A Novel Approach to Improve Word Translations Extraction from Non-Parallel, Comparable Corpora

Yun-Chuang Chiao  
STIM/DSI  
& ERM202 INSERM, Paris  
ycc@biomath.jussieu.fr

Jean-David Sta  
EDF - R & D  
92410 Clamart  
jean-david.sta@edf.fr

Pierre Zweigenbaum  
STIM/DSI  
& ERM202 INSERM, Paris  
pz@biomath.jussieu.fr

Abstract

After an attempt at extracting translational equivalents from domain-specific comparable corpora of limited size, we propose in this paper an approach to prune translation alternatives. In particular, we re-score translation candidates in the target language by applying the same translation algorithm in the reverse direction and re-ranking them according to the harmonic mean score. After a preliminary investigation limited to identifying translations of most common words, we validated in this paper our proposed model on a larger scale, including less frequent terms in both corpora. The results show an improvement in precision of the top candidate translation. The translational equivalents obtained may then be used, e.g., for extending an existing medical lexicon as human translation aid or for query expansion and translation in cross-language information retrieval.

1 Introduction

World alignment is a well-studied problem in Natural Language Processing and has been used in many applications such as translation lexicon acquisition, statistical machine translation, and also cross language information retrieval. Manual alignment of bilingual data is a labor-intensive process and for applications such as bilingual lexicon construction, human compiled dictionaries were often out-of-date as soon as they became available. Recent advances in automatic lexicon extraction and statistical alignment algorithms allow us to build models which can identify translation equivalents at the word or phrase level. Such techniques can be useful especially for technical domains which are in constant evolution, producing new terms. Many of related works on using statistical models for mapping bilingual terms (Hiemstra et al., 1997; Littman et al., 1998) are typically based on parallel texts or 'bitexts' - pairs of texts that are translation of each other. Most of those methods are based on the following assumption: words that are translations of each other are more likely to appear in corresponding parallel text regions than other pairs of words. By using various correlation metrics, these approaches derive co-occurrence patterns of words across languages. The limit is that large-scale parallel corpora are not always available, although (Chen and Nie, 2000)’s experiments reveal a potential solution by automatically collecting parallel Web pages. Therefore, it seems natural to enlarge the scope of corpus resources by looking for non-parallel, comparable corpora. Comparable corpora are a ripe area of investigation in the development of bilingual lexicons (Fung and Yee, 1998; Rapp, 1999; Picchi and Peters, 1998). However, these experiments dealt with very large, ‘general language’ corpora and assume the availability of NLP tools such as POS tagger, morphological analyzer, etc. This paper addresses this issue in the medical domain and describes a model where these resources are assumed unavailable.

After a preliminary investigation limited to identi-
fying translations of most common words from comparable medical corpora, we validated in this paper our proposed model on a larger scale, including less frequent terms in both corpora. We also varied the context window size, with consideration of sentence boundaries, to look at the different types of semantic relationships between words. Once a ranked list of translation candidates is generated for a given source word, we applied the same translation algorithm to each of top ranked candidates. We then re-ranked translation candidates according to the harmonic mean of their original and ‘reverse’ ranks with a given source word.

The translational equivalents obtained may then be used, e.g., for extending an existing medical lexicon as human translation aid or for query expansion and translation in cross-language information retrieval. In the following section, we first recall related work on this topic. After describing the corpora we used and their characteristics, we overview our algorithm, which is validated on all lexicon words. We then provide and discuss the results. Finally, we describe future directions.

2 Background

Compared with other approaches to use comparable corpora for word to word translation, our work is mostly related to research on alignment of non-parallel texts at the word level and research on domain-specific bilingual lexicon acquisition. Previous works in this area are based on the assumption that words which have the same meaning in different languages should have similar context distributions.

Rapp (1999) proposes an approach very similar to the method presented here. He supposes that in any language there is a correlation between the cooccurrences of words which are translations of each other. The main difference between our approach and his model is that he used lemmatized corpora and limited the number of translation candidates considered. Fung and Yee (1998) proposed a method based on the vector space model for translating new words in non-parallel, Chinese English comparable corpora. They claim that the association between words and their ‘context seed words’ are preserved in comparable texts. By designing procedures to retrieve cross-lingual lexical equivalents together, Picchi and Peters (1998) proposed that their system could have applications such as retrieving documents containing terms or contexts which are semantically equivalent in more than one language.

3 Material

The material prepared for the present experiments consists of non parallel medical corpora in French and English and a bilingual, French-English combined lexicon including both general and medical words. The two monolingual corpora have been compiled from Internet catalogs of medical websites, CISMeF (Darmoni et al., 2000) (www.chuerouen.fr/cismef) for the French language and ClinWeb (Hersh et al., 1999) (www.ohsu.edu/clinweb for English. We chose a common domain, corresponding to the subtree under the MeSH concept ‘Pathological Conditions, Signs and Symptoms’ (‘C23’), which is the best represented in CISMeF; automatically downloaded the pages indexed by these catalogs and converted them into plain text.

The ‘French’ corpus contains many foreign words (mainly English and Spanish) since we did not have language filtering on it. Besides, the two corpora should be considered as unrelated rather than comparable from the point of view of their size. The French corpus presents 7,604,381 word tokens and the English one contains 639,662 tokens. However, as explained in Diab and Finch (2000, p. 1501), one does not need to have corpora of the same size for this kind of approach to work.

A combined French-English lexicon of simple words was compiled from several sources: for the medical domain, an online French medical dictionary (Dictionnaire Médical Masson, www.atmedica.com) and the English-French biomedical terminologies in the UMLS metathesaurus (NLM, 2001): MeSH, WHOART and ICPC; for general words, we used the French-English dictionary distributed in the Linux package dictdictionaries. The resulting lexicon contains 22,036 (‘one-word term’) entries, mainly specialized medical words (see excerpt in table 1).

4 Methods

Figure 1 shows a schema of the translation model, which we summarize in the rest of this section. The
Figure 1: Schema for the translation model

| source corpus | source weighted vectors | filtering | source vectors | partial bilingual lexicon | comparing source to target | comparing target to source | target weighted vectors | transferring | target vectors | foie: liver bile dubus ... ...

Table 1: Lexicon excerpt

| basic intuition is that there is a correlation between the context distribution of words which are translations of each other. The approach is an attempt at finding the target words whose distributions are the most similar to that of a given source word. The method depends primarily on co-occurrence information of collocate terms. No morphological analysis is applied to either corpus during the experiment.

We achieve this by approximating the distributional behavior through context vectors and finding a mapping of source and target words which preserves the context mapping as much as possible. Therefore, our goals are: (a) to build a context vector of each word within each of the corpora; (b) to produce translation candidates using similarity score; and (c) to provide an algorithm for re-ranking the possible matched word vectors between corpora.

### 4.1 Computing context vectors

For each word $i$ in source and target language corpora separately, we first create a context vector which consists of its co-occurrence patterns. Stop words are eliminated from both corpora for co-occurrence counting. With a simple heuristic to identify sentences boundaries (with punctuation marks such as !, . or ?), we defined two alternate sliding context windows of 5 and 7 words (table 2) to calculate the cooccurrences of $i$. A lemmatization is applied to each co-occurrence pattern. Since this lemmatizer does not handle gender nor verb inflection, this lemmatization is far from perfect.

The window size parameter allows us to look at different scales (Church and Hanks, 1990). A smaller window size will identify fixed expressions and other relations such as syntactic dependencies; a larger window size will highlight semantic concepts and other relationships that hold over a larger range. In order to take into account word-frequency effects and to emphasize significant word associations, we chose the following $tfidf$ (Sparck Jones, 1979) as weighting measure:

$$tf(idf(i, j)) = tf(i, j)idf(i)$$

$$tf(i, j) = \frac{cooc(i, j)}{\max_{k, l} cooc(k, l)}$$

$$idf(i) = 1 + \log \frac{\max_{k, l} cooc(k, l)}{|k:cooc(i, k)\neq 0|}$$

### 4.2 Transferring context vectors through pivot words

When a translation is sought for a source word, its context vector is translated into the target language, using the bilingual lexicon. Only the words in the bilingual lexicon (the ‘pivot’ words) can be used in the transfer. Several translations may exist; pending a better solution, we decided to insert only the first one into the context vector. The result is a target-language context vector, comparable to ‘native’ context vectors directly obtained from the target corpus (table 3). Since we want to compare transferred context vectors with native context vectors, these two
 sorts of vectors should belong to the same space, i.e., range over the same set of context words. Using the bilingual lexicon, we reduced the context vector space to the set of ‘cross-language seed words’. A word belongs to this set if it occurs in the target corpus, is listed in the bilingual lexicon and its source counterpart(s) occurs in the source corpus. In our experimental setting, 6,210 pivot words are used.

### 4.3 Computing vector similarity

Given a transferred context vector, for each native target vector, a similarity score is computed and target vectors are ranked. The best-ranked target words are considered as translation candidates. The Jaccard (Romesburg, 1990) similarity metric\(^1\) is used for comparing two vectors \(V\) and \(W\) (of length \(n; k, l, m\) range from 1 to \(n\)):

\[
\text{Jaccard}(V, W) = \frac{\sum_k v_k w_k}{\sum_k v_k + \sum_l w_l - \sum_m v_m w_m}
\]

### 4.4 Re-ranking translation candidates

After the translation algorithm generates a ranked list of translation candidates for a given source word, the same model is applied in the reverse direction to find the source counterparts of translation candidates. For each of the top ranked candidates, a return ranked list of its source counterparts is computed. We then re-score the translation candidates according to the following harmonic mean measure\(^2\):

\[
\text{HM}(r_1, r_2) = \frac{2r_1 r_2}{r_1 + r_2}
\]

where \(r_1\) is the original rank of a target word \(i\) given a source word \(j\); \(r_2\) is the rank of word \(j\) for \(i\) obtained by the ‘reverse’ translation module.

-----

\(^1\)In preliminary experiments, we also computed the vector similarity with different measures such as cosine, Dice, city-block and found that the Jaccard metric yielded the best results.

\(^2\)http://mathworld.wolfram.com/HarmonicMean.html
posals of the algorithm (target context vectors with their target word). Test words range over the set of known words (or ‘pivot’ words) \( P \).

We also tested an ‘unknown’ set \( U \) which contains candidate target context vectors of all words in the corpora which are not present in our lexicon. Assuming that we know all the words in our lexicon but one (the test word), we only need to look for a translation among the set \( U \) of ‘unknown’ target words – augmented with the expected translation.

It might be objected that known words (those found in our lexicon) should be expected to be more frequent than ‘new’ or ‘unknown’ words; and that this difference in frequency might favorably bias the discrimination of a test word from actually unknown words. To check this, we compared the distributions of frequency ranks of known and unknown words in the source corpus. These ranks are obtained by ordering all the words in the corpus (recall that stop words have been removed) by descending frequency: from rank 1 (corresponding to the most frequent words – frequency range = 35,000 \( \pm 10\% \)) to rank 89 (words which appear only once in the corpus, or ‘hapaxes’). Figure 2 plots the relative distribution of known words in the different ranks and that of unknown words in these ranks. It shows that although this tendency exists, many known words (‘pivot’ words) are present at large ranks (are rare in the corpus: 12% are hapaxes) and a few words absent from the lexicon are present in the top ranks (are frequent in the corpus). Let us recall too that grammatical words, which are the most frequent in a corpus, have been removed as part of stop words.

Furthermore, we chose two different sets of French test words for which we try to identify the possible translations in our English corpus. The specialized medical set \( M \) consists of 886 medical one-word terms extracted from the SNOMED Micrglossary for Pathology (Côté, 1996). The global, mixed set \( P \) of all lexicon words consists of 6,210 medical and general one-word terms.

5 Results

For each test word in \( P \), we produced a list of its translational equivalents ranked in decreasing order of similarity score. The rank \( R \) of its expected translations provides the basis for evaluation. Sample results are provided in table 4, showing the top-ranked candidate translations for French words nécrose, gène, sclérose and abcès.

5.1 Effects of reverse translation

As we described earlier in this paper, our algorithm first generates a ranked list of translation candidates, then re-scores them by applying the same algorithm to compute the list of source counterparts for each of the top 30 candidates. Table 5 shows the re-ranking results for words in table 4. As we can see in this example, when applying ‘reverse’ translation, the expected translations for sclérose (sclerosis) and abcès (abscess) are ranked first in the re-scored list against 3 and 9 in the original ranked list.

Figures 3 and 4 show a detailed comparison of the translation accuracy, within the top 10 ranked translation candidates, for different test sets \( M \) and \( P \) with different context window sizes. The results are similar for both pivot word set \( P \) and medical word set \( M \), whether with 5 or 7-word windows, reverse translation improves precision. With 5-word window, about 43% of test words are correctly translated when applying reverse translation against 39% before reverse translation for \( M \) set (3); and for \( P \) set (3), 32% after reverse translation against 30% before. With 7-word window, 34% of test words have their expected translation as first ranked can-
Table 4: Example results; Occ = occurrences in corpus; R = original rank of expected target English word before ‘reverse’ translation

<table>
<thead>
<tr>
<th>Fr word</th>
<th>En word</th>
<th>Occ.</th>
<th>R</th>
<th>Top 5 ranked candidate translations, followed by similarity score</th>
</tr>
</thead>
<tbody>
<tr>
<td>nécrose</td>
<td>necrosis</td>
<td>667</td>
<td>1</td>
<td>necrosis .181, chronic .148, renal .142, inflammation .135, infarction .123</td>
</tr>
<tr>
<td>gène</td>
<td>gene</td>
<td>1,813</td>
<td>1</td>
<td>gene .247, mutation .243, recessive .197, protein .194, chromosome .145</td>
</tr>
<tr>
<td>sclérose</td>
<td>sclerosis</td>
<td>347</td>
<td>9</td>
<td>sep .263, lateral .187, sclerosis .178, passe .177, poliovirus .158</td>
</tr>
<tr>
<td>abcès</td>
<td>abscess</td>
<td>564</td>
<td>9</td>
<td>perforation .227, rupture .192, visible .156, invasive .153, impose .149</td>
</tr>
</tbody>
</table>

Table 5: Example results; R = new rank of expected target English word after ‘reverse’ translation

<table>
<thead>
<tr>
<th>Fr word</th>
<th>En word</th>
<th>R</th>
<th>Top 5 ranked candidate translations, followed by harmonic mean score</th>
</tr>
</thead>
<tbody>
<tr>
<td>nécrose</td>
<td>necrosis</td>
<td>1</td>
<td>necrosis 1, chronic 3.74, renal 4.2, inflammation 5.53, infarction 10</td>
</tr>
<tr>
<td>gène</td>
<td>gene</td>
<td>1</td>
<td>gene 1, mutation 1.33, protein 2.66, chromosome 2.85, recessive 3</td>
</tr>
<tr>
<td>sclérose</td>
<td>sclerosis</td>
<td>1</td>
<td>sclerosis 1.5, sep 2, lateral 2.66, passe 8, poliovirus 10</td>
</tr>
<tr>
<td>abcès</td>
<td>abscess</td>
<td>1</td>
<td>abscess 1.8, perforation 2, rupture 4, visible 6, invasive 8, impose 10</td>
</tr>
</tbody>
</table>

Figure 3: Comparison of translation accuracy, before and after reverse translation, between medical word set $M$ and pivot word set $P$, with 5-word context window: $y = \%$ percentage of the words ranked in $x$ rank.

Discussion

The reverse translation method we proposed significantly improves the translation performance. The main reason seems to be that it eliminates general words, which are therefore relatively more ambiguous, from the translation candidate list. For a given French word abcès of which the expected translation in English is abscess, a general English word such as perforation might have a higher similarity score than abscess since general words are assumed to be frequent, then the chances of their co-occurrence with a given word would be higher than that of specific words. With the reverse translation method, the sim-
Figure 4: Comparison of translation accuracy, before and after reverse translation, between $M$ and $P$, with 7-word context window.

Figure 5: Comparison between pivot word set $P$ and medical word set $M$: $y =$ the percentage of the expected translation ranked among top 10 candidates in function of decreasing frequency distributions $x$.

Similarity score for abscess and abcès is highest while it is lower for perforation and abcès because perforation could be similar to other words. We therefore expect that this method could highlight the ‘good’ translation candidate by filtering out those which are ambiguous.

The better results on the smaller, 5-word window may be interpreted as a bonus to syntactic dependencies, which may be expected to be more prevalent at short distance than at a longer distance. This prompts us to investigate an even shorter, 3-word window in our next experiments, as well as the application of a parser to identify actual functional dependencies.

It is not surprising that our method fares better on more frequent words than on less frequent ones since the alignment algorithm is based on the cooccurrence frequency. A more common word is assumed to have a sufficient number of contexts, which is favorable to this kind of measure. However the performance on the medical word set $M$ looks promising, especially when compared with the pivot word set $P$, as we can observe a clear improvement of results whether the word to be aligned is frequent or not in the French corpus. This might be due to the fact that domain-specific terms are generally less ambiguous in corpora of the same domain since their meanings are restricted to that definition. Another reason that might help to explain the less performing results on the pivot word set $P$ might be the lack of general word pairs in our corpora, especially for the English corpus since it contains less word occurrences than the French one.

These better results at higher percentiles might be linked to a better contrast between a specific test word and general unknown words in the test conditions where the candidate set contained unknown and relatively rare words.

On a ‘general-language’ corpus, Rapp (Rapp, 1999) reports an accuracy of 65% at the first percentile by using loglike weighting and city-block metric, whereas neither of these improved our results. A larger size for the corpora (135 and 163 Mwords) and the consideration of word order within contexts may help to explain this difference in accuracy.
7 Conclusion and Perspectives

In summary, these experiments confirm a positive effect of reverse translation on the suggestion of appropriate translational equivalents for medical words. They show that medical words in this corpus are better handled than less specialized words, and show a clear influence of context window size.

Our proposed approach relies on an initial bilingual lexicon to build context vectors. (Chiao and Zweigenbaum, 2003) showed that the performance of such statistical translation model can be improved by adding general words in the domain-specific lexicon. We would like to test our method further by counting only cooccurrences with general words in the context.

The main limitation of the present work lies in the moderate corpus size, which limits the frequency and diversity of its words, so that an insufficient number of word pairs may be aligned by the proposed algorithm. We should investigate the effect of very large corpora, for which the Web has a vast potential of resources.

Further investigations must now obtain better performance for all types of words including the less frequent terms. Several directions are still open for investigation, among which selecting words with the same part of speech as the source word, boosting morphologically similar candidates (‘cognates’) or enlarging the size of corpora.

Acknowledgements

We thank Julien Quint and Salah Ait Mokhtar for their help during this work.

References


Roger A. Côté. 1996. Répertoire d’anatomopathologie de la SNOMED internationale, v.3.4. Université de Sherbrooke, Sherbrooke, Québec.


