MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

Aligning Words in French-English Non-Parallel Medical Texts: Effect of Term Frequency Distributions

Yun-Chuang Chiao, Pierre Zweigenbaum

STIM: DSI, Assistance Publique –Hôpitaux de Paris & ERM202 INSERM, Paris, France
{ycc,pz}@biomath.jussieu.fr http://www.biomath.jussieu.fr/

Abstract

In this paper, we present a method for aligning words based on a statistical model of word distribution similarity. The basis underlying our method is that there is a correlation between the patterns of word cooccurrences in texts of different languages. Using automatically downloaded pages from different medical web sites and a combined bilingual lexicon of general and medical terms as language sources, a similarity score is assigned to each proposed translated pair of words, based on the distributional contexts of these two words. We vary several parameters of the method. Experimental results confirm a positive effect of frequency, show that medical words are better handled than less specialized words, and do not evidence a clear influence of context window size. Future directions for improvement include working with very large, part-of-speech tagged corpora.

Keywords:

Natural Language Processing; Controlled Vocabulary; Multilingualism; Translation; Algorithms.

Introduction

World alignement 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 alignement 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intensive process and for applications such as bilingual lexicon construction, human compiled dictionaries were often out-of-date as soon as they became available.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 medical and technical domains which are in constant evolution, producing new terms as knowledge advances. Many related works on using statistical models for mapping bilingual terms [1, 2] are based on parallel texts or 'bitexts' —pairs of texts that are translation of each other. Most of these 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 some experiments reveal a potential solution by automatically collecting parallel Web pages [3]. Therefore, it seems natural to enlarge the scope of corpus resources by looking for non-parallel, comparable corpora. Comparable corpora are a ripe area for investigation in the development of bilingual lexicons [4, 5, 6]. However, these experiments have 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 an alternative model where these ressources are assumed unavailable.

After a preliminary investigation limited to the identification of the translations of most common words [7] from comparable medical corpora, we validate here our proposed model on a larger scale, including rare terms in both corpora. We also test the influence of the context window size parameter, thus approximating different types of semantic relationships between words. The translational equivalents obtained may then be used, *e.g.*, as human translation aid for extending an existing medical lexicon 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 from the point of view of word frequency distributions, we overview our algorithm, which is validated on all lexicon words. We then provide and discuss the results. Finally, we describe future directions.

Background

Compared with other approaches which 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 to 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 [5] 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 [4] 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 'con-

text seed words' are preserved in comparable texts. By designing procedures to retrieve cross-lingual lexical equivalents together, Picchi and Peters [6] proposed that their system could have applications such as retrieving documents containing terms or contexts which are semantically equivalent in more than one language.

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 Web sites, CISMeF [8] (www.chu-rouen.fr/cismef) for the French language and CliniWeb [9] (www.ohsu.edu/cliniweb) for English. The detailed description of the construction of comparable corpora can be found in previous papers [7, 10].

The 'French' corpus contains many foreign words (mainly English and Spanish) since we did not have language filtering on it. Therefore, 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 by Diab and Finch [11:1501], 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 [12]: MeSH, WHOART and ICPC; for general words, we used the French-English dictionary distributed in the Linux package dictd-dictionaries.

The resulting lexicon contains 22,036 'single-word' entries, mainly specialized medical words, *e.g.*, *anévrisme:aneurysm; assiette:plate; caoutchouc:rubber; champignon:fungus, mycete; dent:tooth; derme:corium, dermis; falaise:cliff*. When the same word has several translations, they are all listed.

Methods

The 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 experience. 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) build a context vector of each word within each of the corpora; and (b) provide an algorithm for ranking the possible matching word vectors between corpora. Additional detail can be found in [10].

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. Using a simple sentence boundary (defined by punctuation marks such as *!, . or ?*), we used two varied sliding context windows of 5 and 7 words (table 1) to calculate the cooccurrences of *i*. A simple lemmatization is applied to each co-occurrence pattern. Since this lemmatizer does not handle gender nor verb inflection, this lemmatization is far from perfect.

Table 1: Example context windows for the word sialophorin*.*

Original text: quantitative or qualitative deficiency of *sialophorin* in some way due to abnormal Wiskott-Aldrich Syndrome protein
7-word win.: quantitative, — qualitative, deficiency, — *sialophorin* — abnormal, wiskott-aldrich, syndrome
5-word win.: qualitative, deficiency, — *sialophorin* — abnormal, wiskott-aldrich

The window size parameter allows us to look at different scales [13]. Smaller window sizes will identify fixed expressions and other relations as syntactic dependencies ; larger window sizes will highlight semantic concepts and other relationships that hold over larger range. The *tf.idf* weighting measure [14] gave better performance in [7] when combined with any similarity formula; we therefore choose it here (table 2).

Table 2: Weighting factor and similarity measure

$$tf.idf(i,j) = tf(i,j)idf(i)$$
$$tf(i,j) = \frac{COOC(i,j)}{\max_{k,l} COOC(k,l)}$$
$$idf(i) = 1 + \log \frac{\max_{k,l} COOC(k,l)}{|k; COOC(i,k) \neq 0|}$$

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

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. 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 (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.

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 [15] similarity metric is used for comparing

two vectors *V* and *W* of length *n* (see table 2; *klm* range from 1 to *n*).

Table 3: Example translated context vector: French word adénose

Word ; Occ. in corpus ; Cooc. ;	Target context vector consists of translated source pivot words
adénose ; 11 ; 40 ;	loose, image, proliferation, hyperplasia, dendritic, fever, lesion, fibrosis, papilloma, calcium, epithelium, inside, adenoma, cell

Experiments

Word alignment generally depends on word frequency distributions in corpora. Words that are translations of each other tend to have similar frequencies in parallel, but also in comparable corpora. Additionally, frequent words have more contexts in a corpus, and therefore provide more alignment information than rare words to the algorithms. One may expect these words to obtain better translations. The present work tests whether less frequent words can still obtain some translational equivalents.

It proceeds by leave-one-out experiments: given a word whose translation is known in advance (this provides a gold standard) but which is assumed to be 'unknown', and given the set of context vectors for other words in both languages, it examines the rank of the expected translation among the ordered proposals of the algorithm (target context vectors with their target word). Test words range over the set of known words (or 'pivot' words) *P*.

We tested two different sets of target context vectors: the complete set *T* contains vectors for all words in the corpora. Among these words, some are present in our lexicon (they are 'known' words), and the rest are considered 'unknown'. 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. With the *U* set, we investigate the utility of our method for the translation of new words. And with the *T* set we test whether the expected translation of the test words can be differentiated from other well-known words of the domain.

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 a corpus by descending frequency: from rank 1 for the most frequent word to rank 89 for single-occurrence words ('hapaxes') in our corpus. Figure 1 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 which we try to align the possible translations in our English corpus. The specialized medical set *M* consists of 886 medical one-word terms extracted from the SNOMED Microglossary for Pathology [16]. The global, mixed set *A* of all words consists of the *M* set plus 6,210 medical and general one-word terms

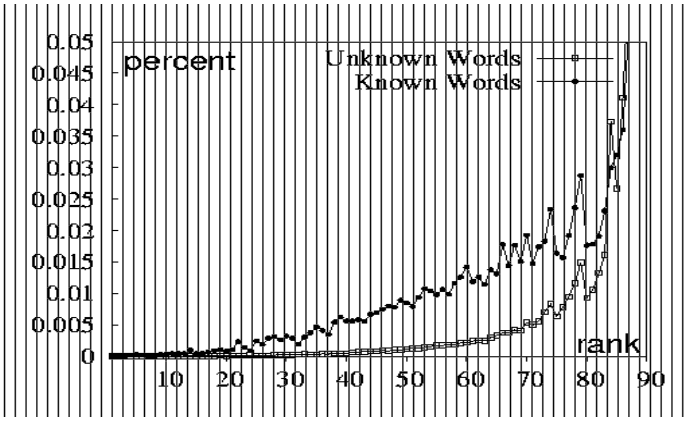

Figure 1 - Word frequency distribution in French corpus

Results

Table 4: Results for French word nécrose and gène; R = rank of expected target English word.

Fr word	En word	R	Top 5 ranked candidate translations, followed by similarity score
nécrose	necrosis	1	necrosis .181, chronic .148, renal .142, inflammation .135, infarction .123
gène	gene	1	gene .247, mutation .243, recessive .197, protein .194, chromosome .145

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* and *gène*.

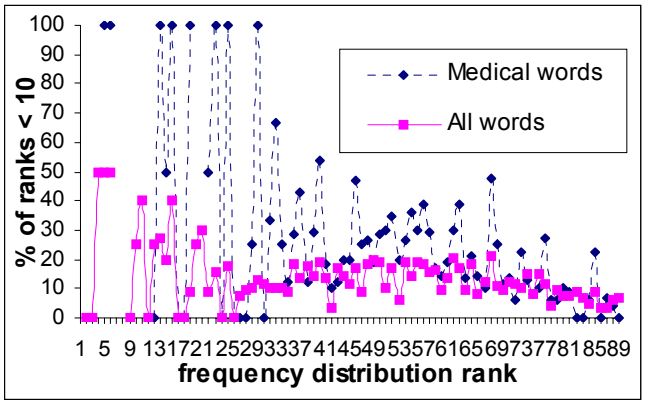

Figure 2 - Comparison between all-words (set A) and medical word (set M): y = percentage of words with expected translation in top 10 positions

Figure 2 presents a comparison of the results with the global word set *A* and with the medical word set *M*, illustrated by the percentage of French words which obtain their correct translation among the top 10 candidates, depending on their word frequency distributions in the corpora. In the higher frequency region, all medical words find their correct translation while about half of the pivot word set *P* is correctly translated. When we look at less frequent words for the two sets, the *M* set still yields a better result than the *P* set. Figure 3 shows the results for the medical words set *M* when using different context window sizes. Except for a slight difference at several low-frequency ranks, where the 7-word window presents better results than the 5-word window, similar results are generally obtained both on a larger segment (7-word window) and on a smaller segment (5-word window).

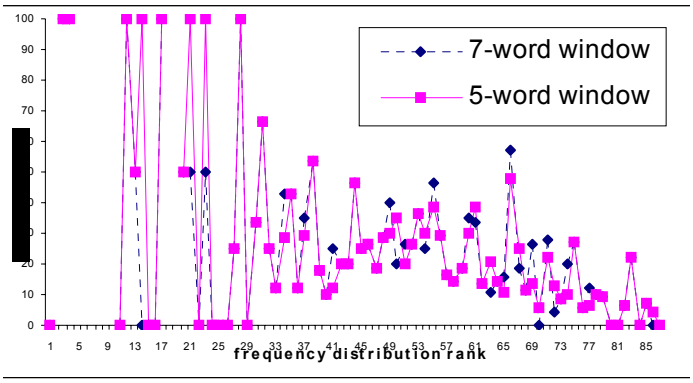

Figure 3 - Comparison between a 7-word window and a 5-word window

We found no significant difference between the two target context vector sets *U* (unknown words) and *T* (all-words). We therefore discarded the corresponding figure from the present paper.

Discussion

It is not surprising that our method fares better on more frequent words than on rare 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 all-words set *A*, 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 all-words set *A* 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 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 the medical words contained in *P* may have more precise definitions than general words in *U*, so that the inclusion of more matching candidates might not improve the overall accuracy rate.

On a 'general-language' corpus, Rapp [5] 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.

Conclusion and Perspectives

In summary, these experiments confirm a positive effect of frequency 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, but do not evidence a clear influence of context window size.

Our proposed approach relies on an initial bilingual lexicon to build context vectors, and we have shown in previous work [7] that the performance can be improved by adding general words in the medical lexicon. We would like to test our method further by counting only coccurrences 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 rare terms. 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 enlarging the size of corpora.

Acknowledgments

We would like to acknowledge Jean-David Sta for his useful help and comments during this work.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] Chiao YC and Zweigenbaum P. The effect of a general lexicon in corpus-based identification of French-English medical word translations. In: Baud R, Fieschi M, Le Beux P, and Ruch P, eds, *Proceedings Medical Informatics Europe*, Amsterdam. IOS Press, 2003:397–402.

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

[9] 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.

[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]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.

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

[13]Church K and Hanks P. Word association norms, mutual information, and lexicography. *Computational Linguistics* March 1990(16(1)):22–9.

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

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

[16]Côté RA. *Répertoire d'anatomopathologie de la SNOMED internationale,* v3.4. Université de Sherbrooke, Sherbrooke, Québec, 1996.

Address for correspondence

Chiao Yun-Chuang

STIM, 91 bd de l'Hôpital, 75634 Paris Cedex 13, France

ycc@biomath.jussieu.fr