The Bilingual Concordancer TRANSSEARCH

Guy Lapalme, Philippe Langlais, Fabrizio Gotti
RALI - DIRO - Université de Montréal
CP 6128 Succ. Centre-Ville
Montréal, Québec, Canada, H3C 3J7
{lapalme,felipe,gottif}@iro.umontreal.ca

Abstract

TRANSSEARCH is a web-based translation search engine. When a user submits a translation query, the system replies with a set of sentence pairs whose source sentence contains the query. The source expression is highlighted and, with the help of statistical word alignment techniques, the corresponding target expression is also identified. When many sentences share the same translations, the translations are grouped and sorted in decreasing order of frequency to give the translators a variety of different translations whose contexts can be further explored.

1 Introduction

Despite the impressive amount of studies devoted to improving the state of the art in machine translation, translation memories remain the preferred solution for human translators when publication quality is of concern. This demonstration presents the translation search engine TRANSSEARCH. This web-based commercial application, aimed at language professionals, relies on a sentence-aligned bitext and statistical word alignment techniques. It is currently used by more than 3000 professional translators.

TRANSSEARCH is a web-based translation memory, developed by RALI\(^1\) and commercialized by Terminotix\(^2\), a Montréal-based company specializing in computer-aided translation tools. TRANSSEARCH is a translation search engine very popular among professional translators (Macklovitch et al., 2000; Macklovitch et al., 2008).

In the production version of TRANSSEARCH, when a user submits a translation query, the system replies with the set of sentence pairs whose source sentence contains the query. The translator then has to mine this material to discover the corresponding translation in each target sentence. This demonstration presents a new version of this application that exploits statistical word alignment techniques to turn TRANSSEARCH into a translation search engine.

Colloquial expressions, especially adverbial phrases such as in keeping with, are difficult to translate correctly because they depend on the context of use and are hard to find in bilingual dictionaries. Translators want to see many contexts of use to determine the most appropriate translation for their case. Bitexts of aligned sentences of previously translated texts are a useful source of inspiration to find solutions to difficult translations problems.

Figure 1 is a screenshot of TRANSSEARCH after the user has submitted the query in keeping with. TRANSSEARCH found 668 occurrences of the query for which 169 different translations were discovered by the system in the proceedings of the Canadian parliament (Hansards). The system presents the distribution of the translations of the query on the left. Each translation is presented with its frequency of occurrence. For each translation, the user can consult its contexts of occurrence on the right. For instance, the translation conforme à occurred 146 times in the corpus. To see the occurrences of another translation, e.g. fidèle à for

\(^1\)http://rali.iro.umontreal.ca
\(^2\)http://www.terminotix.com
Figure 1: Searching for two French translations of *in keeping with* with TRANSSEARCH. The top part shows an excerpt of the results for *conforme à* and the bottom part for *fidèle à*. In each screen, the list on the left indicates identified translations with their frequency of occurrence and the panel on the right shows the corresponding contexts in which this translation was used. Clicking on a different translation updates the panel on the right with the new contexts. Sentences that appear over a darker background are sentences in their original language while the sentence over a white background are their translations.
which there were 11 occurrences, one only clicks on it and the panel on the right is updated immediately with the occurrences of this new translation.

2 Underlying Technology

TRANSSEARCH must first identify the set of sentence pairs containing the query (Step 2 in Figure 2), this is a straightforward application of search engine methods provided each pair is considered a document to be indexed. Given the fact that the system is aimed at language professionals, TransSearch carefully handles morphological variants of French and English query words. For example, the query go+ will search for go, but also for goes, went, going and gone. This is the output of the current version of TRANSSEARCH.

![Figure 2: Processing steps of TRANSSEARCH starting from the translation memory of bilingual pairs (Step 1) to the display of groups of translations (Step 6).](image)

Once the sentence pairs are retrieved, the translation must be identified in the sentence written in the target language. This will be the French sentence if the query is in English. This process is called translation spotting, relabeled here as transpotting (Step 3 in Figure 2). We call transpot the target word-tokens automatically associated with a query in a given pair of sentences. For instance, in Figure 1, conforme à and respecte are 2 out of 169 distinct transpots displayed to the user for the query in keeping with. Transpotting is performed using word alignment techniques that combine HMM and IBM translation models. TransSearch also assumes that, since the query is a sequence of contiguous words, so should be the transpot. This hypothesis simplifies transpotting, which must be performed on the fly when the user queries the system. Because between 20% and 40% of the identified transpots are not acceptable translations, they are filtered (Step 4 in Figure 2) in order to keep only correct and diversified translations. Filtering is performed using a supervised classification process. (Bourdaillet et al., 2010) describes in detail the many alternatives that have been tested and evaluated for the transpotting and filtering processes.

Once translations have been identified using these methods, it becomes possible to merge the translations (Step 5 in Figure 2) that are identical or at least very close, for example if they differ only by inflection, conjugation or by a few words. Figure 1 shows some flexional variations of conforme à. Merging is very important for the user because it groups uninteresting variations of the same translations and, in some cases, hides deficiencies of the word alignment process. This can be seen in the third context of the top at Figure 1 in which conforme à should have been underlined instead of only conforme.

TRANSSEARCH displays each translation only once (Step 6 in Figure 2), but allows the user to view all its contexts of occurrence. The translations are sorted in decreasing order of their frequency, making the assumption that the most frequent translations are the most likely ones. They are displayed in a web interface, loosely inspired from the one pioneered by Linear B (Callison-Burch et al., 2005). Similar interfaces have been developed for other web based concordancers such as Linguee and Tradooit.

CSS and Javascript are used to selectively display translation pairs. The list of transpots is presented to the user who can then browse the associated pairs of sentences by clicking on each translation. Both the query and its translation are highlighted in all

---

3The user does not have to specify the language of the query, TRANSSEARCH assumes that it is the language in which the query was found most often.


pairs of sentences. TRANSSEARCH is also available as a web service easily invoked from other applications. Terminotix has developed a TRANSSEARCH toolbar to be used from within Microsoft Word. See (Bourdaillet et al., 2010) for details of the merging processes and the evaluations by both machine and man that have been carried out in the course of developing TRANSSEARCH.

TRANSSEARCH has also been used in studies of linguistic phenomena such as discourse connectives (Meyer et al., 2011) and idiomatic expression (Huet et al., 2010).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>docs</th>
<th>M pairs</th>
<th>M words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hansard</td>
<td>2 854</td>
<td>10.5</td>
<td>339.6</td>
</tr>
<tr>
<td>Senate</td>
<td>1 025</td>
<td>1.2</td>
<td>47.6</td>
</tr>
<tr>
<td>Canadian courts</td>
<td>13 068</td>
<td>3.2</td>
<td>142.6</td>
</tr>
<tr>
<td>Total</td>
<td>16 947</td>
<td>14.9</td>
<td>528.2</td>
</tr>
</tbody>
</table>

Table 1: Size of the Transbases used by TRANSSEARCH

TRANSSEARCH transbases (translation data bases of French and English sentence pairs) contain more than half a billion words (Table 1). They contain the Canadian Hansards (Canadian parliamentary debates) since 1986, but also from the debates of the Senate of Canada since 1995. As TRANSSEARCH is also used by the translators of the Canadian Translation Bureau that produces the Hansards, these databases are updated daily when new texts are produced and translated. TRANSSEARCH also contains more than 13K decisions from Canadian courts since 1968.

3 Conclusion

This paper describes the bilingual concordancer TRANSSEARCH which features word alignment. Interestingly, this transforms the nature of the application: it now behaves like a translation finder with a concordancer feature. The application goes beyond a bilingual dictionary thanks to its ability to find phrase translations while providing their contexts of occurrence. TRANSSEARCH can be accessed on a 5-day trial basis at:

http://transsearch3.com

Acknowledgments

We thank Julien Bourdaïlet and Stéphane Huet for their research contributions as postdoctoral fellows at RALI. We also thank Gilles Gamas, Micheline Cloutier and Jean-François Richard from Terminotix who supported this research via their contribution to a Collaborative Research and Development (CRD) grant from the Natural Science and Engineering Research Council (NSERC) of Canada. We thank also Jacques Steinlin and Elliott Macklovitch for their collaboration to this work.

References


