The New Navigators: from Professionals to Patients
R. Baud et al. (Eds.)
IOS Press, 2003

The Effect of a General Lexicon in Corpus-Based Identification of French-English Medical Word Translations

Yun-Chuang Chiao, Pierre Zweigenbaum

STIM/DSI, Assistance Publique – Hôpitaux de Paris, France
{ycc,pz}@biomath.jussieu.fr http://www.biomath.jussieu.fr/

Abstract:

We present a method, based on the similarity of word distribution across languages, of finding 'new' words' translations in French-English comparable medical texts, starting from a partial bilingual medical lexicon. In this paper, we test the influence of adding general-language words to this initial lexicon. Our experimental results show that all test words are correctly translated within the top 25 candidates; and that the addition of general words to the lexicon helps to improve translation accuracy for medical words.

Keywords:
Natural Language Processing; Controlled Vocabulary; Multilingualism; Translation; Algorithms

1 Introduction

In recent years, with a rapid expansion of online information available on medical web sites in different languages, it becomes more common that non-native speakers explore documents in several languages. In this situation, cross-language information retrieval [1] is getting more focus and interest than ever to enable transparent access through a single query to information provided on the Web in different languages. One of the issues that have to be addressed is that of query translation, which relies on some form of bilingual lexicon. It generally assumes that a large, bilingual lexicon is available for each language pair. Such a lexicon can never be expected to be complete, especially in a rapidly evolving domain such as medicine, and one must be able to cope with 'unknown' words. This is the subject of the present work.

Corpus-based methods have been proposed to find translations for unknown words: they rely on parallel or comparable corpora. Parallel corpora are sets of texts that are translations of each other; they have been used in many experiments for training statistical models to produce bilingual term equivalents [2, 3]. The limit is that large-scale parallel corpora are not always available, although [4]'s experiments reveals 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" [5]. Works on word translation identification in comparable corpora are based on the assumption that words which have the same meaning in different languages should have similar context distributions [6, 7, 8, 9]. However, previous experiments have dealt with very large, 'general language' corpora and words, and less attention has been paid to the problem of acquiring domain-specific translation lexicons given specialized comparable corpora of limited size. The present work addresses this issue in the medical domain. After an attempt presupposing an initial medical lexicon for identifying word translations [10] from comparable medical corpora, we investigate the effect on our translation method of adding a general lexicon. The translational equivalents

obtained may then be used, *e.g.*, for extending an existing medical lexicon or for query expansion and translation in cross-language information retrieval.

We first give a detailed description of data collection and the proposed method. We then provide and discuss experimental results on a test set of French medical words.

2 Data collection

We took advantage of the existence of MeSH-indexed Internet catalogs of medical web sites, such as CISMeF [11] (www.chu-rouen.fr/cismef) for the French language and CliniWeb [12] (www.ohsu.edu/cliniweb for English, to build comparable corpora. 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, and automatically dowloaded the pages indexed by these catalogs. Once converted into plain text, they yielded a 591,594-word French corpus (39,875 unique words after a simple lemmatization; see below) and a 608,320-word English corpus (32,914 unique words after lemmatization). Although the total number of occurrences is about the same in both corpora, there are more different word types in the French corpus; this can be attributed to the imperfect lemmatization and to the presence of foreign words (mainly English and Spanish). However, as explained by Diab and Finch [9, 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 [13]: MeSH, WHOART and ICPC; for general words, we used the French-English dictionary distributed in the Linux package dictd-dictionaries.

The resulting lexicon contains 22036 ('simple-word') entries, mainly specialized medical words, *e.g.*, *abasie: abasia; abattement: prostration; abdomen: abdomen, belly; abeille: bee; abducteur: abducens, abducent; accuser: accuse*. When the same word has several translations, they are all listed.

3 Methods

The basis of the method is to find the target words whose distributions are the most similar to that of a given source word. Figure 1 shows a schema of the method, which we summarize in the rest of this section. Additional detail can be found in [10].

3.1 Computing context vectors

For each occurrence of word *i* in source and target language corpora, we create a vector whose size depends on the number of 'pivot words' (see below). A sliding context window of 7 words as showed in table 1 is used to calculate the cooccurrences of *i*. Stop words are removed and a simple lemmatization is applied to each word in the context windows. Since this lemmatizer does not handle gender nor verb inflection, this lemmatization is far from perfect. Each word *i* in a context vector is assigned a weight of association with word *j*. Besides simple cooccurrence count *cooc*(*i*,*j*), we tested two weighting factors (table 2a): mutual information (*MI*) [14], and *tf.idf* [15]. In the formulas, *occ*(*i*) is the number of occurrences of word *i* in the corpus.

Figure 1: Schema for the translation model

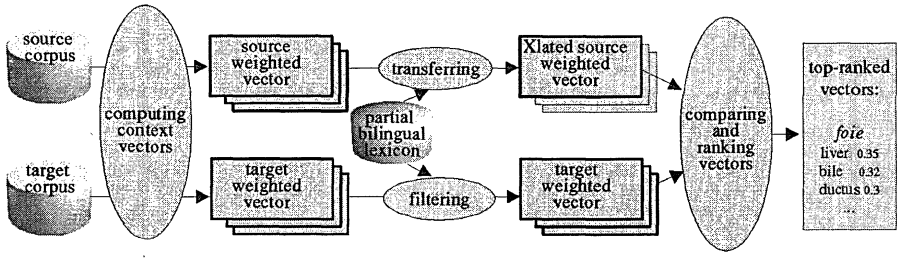


Table 1: Example of context window for the word asphyxiation.

original text	colorless	odorless	gas	can	cause	*asphyxiation*	in	poorly	ventilated	spaces
7-word window	colorless	odorless	gas	—	—	*asphyxiation*	—	poorly	ventilated	space

Table 2: Weighting factors and similarity measures

$MI(i,j) = cooc(i,j) \log \frac{cooc(i,j)}{occ(i)occ(j)}$	$Jaccard(V,W) = \frac{\sum_k v_k w_k}{\sum_k v_k^2 + \sum_l w_l^2 - \sum_m v_m w_m}$
$tf.idf(i,j) = tf(i,j)idf(i)$ $tf(i,j) = \frac{cooc(i,j)}{\max_{k,l} cooc(k,l)}$ $idf(i) = 1 + \log \frac{\max_{k,l} cooc(k,l)}{\lvert k; cooc(i,k) \neq 0 \rvert}$	$\cos(V,W) = \frac{\sum_k v_k w_k}{\sqrt{\sum_k v_k^2}\sqrt{\sum_l w_l^2}}$
(a) Weighting factors	(b) Similarity measures

3.2 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. 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 pivot words' (table 3). 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. Besides the lexicon of medical words M used in our previous experiment [10], we use the combined lexicon C which contains both medical and general words (n=6,243). We test here the impact of including general words in the context vectors on the performance of our methods.

Table 3: Example context vector for asphyxiation*; Oc = number of occurrences in the corpus; Co = number of cooccurrences. The value for each context word is its MI score.*

Word	Oc	Co	*asbestos*	*asthma*	*baby*	*bottle*	*gas*	*material*	*odorless*	*space*
asphyxiation	2	12	.0023	.00017	.00018	.00021	.00018	.00014	.00025	.00013

3.3 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. Two similarity metrics are used for comparing two vectors V and W (of length n): Jaccard [16] and cosine [17], each computed with any of the three different weighting factors (table 2b where k,l,m range from 1 to n). Cosine measures the angle of two vectors, and is maximal (=1) if the vectors are identical.

3.4 Experiments

To test the method in a setting where a sufficient number of contexts are available, we selected word-pairs among cross-language pivot words which are frequent in both corpora. This provides us with a test set of 97 French words of which we know the correct translation.

We tested two different sets of candidate target context vectors: the set of cross-language pivot words P and that of unknown U words which are not listed in the lexicon. With set P, we test whether the expected translation of the test word can be differentiated from other well-known words of the domain. With set U, we investigate the utility of our method for the translation of new words.

4 Results

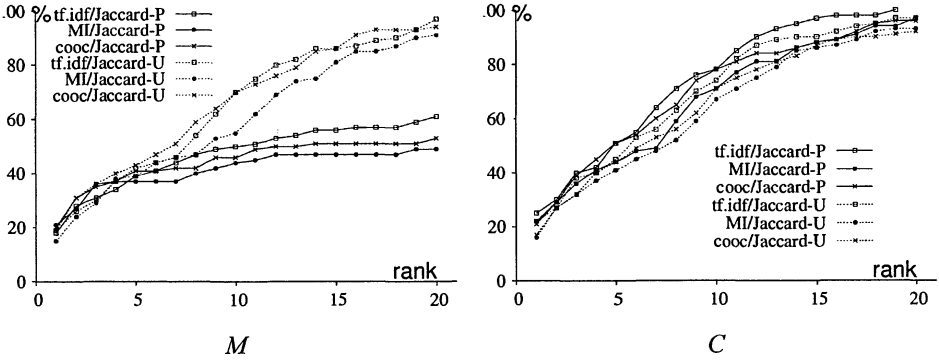
For each test word, we produced the list of its translational equivalents ranked in decreasing order of similarity score. The rank R of the expected translation provides the basis for evaluation. Sample results are provided in table 4, showing the top ranked candidate translations for French word *nausée*. Figure 2M presents the percentile rank distribution for both pivot and unknown words sets with the specific lexicon M. At the first percentile, the results for sets P and U are similar: whether with *MI* or *tf.idf*, 23% and 22% of test words are correctly translated against 15% and 19% for Cosine. All French test words find their correct translation among the top 25 and 29 candidates when using Jaccard with any of the three weighting factors for the unknown word set U. Figure 2C shows the results for the combined lexicon C. For the pivot word set P, 25% of the French test words have their expected translation as the first ranked word against 22% for the unknown word set U. All test words find their correct translation among pivot word set P in the first 19 and 23 candidates using Jaccard combined with any of the three weighting factors. For the unknown word set U, all test words have been correctly translated at the 24th percentile with the combination *cooc*/Jaccard.

Table 4: Results for French word nausée *(set P, lexicon C); R = rank of expected target English word.*

Meas.	Weight	R	Top 5 ranked candidate translations, followed by similarity score
Cosine	*cooc*	1	nausea .86, depigmentation .53, parturition .41, aura .39, dysuria .37
	MI	1	nausea .75, depigmentation .52, chlordiazepoxide .50, parturition .36, aura .35
	tf.idf	1	nausea .86, depigmentation .50, parturition .42, aura .38, diathesis .36
Jaccard	*cooc*	1	nausea .73, abdominal .08, constipation .08, vertigo .08, anorexia .07
	MI	1	nausea .60, constipation .10, vertigo .10, neuroleptic .09, anorexia .09
	tf.idf	1	nausea .76, abdominal .10, vertigo .10, constipation .09, anorexia .08

If we compare the results of M with that of C, we find that on the one hand, the C lexicon yields better results than M for the pivot word set P. On the other hand, the difference between the results of M and C for the unknown word set U is not significant.

Figure 2: Comparison of the percentile rank between pivot word P *and unknown word* U *sets with each lexicon (*M = *medical,* C = *combined): y = percentage of the words ranked in the top* x *ranks.*



5 Discussion and conclusion

At low percentiles, the results on set *P* are better than those on set *U* with both lexicons *M* and *C*. However, the performance on set *U* with *M* (figure 2M) continues to increase, up to 94% at the 20th percentile (*tf.idf*/Jaccard) against 61% for set *P*. 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 for set *U* where the test words were frequent in both corpora and the candidate set contained unknown and relatively rare words. This might also be consistent with the fact that medical words contained in *P* have more precise definitions than general words in *U*, so the inclusion of more matched candidates might not improve the overall accuracy rate.

On a 'general-language' corpus, Rapp [7] 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.

Figure 2C shows that the performance of our algorithm does improve with a larger vector space (*C*) where general words are taken into consideration. This leads us to assume that general words in the context might be useful for the disambiguation of well-known words in the present setting. These results seems promising enough to proceed further with this combined lexicon toward application to disambiguation of query translation. Also, other window sizes should be tested beyond the 7-word window used here.

Further investigations must now obtain better performance at lower percentiles. We have proposed to filter and rerank translation candidates by reverse translation [10]. Several other 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 using part-of-speech-tagged corpora.

Also, the present tests were performed on frequent words, and we must now experiment with rare words. This is a situation where the combination of both general and medical words might prove particularly useful, increasing their chances of cooccurrence with the rare words.

6 Acknowledgments

We thank Jean-David Sta, Julien Quint and Salah Ait Mokhtar for their help during this work.

7 Bibliography

[1] Grefenstette G. The problem of cross-language information retrieval. In: Grefenstette G, ed, *Cross-Language Information Retrieval.* Kluwer Academic Publishers, London, 1998:1–9.

[2] Hiemstra D, de Jong F, and Kraaij W. A domain specific lexicon acquisition tool for cross-language information retrieval. In: Proceedings of RIAO97, Montreal, Canada. 1997:217–32.

[3] Littman M, Dumais S, and Landauer T. Automatic cross-language information retrieval using latent semantic indexing. In: Grefenstette G, ed, *Cross-Language Information Retrieval.* Kluwer Academic Publishers, London, 1998:51–62.

[4] Chen J and Nie JY. Parallel web text mining for cross-language IR. In: Proceedings of RIAO 2000: Content-Based Multimedia Information Access, (vol1), Paris, France. C.I.D., April 2000:62–78.

[5] Laffling J. On constructing a transfer dictionary for man and machine. *Target* 1992;4(1):17–31.

[6] Fung P and Yee LY. An IR approach for translating new words from non-parallel, comparable texts. In: Proceedings of the 36th ACL, Montréal. August 1998:414–20.

[7] Rapp R. Automatic identification of word translations from unrelated English and German corpora. In: Proceedings of the 37th ACL, College Park, Maryland. June 1999.

[8] Picchi E and Peters C. Cross-language information retrieval: A system for comparable corpus querying. In: Grefenstette G, ed, *Cross-Language Information Retrieval.* Kluwer Academic Publishers, London, 1998:81–90.

[9] Diab M and Finch S. A statistical word-level translation model for comparable corpora. In: Proceedings of RIAO 2000: Content-Based Multimedia Information Access, Paris, France. C.I.D., April 2000:1500–8.

[10]Chiao YC and Zweigenbaum P. Looking for French-English translations in comparable medical corpora. *J Am Med Inform Assoc* 2002;8(suppl):150–4.

[11]Darmoni SJ, Leroy JP, Thirion B, et al. CISMeF: a structured health resource guide. *Methods Inf Med* 2000;39(1):30–5.

[12]Hersh W, Ball A, Day B, et al. Maintaining a catalog of manually-indexed, clinically-oriented World Wide Web content. *J Am Med Inform Assoc* 1999;6(suppl):790–4.

[13]National Library of Medicine, Bethesda, Maryland. UMLS Knowledge Sources Manual, 2001. www.nlm.nih.gov/research/umls/.

[14]Schiffman B and McKeown KR. Experiments in automated lexicon building for text searching. In: Proc 18th COLING, 2000.

[15]Sparck Jones K. Experiments in relevance weighting of search terms. *Inform Proc Management* 1979;15:133–44.

[16]Romesburg HC. *Cluster Analysis for Researchers.* Krieger, Malabar, FL, 1990.

[17]Losee RM. *Text Retrieval and Filtering: Analytic Models of Performance,* (vol3) of *Information Retrieval.* Kluwer Academic Publishers, Dordrecht & Boston, 1998.

Address for correspondence

Chiao Yun-Chuang,
STIM, 91 bd de l'Hôpital, 75634 PARIS cedex 13, France
ycc@biomath.jussieu.fr