

Contribution to Terminology Internationalization by Word Alignment in Parallel Corpora

Louise Deléger, MS¹

Magnus Merkel, PhD²

Pierre Zweigenbaum, PhD^{1,3,4}

¹INSERM, U729, Paris, F-75006 France;

²Dept of Computer and Information Science, Linköping University, Sweden;

³Assistance Publique – Paris Hospitals, Paris, F-75674 France; ⁴INALCO, CRIM, Paris, F-75343 France

Background and objectives. *Creating a complete translation of a large vocabulary is a time-consuming task, which requires skilled and knowledgeable medical translators. Our goal is to examine to which extent such a task can be alleviated by a specific natural language processing technique, word alignment in parallel corpora. We experiment with translation from English to French. Methods.* Build a large corpus of parallel, English-French documents, and automatically align it at the document, sentence and word levels using state-of-the-art alignment methods and tools. Then project English terms from existing controlled vocabularies to the aligned word pairs, and examine the number and quality of the putative French translations obtained thereby. We considered three American vocabularies present in the UMLS with three different translation statuses: the MeSH, SNOMED CT, and the MedlinePlus Health Topics. **Results.** We obtained several thousand new translations of our input terms, this number being closely linked to the number of terms in the input vocabularies. **Conclusion.** Our study shows that alignment methods can extract a number of new term translations from large bodies of text with a moderate human reviewing effort, and thus contribute to help a human translator obtain better translation coverage of an input vocabulary. Short-term perspectives include their application to a corpus 20 times larger than that used here, together with more focused methods for term extraction.

INTRODUCTION

The need for terminology internationalization keeps growing as, on the one hand, controlled medical vocabularies evolve into international standards and, on the other hand, the health systems in more countries reach a level of information technology where the requirement for interoperable vocabularies makes them consider using such international standards. The first condition for adoption of an international vocabulary, however, is the existence of a national-language translation of this vocabulary. Creating a complete translation of a large vocabulary is a time-consuming task, which requires skilled and knowledgeable medical translators. Our goal is to examine to which extent

such a task can be alleviated by a specific natural language processing (NLP) technique, word alignment in parallel corpora.¹ We experiment with translation from English to French, an instance of the prototypical situation where an international terminology in English needs to be translated into the local language.

Previous work has addressed the use of text corpora to extend controlled medical vocabularies, starting with the search for new terms in monolingual corpora.^{2,3} Work more relevant to our current goals looked for translational equivalents of source words by morphological decomposition^{4,5,6} or transduction⁷. Comparable corpora, *i.e.*, text corpora addressing the same general topic in two different languages, were also used⁸ to collect word translations. Already parallel medical vocabularies such as ICD-10 can be used to derive word translations.^{9,10} While the previous methods focused on single-word translation, the translation of some multiword expressions can be obtained by compositional translation.¹¹ Word alignment in both parallel and comparable corpora were also used to extend the German version of the MeSH and help cross-lingual information retrieval.¹²

Our method relies on parallel corpora, and addresses the translation of both single- and multi-word expressions. The present paper extends our previous work¹³ which was focused on the MeSH thesaurus. To check whether differences are encountered depending on the size and degree of completion of the translation of a given source vocabulary, it considers three American vocabularies present in the UMLS Metathesaurus[®] at three different degrees of completion of translation: the MeSH, where each descriptor has at least one French translation; SNOMED CT, where only a part of the concepts have French terms (outside the UMLS); and the MedlinePlus Health Topics, for which there is no French equivalent. It also examines how human revision work can be reduced.

MATERIAL AND METHODS

Material

The corpus processed in the present work was built

from the “Health Canada/Santé Canada” Web site. This is a governmental Web site mainly intended for the general public, with large sections more targeted to health care professionals. Its distinctive property is that it is completely bilingual, each English page pointing to its French translation and vice-versa. We downloaded a large part of the site in January 2005, using the wget program in recursive mode (Google announced about 100,000 pages for this site at this period). We only kept the HTML pages. We detected the parallel pages by parsing their hyperlinks, selecting those whose anchor text (or image name) contained the words “English” or “Français”. We converted these pages into XHTML (using the W3C program tidy), then into lightly marked-up text (through an XSLT program). The texts were finally segmented into sentences.

The English and French MeSH were obtained from the UMLS 2005AC; work on SNOMED CT and MedlinePlus Health Topics was done after the release of UMLS 2006AA, and used this newer material. The partial French translation of SNOMED International was obtained courtesy of the Secrétariat Francophone International de Nomenclature Médicale (Sherbrooke, Canada).

Alignment

Alignment was first performed at the sentence level, using Melamed’s GMA program (<http://nlp.cs.nyu.edu/GMA/>),¹⁴ which relies both on statistical and on linguistic techniques, with an emphasis on the geometric properties of series of aligned words in parallel texts. Sentence alignment needs to detect the places where there isn’t a one-to-one correspondence between source and target sentences.

The next step, which is a more challenging task, aimed at aligning words within sentences. A word is often translated with several ones, or can be omitted in the corresponding sentence (this is typically the case for grammatical words that are specific to a language). Parallel sentences, though being translations of each other, can differ considerably in terms of structure. In that case even a human has trouble determining which words should be paired together. The results we expect are therefore on a lower level than from the previous sentence alignment task. Besides, we can question whether we really want a word-to-word alignment. The objective of this work is to obtain medical terms. A term can be either a single word or a multi-word unit. A common approach is to first extract candidate terms using a separate tool — a term extractor — and then to proceed to their alignment.¹⁵

An originality of the present work is that we do not separate the detection of candidate terms from the alignment process. In other words we use a tool that is able to detect multi-word units and to align both single words and multi-word expressions.

We used the I*Tools suite¹⁶ (developed at Linköping University, Sweden) to perform word alignment. We chose these tools partly because they are based on a hybrid approach, using both statistical and linguistic techniques. On the one hand, statistical methods look for pairs of words that are frequently found in corresponding sentences (i.e. sentences aligned during the previous step). On the other hand, linguistic techniques spot word pairs that have similar spelling — “cognates” (e.g. hospital/hôpital) — or that are found in bilingual dictionaries; they can also check that the parts-of-speech of source and target words are similar or that these words play the same syntactic function in the two aligned sentences. The I*Tools also align multi-word units, which suits our terminological purpose. A pre-processing step is required to provide input for the linguistic processing: the corpus is tagged and lemmatized (using the Treetagger, <http://www.ims.uni-stuttgart.de/projekte/corplex/>) and syntactically annotated (with the syntactic analyzer SYNTAX¹⁷). The alignment process with the I*Tools can be divided into three steps: training, automatic alignment and review of the results, each one corresponding to a specific tool of the suite: I*Link, I*Trix and I*View. Training and review are both done with graphical, interactive tools that are fast to work with. In summary, the word alignment process is as follows:

1. *training* of the automatic word aligner on a set of sentences randomly taken from our corpus, by interacting manually with I*Link;
2. *automatic word alignment* with I*Trix. If results seem poor, a first review (with I*View) may also be done followed by a second run of I*Trix;
3. *selection* of medical terms, as explained below;
4. *filtering* and *human review*, in an order to be discussed below.

Term selection and filtering

An input English source vocabulary can then be joined to the set of aligned term pairs, thus providing candidate French translations for the input English terms. (a) Human review (with I*View) could be performed at this stage; however, it is more economical to first

apply more automatic filters to these term pairs. We checked and removed upper/lowercase duplicate term pairs (this step corresponds to the label “case” in table 2). Since most input terms are nouns or noun phrases, we removed verb entries, which most often correspond to words that are ambiguous between noun and verb (label “verb” in table 2). This was possible since our corpus was tagged with part-of-speech. As we are looking for new translations and as we assume that existing translations are correct, we also removed known term pairs: by searching them either in the Metathesaurus (label “UMLS” in table 2), or (for SNOMED CT) in the partial French translation of SNOMED International (label “French” in table 2). (b) Human review can then be performed, with negligible or no loss compared to point (a). Finally, the quality score computed by the automatic word aligner can be used to eliminate term pairs with a low confidence score. (c) Human review, performed at this third point, will have even less term pairs to consider, but some correct term pairs may be lost in this last filtering step. During human review, we also found words that were not translated into French, i.e. pairs of identical English terms (label “English” in table 2).

Evaluation

The aligned word pairs obtained by this method are proposed as French translations of English medical terms. These translations will need to be assessed by terminology maintainers, who are the only ones who can ultimately decide upon their relevance for these controlled vocabularies. In previous work,¹³ we extracted a random, 79-term sample of proposed, new translations for MeSH terms and submitted it to the head of the CISMef team, who are using the French MeSH to index Web documents. 64 terms (81%) were judged useful. The present paper focuses instead on the following dimensions:

- How many terms must be subjected to human review? Note that this review is performed by a language engineer (LD). Only after this linguistic validation will a medical vocabulary maintainer receive terms for consideration.
- How many new translations are obtained?
- How does the number of new translations relate to the number of source terms? Is there a difference in yield between the three source vocabularies?

RESULTS

The whole corpus was converted into text format and aligned at the level of sentences. Due to the size of

Table 1: Decrease in term pairs submitted to revision

#	MeSH	SNOMED CT	MedlinePlus
terms	567,655	850,521	1,434
term (a)	10,392	14,558	1,188
... (b)	6,529 -37%	9,148 -37%	737 -38%
pairs (c)	4,357 -58%	6,009 -59%	524 -56%

Table 2: Filtered and validated term pairs (strategy b). *rem.* = pairs removed; *left* = pairs left; *pairs* = selected term pairs; what follows is the reason for removing term pairs. *case* = duplicate pairs modulo case; *verb* = term pair including a verb; *French* = French term already in French version of the vocabulary; *UMLS* = French term already in another vocabulary of the UMLS Metathesaurus; *incor.* = incorrect term pair (incorrect or partial translation); *English* = term pair made of two (identical) English terms; *concepts* = number of concepts.

#	MeSH	SNOMED CT	MedlinePlus
terms	567,655	850,521	1,434
	rem.	left	rem.
pairs	10,392	14,558	1,188
case	1,693	8,699	1,894
verb	714	7,985	2,238
French	1,315	6,670	879
UMLS	141	6,529	399
incor.	4,452	2,077	6,065
English	262	1,815	285
concepts	978	1,444	122

the corpus, subsequent processing was only performed on 540 document pairs, totalling 1.3 million words and 46,100 sentence pairs, that is 5% of the whole corpus. Training (1) was performed on a 600-sentence subset, then 91,171 word alignments (including both single- and multi-word terms) were obtained (2).

Human review. The English terms (“strings”) of the MeSH, SNOMED CT and MedlinePlus Health Topics were projected to these alignments, selecting respectively 10,392, 14,558 and 1,188 term pairs (see table 1). Human validation, if performed at this stage (a), would have to review that many terms. However, as explained in the Methods section, this can be reduced through term filtering steps (b) and (c), avoiding 37% to 59% of the human review. Not on the table, missed correct term pairs (false negatives) vary from 8.7% (MedlinePlus) to 12.8% (MeSH and SNOMED).

New translations. Table 2 shows how each filtering step affects the number of term pairs in strategy (b). For each source vocabulary, the first column displays the number of removed term pairs, whereas the sec-

Table 3: Overlap between the 3 vocabularies

	MeSH SNOMED CT	MeSH MedlinePlus	SNOMED CT MedlinePlus	All
terms	578	92	68	42
concept	283	54	44	28

Table 4: Example new translations

English	French	Valid
bone cancer	cancer des os	Yes
breast milk	lait maternel	Yes
reproduction rights	droits de reproduction	No

ond column contains the number of term pairs left after each step. The last two rows correspond to human review. The very last row provides the final number of different candidate translations after human validation. Table 3 shows the overlap between the three vocabularies for the candidate translations, in terms of both terms and concepts.

Example new translations, together with their evaluation, are listed in table 4. In the third example, the source term is polysemous, and the proposed translation corresponds to its common, non-medical sense.

Relation to the number of source terms. The number of translations obviously depends on the size of the source vocabulary. Table 5 displays the ratio of new translations over the number of source terms. It shows that the main difference between our three source vocabularies occurs at the term projection stage (a): our two large vocabularies, MeSH and SNOMED CT, map to less than 2 aligned word pairs per 1,000 input terms (0.018 and 0.017), whereas the much smaller vocabulary, MedlinePlus Health Topics, maps to 830 aligned word pairs per 1,000 inputs terms. The subsequent filtering and validation operations, in contrast, operate at very similar levels of reduction (0.62–0.63 filtering and 0.25–0.31 validation rates), despite the difference of an order of magnitude in their inputs.

DISCUSSION

Aligned word pairs included both single- and multi-word expressions, so that both single- and multi-word English medical terms could be proposed French translations. We relied on the automatic detection of multi-word units by the I*Tools to spot the multiword terms candidate to translation. This may miss occurrences of the multiword terms that were present in our input controlled vocabularies, thus missing potential transla-

Table 5: Ratio of discovered term pairs over initial number of terms (strategy b). Each row provides a ratio relative to the previous row, except the last row (*cumulated ratio*) which shows the ratio of the number of validated term pairs over the initial number of terms.

#	MeSH	SNOMED CT	MedlinePlus
terms	567,655	850,521	1,434
	# ratio	# ratio	# ratio
pairs	10,392 0.018	14,558 0.017	1,188 0.83
filtered	6,529 0.62	9,148 0.63	737 0.62
validated	1,815 0.28	2,798 0.31	183 0.25
cum. ratio	0.0032	0.0033	0.13

tions. To avoid this situation, a different method would consist in first spotting all input terms in the English sentences (e.g., with a tool such as MetaMap¹⁸), and then taking these into account in the alignment process. Restricting the corpus to the matching sentences only would also speed up the alignment process and further reduce the amount of human review. We plan to examine this method in further work.

Although human review keeps about one third of the proposed translations (see table 2), we believe this method still provides a useful service in identifying a number of attested translations that the human translator might not have thought of.

We used 1.3 million words from a total 27.7 millions, i.e., about 5% of our corpus. This allows us to expect to obtain many more different translations by exploiting the rest of the corpus. Indeed, we cannot expect a direct proportion, i.e., 20 times more different word pairs, because of the properties of text corpora (Zipf law, LNRE distributions), so that the percentage of new pairs should be gradually decreasing. We compared the word pairs obtained by processing two subsequent batches of documents of the corpus, and observed that in the second one, 80% of the extracted candidate pairs were new. So in total, the number of additional translations when processing the rest of our corpus should be significant, and a larger proportion of our input vocabularies should be covered.

Another important factor is the relation between the concept types in the source vocabularies and the nature of the documents in our corpus. Déjean et al.¹² expanded the German version of the MeSH through a corpus of article abstracts from the Springer Verlag Web site of about 1 million words. Given the function of the MeSH thesaurus, article abstracts should be

particularly suited for this task. The part of their work which deals with parallel abstracts obtained 1,400 new German terms for the MeSH, which is comparable to our results. In contrast, the Health Canada/Santé Canada Web site devotes large sections to the general public, so that it is not surprising that the MedlinePlus Health Topics should obtain a good coverage. Further experimentation will tell us whether the MeSH and SNOMED CT will find significantly more occurrences of their terms in the rest of the corpus.

CONCLUSION

Given a corpus built from a bilingual, English-French, health-oriented Canadian Web site, we were able to identify a large number of translations of English medical terms originating in several controlled vocabularies included in the UMLS. We selected controlled vocabularies with different levels of translation into French, and showed that for each of them, we could identify additional relevant French translations that were not present in other French vocabularies of the UMLS. We also showed how to reduce the number of terms that must be submitted to human validation. We must stress the fact that this method is directly applicable to other language pairs, for instance English-Spanish, English-German, etc., provided parallel corpora can be built, *e.g.*, from parallel Web pages.

The results reported here were obtained given a single, albeit large, corpus built from one Web site. It is not unreasonable to assume that by extending the number and variety of sources, a larger coverage of the source medical terms should be obtained. For instance, using a corpus of parallel article abstracts¹² from Medline may be particularly appropriate for extending the MeSH thesaurus. Orienting the alignment process by first identifying the occurrences of multiword terms from the source vocabulary, as mentioned in the Discussion section, might also increase the number of proposed translations. Finally, this method must be considered as one among different methods, presented in the Background section, which propose translations for medical terms. Research into each of these methods, as that presented in this paper, should be complemented by research into their combination.¹²

REFERENCES

1. Gale W and Church KW. Identifying word correspondences in parallel texts. In: Proceedings of the 4th Darpa Workshop on Speech and Natural Language, Pacific Grove, CA, USA. 1991:152–7.
2. Hersh WR, Campbell EH, Evans DA, and Brownlow ND. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. *J Am Med Inform Assoc* 1996;3(suppl):159–63.
3. Nelson SJ, Kuhn T, Radzinski D, et al. Creating a thesaurus from text: A “bottom-up” approach to organizing medical knowledge. *J Am Med Inform Assoc* 1998;5(suppl):1046–.
4. Lovis C, Baud R, Rassinoux AM, Michel PA, and Scherrer JR. Medical dictionaries for patient encoding systems: a methodology. *Artif Intell Med* 1998;14:201–14.
5. Markó K, Baud R, Zweigenbaum P, et al. Cross-lingual alignment of medical lexicons. In: Zweigenbaum P, Schulz S, and Ruch P, eds, Proc LREC Workshop Acquiring and representing multilingual, specialized lexicons: the case of biomedicine, Genova, Italy. ELDA, 2006:5–8.
6. Namer F and Baud R. Predicting lexical relations between biomedical terms: towards a multilingual morphosemantics-based system. In: *Stud Health Technol Inform*, (vol116). 2005:793–8.
7. Claveau V and Zweigenbaum P. Translating biomedical terms by inferring transducers. In: Silvia Miksch, Jim Hunter EK, ed, Proc 10th Conference on Artificial Intelligence in Medicine Europe, (vol3581) of *LNCS*, Berlin / Heidelberg. Springer, 2005.
8. Chiao YC and Zweigenbaum P. Looking for French-English translations in comparable medical corpora. *J Am Med Inform Assoc* 2002;8(suppl):150–4.
9. Baud RH, Lovis C, Rassinoux AM, Michel PA, and Scherrer JR. Automatic extraction of linguistic knowledge from an international classification. In: Cesnik B, Safran C, and Degoulet P, eds, Proc 9th World Congress on Medical Informatics, 1998:581–5.
10. Nyström M, Merkel M, Ahrenberg L, et al. Creating a medical English-Swedish dictionary using interactive word alignment. *BMC* 2006. *Submitted*.
11. Ozdowska S, Névél A, and Thirion B. Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés. In: Proc 6èmes rencontres TIA.
12. Déjean H, Gaussier E, Renders JM, and Sadat F. Automatic processing of multilingual medical terminology: applications to thesaurus enrichment and cross-language information retrieval. *Artif Intell Med* February 2005;33(2):111–24.
13. Deléger L, Merkel M, and Zweigenbaum P. Enriching medical terminologies: an approach based on aligned corpora. In: Medical Informatics Europe, 2006. *To appear*.
14. Melamed ID. Bitext maps and alignments via pattern recognition. *Computational Linguistics* 1999;25(1):107–30.
15. Daille B, Éric Gaussier, and Langé JM. Towards automatic extraction of monolingual and bilingual terminology. In: Proc 15th COLING, Kyoto, Japan. August 1994:515–21.
16. Ahrenberg L, Merkel M, and Petterstedt M. Interactive word alignment for language engineering. In: Copestake A and Hajic J, eds, Proc EACL 2003, Budapest. 2003:49–52.
17. Bourigault D, Fabre C, Frérot C, Jacques MP, and Ozdowska S. Syntex, analyseur syntaxique de corpus. In: TALN, Dourdan, France. 2005.
18. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *J Am Med Inform Assoc* 2001;8(suppl):17–21.