Parallel Corpora as a Linguistic Recourse and their Applications

Natallia RUBASHKO
Belarusian State University, Minsk, roubashko@bsu.by

Abstract: *This paper focuses on investigation of the parallel corpora role as a linguistic recourse. The applications based upon parallel corpora are growing in number: multilingual lexicography and terminology, machine and human translation, cross-language information retrieval, etc.*

***Keywords*:** linguistic resource, parallel corpora, extraction of linguistic meta-knowledge, multilingual text alignment.

1. INTRODUCTION

With the rising importance of multilingualism in language industries, parallel corpora, consisting of source texts along with their translations into other languages, have become key resources for the development of natural language processing tools [1].

Parallel corpus means a text which is available in two (or more) languages: it may be an original text and its translation, or it may be a text which has been written by a consortium of authors in a variety of languages, and then published in various language versions. A corpus of this type of text is sometimes called a “comparable corpus”, though this term is also used (confusingly) for a corpus of similar but not necessarily equivalent texts. Another term is *bitext* [2].

Parallel corpora are a valuable source of a kind of linguistic metaknowledge, which forms the basis of techniques such as tokenization, POS-tagging, morphological and syntactic analysis, which in turn can be used to develop different applications [3].

Parallel corpora have taken on an important role in machine translation and multilingual natural language processing. They represent resources for automatic lexical acquisition, they provide indispensable training data for statistical translation models and they can provide the connection between vocabularies in cross-language information retrieval.

For these reasons, parallel corpora can be thought of as a critical resource.

The article deals with:

1. Techniques and methodology for the alignment of parallel texts at various levels such as sentences, clauses, or words.
2. Applications of parallel texts in fields such as translation, terminology and information retrieval.

2. ALIGNMENT METHODOLOGY

In order to extract information from parallel text, it is first necessary to **align** the two texts at some global level, typically paragraph or sentence. By "align" is meant the association of chunks of text in the one document with their translation or equivalent text in the other document. Some approaches to alignment involve the use of some sort of traditional analysis of the texts (e.g. parsing, tagging, or the use of bilingual lexicons), while others take an entirely automatic approach.

Most alignment programs base on the simple assumption that there is a significant correlation in the **relative length** of texts which are translations of each other.

Most of the approaches have in common a technique which involves identification of **anchor points** and verification of the comparability of the textual material between the anchors. These anchors can, in the simplest case, be structural, when sentence boundaries are taken to make an initial segmentation. Then, certain types of alignment across sentence boundaries are permitted and quantified (e.g. where two sentences in one text are merged in the translation, or vice versa), with all possible alignments being compared using dynamic programming techniques.

Alternatively, and quite commonly, translation **word-pairs** are taken as the anchor points. This alignment at the word level is often an end-goal in itself [2].

Apart from automatic estimation of translation pairs, a number of sentence alignment algorithms rely on **machine-readable dictionaries** as a method for finding lexical anchor points. This technique of course relies on the availability of a suitable dictionary, not to mention the need for efficient lemmatization in the case of highly inflected languages. Again, this seems like a vicious circle if the aim of the alignment is to extract vocabulary; but aligned parallel corpora can be used for the extraction not of everyday vocabulary, but of domain-specific lexical pairings, notably novel terminology and, especially where different writing systems are involved, transliterations of proper names.

There are several tasks concerning text aligment:

1) *Sentence alignment task.* Two subtasks have been defined: witn segmented corpus and raw corpus.

The main approach for solution is based on the assumption that in order for the sentences in a translation to correspond, the words in them must also correspond. Only internal information is used. In other words, all necessary information (and in particular, lexical mapping) is derived from the to-be-aligned texts themselves.

2) *Word and expression alignment task*. Evaluation of words alignment between parallel texts presents more difficulties than the evaluation of sentence alignment, given the differences in word order between languages, the difference in part-of-speech and syntactic structure between the source and its translation, the discontinuity of multi-token expressions, etc. As a result, research is less advanced than in the area of sentence alignment [1].

There are a variety of techniques, including bootstrapping and relaxation, that perform the two types of alignment at the same time. However, in a sentence alignment, word alignment is not the primary goal and as such, it is at best a by-product with no inherent significance On the other hand, when the primary goal is word alignment, one can no longer settle for rough and partly erroneous alignments.

3) *Clause and sentence structure alignment task.* Another research area that seems to be growing rapidly is the alignment of linguistic segments with a span longer than the word or the expression but shorter than the sentence.

These include clauses, syntax tree fragments, and skeleton sentences. An alignment of this type would be very useful for a variety of applications including example-based translation, language teaching, and comparative linguistics. But the problem is extremely hard to solve due to the problems in detecting clause boundaries in each language, difficulty in coming up with even a partial syntactic analysis, and substantial structural differences across languages, even related ones.

Classical methods for parallel text alignment consider one specific level (e.g. sentences) at which two or more versions of a text are synchronized. This may lead lo some problems when these documents are particularly long since alignment errors at some point in the text may, in the absence of any other linguistic information, propagate for some time without any chance of recovery.

Many existing translators' tools and machine translation strategies depend on aligned text segments. An alignment does not permit crossing correspondences, so it is a special case of the more general correspondence relation.

3. EXTRACTION OF BILINGUAL VOCABULARY AND TERMINOLOGY

Parallel bilingual texts are a valuable source of information to advanced language learners, particularly in the area of lexis, subtle lexical dependencies. Typically this information is either not available or sporadically available only in very large dictionaries.

Algorithms for bilingual lexicon extraction from parallel corpora exploit the following characteristics of translated, bilingual texts:

– words have *one sense* per corpus;
– words have *single translation* per corpus;
– *no missing translations* in the target document;
– *frequencies* of bilingual word occurrences are *comparable;*
– *positions* of bilingual word occurrences are *comparable.*

Most translated texts are domain-specific, thus their content words are usually used in one sense and are translated consistently into the same target words. Pairs of sentences from both sides of the translated documents contain the same content words, and each word occurs in approximately the same sentences on both sides. Once the corpus is aligned sentence by sentence, it is possible to learn the mapping between the bilingual words in these sentences.

To identify likely word-pairs is to find word pairs which are most probably alignable on the basis of similar **distribution.** This distribution is defined in terms of text sectors, and Dice's coefficient is used to quantify the probability. Dice's coefficient (1) is a simple calculation which compares *c,* the number of times the two candidate words occur in the same sector with *a* and b, the number of times the source or target words occur independently.

$$Dice = \frac{c}{a+b} \qquad (1)$$

The algorithm is iterative in that the sentences containing high-scoring word pairs are established as anchors which allow the text to be split up into smaller segments, affording more and more results.

In addition to the Dice coefficient, other similarity measures widely used in Information Retrieval (IR) could also be applied in this case, including the Jaccard coefficient (2) and the Cosine coefficient (3):

$$Jaccard = \frac{c}{a+b-c} \qquad (2)$$

$$Cosine = \sqrt{\frac{c}{ab}} \qquad (3)$$

One major drawback to all the techniques described above is the necessary assumption that word equivalents are on a 1:1 basis. Apart from the fact that this is generally not always true in languages (even making the prior assumption that we know what a word is!), it is especially unhelpful if the aim is to identify bilingual terminology. It is in this endeavour that parallel corpus work comes to the fore: often, the goal of extracting parallel vocabulary is undermined somewhat by the existence of machine-readable bilingual dictionaries. Specialist terminology, however, is almost always absent from such resources, and bilingual parallel corpora become the primary -perhaps only - source of such material.

Searching for multiword terms in a parallel corpus introduces a further aspect of word distribution which can be addressed by statistical means: considering the corpora independently, we can search for likely terms by looking for **collocations,** i.e. sequences of words which co-occur frequently and - if we are lucky - tend not to occur on their own.

There is a range of measures for (monolingual) collocations. The z-score is perhaps the most familiar: it quantifies the collocational force of one word w_i with respect to another w_j as in (4):

$$z = \frac{O-E}{\sigma} \qquad (4)$$

where *O* is the observed frequency of w_i co-occurring with w_j (in close proximity, or contiguous with it, as the case may be), *E* the expected frequency of w_i, and σ is the standard deviation of occurrence of w_i in the whole text as given by (5):

$$\sigma = \sqrt{N(p(1-p))} \qquad (4)$$

where *p* is the probability of occurrence of w_i, and *N* is the total number of word tokens in the text.

Once candidate terms have been determined monolingually, the attention can be turn to identifying their translation equivalents in the parallel corpus. Many of the experiments reported seem to take the same basic approach, namely identifying possible terms monolingually and then searching for their translation in an aligned parallel corpus [2].

Besides bilingual vocabulary and terminology, aligned parallel corpora have been used to extract translation templates, especially for the purposes of Example-based Machine Translation and the related translator's tool, Translation Memory.

The basic idea is to reuse examples of already existing translations as the basis or model for a new translation. In its basic form, the examples are stored as pairs of aligned text fragments, usually sentences, though in some implementations stored examples are annotated with POS tags or other information, or even stored as linked tree structures. Translation proceeds by first matching the input with a suitable example, then adapting the example to the new case.

There are certain assumptions about the nature of parallel corpora.

1) *Words have one sense per corpus.* This is the basic assumption underlying the "sublanguage" approach to natural language processing. It is often true, especially for words which have terminological status; but homonymy is not avoidable, even in narrow domains.

2) *Words have a single translation per corpus.* This is a much less safe assumption, which is particularly undermined by the fact that inflectional morphology and compounding methods differ from language to language. The assumption of 1:1 word correspondence is of course naive, bearing in mind polysemy and homonymy, and the converse problem of translation divergence. The assumption is undermined further by the fact that local syntactic conditions might result in inflectional morphology in one language but not the other: in particular, the distribution of singular and plural can differ widely between otherwise closely related languages, without even considering grammatical case and gender. Where possible, this can be overcome by subjecting the corpora to a process of **lemmatization.**

3) *There are no missing translations in the target document.* This is a somewhat safer assumption, but not entirely so. It is not unusual to find that some portion of the source text has been omitted in the target text, either through carelessness, or because it does not apply to the target-language readership. Interestingly, one off-shoot of work on alignment has been the development of tools to help translators check for missing text in translations.

4) *The frequencies of words and their translations are comparable.* The main problem with this assumption is again the fact that a single word in one language can have a variety of translations in the other just because of grammatical inflection.

5) *The positions of words and their translations are comparable.* This seems to be the most fundamental of assumptions in alignment. The extent to which it is true depends on the granularity of the alignment. Clearly, at sentence level it is hardly true at all: word-order is a fundamental difference between many languages, not just the obvious case of, say, adjectives preceding or following the noun, but also the relative order of main and subordinate clauses *(A because B* vs. *B and so A,* for example). But as the size of the text element being considered grows, this effect becomes minimised. For some language pairs, there remains a certain amount of "scrambling".

4. TRANSLATION

Most researchers have given up on the idea of a fully automated, high-quality machine translation, at least as a short- or middle-term goal. Today's main research trends are spread across a continuum, with machine-assisted human translation at one end and human-assisted machine translation at the other.

Parallel corpora can be a valuable tool and resource. If the enormous volume of translations could be put to use in a systematic way, it would supply translators with a solution to far more problems than all the dictionaries in the world. The answer to many translation problems could no doubt be found in the huge bulk of already existing translations.

5. CROSS-LANGUAGE INFORMATION RETRIEVAL

Cross-language information retrieval aims to find all of the corresponding documents in one or more other languages from that of the query in one language. Cross-language queries are necessary when there are many users who are bilingual enough to understand documents written in another language but not necessarily capable of expanding queries with the appropriate keywords. Various techniques have been proposed including the fully machine-based translation of queries, or simply their word-for-word translation into the other language with the help of bilingual dictionaries. However, none of the techniques are fully satisfying due to the shortcomings of machine translation and the imperfections of bilingual dictionaries, which do not necessarily cover all areas of the document base. This is another case where parallel texts, if they exist, can be put to valuable use even though only for pan of the document base.

6. CONCLUSIONS

The article has looked at a range of issues related to bilingual parallel corpora.

The early stages of multilingual alignment systems paralleled the increasing interest which the computational linguistics research community paid to the use of corpora in exploring the reality of languages as they are expressed in speech or text. As a result, progress in this area has gone through the same exploratory realms as other techniques in the field. Indeed, the first attempts to put into correspondence a text and its translation were based on the idea that source and target texts had to be aligned at the level of granularity of sentences.

7. REFERENCES

[1] N. Rubashko. To the Problem of Creation of Parallel Corpus for Belarusian and Russian Languages. *Proceedings of the International Conference "Information Systems and Technologies (IST'2002)",* Minsk, Belarus, 5-8 November 2001, V.1. pp.148-152. (in Russian)

[2] Jean Véroni. *Parallel text processing: alignment and use of translation corpora*. Kluwer Academic Publisher, 2000. p. 413. – available at http://books.google.by/

[3] I. Dan Melamed *Empirical methods for exploiting parallel texts*. Cambridge, MA: The MIT Press, 2001, xi+195 pp. – available at http://books.google.by/