be to find the equivalent, with basically the same methodology explained in this paper.

The process of disambiguation requires more input than just the single term that we are trying to translate. For this we need the term and a collection of documents of the proper technical domain where this term occurs. The purpose of this collection of documents in the source language is to see with which other terms the term that we are trying to translate is related. That is to say, in which context of occurrence it is usually found. To give an example, the word *light* may have different meanings depending on the context. But if we provide examples of documents of the domain we are studying, our algorithm will be able to find some of the multi-word expressions where this word *light* is contained, for example, *light beam*; *light source*; *light spot*; *incident light*; etc. If we loop through all these examples applying the same methodology that we have explored in this paper, then we will find that there is one element that is recurrent across all the set of candidates that each of these terms will have generated, and that element will be *luz*, which is the correct Spanish equivalent in the technical domain in question.

There are also other possibilities, like seeing if the immediate context (10 or 15 words at each side) of a certain equivalent candidate that we are studying is or not in the target language. If it is, then the probability of being the correct translation increases. Another possibility is to observe which is the relation that holds between the frequency of the input term in a reference corpus of the source language and the frequency of a translation candidate in a reference corpus of the target language. For instance, if we started with an English term as input such us *pulmonary sarcoidosis*, we obtain the following ordered set of equivalent candidates: {*pulmonar*; *sarcoidosis pulmonar*; *sarcoidosis*; *pulmonary*; *universitario*; *pulmonary sarcoidosis*; *dermatol*; *hospital*; *therapy*; *hospital universitario*; *disease*}. We can expect that the input term and the correct translation may have a similar frequency of occurrence in their respective reference corpora, null frequency in this case. Thus we will be able to eliminate false candidates like *disease*; *hospital*; *therapy* and *hospital universitario* because they have a much greater frequency.

The final move, and the most computationally expensive, would be the following: given that we have obtained a set of equivalent candidates in the target language for each input term in the source language, in order to find the correct translation among that set we should then reiterate the process for each of the candidates but now in the opposite direction. How can we automatically infer that *sarcoidosis pulmonar* is the correct translation? Because if we start the process again with this Spanish term as input to obtain the English equivalent, then we observe that our initial term, *pulmonary sarcoidosis*, is one of the candidates, which does not happen with the rest of the Spanish candidates that we had initially obtained for *pulmonary sarcoidosis.*

The combination of these as well as other strategies has given better results than the one reported in this paper, but at the cost of a more complex procedure and longer processing time. Being therefore a different approach, it justifies a new paper and not just an extension of the present one.

References

Brown, P. F. et al. (1990). "A Statistical Approach to Machine Translation". *Computational Linguistics* 16. 79-85.

Fung, P. (1995). "Compiling Bilingual Lexicon Entries from a Non-Parallel English-Chinese Corpus". In *Proceedings of the Third Workshop on Very Large Corpora*. 173-183.

Fung, P. (1998). "A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora". In *Proceedings. of the Third Conference of the Association for Machine Translation in the Americas*. 1-16.

Fung, P.; McKeown, K. (1997). "Finding Terminology Translations From Non-Parallel Corpora". In *The 5th Annual Workshop on Very Large Corpora*. Hong Kong. 192-202.

Gale, W.; Church, K. (1991). "Identifying word correspondences in parallel texts". In *Proceedings of the DARPA Workshop on Speech and Natural Language.*

Gale, W.; Church, K. (1993). "A Program for Aligning Sentences in Bilingual Corpora". *Computational Linguistics* 19 (1). 75-102

Karolinska Institute (online). *Alphabetic List of Specific Diseases/Disorders*. *http://www.mic.ki.se/Diseases/Alphalist.html* [Acces date: 23 January 2008].

Medline Plus (online). U.S. National Library of Medicine and the National Institutes of Health *http://medlineplus.gov* [Access date: 26 March 2008].

Mosby (2001). *Diccionario Mosby de Medicina, Enfermería y Ciencias de la salud*. 5th edition. Harcourt: Madrid. Spanish version of the original in English: *Mosby's Medical, Nursing, and Allied Health Dictionary*. Mosby-Year Book, Inc.

Rapp, R. (1999). "Automatic Identification of Word Translations from Unrelated English and German Corpora". In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*. 519-526.

Resnik, P. (1999). "Mining the Web for Bilingual Text". In *37th Annual Meeting of the Association for Computational Linguistics* (ACL'99), College Park, Maryland, June 1999.