Automatic Evaluation of Alignments without using a Gold-Corpus - Example with French-Japanese Aligned Corpora

Raoul Blin
CNRS-CRLAO
105 Bd Raspail, 75006 Paris, France
blin@ehess.fr

Abstract

This paper presents a test designed to automatically evaluate the alignment quality of a bilingual aligned corpus without comparing it to a reference/gold corpus. We justify bypassing a reference corpus and present the advantages of the test, particularly for language pairs where few bilingual aligned corpora exist. The test, which is based on the observation of single-translation words, is demonstrated using three bilingual French-Japanese corpora : a manually translated and aligned corpus, a freely translated and semi-manually aligned corpus, and a freely translated and automatically aligned corpus. This paper shows that the results achieved validate the proposed evaluation technique.

Keywords: Evaluation, Alignment, Japanese, French, OPUS

1. 1 Introduction

There is a great and undoubtedly growing need for bilingual and multilingual aligned corpora. Aligning corpora semi-manually is costly in terms of time and human resources, hence the need for automatic alignments. This poses the question of how we can evaluate the quality of such alignments. Reviewing the literature on this subject (for example (Langlais, Véronis & Simard, 1998)) along with subsequent research (for example Cherry & Lin (2003), Matusov, Zens & Ney (2004), Li, Sun & Xue (2010), etc.) shows that the evaluation method generally adopted involves comparing the aligned corpus to a gold corpus. The gold corpus itself is a semi-manually aligned subset of the corpus being evaluated. Using a gold corpus both impedes and is contradictory to the production of automatically aligned corpora. Indeed, as we have already seen, semi-manually aligning a corpus is costly. It would therefore be more efficient to evaluate alignments without relying on gold corpora.

Existing methods that bypass gold corpora are problematic for several languages. Firstly, manual evaluation is extremely costly if scientific procedures are to be respected (double or triple evaluation by experienced bilingual evaluators followed by calibration and re-evaluation). Secondly, such an evaluation can only be conducted on samples, whereas automatic evaluation is assumed to encompass the entire corpus. The use of identical words (Simard, Foster & Isabelle, 1993 ; Zhang et al., 2005) is only possible for languages employing the same writing system. Lastly, hapaxes (Lardilleux & Lepage, 2008) require large bilingual corpora which unfortunately are lacking in many language pairs, including Japanese-French, our focus here.

In this paper, we suggest a simple method to *evaluate* the quality of alignment by observing single-translation words. Single-translation words (abbreviated henceforth as *stwords*) are words with only one possible translation (Langé & Gaussier, 1995) but are not necessarily monosemic. For example, although “John” denotes many men in the world, it is always translated as “jon” in Japanese (nevertheless, see the discussion in section 2.2). It can therefore be considered an stword. We must insist here that the method is not designed to make alignments, even though it can be used for this purpose. This method can be applied to any kind of bisegment: words, chunks, sentences and so on. In section 2, we set out the principles of the method; section 3 describes the results of an experiment conducted on two French-Japanese aligned corpora; and section 4 consists of a discussion of these results.

We originally used this method to evaluate the very small number of existing Japanese-French bilingual aligned corpora, hence the focus on this language pair here. The linguistic particularity of this pair is that both languages possess different syntactic properties, including word order, and use different writing systems. They also represent a very common situation: namely, that each language has long been well-endowed with monolingual corpora but has few bilingual alignments (compared with Japanese-English or European language pairs, for example). The same applies to most Japanese-European language pairs (except English), where bilingual aligned corpora are rare. In section 4 we suggest ways to apply our procedure to other language pairs.

2. Evaluation Based on the Observation of Single-translation Words

2.1 The Base

The purpose of the tests is to evaluate (not to make) alignments of texts which do not necessarily use same characters, same syntactic structures and which can be short. Also, the texts are neither (necessarily) annotated nor tokenized.

We use two tests, both based on the observation of stwords. Let us consider an aligned bitext. The source text contains stwords; the target text, the translations of these stwords. Ideally, if the corpus has been correctly aligned, any source segment with a given number (N1) of occurrences of an stword should be aligned with a target segment containing the same number (N2) of translations of this stword. The difference between the two numbers is

the evaluation criteria. N1>N2 means that at least one sub-segment that contains an occurrence of the translation of the stword is missing in the target segment, and that this sub-segment belongs to another target segment which has been wrongly aligned with another source segment. Conversely, N1<N2 means that a sub-segment (containing a translation of the stword) of another segment is incorrectly aligned with the current source segment. Accordingly, the test consists in counting the number of lines for which N1==N2. The higher the result of this first test, the better the alignment.

Unfortunately, this test is not sufficiently reliable. One reason for this is that many languages substitute special structures (pronouns, etc.) for repeated occurrences of phrases (for example, pronouns are substituted for repeated noun phrases in French). Some languages, like Japanese, elide the repeated phrases. Substitution and elision depend on the distance between occurrences of the phrase. Defining this distance falls outside the scope of this paper. The only thing we can say for sure is that the first occurrence of a proper noun (PN) within a text is systematically translated as word for word. For our purposes, it is these first occurrences of a PN which are reliable. Accordingly, a second test must be conducted. Let us call FO^{src} the first occurrence of a PN in the source text, and $NLFO^{src}$ the number of bisegments in which the source segment contains an FO^{src}. Among these bisegments, let us call $NLFO^{tgt}$ the number in which the target segment contains at least one occurrence of a translation of an FO^{src}. We then obtain ($NLFO^{tgt}$ / $NLFO^{src}$), which can be considered the recall value.

We need to use stwords that can be extracted without morphosyntactic preprocessing of the bitext. For obvious reasons, we want to avoid manual preprocessing. There are two reasons why we also want to avoid automatic preprocessing. The first is that automatic analyzers can induce wrong POS tagging and consequently errors while evaluating the alignment. The second is that, for most languages, automatic parsing is performed by statistical analyzers that require large training (specialized) monolingual corpora. While this is not a problem for languages like Japanese or French, for others it can be difficult to find efficient analyzers and build suitable corpora.

The reliability of the evaluation depends on the number of stwords in a text, and on the number of segments which contain (first) occurrences of PN. We will provide these values too.

2.2 Proper Nouns as Single-translation Words

Unfortunately, “perfect” stwords do not exist in the case of Japanese-French. However, it is assumed that PNs are very similar to stwords. Let us consider the advantages and disadvantages of PNs with regard the proposed test.

The first advantage is the existence of bilingual Japanese-French lexicons of PNs. Retrieving or building such resources is simple for a wide number of languages.

The second advantage is that for many languages, PNs are morphologically invariable. This implies that finding them in a text should be easy, even without morphological preprocessing. In the case of French and Japanese, there are graphic variations but these are easily predictable, few in number, and can be exhaustively (and automatically) listed. For example, in French, *Tokyo* can be written *Tôkyô* and sometimes *Toukyou*. In Japanese, it is usually written 東京, or perhaps using hiragana (とうきょう) or katakana (トウキョウ). Some PNs can also be written using the Latin alphabet.

The third advantage of PNs is that, provided certain rules are respected, as detailed below, PNs are easy to spot in French and Japanese, and morphosyntactic analysis is not required.

Unfortunately, for our purposes, PNs also have certain drawbacks.

1) Some common nouns are derived from PNs but not their translation. For example, the French CN *Français* ('French people') is translated by a derived PN in Japanese (ex.: *nihon-jin*, 'Japan-people'). The number of occurrences of PNs differs in bisegments that contain such word pairs. For example, consider the following French and Japanese sentences that correspond to the sentence "[French people]$_{CN}$ leave France$_{PN}$". The French sentence contains only one occurrence of the PN/stword *France*, whereas the Japanese sentence contains two occurrences of its translation, *huransu*, because *huransu* also appears in a derived CN.

Les Franç-ais$_{CN}$ quittent la France$_{PN}$.
<->
huransu$_{PN}$ - jin wa huransu$_{PN}$ wo deteiku.
France - peo. TOP France OBJ leave

To avoid such a shift without morphosyntactically preprocessing the text, the simplest solution would be to use the French corpus as the source corpus. This is convenient for other reasons, as set out below. We studied other strategies but rejected them on the grounds that they required morphosyntactic preprocessing of the text or lexicon.

2) Many of the Chinese characters used in Japanese can be both PNs or graphic components of morphs. For example, the character 順 *jun* can occur as a PN (the given name “Jun”) or as a graphic (as opposed to semantic) component in the CN 順序 (*junjo*, “order”), among others. However, these two occurrences are unrelated. Such ambiguity most often arises with PNs consisting of one Chinese character. Morphological preprocessing would eliminate such errors but might produce others. Given this, we preferred to simply eliminate all one-Chinese-character PNs from our lexicon.

3) The same ambiguities can arise in French too. For example, *Violette* is ambiguous because it can refer to the flower *violette* or to the girl’s name. Once again, in order to avoid errors without resorting to morphological parsing, we simply retrieved the 13,000 French morphemes that can occur as something other than PNs.

4) Use of PNs can differ by language. While French substitutes a pronoun for a repeated PN, Japanese allows for repetition. To account for this phenomenon, we added a second test based on the first occurrence of a PN in a text. In contrast, in Japanese, people are frequently designated by their name, whereas French uses a pronoun

(*tu*, *vous*, etc.). The number of anthroponyms may thus be lower in French than in Japanese, regardless of the quality of the alignment. This is the case in the following dialog with the interlocutor Tanaka.

Ja: *tanaka san wa dô omoimasuka?*
Tanaka Mr TOP how think?
(lit.: 'what Mr Tanaka thinks?')

Fr: *Qu'en pensez-vous?*
What think you?
'What do you think?'

There does not appear to be any way to avoid such errors, or to quantify the shift caused by this phenomenon. This shift will no doubt differ significantly depending on the type of corpus used, with dialogs being particularly affected.

Metonymy also affects results. For example, French frequently designates a governing body by the place where it is located. For example, “Paris refused”, meaning that “France/the French government (located in Paris) refused”. Such use of metonymy is not observed in Japanese, where the above sentence would no doubt be translated as *huransu ga kotowatta*, “France refused”. In this case, if *Paris* occurs in the French segment, its translation will not appear in the corresponding Japanese segment.

4) Some idioms use PNs but cannot be translated word for word. *Doux Jésus!* (meaning “Jesus Christ”) is translated as *masaka!* which is not a PN. Such idioms are well known. To avoid errors, any segments in which they occur can simply be excluded.

3. Experiment

We applied the proposed tests to three aligned bitexts (see quantified descriptions in Table 1). A manual evaluation of the alignments suggests that they differ significantly in quality. If the test is efficient, it should reveal this difference. The first corpus (OPUS-fj) is the Japanese-French sub-corpus of the OPUS project (Tiedemann & Nygaard, 2004; Tiedemann, 2012). This is actually the only freely available, *large-scale* aligned Japanese-French corpus. Several alignment units are used: graphic lines, sentences, syntactic phrases. We assume this corpus to be representative of the attempts to build large-scale aligned corpora automatically. To our knowledge, OPUS-fj has never been evaluated. A manual evaluation suggests that the alignments and translation are of poor quality. For example, in the large bitext OpenSubtitles2013, a one-line offset clearly appears. The second corpus (PUD[1]) is the test-corpus used for CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies. It is manualy aligned and translated by professionals. The alignment unit is the sentence. Unfortunately, it is small (only 1,000 sentences). We then use a third corpus: ALIGNJaFr_BABT-0.2_specialEval. It is composed of (semi-)professional translations that have been aligned automatically and manually corrected. Depending on the bitext of the corpus, the alignment unit may be the sentence or syntactical phrases. Assuming the test is efficient, the best score should be obtained with PUD, followed by ALIGN. The worse score should be obtained with OPUS-fj.

We used a lexicon (NP-fj.v0.2) made of 172K pairs of Japanese and French PNs, extracted from JaLexGram-v0.25 . 858 PNs written with only one character were excluded. We used Lefff (Sagot 2010). to exclude 731 Japanese PNs that can be mistaken for other part of speech.

In both tests, French is the source language and Japanese the target language. To count the PNs in a bisegment, we count all the PNs from the French segment, look for their translations in the list of PNs, and then count the occurrences of those translations in the target segment. In order to extract the PNs from the French text, we locate any words that begin with a capital letter and retain those that appear on the list of PNs.. When extracting a translation in Japanese, we simply look for a substring equal to this translation.

3.1 Test 1 : Observation of all Occurrences of Stwords

To evaluate the quality of alignment, we focus on the bisegments in which the source segment contains at least one PN. Let us call nbbiseg the number of such bisegments in a text. For each bisegment, let's call $PNsrc_i$ a PN which occurs in the source segment, $PNtgt_i$ its translation, occ(W) the number of occurrences of a word W in the segment (not the bisegment) where it occurs. For example $occ(PNsrc_i)$ is the number of occurrences of $PNsrc_i$ in the source segment. A “good” $PNsrc_i$ is such that $occ(PNsrc_i) = occ(PNtgt_i) =/= 0$.

We then calculate the proportion of “good” (source) segments which include only good PNsrc. The result can be interpreted as the recall score:

$$\frac{\text{number of “good” segments}}{\text{nbbiseg}}$$

	OPUS-fj	ALIGN	PUD
Nb of words (French)	12,672,676	202,687	20,543
Nb of bisegments	1,868,319	10,821	1,000
% of src segments with PN(s)	1.94	19.72	19,08
% good segments	56.40	72.54	91.62

Table 1: Results of test 1

Because ALIGN and PUD have been manually aligned, they should provide similar scores. The scores for ALIGN are lower for many reasons relating to translation rather than alignment. ALIGN-French frequently uses metonymy. For example, the 87 occurrences of “Washington” all refer in fact to the United States, not to the city. In ALIGN-Japanese, this word has therefore been translated as *amerika* or *beikoku* , not as *washinton*.

[1] Files for Japanese and French are downloadable from https://github.com/UniversalDependencies/

3.2 Test 2 : Observation of the First Occurrence of each Stword

The second test is similar to the previous one but only takes into account the first occurrence of a PN in each text (see the explanation of the counting method in section 2.1). nbbisegO is the number of bisegments in which the source segment contains only first occurrences of PNs. There is a slight difference in the definition of "good PNs": in test 2 a good $PNsrc_i$ is such that $occ(PNsrc_i) \geq occ(PNtgt_i) > 0$. We thus accept that a PN is translated only once in the target segment (the other occurrences can be elidated or replaced by pronouns). We then provide the percentage of "good" segments:

$$\frac{\text{number of "good" segments}}{\text{nbbisegO}}$$

	OPUS-fj	ALIGN	PUD
% of segments with at least one first occ.	0.15	4.97	11.49
% good segments	50.31	82.71	89.57

Table 2: Results of test2

3.3 Synthesis

We combine the above results in two scores. For each corpus, the first score is the average of test 1 and test 2. To emphasize the results obtained with the first occurrences of stwords when manipulating languages which do not repeat PNs, we provide a second score : (SPN+(2*SPNO))/3.

	OPUS-fj	ALIGN	PUD
% good segments (average)	53.35	77.62	90.59
Average (emphasize test 2)	52.34	79.32	90.25

Table 3: Synthesis of test 1 and 2.

For both tests using OPUS-fj, the values significantly differ depending on the sub-corpus. However, the overall score of OPUS-fj was dragged down by the low score of large subcorpora like Open Subtitle. Perhaps some sub-bitexts in OPUS-fj have been automatically (incorrectly) translated (see the discussion of this problem in automatically building corpora in Ruopp & van der Meer (2015)).

3.4 Reliability of Extraction Method

One particularity of our method is that it does not involve preprocessing the corpora. We carried out a qualitative comparison of this method with manual extraction and automatically POS-tagged texts. For this purpose, we compared extraction procedures using our method with a French corpus POS-tagged with TreeTagger (Schmid,1995) and a Japanese corpus POS-tagged with Mecab (Kudo, 2006) with the dictionnary mecab-jumandic[2]. Both TreeTagger and Mecab are commonly used in NLP. We used a test-corpus consisting of 100 sentences randomly extracted from ALIGN (version 1).

[2] mecab-jumandic 5.1.20070304-7

French: there were no errors using our extraction method of *known PNs*. But compound and unknown PNs are not take into account. TreeTagger take into account unknow PNs but not compound PNs. It encountered several errors. For example, most of the non-PN words positioned at the beginning of a sentence with a capital letter were wrongly interpreted as PNs, including adverbials like *malgré* "despite". Thus, preprocessing with Treetagger do not necessarily improve extraction from French.

Japanese: there were no errors using our extraction method but many occurrences of PNs were overlooked. By excluding PNs with one character, we missed 16 occurrences of translations. On the other hand, Mecab correctly analyzed all these short PNs. In addition, as we predicted, with both methods country names (ex.: *huransu* "France") were retrieved from the derived CN (*furansu-jin* "French people"). Mecab made errors on 7 nominal morphs which have been analyzed as PNs. It does not take into account compound PNs.

4. Discussion

As we can see in Table 3, for both tests best values were obtained with the manually aligned corpus PUD, followed by ALIGN, and then by the automatically-aligned corpus OPUS-fj. These results are in line with our expectations. We therefore assume that despite their simplicity, the two tests provide a reliable measure of alignment quality, even without using gold corpus.

We provided three scores that can be used differently. Test 2 produces the most reliable score because it is less sensitive to the syntactic and pragmatic differences between French and Japanese. However, it uses only part of the stwords. Its efficiency is therefore low when it is applied to corpora containing few stwords. For corpora of this kind, we prefer test 1, despite it being less reliable. Some people may prefer to have a single score rather than manipulate two scores. In such situations, we suggest a simple synthesis obtained by calculating the average of both tests. Of the two resulting scores obtained with this method, we suggest using the score that emphasizes the more reliable test 2.

The evaluation of alignments by hand or using gold corpora is reliable enough to be self-sufficient. However, the test proposed here no doubt has some weak points. Reliability depends on the frequency of the PNs and on the exhaustivity of the lexicon. While the test provides good information, it may not be reliable enough. It should be used in conjunction with other tests.

We are currently exploring additional ways of evaluating alignments without using a gold standard corpus or resorting to manual evaluation. One method is based on words written with Latin characters. These are easy to extract from Japanese and make it possible to evaluate the alignment with Japanese as the source language. The other evaluation method we are currently exploring is based on automatic translation created by systems trained on the corpus to be evaluated. Unfortunately, it is only applicable to large corpora. Ultimately, we expect to provide a global score based on the three tests.

Although we evaluated the test using French-Japanese aligned corpora, it can be applied to many other language pairs where the target language is Japanese. To avoid pre-processing the source language, the target language has to conform to at least two requirements. First, words need to be graphically separated. Second, PNs must be graphically distinguished from other words. For example, in French, PNs are marked by a capital letter. Such a mark is inefficient in German, where CNs also begin with a capital. Of course, a bilingual lexicon would also be necessary.

5. Bibliographical References

Cherry, C. and Lin, D. (2003). A Probability Model to Improve Word Alignment. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo (JA), pp.88-95.

Kudo, T. (2006). MeCab: yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net

Langlais, P. Véronis, J. and Simard, M. (1998). Methods and practical issues in evaluating alignment techniques ». In Proceeding of 17th international conference on Computational linguistics, Montréal (CA), pp. 711--717.

Lardilleux, A. and Lepage, Y. (2008). A truly multilingual, high coverage, accurate, yet simple, sub-sentential alignment method », in The 8th conference of the Association for Machine Translation in the Americas (AMTA 2008), Waikiki, Honolulu, United States, pp. 125-132.

Langé, J.-M. and Gaussier, É. (1995). Alignement de corpus multilingues au niveau des phrases », *TAL*, 36(1-2).

Li, P. and Sun, M. and Xue, P. (2010). Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm. In Coling 2010, Beijin, pp.710-718.

Matusov, E. and Zens, R. and Ney, H. (2004). Symmetric Word Alignments for Statistical Machine Translation. In Proceedings of Coling 2004, Geneva (CH), pp. 219--225.

Ruopp, A. and von der Meer, J. (2015). TAUS Moses MT Marquet Report, TAUS, 2015.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In Proceedings of the 7th international conference on Language Resources and Evaluation, Istanbul, Turkey.

Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German. In Proceedings of the ACL SIGDAT-Workshop, pp. 47–50.

Simard, M. Foster, G. F. and Isabelle, P. (1993). Using Cognates to Align Sentences in Bilingual Corpora. In Proceedings of the 1993 Conference of the Centre for Advanced Studies on Collaborative Research: Distributed Computing - Volume 2, Toronto, Ontario, Canada, pp. 1071–1082.

Tiedemann, J. and Nygaard, L. (2004). The OPUS corpus - parallel & free. In Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 23-25.

Zhang, Y. and Liu, Q. and Ma, Q. and Isahara, H. (2005). A Multi-aligner for Japanese-Chinese Parallel Corpora. In The Tenth Machine Translation Summit Proceedings, 133-140.

6. Language Resource References

ALIGNJaFr_BABT-0.2_specialEval, https://sharedocs.huma-num.fr/wl/?id=kPbOl7luxaq9GtVPGqhg9jRDB3DGKKoq

JaLexgram-v0.25 (2017). https://sharedocs.huma-num.fr/wl/?id=5Y12iTTm0zyVCQ7cmRUcb8mo2FP4U VsZ

NP-fj.v0.2.csv (2017). https://sharedocs.huma-num.fr/wl/?id=t7iRrYjqzIdYrIs50yqc5eaNDmJbAtQ8

OPUS-jafr-20170906.1 (2017).https://sharedocs.huma-num.fr/wl/?id=XmI18ZR6CXBC9XZSHiJEHB4VQ5IrKS8V

Acknowledgements

Thanks to Yves Lepage, Waseda University (tests on OPUS-fj with Moses and Giza++).