Looking for candidate translational equivalents in specialized, comparable corpora

Yun-Chuang Chiao and **Pierre Zweigenbaum**
STIM/DSI, Assistance Publique – Hôpitaux de Paris &
Département de Biomathématiques, Université Paris 6

Abstract

Previous attempts at identifying translational equivalents in comparable corpora have dealt with very large 'general language' corpora and words. We address this task in a specialized domain, medicine, starting from smaller non-parallel, comparable corpora and an initial bilingual medical lexicon. We compare the distributional contexts of source and target words, testing several weighting factors and similarity measures. On a test set of frequently occurring words, for the best combination (the Jaccard similarity measure with or without $tf.idf$ weighting), the correct translation is ranked first for 20% of our test words, and is found in the top 10 candidates for 50% of them. An additional reverse-translation filtering step improves the precision of the top candidate translation up to 74%, with a 33% recall.

1 Introduction

One of the issues that have to be addressed in cross-language information retrieval (CLIR, Grefenstette (1998b)) is that of query translation, which relies on some form of bilingual lexicon. Methods have been proposed to acquire a lexicon from corpora when such a lexicon does not exist or is not complete enough (Fung and McKeown, 1997; Fung and Yee, 1998; Picchi and Peters, 1998; Rapp, 1999). The present work addresses this issue in a specialized domain: medicine. We aim at identifying French-English translation candidates from comparable medical corpora, extending an existing specialized bilingual lexicon. These translational equivalents may then be used, *e.g.*, for query expansion and translation.

We first recall previous work on this topic, then present the corpora and initial bilingual lexicon we start with, and the method we use to build, transfer and compare context vectors. We finally provide and discuss experimental results on a test set of French medical words.

2 Background

Salton (1970) first demonstrated that with carefully constructed thesauri, cross-language retrieval can perform as well as monolingual retrieval. In many experiments, parallel corpora have been used for training statistical models for bilingual lexicon compilation and disambiguation of query translation (Hiemstra et al., 1997; Littman et al., 1998). A limiting factor in these experiments was an expensive investment of human effort for collecting large-size parallel corpora, although Chen and Nie (2000)'s experiments show a potential solution by automatically collecting parallel Web pages.

Comparable corpora are "texts which, though composed independently in the respective language communities, have the same communicative function" (Laffling, 1992). Such non-parallel texts can become prevalent in the development of bilingual lexicons and in cross-language information research as they may be easier to collect than parallel corpora (Fung and Yee, 1998; Rapp, 1999; Picchi and Peters, 1998). Among these, Rapp (1999) proposed that in any language there is a correlation between the cooccurrences of words which are translations of each other. Fung and Yee (1998) demonstrated that the associations between a word and its context seed words are preserved in comparable texts of different languages. By designing procedures to retrieve crosslingual 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 Collecting comparable medical corpora

The material for the present experiments consists of comparable medical corpora in French and English and a French-English medical lexicon (Fung and Yee (1998) call its words 'seed words').

3.1 'Signs and Symptoms' Corpora

We selected two medical corpora from Internet catalogs of medical web sites. Some of these catalogs index web pages with controlled vocabulary keywords taken from the MeSH thesaurus (www.nlm.nih.gov/mesh/meshhome), among which CISMeF (French language medical web sites, www.chu-rouen.fr/cismef) and CliniWeb (English language medical web sites, www.ohsu.edu/cliniweb). The MeSH thesaurus is hierarchically structured, so that it is easy to select a subfield of medicine. We chose the subtree under the MeSH concept 'Pathological Conditions, Signs and Symptoms' ('C23'), which is the best represented in CISMeF.

We compiled the 2,338 URLs indexed by CISMeF under that concept, and downloaded the corresponding pages, plus the pages directly linked to them, so that framesets or tables of contents be expanded. 9,787 pages were converted into plain text from HTML or PDF, yielding a 602,484-word corpus (41,295 unique words). The initial pages should all be in French; the additional pages sometimes happen to be foreign language versions of the initial ones. In the same line, we collected 2,019 pages under 921 URLs indexed by CliniWeb, and obtained a 608,320-word English medical corpus (32,919 unique words).

3.2 Base bilingual medical lexicon

A base French-English lexicon of simple words was compiled from several sources. On the one hand, an online French medical dictionary (Dictionnaire Médical Masson, www.atmedica.com) which includes English translations of most of its entries. On the other hand, some international medical terminologies which are available in both English and French. We obtained these from the UMLS metathesaurus, which includes French versions of MeSH, WHOART, ICPC and their English counterparts (www.nlm.nih.gov/research/umls). The resulting lexicon (see excerpt in table 1) contains 18,437 entries, mainly specialized medical terms. When several translations of the same term are available, they are all listed.

abarognosie	abarognosis
abarthrose	abarthrosis
abarticulaire	abarticular
abasie	abasia
abattement	prostration
abaxial	abaxial
abcédé	abscessed
abcès	abscess
abdomen	abdomen, belly
abdominal	abdominal
abdomino-génital	abdominogenital
abdomino-thoracique	abdominothoracic
abdomino-vésical	abdominovesical
abducteur	abducens, abducent

Table 1: Lexicon excerpt

4 Methods

The basis of the method is to find the target words that have the most similar distributions with a given source word. We explain how distributional behavior is approximated through context vectors, how context vectors are transferred into target context vectors, and how context vectors are compared.

4.1 Computing context vectors

Each input corpus is segmented at non-alphanumeric characters. Stop words are then removed, and a simple lemmatization is performed. For English, we used a list of stop words that we had from a former project. For French, we merged Savoy's online stop words list (www.unine.ch/info/clef) with a list of our own. The S-stemmer algorithm (Harman, 1991) was applied to the English words. Another simple stemmer was used for French; it handles some *-s* and *-x* endings.

The context of occurrence of each word is then approximated by the bag of words that occur within a window of N words around any occurrence of that 'pivot' word. In the experiments reported here, N was set to 3 (*i.e.*, a seven-word window) to approximate syntactic dependencies. The context vector of a pivot word j is the vector of all words in the corpus,[1] where each word i is represented by its number of occurrences occ_i^j in that bag of words.

A context vector is similar to a document (the document that would be produced by concatenating the windows around all the occurrences of the given pivot word). Therefore, weights that are used for words in documents can be tested here in order to eliminate word-frequency effects and to emphasize significant word pairs. Besides simple context frequency occ_i^j, two additional, alternative weights are computed: $tf.idf$ and log likelihood.

[1] We shall see below that actually, only a subset of the corpus words will be kept in each vector.

The formulas we used to compute $tf.idf$ are the following: the normalized frequency of a word i in a context j is $tf_i^j = \frac{occ_i^j}{\max occ}$ where occ_i^j is the number of occurrences of word i in the context of j and $\max_{occ} = \max_{ij} occ_i^j$ is the maximum number of cooccurrences of any two words in the corpus; $idf_i = 1 + \log \frac{\max_{occ}}{occ_i}$ (Sparck Jones, 1979) where occ_i is the total number of contexts in which i occurs in the corpus.

For the computation of the log likelihood ratio, we used the following formula from Dunning:[2]

$loglike(a,b) = \sum_{ij} \log \frac{k_{ij}N}{C_i R_j} = k_{11} \log \frac{k_{11}N}{C_1 R_1} + k_{12} \log \frac{k_{12}N}{C_1 R_2} + k_{21} \log \frac{k_{21}N}{C_2 R_1} + k_{22} \log \frac{k_{22}N}{C_2 R_2}$;

$C_1 = k_{11} + k_{12}$, $C_2 = k_{21} + k_{22}$, $R_1 = k_{11} + k_{21}$, $R_2 = k_{12} + k_{22}$, $N = k_{11} + k_{12} + k_{21} + k_{22}$;

k_{11} = # cooccurrences of word a and word b,

$k_{12} = occ_a - k_{11}$, $k_{21} = occ_b - k_{11}$,

k_{22} = corpus size – k_{12} – k_{21} + k_{11}.

At the end of this step, each non-stop word in both corpora has a weighted context vector.

4.2 Transferring context vectors

When a translation is sought for a source word, its context vector is transferred into a target language context vector, relying on the existing bilingual lexicon. Only the words in the bilingual lexicon can be used in the transfer. When several translations are listed, only the first one is added to the target context vector. The result is a target-language context vector which is comparable to 'native' context vectors directly obtained from the target corpus.

Let us now be more precise about the context-word space. Since we want to compare context vectors obtained through transfer 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. A (target) word belongs to this set iff (i) it occurs in the target corpus, (ii) it is listed in the bilingual lexicon, and (iii) (one of) its source counterpart(s) occurs in the source corpus. This set corresponds to the 'seed words' of Fung and Yee (1998). Therefore, the dimension of the target context vectors is reduced to this set of 'cross-language pivot words'. In our experimental setting, 4,963 pivot words are used.

4.3 Computing vector similarity

Given a transferred context vector, for each native target vector, a similarity score is computed; a ranking list is built according to this score. The target words that 'own' the best-ranked target vectors are the words in the target corpus whose distributions with respect to the bilingual pivot words are the most similar to that of the source word; they are considered candidate translational equivalents.

We used several similarity metrics for comparing pairs of vectors V and W (of length n): Jaccard (Romesburg, 1990) and cosine (Losee, 1998), each combined with the three different weighting schemes. With k, l, m ranging from 1 to n:

$$Jaccard(V,W) = \frac{\sum_k v_k w_k}{\sum_k v_k^2 + \sum_l w_l^2 - \sum_m v_m w_m}$$

$$\cos(V,W) = \frac{\sum_k v_k w_k}{\sqrt{\sum_k v_k^2}\sqrt{\sum_l w_l^2}}$$

4.4 Experiments

The present work performs a first evaluation of this method in a favorable, controlled setting. It tests, in a 'leave-one-out' style, whether the correct translation of one of the source (French) words in the bilingual lexicon can be found among the target (English) words of this lexicon, based on context vector similarity. To make similarity measures more reliable, we selected the most frequent words in the English corpus ($N_{occ} > 100$) whose French translations were known in our lexicon. Among these, we chose the most frequent ones ($N_{occ} > 60$) in the French corpus. This provides us with a test set of 95 French words (i) which are frequent in the French corpus, (ii) of which we know the correct translation, and (iii) such that this translation occurs often in the English corpus. For each of the French test words, we computed a weighted context vector for each of the different weighting measures (occ_i^j, $tf.idf$, log likelihood). Then, using the above-mentioned similarity measures (cosine, Jaccard), we compared this weighted vector with the set of cross-language pivot words's context vectors computed from the English corpus. We then produced a ranked list of the top translational equivalents and tested whether the expected translation can be differentiated from other well-known domain words. For the evaluation, we computed the rank of the expected translation of each test word and synthesized them as a percentile rank distribution.

5 Initial Results

Table 2 shows example results for the French words *anxiété* and *infection* with different weightings and similarity measures. For reasons of space, we only

[2] Posted on the 'corpora' mailing list on 22/7/1997 (helmer.hit.uib.no/corpora/1997-2/0148.html).

Meas.	Weight	Fr word	En word	R	Top 5 ranked candidate translations
Cos.	occ_i^j	anxiété	anxiety	1	anxiety .55, depression .45, medication .36, insomnia .36, memory .34
Cos.	$tf.idf$	anxiété	anxiety	1	anxiety .54, depression .41, eclipse .33, medication .29, psychiatrist .29
Cos.	loglike	anxiété	anxiety	1	anxiety .56, depression .43, eclipse .37, psychiatrist .36, dysthymia .33
Jac.	occ_i^j	anxiété	anxiety	2	memory .21, anxiety .21, insomnia .19, confusion .19, psychiatrist .18
Jac.	$tf.idf$	anxiété	anxiety	1	anxiety .21, psychiatrist .17, confusion .15, memory .14, phobia .14
Jac.	loglike	anxiété	anxiety	1	anxiety .26, psychiatrist .19, memory .15, phobia .14, depressed .14
Cos.	occ_i^j	infection	infection	2	infected .55, infection .52, neurotropic .47, homosexual .43
Cos.	$tf.idf$	infection	infection	3	infected .56, neurotropic .49, infection .48, aids .45, homosexual .41
Cos.	loglike	infection	infection	2	infected .67, infection .55, neurotropic .53, aids .48, homosexual .48
Jac.	occ_i^j	infection	infection	1	infection .33, aids .21, tract .17, positive .16, prevention .15
Jac.	$tf.idf$	infection	infection	1	infection .27, aids .24, positive .17, hiv .15, virus .15
Jac.	loglike	infection	infection	1	infection .38, aids .27, tract .18, infected .18, positive .17

Table 2: Example results; R = rank of expected target English word for source French word

print out the top 5 ranked words. *Rank* refers to the performance of our program, with a 1 meaning that the correct translation of the input French word was found as the first candidate.

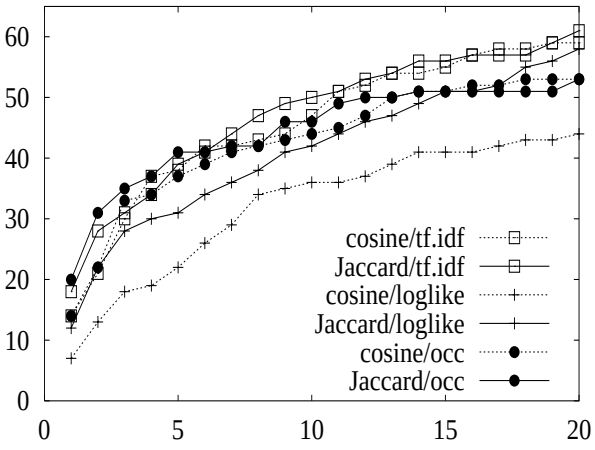

Figure 1: Percentile rank of the measures.

A percentile rank (figure 1) showed that using the combination of occ_i^j and Jaccard, about 20% of the French test words have their correct translation as the first ranked word. If we look at the best ranked words, we find that they have a strong thematic relation: *e.g., anxiety, depression, psychiatrist, phobia*, or *infection, infected, aids, homosexual*.

6 Discussion and Improvement Directions

As the percentile rank figure showed, the combination of context frequency weighting (occ_i^j) and Jaccard gives an accuracy of about 20% for correct translation which is followed by $tf.idf$/Jaccard measures. However, if we look among the top 20 ranked words, we can find that the $tf.idf$/Jaccard and $tf.idf$/cosine have better performance: more than 60% of the words find their correct translations within the top 20 words, which is much better than occ_i^j/Jaccard and occ_i^j/cosine. It seems that the loglike weighting factor did not help to improve the translation performance; this is true when we combined it with the cosine measure, but with Jaccard, we can see an improvement at the 20th percentile.

In some cases where the correct translation was badly ranked, the French test words have different usages, which induces an important context diversity. For instance, for the French word *chirurgie* whose expected translation is *surgery*, we have as top ranked words *pain, breast, desmoplasia, procedure, metastatic...*, and for *médecine* (*medicine*), we have *information, clinician, article, medical....* For common words like, *e.g.*, *analyse*/*analysis* and *sang*/*blood*, we have *girdle, sample, statistic...* for *analysis* and *output, collection, calorimetry...* for *blood* as best ranked translations.

As an attempt to improve the precision of the French-English translation method, the same model was applied in the reverse direction to find the French counterparts of the 10 top-scoring English candidates. We then kept only those English candidates that had the initial French source word among their top 10 reverse translation candidates. In the present settings, only 42 of the 95 French source words remained, 38 of which kept exactly one English candidate; among these, 27 are the expected translation, and 1 is an adjective derived from the expected translation (*estomac*/*gastric*). The other

4 words still have multiple translation candidates, which can be ordered according to their combined similarity scores: for 2 of them, the top ranked candidate is then correct, and 1 is a derived adjective (*thérapie*/*therapeutic*).

Altogether, if we propose the top ranked remaining candidate according to this scheme, recall/precision reach .31/.69, or .33/.74 if derived adjectives are considered acceptable. This result is really encouraging as it shows that the reverse application of the translation method to the English candidate words improves its effectiveness.

As a comparison, on a 'general language' corpus, Rapp (1999) reports an accuracy of 65% at the first percentile by using loglike weighting and city-block metric.[3] This difference in accuracy may be accounted for by the larger size of the corpora (135 and 163 Mwords), the use of a general English-German lexicon (16,380 entries), and the consideration of word order within contexts. In Fung and McKeown (1997), a translation model applied to a pair of unrelated languages (English/Japanese) with a random selection of test words, many of them multi-word terms, gives a precision around 30% when only the top candidate is proposed.

Our bilingual lexicon does not include general French and English words. This implies that some contexts are ignored: all cooccurrences of a specialized word with a general word are lost in our case. We therefore plan to explore the effectiveness of incorporating a general lexicon, as well as applying POS-tagging to the corpus. An additional difference with Fung and Yee (1998) is that they look for translational equivalents only among words that are unknown in both corpora. This additional condition might also help to improve our current results.

7 Acknowledgements

We thank Jean-David Sta, Julien Quint and Benoît Habert for their help during this work.

References

Jiang Chen and J-Y. Nie. 2000. Parallel web text mining for cross-language IR. In *Proceedings of RIAO 2000: Content-Based Multimedia Information Access*, volume 1, pages 62–78, Paris, France, April. C.I.D.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-paralle corpora. In *Proceedings of the 5th Annual Workshop on Very Large Copora*, volume 1, pages 192–202, Hong Kong.

Pascale Fung and L. Y. Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36^{th} ACL*, pages 414–420, Montréal, August.

Gregory Grefenstette. 1998a. *Cross-Language Information Retrieval*. Kluwer Academic Publishers, London.

Gregory Grefenstette. 1998b. The problem of cross-language information retrieval. In *Cross-Language Information Retrieval* (Grefenstette, 1998a), pages 1–9.

D. Harman. 1991. How effective is suffixing. *Journal of the American Society for Information Science*, 42:7–15.

D. Hiemstra, F. de Jong, and W. Kraaij. 1997. A domain specific lexicon acquisition tool for cross-linguage information retrieval. In *Proceedings of RIAO97*, pages 217–232, Montreal, Canada.

J. Laffling. 1992. On constructiong a transfer dictionary for man and machine. *Target*, 4(1):17–31.

M.L. Littman, S.T. Dumais, and T.K. Landauer. 1998. Automatic cross-language information retrieval using latent semantic indexing. In Grefenstette (Grefenstette, 1998a), chapter 5, pages 51–62.

Robert M. Losee. 1998. *Text Retrieval and Filtering: Analytic Models of Performance*, volume 3 of *Information Retrieval*. Kluwer Academic Publishers, Dordrecht & Boston.

E. Picchi and C. Peters. 1998. Cross-language information retrieval: A system for comparable corpus querying. In Grefenstette (Grefenstette, 1998a), chapter 7, pages 81–90.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37^{th} ACL*, College Park, Maryland, June.

H. Charles Romesburg. 1990. *Cluster Analysis for Researchers*. Krieger, Malabar, FL.

Gerald Salton. 1970. Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194.

Karen Sparck Jones. 1979. Experiments in relevance weighting of search terms. *Information Processing and Management*, 15:133–144.

[3] The city-block metric is computed as the sum of the absolute differences of corresponding vectors positions.