Searching Parallel Corpora for Contextually Equivalent Terms

Caroline Barrière
Centre de Recherche Informatique de Montréal (CRIM) [i]
Montréal, Québec, Canada
caroline.barriere@crim.ca

Pierre Isabelle
Institute for Information Technology
National Research Council Canada
Gatineau, Québec, Canada
pierre.isabelle@nrc-cnrc.gc.ca

Abstract

In this paper, we show how a large bilingual English-French parallel corpus can be brought to bear in terminology search. First, we demonstrate that the coverage of available corpora has become substantially more extensive than that of mainstream term banks. One potential drawback in searching large unstructured corpora is that large numbers of search results may need to be examined before finding a relevant match. We argue that this problem can be alleviated by *contextualizing* the search process: instead of looking up isolated terms one searches for terms appearing in a context that is similar to that of the term to be translated. We present an experiment on context-based re-ranking and report highly positive results. We conclude that translators will increasingly rely on very large scale corpora for searching term equivalents.

1. Introduction

Isabelle (1992) has often been cited for the observation that bilingual corpora contain more solutions to more translation problems than any other existing resource. As a matter of fact, in recent years translators have been increasingly turning to corpus-based tools: translation memories, bilingual concordancers and, more recently, statistical machine translation.

Yet, it seems that for the specific task of looking up the translation of technical terms, translators continue to prefer traditional term banks and corpora remain to this day at best a secondary resource.

Term banks offer at least two important advantages: 1) their content is edited and validated by experts, so that translators feel they can mostly trust it; and 2) each database record is manually assigned domain categories that help users select contextually appropriate translations for polysemous terms. But term banks also have some significant drawbacks. Most importantly: a) their development is extremely time-consuming and costly; and b) their coverage is always incomplete, especially for recently-introduced (and thus lesser known) terms.

In this paper, we first argue that corpus search can circumvent some of the drawbacks of traditional term banks. Clearly, development time and cost are much less of an issue with corpora since they are mostly a product of the normal activity of translators. But until recently, no corpora were available whose terminological coverage would rival that of the mainstream general-purpose term banks. In this paper, we show that this is no longer true: the larger existing corpora now cover substantially more technical terms than the larger existing term banks.

Second, we argue that some of the known advantages of term banks can be emulated in the world of corpora. Note, however, that we will not address the issue of trust in the present paper. We will focus mainly on the issue of knowledge organization and argue that even though corpora lack an explicit semantic backbone such as domain categories, it is still possible to efficiently retrieve **contextually-relevant** matches from them.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 105–112
Leuven, Belgium, May 2011

Section 2 describes the experimental set-up: a) the English/French term pairs that we will be using as test data and b) the parallel corpus that we will be searching. Section 3 presents a comparison of the terminological coverage of our corpus with that of a range of alternative resources: large-scale terms banks, Wikipedia and smaller-scale corpora. Section 4 defines the notion of contextual search and describes the algorithms we will be testing. Section 5 reports results of our contextual search techniques. Finally Section 6 situates our work with respect to some related efforts.

2. Experimental Set-up

We first describe the term pair data, and then the bilingual corpus used in our experiments.

2.1 Gathering experimental data

To simulate translation contexts in specialized domains, we use a set of articles from eight scientific journals[1] for which both an English and a French abstract are provided (the second being a translation of the first). Each article contains 2 to 5 author-provided "keyphrases" (invariably technical terms) that have also been translated by the journal editors. The alignment of these keyphrases in the electronic files is coherent enough that we were able to automatically extract a large set of keyphrase pairs.

We restricted the set of term pairs to those that explicitly appear in the abstracts, as these abstracts will constitute our test set for contextual search (Section 4). Table 1 shows for each journal the number of abstracts and term pairs extracted (before and after filtering). The table also includes examples of term pairs as well as journal name codes provided for future concise reference.

As most of these abstracts were translated from English into French, we assume in the rest of this paper that the relevant task is to find French translations for English terms.

2.2 Gathering a bilingual resource

A very large parallel English-French corpus, called the Giga-F/E (Callison-Burch 2010) has recently been made available to the scientific community. It contains more than 1 billion words (half a billion in each language) in parallel sentences that were automatically extracted by crawling more than 1 TB of Canadian and European Union sources. Document-level alignments were first produced, using URL-based heuristics, and then Moore's sentence aligner (Moore 2002) was used to automatically align sentences. Duplicates sentences were removed, as were pairs in which the English and French sides were identical. The research community in machine translation has already been making extensive use of Giga-F/E, in particular on the occasion of the WMT-ACL-2010 shared evaluation tasks (Callison-Burch & al 2010).

For our experiments, we first indexed Giga-F/E using the Lemur toolkit[2], and then used Lemur's Java API to write software code for retrieving pairs of sentences whose English side contains the terms we are interested in.

3. Estimating terminological coverage

We now compare the terminological coverage of Giga-F/E with that of the Grand dictionnaire terminologique (GDT) from the Office québécois de la langue française.[3] We use GDT as a representative example of traditional term banks. We will also compare the coverage of Wikipedia, which constitutes a particularly fast-growing freely-available online multilingual resource.

Note that for that purpose we resort to a weak notion of "coverage": the resource (term bank or corpus) is considered to cover a test term inasmuch as it contains any word-level match whatsoever for that term. In so doing we acknowledge counting contextually inappropriate matches such as for example the term *stress* in its phonetic sense matching the term *stress* in its physics-of-materials sense. Clearly we are getting at an upper bound on real coverage. The key point here is that we can assume that the relative "weak" coverage of two different resources constitutes an unbiased estimator for their relative real coverage.

[1] Thanks to the Canadian Institute for Scientific and Technical Information (CISTI) for providing electronic copies of these journals for our research.

[2] See http://www.lemurproject.org/

[3] See http://www.oqlf.gouv.qc.ca/ressources/gdt.html

Table 1 – Scientific journals and term pairs

Code	Journal title	Nb abstracts	Nb term pairs	Nb terms after filtering	Examples of term pairs
bcb	Biochemistry and Cell Biology	472	736	725	cinetics / cinétique fluorescence spectroscopy / spectrofluoroscopie chemokines / chémokines
cgj	Canadian Geotechnical Journal	618	1047	1047	jacking test / essai de vérinage loose fill / remblai meuble interface friction angle / angle de frottement à l'interface
cjb	Canadian Journal of Botany	1002	1925	1905	tannin / tannin proximate analysis / analyse de proximité anatomical pattern / architecture anatomique
cjc	Canadian Journal of Chemistry	1341	1846	1820	octadecasil / octadecasil L-valine / L-valine intramolecular transglycosylation / transglycosylation intramoléculaire
cjce	Canadian Journal of Civil Engineering	608	1239	1238	urban boundary / frontière urbaine deformability / état de deformation frazil ice / glace frasil
cjm	Canadian Journal of Microbiology	852	1456	1431	metabolism / métabolisme fungicides / fongicides chitinases / chitinase
cjpp	Canadian Journal of Physiology and Pharmacology	846	1504	1470	b-adrenoceptor / adrénorécepteur-b pirarubicin / pirarubicine mesenteric bed / lit mésentérique
gen	Genome	854	1191	1173	wild radish / ravenelle 2n pollen / pollen 2n oilseed rape / colza
	Total	6593	10944	10809	

3.1 Giga-F/E terminological coverage

To ensure that the parallel texts we use as test data are not contained as such in our corpus, we explicitly removed from Giga-F/E all sentence pairs that appeared in our 6593 pairs of abstracts. We can assume that Giga-F/E does not contain the full text of the abstracted articles, since the normal practice in Canada for such scientific papers is to only translate their abstracts.

We then search the English sentences of Giga-F/E for the English side of each of our 10809 English-French term pairs. Table 2 shows the number of sentences (up to a maximum of 5000) containing one or more matches. The first row indicates the number of sentences returned: for example, 7.6% of English terms are found in a single sentence while 11.8% appear in 21 to 50 sentences.

First and foremost, observe that even though we are searching for terms extracted from highly technical journals, most of them are indeed found in the Giga-F/E. No less than 97.4% appear in at least one sentence of the corpus, and 89.8% appear in at least two sentences. Moreover, the split per journal indicates that the level of coverage varies only very slightly across domains. Only the chemistry domain (code *cjc*) seems to be somewhat at a disadvantage.[4]

[4] Note however, that our sample of journals is limited to the areas of engineering and health science, so that we do not have a comparison for other areas such as social studies, for example.

Table 2 - Coverage of English terms in Giga-F/E

Journal	Nb. of terms	1	2	3-5	6-10	11-20	21-50	51-100	101-200	201-500	501-1000	1001-2000	2001-5000	Total
bcb	725	38	50	66	46	71	72	72	66	83	120	7	8	699
cgj	1047	68	63	100	82	85	139	90	57	89	268	1	0	1042
cjb	1905	152	137	174	161	162	235	172	143	171	317	16	9	1849
cjc	1820	204	205	201	146	170	200	130	122	135	219	2	3	1737
cjce	1238	104	78	107	101	105	111	76	72	118	358	2	3	1235
cjm	1431	96	102	151	103	114	171	143	107	138	235	14	8	1382
Ccpp	1470	68	97	112	119	127	209	145	144	166	231	8	7	1433
gen	1173	89	96	121	108	108	138	119	108	93	88	49	31	1148
ALL	10809	819	828	1032	866	942	1275	947	819	993	1836	99	69	10525
	%	**7.6**	**7.7**	**9.5**	**8.0**	**8.7**	**11.8**	**8.8**	**7.6**	**9.2**	**17.0**	**0.9**	**0.6**	**97.4**

3.2 Comparing coverage with other resources

Table 3 compares the coverage of the same English terms in Giga-F/E, GDT, and Wikipedia. For Wikipedia, the search was performed in October 2010. The GDT version available to us with a programmatic interface is unfortunately older, dating back from 2005.

Table 3. Coverage of GDT-2005 and Wikipedia

Jrnal code	Nb terms	Found in Giga	Found in GDT	Found in Wikipedia
bcb	725	699	254	559
cgj	1047	1042	473	522
cjb	1905	1849	516	1221
cjc	1820	1737	490	1008
cjce	1238	1235	581	694
cjm	1431	1382	415	883
cjpp	1470	1433	537	1048
gen	1173	1148	279	697
Total	10809	10525	3545	6632
%		**97.4%**	**32.8%**	**61.4%**

At 61.4%, the coverage of Wikipedia remains far below that of Giga-F/E while the GDT is getting a meagre 32.8%.

In order to assess the coverage of up-to-date resources for which no programmatic interfaces are available, we did a sampling of 160 terms (20 per journal). We then manually searched for those terms in: a) the current online version of GDT; b) the current online version of Termium[5]; and c) in the large Hansard corpus available through the TransSearch bilingual concordancer[6]. Results are presented in Table 4. The coverage of current version of GDT appears significantly improved, reaching 53.8%. Termium comes out somewhat ahead with 63.1%. In spite of the large size of the Hansard corpus (tens of millions of words), its terminological coverage of 30% is substantially inferior to that of either term bank. But Giga-F/E's coverage turns out to be strikingly superior to that of any of the other three resources.

4. Contextual search in bilingual resources

Term banks possess an explicit semantic structure that is meant to facilitate information access. Each database record is intended to represent a unique concept and is explicitly associated with one or more domain codes and with one or more terms in each of the bank's languages. As a result, polysemous terms will appear in different database records, each associated with different semantic domains. When a user queries the bank with a polysemous term, all relevant records are

[5] The other mainstream Canadian term bank, http://www.termium.com/

[6] http://www.tsrali.com

retrieved and the user is expected to manually select the one record (and translation) whose domains are most appropriate to the context in which the query term was found.

Table 4. GDT-2010, Termium, and TransSearch coverage of specific term samples

Jrnal Code	Nb terms	Found in Giga	Found in GDT	Found in Termium	Found in TSearch
Bcb	20	20	12	16	6
cgj	20	20	14	18	10
cjb	20	20	10	12	7
cjc	20	20	11	11	5
cjce	20	20	14	14	8
cjm	20	19	8	9	2
cjpp	20	20	12	11	5
gen	20	20	5	10	5
Total	160	159	86	101	48
%		**99.4%**	**53.8%**	**63.1%**	**30%**

Bilingual corpora such as Giga-F/E do not possess any such a priori semantic organization. For example, in bilingual concordancers such as TransSearch or WebiText[7], the output of a match is not a set of domain-annotated records but just a (possibly large) set of parallel sentence pairs. The user needs to examine sentence pairs until at least one contextually appropriate match will be found. The question therefore arises of how difficult and time-consuming it will be for users of very large scale corpora to narrow down on contextually appropriate examples.

One of our core claims in this paper is that it is possible to help the corpus user in this task by an automatic assessment of the relevance of each corpus match. This can be done using well-known techniques from the domain of information retrieval (IR, Salton (1989)) to measure the level of similarity between the context around a corpus match and the context around the query term. The relevant notion of context can vary from a small window of words around the term, complete sentences, paragraphs or even whole documents. One needs to determine what works best in practice. In this paper we use the following definitions: 1) for the source term, the context is taken to be the complete abstract in which the term appears; and 2) for corpus matches, the context is limited to the sentence in which the match is found. We will show that this particular configuration does lead to successful re-ranking of corpus matches. But we will leave open the possibility that some other context windows could yield even better results. One reason is that the current version of Giga-F/E does not make it possible to experiment with larger context windows: all we are given is a set of sentence pairs without any document structure.

[7] http://www.webitext.ca/

4.1 From sentences to tokens

Both the abstract and the sentences from Giga-F/E need to be tokenized before we can apply our similarity checks. We tested 2 token sizes: single words and bigrams.

4.2 Weighting the tokens

As is standard practice, we filter out stop words, and then for remaining words, we tested three different weighting systems:

(a) uniform weights on all tokens in abstract and corpus sentences;
(b) using a decay factor that makes the weight of each token inversely proportional to its distance from the term of interest; in practice we limited the context to a window of +/- 10 token and took each intervening token to reduce the weight by an extra 10%;
(c) a combination of the two strategies above: uniform weighting for the abstract and decay factor for the sentences of Giga-F/E.

4.3 Similarity measures

We tested four standard similarity measures: cosine, overlap, Jacquard and Dice. These measures, combined with the tokenization and the token weighting system, provide a range of different similarity scores.

5. Results from contextualisation

We define a baseline performance as follows. We query Giga-F/E (using Lemur) with the Eng-

lish side of each term pair from our test set (a set of term of E/F term pairs – see section 2.1). Then, we calculate the average rank of the first returned sentence pair that contains the exact same translation as in the reference term pair. In so doing, we are ignoring the possibility that some of the higher ranking examples might have been an inflectional variant, a spelling variant or an otherwise different but acceptable translation. Thus, our baseline constitutes a lower bound on the real performance of the search process. The performance of our context-based re-ranking will be measured in the same way and the comparison should be significant and unbiased.

Table 5 – Ranking results

Method	Token size	Weight	Rank
Dice	Word	uniform	**1.92**
Cosine	Bigram	uniform	1.95
Overlap	Bigram	uniform	1.97
Jacquard	Bigram	uniform	2.00
Cosine	Word	uniform	2.00
Dice	Bigram	uniform	2.01
Overlap	Word	uniform	2.17
Cosine	Word	combination	2.17
Jacquard	Word	uniform	2.17
Cosine	Word	decaying	2.20
Dice	Word	combination	2.21
Dice	Word	decaying	2.22
Overlap	Word	combination	2.24
Overlap	Word	decaying	2.25
Jacquard	Word	combination	2.27
Jacquard	Word	decaying	2.27
Overlap	Bigram	combination	2.68
Cosine	Bigram	combination	2.69
Overlap	Bigram	decaying	2.69
Cosine	Bigram	decaying	2.71
Jacquard	Bigram	combination	2.72
Jacquard	Bigram	decaying	2.73
Dice	Bigram	combination	2.75
Dice	Bigram	decaying	2.77

Over all 10525 terms, we get a baseline figure of 4.80. In other words, the right equivalent would typically be found while examining the fifth sentence pair that was automatically extracted from our huge corpus. This is an interesting result in itself: even without any contextual re-ranking, the search process tends to bring correct answer close to the surface.

To measure the impact of contextual re-ranking we tested some 24 combinations of values for the parameters described above (2 types of tokens, 3 weighting systems, 4 metrics). Table 5 shows the results.

Our best combination turns out to be the use of the dice similarity metric on unigrams with uniform weights. With this combination the average rank gets 2.5 times closer to the top: from the baseline 4.80^{th} to 1.92^{nd}.

5.1 Polysemy

There is a simple reason why our baseline rank is rather good: many terms happen to be monosemous. For all those, contextualization techniques are rather irrelevant. In order to test the effect of contextualization where it is most relevant, we separated polysemous terms from monosemous ones. To that end, we used the *disambiguation page* facility of Wikipedia. Whenever a polysemous term receives several entries in Wikipedia, the user is invited to pick the appropriate one on a disambiguation page that list all available entries, each accompanied with some semantic cues. For our experiment we simply assumed that Wikipedia entries that lead to a disambiguation page are polysemous while the rest are not. By that metric, 731 of the 6632 terms found in Wikipedia are polysemous. Table 6 shows the detailed results.

Since Wikipedia is far from complete (cf Table 3), many of the terms it records as monosemous will prove to be polysemous in actual fact. Thus, the proportion 731/6632 should be considered as a lower bound on the proportion of terms that are polysemous in our test set. Recalculating our baseline on this subset of 731 polysemous terms, we find that the average rank falls from 4.80^{th} down to 11.57^{th}. Unsurprisingly, finding a good match is substantially harder for polysemous terms. Next, running our contextualized search process on the same subset, we obtain the results of Table 7.

Table 6 – Terms found in Wikipedia

Code	Nb terms	Wiki pages	Ambig. Terms
Bcb	725	559	66
Cgj	1047	522	107
Cjb	1905	1221	97
Cjc	1820	1008	93
Cjce	1238	694	152
Cjm	1431	883	68
Cjpp	1470	1048	80
Gen	1173	697	68
Total	10809	6632	731

Table 7 – Ranking results on polysemous terms

Method	Token size	Weight	Rank
Dice	Word	uniform	**4.30**
Overlap	Bigram	uniform	4.43
Jacquard	Bigram	uniform	4.46
Dice	Bigram	uniform	4.46
Cosine	Bigram	uniform	4.51
Cosine	Word	uniform	4.89
Dice	Word	decaying	4.92
Cosine	Word	decaying	4.96
Cosine	Word	combination	5.06
Dice	Word	combination	5.13
Jacquard	Word	combination	5.19
Overlap	Word	combination	5.22
Jacquard	Word	decaying	5.28
Overlap	Word	decaying	5.28
Overlap	Word	uniform	5.98
Jacquard	Word	uniform	5.99
Dice	Bigram	decaying	6.66
Dice	Bigram	combination	6.79
Overlap	Bigram	decaying	6.85
Overlap	Bigram	combination	6.88
Cosine	Bigram	decaying	6.89
Jacquard	Bigram	decaying	6.94
Jacquard	Bigram	combination	6.99
Cosine	Bigram	combination	6.99

The best combination of parameters remains the same. On average, contextualization pushes the first good sentence pair up substantially closer to the top, from the 11.57th to the 4.30th position. Again, we are not claiming that we have found the best possible contextualization technique but only that contextualization techniques in general do lead to significantly improved search results.

6. Related work

Our core findings are the following: (1) the terminological coverage of very-large-scale corpora is now becoming far superior to that of existing term banks; and (2) there are some simple techniques that can be used to focus the search process on the more contextually-relevant examples within such massive corpora.

Concerning the first point, we are not aware of any previous work comparing the terminological coverage of bilingual corpora to that of term banks, as we have done in section 3. While many corpus linguists may not be surprised by the results we are reporting, we suspect that the opposite will be true for translators and terminologists.

Our finding that context-sensitive matching can be useful in terminology search is a rather direct echo to some recent findings in machine translation. For example, Carpuat & Wu (2007) show that applying context-based word sense disambiguation techniques to the phrase pairs of a phrase-based SMT system at decoding time will improve the BLEU score of the resulting translations. Foster & Kuhn (2007) describe mixture-model adaptation techniques for SMT systems. Their main idea is to split a training corpus into a number of more homogeneous components and to assign each component (i.e. each "context") a weight that depends on its similarity with the text to be translated. Here again, improved BLEU scores are reported.

Like Carpuat & Wu, we focus on translation selection for basic translation units: in their case, the phrases of a phrase-based SMT system and in our case, technical terms. Unlike Carpuat & Wu, but like one of the options considered by Foster & Kuhn, we use an IR-style similarity metric for scoring contexts. But the main difference with the above-cited works is that our study is entirely focused on the more specific problem of translating technical terms. Consequently, our results are of direct relevance not only to machine translation but also to any corpus-based

terminology search tool, such as bilingual concordancers.

Barrière (2010) explores a different version of this idea of context-sensitive terminology search: while we apply it to very large corpora, she is rather applying it to large term banks in which relevant contexts are captured through the *domain codes* used in term banks rather than the words found in corpus examples. The goal is then to re-rank the term bank records rather than corpus examples. The benefits reported by Barrière for contextualized search in term banks are similar to those reported here.

7. Conclusions

We have shown that the terminological coverage of the largest parallel corpora is now becoming substantially better than that of mainstream term banks. We have also shown that it is possible to use simple techniques from the field of information retrieval to quickly narrow down on contextually-relevant examples in very large corpora. As far as we know, no search tools of that kind are widely available yet. But given their huge potential for language workers, we predict that this situation will change very soon.

This will raise the question of the division of labour between terms banks and corpus search tools. While these two different resources might conceivably continue to evolve separately and compete for the attention of the users, a more interesting possibility lies in the development of approaches and tools that will bring progressively higher levels of integration between them. Clearly, corpus analysis techniques can help speed up the development of larger-coverage term banks. But one can also ask whether or not it is possible to leverage the contents and organisational structure of term banks in improving corpus search.

Acknowledgement

We thank the Office Québécois de la Langue Française for providing us with the GDT data for research purposes.

References

1. Barrière C. (2010) Recherche contextuelle d'équivalents en banque de terminologie, 17ième conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010), July 2010, Montreal.
2. Brown, R.D. (2005) Context-Sensitive Retrieval for Example-Based Machine Translation, MT Summit X, Phuket, pp. 12-16.
3. Callison-Burch C. (2009) Corpora released at http://www.statmt.org/wmt09/translation-task.html
4. Callison-Burch C., Koehn P., Monz C., Peterson K., Przybocki M. and Zaidan O. (2010), Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation, Proceedings of ACL 2010 Fifth Workshop on statistical machine translation, July 2010, Uppsala.
5. Carpuat M., Wu D. (2007) Improving Statistical Machine Translation using Word Sense Disambiguation. In Proceedings of EMNLP-CoNLL, Prague, June 2007.
6. Foster G., Kuhn R. (2007) Mixture-Model Adaptation for SMT, Proceedings of the ACL Workshop on statistical machine translation, Prague, June 2007.
7. Isabelle P.: Bi-Textual Aids for Translators, Proceedings of the Eight Annual Conference of the UW Centre for the New OED and Text Research, University of Waterloo, Waterloo, Canada, 1992.
8. Macklovitch E., Lapalme G., Gotti, F. (2008) TransSearch: What are translators looking for? AMTA'2008 – The Eight Conference of the Association for Machine Translation in the Americas, Waikiki, Hawai.
9. Moore R.C (2002) Fast and Accurate Sentence Alignment of Bilingual Corpora, AMTA'02 Proceedings of the 5th Conference of the Association for machine Translation in the Americas, Tiburon, CA, USA.
10. Salton G. (1989) Automatic Text Processing, Addison-Wesley.

[i] This research was performed while Caroline Barrière was working at the Institute for Information Technology, National Research Council of Canada.