Performance of four sentence aligners on English-Polish bitexts

Grzegorz Krynicki

Faculty of English, Adam Mickiewicz University, Poznań, Poland

krynicki@wa.amu.edu.pl

ABSTRACT

Collections of mutual translations gain in their value if aligned at the sentence level. As such they can be used e.g. in Statistical Machine Translation, Translation Studies and Lexicography. In this study, four leading systems of automatic sentence alignment are tested on four English-Polish translationally equivalent documents. Their performance is evaluated in terms of precision, recall and F-measure as well as in terms of their coverage of the source and target text. Some of the advantages and disadvantages of these methods with respect to different applications are discussed.

STRESZCZENIE

Zbiory tekstów będących wzajemnymi ekwiwalentami tłumaczeniowymi zyskują na wartości jeśli są dopasowane na poziomie zdania. Jako takie mogą być wykorzystane w statystycznym tłumaczeniu maszynowym, przekładoznawstwie czy leksykografii. W niniejszej pracy przedstawiono test czterech wiodących systemów dopasowujących automatycznie zdania w polskich i angielskich tekstach będącymi wzajemnymi odpowiednikami tłumaczeniowymi. W teście porównano skuteczność tych systemów pod względem dokładności, zwrotu oraz miary F. Przedstawiono niektóre z zalet i wad tych systemów względem różnych zastosowań.

1. Introduction

The purpose of this study is to present the results of the comparison of four leading systems of automatic sentence alignment with respect to their accuracy of alignment of four English-Polish bitexts. The systems selected were *moore* [1], *hunalign* [2], *bleualign* [3] and *gargantua* [4]. The accuracy was expressed in terms of precision, recall and F-measure (combining both

precision and recall). The bitexts were short articles from New York Times published in 2010 along with their professional translations found on Polish web portals. The parallel corpus from which they were extracted and which was used to train *moore* and *gargantua* was PHRAVERB [5][1].

Similar studies were conducted for pairs of languages like Urdu-English and French-English [6], German-English and French-English [4] or French-English [7]. The following study confirms high accuracy of *bleualign* in terms of recall and F-measure and high accuracy of *moore* in terms of precision.

2. Terminology

We will assume that a *bead* is the smallest unit of sentence alignment, be it automatic or manual. The bead consists of a *source* and a *target segment*, also referred to as two *sides* of the bead. The segment in the bead may consist of one or more sentences. The bead segments may be mutual translational equivalents (i.e. all component sentences are equivalent) in which case the bead represents a *correct alignment*. Two types of *incorrect alignments* are distinguished, *incomplete assignment* (at least one pair of sentences in the bead are equivalent and at least one sentence has no equivalent on the other side) and *complete misassignment* (no equivalent sentences in the bead). The segment in a bead may be empty (0 sentences) if the segment on the other side contains one or more sentences. Such a bead represents correct alignment if these sentences have no equivalents on the other side of the bitext. The *gold standard* for the alignment of a bitext is such alignment of all the sentences from that bitext that contains only correct alignments. The gold standard needs to be prepared manually. By *structural fidelity* of a bitext we will mean the proportion of 1-to-1 beads relative to the total number of beads in the gold standard obtained for that bitext.

3. Corpus

The corpus used as a basis for this study was PHRAVERB [5]. It is a unidirectional English-to-Polish parallel corpus compiled to analyse Polish equivalents of English phrasal verbs. It consists of 408 English press articles and their Polish translations collected in 2006-2011. English articles were mostly published in New York Times and their translations were published on Polish web portals. The corpus currently amounts to 926 725 tokens.

From the corpus 4 bitexts were manually extracted for the purpose of testing the alignment systems. The corpus was also used for training *moore* and *gargantua*. The corpus and the test bitexts were downcased before they were fed into the aligners.

[1] The author expresses his gratitude to dr Magdalena Perdek for making the corpus available for this study.

4. Test bitexts and gold standard

From the PHRAVERB corpus, 4 articles were randomly selected (a-d, referenced at the end of the paper). In total, they included 162 English and 184 Polish sentences and 8006 tokens.

The gold standard of alignment was carefully created by hand as a point of reference for later evaluation of automatic alignments. The bitexts represented a wide spectrum of structural fidelity (see Table 2). The gold standard contained 157 links: 123 1-to-1, 26 1-to-2, 4 2-to-1, 2 0-to-1, 1 1-to-3 and 1 2-to-2.

5. Automatic aligners

Four publically available automatic aligners were chosen for this study: *moore, hunalign, bleualign* and *gargantua*.

moore and *gargantua* work in an unsupervised fashion, i.e. they infer the alignment model directly from the data set to be aligned. Due to this property they are applicable to parallel corpora for any pair of languages. *hunalign* and *bleualign* rely on language-specific resources. The former resorts to a bilingual dictionary and morphological rules. The latter uses machine translation as an intermediary between the source and the target text.

5.1. *moore*

The algorithm this aligner implements was presented in [1]. The program is written in PERL and has been released under the Microsoft Research end user license that allows free usage for research or teaching purposes.

The algorithm combines a sentence-length-based method with a word-correspondence-based method. The lexical model is based on IBM Translation Model 1 [8] and is trained in the second pass of the alignment. *moore* is intended to provide high precision of alignment at the cost of generating only 1-to-1 beads.

In this study, the training data for *moore* included the whole PHRAVERB corpus. The probability threshold above which the 1-to-1 sentences from the input corpus were returned was set to the default of 0.5.

5.2. *hunalign*

The second algorithm tested in this study was presented in [2]. Its implementation in C++ has been released under GNU LGPL.

Similarly to the Moore's algorithm, *hunalign* utilizes an alignment algorithm based on both sentence length and lexical similarity. Hunalign however uses a word-by-word bilingual dictionary instead of the IBM Translation Model 1.

In this test, *hunalign* was used with a dictionary of approx. 100 000 Polish-English pairs of single-word equivalents compiled from various electronic bilingual Polish-English dictionaries. Although *hunalign* is capable of using morphological information (affixation rules, POS tags), its format and tagsets were not compatible with the resources available to the author and this functionality was not taken advantage of.

hunalign is the only the aligner of the four that is capable of generating many-to-many alignments.

5.3. *bleualign*

The algorithm was proposed in [3]. It is programmed in Python and is available under GNU GPL.

In the first pass, *bleualign* computes the alignment between the translated source text and the target text by scoring similarity between sentence pairs by means of the BLEU metric [9]. Then, an optimum path for 1-to-1 alignments is found through dynamic programming. Finally, other 1-to-1, many-to-1 and 1-to-many alignments are added.

For this study, the external MT system used was the English-to-Polish online version of *Translatica* developed by *Poleng Ltd*.

5.4. *gargantua*

The system was presented in [4]. It is implemented in C++ and released as open source.

Its alignment model is similar to *moore*, but it introduces differences in pruning and search strategy. As in *moore*, the first pass is based on sentence-length statistics. However, it replaces the second pass of the *moore* algorithm with a two-step clustering: first, all 1-to-1 alignments are obtained and then these beads are merged with unaligned sentences to build 1-to-many and many-to-1 alignments, thus solving the problem of low recall in *moore.*

Of all four aligners, this one is probably the most cumbersome to install and the only one that does not run on MS Windows. It is also the only system that can take advantage of structural anchors that appear in *Europarl* documentation.

6. Evaluation criteria for sentence alignment

The performance of sentence alignment can be evaluated by means of measures of relevance known in the field of information retrieval as precision, recall and F-measure [7]. These metrics have been reported to have relatively high correlation with human judgements [10]. They combine the results of the aligner on correct links relative to test and reference links in the following way:

$$precision = \frac{correct\ links}{test\ links}, \quad recall = \frac{correct\ links}{reference\ links},$$

$$F = \frac{2 \times (recall \times precision)}{recall + precision}.$$

Correct links is the number of correct links among those proposed by the aligner; *reference links* is the number of links in correctly aligned texts (the gold standard); *test links* is the number of all links proposed by the aligner. F-measure is a harmonic mean of precision and recall. An in-depth discussion of these measures in the context of alignment is presented in [10] and [11].

7. Results of the evaluation

The table below includes the comparison of the results produced the aligners. The best aligner in terms of F-measure and recall was *bleualign* (0.8228 and 0.8469 respectively). The best alignment in terms of precision was generated by *moore* (0.9625).

In all cases the number of reference links was the same (157) as one gold standard was used as reference for evaluating all automatic alignments.

Table 1. *Accuracy of four automatic aligners. Bolded numbers indicate highest scores.*

	moore	*hunalign*	*bleualign*	*gargantua*
correct links	77	122	130	126
test links	80	159	149	157
precision	**0.9625**	0.7673	0.8725	0.8025
recall	0.4873	0.7722	**0.8228**	0.8025
F-measure	0.6471	0.7697	**0.8469**	0.8025

The results presented above are relatively low compared to the results reported in the literature (e.g. [7] p. 14, [2] p. 5). The reasons may include the fact that these studies are largely based on official EU documents, which have more rigorous structure and are easier to align than press articles. An important reason may also be the relatively low average structural fidelity of test bitexts selected for this study.

Quality of sentence alignment is strongly dependent on the number of deletions, insertions and free translations ([12] p. 1, [1] p. 9). Intuitively speaking, the greater the proportion of 1-to-1 sentence correspondences in a bitext, the easier the job of the automatic aligner. Structural fidelity of each bitext, i.e. the ratio of 1-to-1 to all beads in the gold standard alignment for that particular bitext, is presented in Table 2.

Table 2. *Structural fidelity (SF) of each test bitext (a-d) vs. F-measure for the alignments generated by the aligners on each of these bitexts.*

Bitext	SF	*moore*	*hunalign*	*bleualign*	*gargantua*	*avr. F*
a	0.6111	0.4490	0.6133	0.6000	0.5676	0.5575
b	0.6977	0.2174	0.6173	0.8831	0.7500	0.6169
c	0.8919	0.7458	0.8378	0.9211	0.8649	0.8424
d	0.9091	0.9398	1.0000	0.9639	1.0000	0.9759

The low result of *moore* on the bitext *b* is a result of low recall, in fact *moore* performed with 100% precision on that bitext. A result of 1.0 indicates the alignment identical to the gold standard. Average F-measure values are strictly increasing with the SF arguments. However, a greater number of data points would be necessary to reliably model the relationship between them or calculate the correlation coefficient.

8. Examples of alignments and alignment errors

Table 3 contains an example of several sentences aligned manually and Table 4 – the alignment of these sentences proposed by 3 automatic aligners. The bitext exemplifies a case of translator's decision not to render source-text material judged to be redundant or untranslatable. In this case, the elements omitted in the translation are comments on the form of the English word 'babble' as well as its folk and scientific etymology. The omission is justified considering different form and etymologies of the Polish equivalent 'gaworzenie'.

Table 3. *Gold standard alignment.*

[1] During the second year of life, toddlers shape their sounds into the words of their native tongues.	[1] Dopiero w drugim roku życia zaczynają składać dźwięki w słowa swojego własnego, ojczystego języka.
[2] The word "babble" is both significant and representative — repetitive syllables, playing around with the same all-important consonants.	0
[3] (Indeed, the word seems to be derived not from the biblical Tower of Babel, as folk wisdom has it, but from the "ba ba" sound babies make.)	0
[4] Some of the most exciting new research, according to D. Kimbrough Oller, a professor of audiology and	[2] - Wśród najnowszych, szalenie interesujących badań w tej dziedzinie - mówi D. Kimbrough Oller, profesor

speech-language pathology at the University of Memphis, analyzes the sounds that babies make in the first half-year of life, when they are "squealing and growling and producing gooing sounds."	audiologii i patolog mowy i języka z University of Memphis - są takie, które za przedmiot analizy obierają dźwięki produkowane przez dziecko w pierwszym półroczu jego życia, kiedy wydaje ono z siebie "piski, pomruki i dźwięki przypominające gruchanie".
[5] These sounds are foundations of later language, he said, and they figure in all kinds of social interactions and play between parents and babies — but they do not involve formed syllables, or anything that yet sounds like words.	[3] Jak wyjaśnia profesor, wszystkie one stanowią fundament mającego się rozwinąć języka, figurując w różnych typach zabaw i społecznych interakcji, jakie zachodzą pomiędzy rodzicami i dziećmi. [4] Na tym etapie dziecko nie wypowiada jednak uformowanych sylab ani niczego, co przypominałoby słowa.
[6] "By the time you get past 6 months of age, babies begin to produce canonical babbling, well-formed syllables," Professor Oller said.	[5] - Mniej więcej wtedy, gdy dziecko kończy sześć miesięcy, pojawia się u niego klasyczne gaworzenie i zdolność do artykulacji dobrze uformowanych sylab - tłumaczy profesor Oller.

For the above passage, of all 3 systems, the least accurate alignment was proposed by *hunalign* and the most accurate by *bleu*. *moore* is not included in the list as it skipped the passage in its output. *hunalign* turned out to be particularly sensitive to structural incompatibilities between the source and target text. In this case, after it fixed the first incorrect alignment (2, 3 - 2), it made errors in the next 8 beads before it came back on the right track (only first 3 of these 8 are listed in the Table 4).

Table 4. *Alignment of the passage from Table 3 proposed by 3 automatic aligners. Incorrect beads are bolded.*

alignment	beads				
gold standard	1 - 1	2 - 0	3 - 0	4 - 2	5 - 3, 4
hunalign	1 - 1	**2, 3 - 2**		**4 - 3, 4**	**5 - 5**
bleu	1 - 1			4 - 2	5 - 3, 4
gargantua	1 - 1	2 - 0	**3, 4 - 2**		5 - 3, 4

The correct alignment was produced only by *bleu*. However, sentences omitted in the translation were also omitted from the alignment output, which affects the integrity of the source text.

gargantua made one mistake in the bead 3, 4 - 2 (it should be 3 - 0 and 4 - 2). This error, however, seems to be of lesser gravity than 4 – 3, 4 and 5 - 5 produced by *hunalign* as the former (i.e. 3, 4 - 2) is a case of incomplete assignment while the latter are cases of complete misassignment. In the statistics for the whole study both errors were counted in the same way.

9. Conclusions and future work

One of potentially problematic features of *moore* and *bleualign* is that they do not output beads/sentences they have problem aligning: *moore* skipped 78 beads out of 157 reference beads; *bleualign* skipped 9 out of 162 English sentences and 26 out of 184 Polish sentences.

The possibility of excluding dubious matches may be a valuable functionality in view of, on one hand, the predominant applications of these systems, namely SMT and bilingual terminology extraction, and on the other, the growing availability of bilingual data (c.f. [7] p. 11). From that perspective, having parallel training material of highest precision possible is a priority and leaving out risky alignments may be a negligible loss.

However, for automatic applications that analyse the context of aligned bitexts beyond the sentence level [13], as well as for many human applications – aligners like *moore* and *bleualign* have limited applicability. In translation studies and lexicography, parallel corpora and parallel concordances are used to analyse meanings of words and phrases with their translations in consideration of their pragmatic context. The categories of anaphora, tenses, pronouns or coherence span across sentence boundaries. Resolution of lexical and structural ambiguity often requires a wider context as well. To correctly render these phenomena in translation, the integrity of the source text is required just as the integrity of both source and target is required for later study of that translation.

Therefore, if the integrity of the bitexts needs to be preserved – *gargantua* or *hunalign* are recommended. If the number of the output alignments is predicted to be sufficient and the precision is a priority, the *moore* system seems to be the best choice. *bleualign* attempts to find a compromise between these approaches but the necessity to obtain machine translation of one side of the bitext may pose a practical problem.

Future work will allow estimation of alignment accuracy by means of the quality of translation produced by SMT systems trained on the resulting aligned corpus [4]. A larger number of bitexts representing a greater variety of genres would give the above findings better generalization power.

References

[1] Moore, Robert. 2002. Fast and accurate sentence alignment of bilingual corpora. [In:] *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas, Tiburon, California*. Heidelberg, Germany: Springer-Verlag. Pp. 135-244.

[2] Halácsy, Péter – Dániel Varga – András Kornai – Viktor Nagy – László Németh – Viktor Trón. 2005. Parallel corpora for medium density languages. [In:] *Proceedings of RANLP*. Borovets, Bulgaria. Pp. 590-596.

[3] Sennrich, Rico and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. [In:] *Proceedings of AMTA 2010*. Denver, Colorado.

[4] Braune, Fabienne – Alexander Fraser. 2010. Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora. [In:] Chu-Ren Huang and Dan Jurafsky (Eds.) *COLING Posters*. Chinese Information Processing Society of China. Pp. 81-89.

[5] Perdek, Magdalena. 2012. Lexicographic potential of corpus equivalents: The case of English phrasal verbs and their Polish equivalents. [In:] Fjeld Ruth Vatvedt and Julie Matilde Torjusen (Eds.) *Proceedings of the 15th EURALEX International Congress*. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo. Pp. 376-388.

[6] Abdul-Rauf, Sadaf – Mark Fishel – Patrik Lambert – Sandra Noubours – Rico Sennrich. 2012. Extrinsic Evaluation of Sentence Alignment Systems. [In:] *Proc. of the LREC Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (CREDISLAS)*. Istanbul, Turkey. Pp. 6-10.

[7] Yu, Qian – François Yvon – Aurélien Max. 2012. Revisiting sentence alignment algorithms for alignment visualization and evaluation. [In:] *Proceedings of the 5th Workshop on Building and Using Comparable Corpora, Istambul, Turkey, 2012*. Pp. 10-16. (URL http://perso.limsi.fr/yvon/publications/sources/Yu12revisiting.pdf)

[8] Brown, Peter F. – Stephen A. Della Pietra – Vincent J. Della Pietra and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2). Pp. 263–311.

[9] Papineni, Kishore – Salim Roukos – Todd Ward – Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *ACL*. Pp. 311-318.

[10] Melamed, I. Dan – Ryan Green – Joseph P. Turian. 2003. Precision and recall of Machine Translation. *Proteus technical report 03-004, paper presented at NAACL/HLT 2003*. Edmonton, Canada.

[11] Véronis, Jean. 2000. Evaluation of parallel text alignment systems: the ARCADE project. (citeseer.ist.psu.edu/vronis00evaluation.html)

[12] Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. [In:] *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Pp. 9–16.

[13] Popescu-Belis, Andrei – Bruno Cartoni – Andrea Gesmundo – James Henderson – Cristina Grisot – Paola Merlo – Thomas Meyer – Jacques Moeschler – Sandrine Zufferey. Improving MT coherence through text-level processing of input texts: the COMTIS project. [In:] *Proceedings of Tralogy, Session 6, Traduction et traitement automatique des langues*. (URL: http://lodel.irevues.inist.fr/tralogy/index.php?id=78)

Test bitexts

[a] Carr, David. A News Corp. Newspaper, but Not in Print. New York Times. November 21, 2010. http://www.nytimes.com/2010/11/22/business/media/22carr.html Translation: Nadchodzi "The Daily" – gazeta na iPada. 4 grudnia 2010. http://fakty.interia.pl/new-york-times/news/nadchodzi-the-daily-gazeta-na-ipada,1566806,6806

[b] Klass, Perri. 2010. Understanding ‘Ba Ba Ba’ as a Key to Development. October 11, 2010. http://www.nytimes.com/2010/10/12/health/12klass.html Translation: Gaworzenie – klucz do rozwoju dziecka. 21 października 2010. http://fakty.interia.pl/new-york-times/news/gaworzenie-klucz-do-rozwoju-dziecka,1547733,6806

[c] Bernstein, Nina. 2010. An Agent, a Green Card, and a Demand for Sex. New York Times. March 21, 2008. http://www.nytimes.com/2008/03/21/nyregion/21immigrant.html?page wanted=all Translation: Zielona karta za seks. http://wredna-dama.pardon.pl/dyskusja/804936//3

[d] Sweig, Julia E. 2010. Absent at the Creation. New York Times. October 1, 2010. http://www.nytimes.com/2010/10/02/opinion/02iht-edsweig.html Translation: Nieobecni przy poczęciu. 5 października 2010. http://fakty.interia.pl/new-york-times/news/nieobecni-przy-poczeciu,1540649,6806