First Steps Towards Coverage-Based Sentence Alignment

Luís Gomes[1,2], Gabriel P. Lopes [1,2]
[1]NOVA LINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal
[2]ISTRION BOX Translation and Revision, Lda, Portugal
luis.gomes@istrionbox.com gabriel.lopes@istrionbox.com

Abstract

In this paper, we introduce a coverage-based scoring function that discriminates between parallel and non-parallel sentences. When plugged into Bleualign, a state-of-the-art sentence aligner, our function improves both precision and recall of alignments over the originally proposed BLEU score. Furthermore, since our scoring function uses Moses phrase tables directly we avoid the need to translate the texts to be aligned, which is time-consuming and a potential source of alignment errors.

Keywords: Phrase-Table Coverage, Sentence Alignment, Language Independent

1. Introduction

Sentence alignment is the task of finding corresponding sentences between texts that are translations of each other (parallel texts). It is a fundamental task for corpora-based approaches to Machine Translation, such as Statistical Machine Translation and Example-Based Machine Translation, as well as for creating Translation Memories to be used in Computer-Assisted Translation.

Some parallel texts such as those from the European Parliament and the Canadian Hansard are relatively easy to align because the translations are very clean, i.e. there is almost a one-to-one correspondence between sentences in both languages. Furthermore, those texts have markup that can be used as anchor points to constrain the alignment, allowing accuracies above 95% even with alignment methods based exclusively on sentence length proportionality, such as the ones proposed by Brown et al. (1991) and Gale and Church (1993) (the former measures sentence length in terms of tokens and the latter in terms of characters).

However, for texts available in less-friendly formats, such as PDF, from which we cannot avoid extracting some noise intermixed with the text (such as figure and table captions, page headers and footers, etc) we need more robust aligners that take into account the actual text within sentences and not only their lengths.

In this paper, we improve over Bleualign (Sennrich and Volk, 2010), a state-of-the-art sentence aligner with top-performance on noisy texts (Sennrich and Volk, 2010; Abdul-Rauf et al., 2012).

Our main contribution is a new scoring function that discriminates parallel and non-parallel sentences based on the ratio of text covered by bilingual phrase-pairs from a Moses phrase table (Koehn et al., 2007).

2. Previous Work

The general idea of Bleualign is to automatically translate one of the texts and then align the translation with the other text using the sentence-wise BLEU score (Papineni et al., 2002) as indicator of sentence similarity.

Besides the good performance on noisy texts, Bleualign appealed to us because, in a way, it takes advantage of previously acquired translation knowledge that is encoded within the MT system used to translate the texts to be aligned. By contrast, other aligners such as the Microsoft Bilingual Aligner (Moore, 2002), Hunalign (Varga et al., 2005) and Gargantua (Braune and Fraser, 2010) are autonomous and automatically infer a word-based translation lexicon from the texts being aligned as they proceed. As a consequence, the performance of these methods degrades when the texts to be aligned are short as there is less data to support statistical inference of a bilingual lexicon.

In our view, standalone aligners are more suited for scenarios where no parallel corpora exists for the language pair under consideration. But today, given the ubiquity of corpora-based MT research and application in academia and industry alike, the scenarios where one does not have a parallel corpus or working MT system are becoming less frequent. Our point is that we should take advantage of existing parallel corpora when available.

Some aligners, such as Champollion (Li et al., 2010) and Hunalign (Varga et al., 2005), are able to use external (word-based) lexica, but there are no guidelines how to produce these lexica, nor how size and quality of the lexica relates to quality of alignments. Furthermore, some alignments are not truly between words but instead between (possibly discontiguous) phrases such as the alignment of English "not" with French "ne . . . pas" or German "Mitgliedstaaten" with French "États membres" (member States). We also observe that longer and less frequent phrases such as German "die Europäische Zentralbank" and French "la Banque centrale européenne" (the European Central Bank) tend to be much more reliable sentence-parallelism indicators than single words. Hence our motivation for using a phrase table instead of a word-to-word bilingual lexicon.

We introduce a new scoring function (which replaces BLEU score in Bleualign) that forgoes the need to translate texts in order to compare and align them. Instead, we match bilingual phrase pairs from a Moses phrase table in sentences to be aligned, and we measure the portion of text that is covered by those matches.

3. Coverage-Based Alignment

The alignment hypothesis space for a pair of short parallel texts is represented in Figure 1. In this representation, each point in the x and y axes corresponds to a character

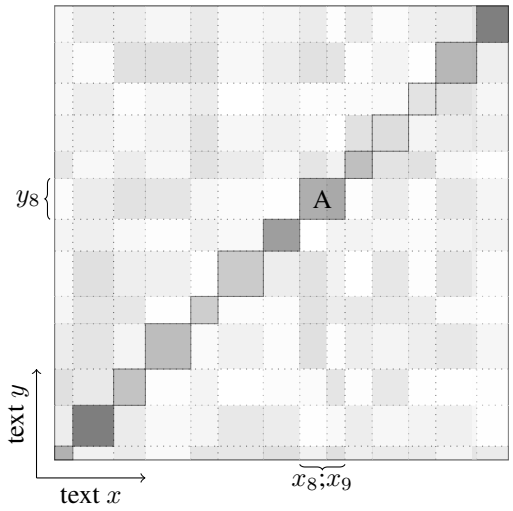

Figure 1: Sentence alignment hypothesis space (dotted lines represent sentence boundaries and darker shades indicate greater coverage).

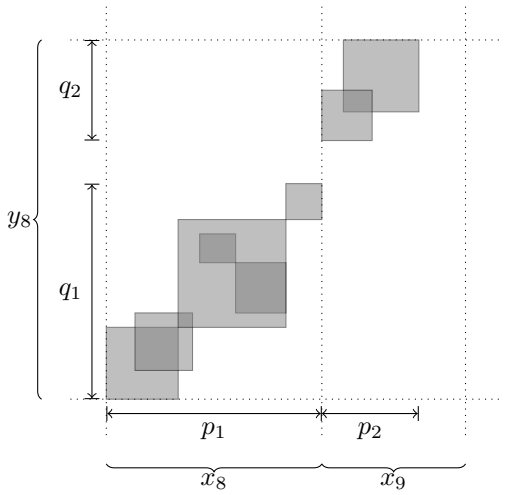

Figure 2: Closeup view of 2:1 aggregate configuration from rectangle **A** in Figure 1 showing matched phrase-pairs (each rectangle represents a matched pair of phrases from the Moses phrase table).

offset in the respective text and dotted lines mark sentence boundaries. Every pair of sentences is filled with a shade of gray proportional to the coverage measured within that particular pair of sentences; darker shades indicate greater coverage.

The chain of rectangles running from the bottom left (the start of both texts) to the upper right (the end of the texts) represents correctly aligned sentences. As we can see, the coverage score is much higher (darker) for parallel sentence pairs than non-parallel sentence pairs, which is indicative of its discriminative power towards parallel vs. non-parallel sentences.

Rectangle A is enlarged in Figure 2 and represents a 2:1 alignment between sentences x_8, x_9 and y_8. To obtain such aggregate alignments, we must consider a number of hypothetical *aggregate configurations* as shown in Figure 3.

The sentence alignment algorithm employs dynamic programming to find a non-overlapping monotonic chain of parallel segments that maximizes the total sum of coverage scores of all chained segments (or BLEU scores in the original Bleualign implementation). Aligned segments will have one of several possible configurations, the most common being 1:1 (one sentence aligned with one sentence), 1:0 and 0:1 (when sentences have no corresponding translation in the other text). Aggregate configurations such as 1:2, 2:1, 2:2, 2:3, etc, are also considered. This dynamic programming framework has been adopted by most aligners since the early lenght-based methods (Gale and Church, 1993; Brown et al., 1991).

We have not made any changes to the Bleualign implementation other than replacing the scoring function with our own.

3.1. Coverage-Based Scoring Function

Intuitively, the coverage score should be 1 (maximum) when all tokens in a given pair of sentences are covered by bilingual phrase pairs from the phrase table. Conversely, the score should be 0 if no bilingual phrase pair from the

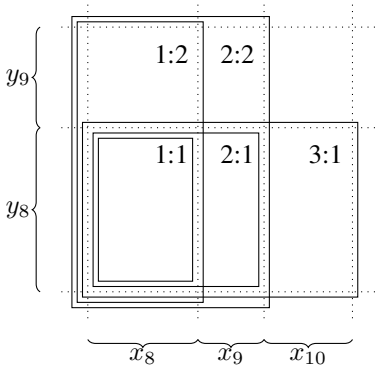

Figure 3: Aggregate configurations that are considered by the aligner (for simplicity we only represent some configurations 1:1, 1:2, 2:1, 2:2 and 3:1).

phrase table appears in the given sentences.

Our scoring function is defined (Equation 1) as the product of two coverage ratios, each indicating how much text is covered on the sentences to be aligned. The ratios are computed in terms of characters, ignoring whitespace.

$$C = \frac{\sum_i p_i}{\sum_j |x_j|} * \frac{\sum_i q_i}{\sum_k |y_k|} \quad (1)$$

Following the example presented in Figure 2, p_i and q_i are continguous segments covered by bilingual phrase matches (represented as gray rectangles). This example shows how coverage is computed for an 2:1 aggregate configuration aligning two sentences x_8 and x_9 with y_8. It is easy to see that the 1:1 configuration aligning x_8 with y_8, has a lower score because segment q_2 is not covered in this configuration.

3.2. Phrase-Table Pruning

The Moses phrase table obtained from German-French Europarl contains approx. 55 million entries. Our prototype implementation (in Python) is not very efficient in terms of memory usage and we had to prune the phrase table heavily. We used two simple statistical filters, one based on phrase frequency and the other based on probability. We thus filtered all (x, y) phrase pairs with joint frequency $F(x, y)$ lower than 5 occurrences and symmetric conditional probability (Equation 2) lower than 0.001.

$$SCP(x, y) = \frac{F(x, y)^2}{F(x)F(y)} \quad (2)$$

The pruned table contains approx. 710 thousand entries. We are aware of more sophisticated phrase-table pruning techniques, such as the one based on statistical significance proposed by Johnson et al. (2007) or the one based on relative entropy proposed by Zens et al. (2012), but we did not try them. Our intuition is that the pruning technique employed is not as critical for sentence alignment as it is for translation, thus we decided to keep it as simple as possible.

4. Evaluation

To evaluate the performance of our sentence aligner we follow the established practice: we compute precision, recall and F_1 metrics for automatically generated alignments with respect to a gold-standard alignment. As usual, precision is defined as the ratio of correct alignments over the number of generated alignments; recall is the ratio of correct alignments over the number of alignments in the gold standard; and F_1 is the harmonic mean of precision and recall. Some aligned segments may be partially correct, as for example when the aligner proposes a 2:2 segment where the reference contains instead two 1:1 alignments. Following the practice from the original Bleualign evaluation (Sennrich and Volk, 2010), we report precision, recall and F_1 according to a *strict* and *lax* criteria. In *strict* mode, only segments that match exactly the reference alignment are considered correct. In *lax* mode, segments will be considered correct if they intersect reference alignments on both language sides. For all application purposes, the *strict* scores are the ones that count. The *lax* scores give us a hint of how close or far the generated alignments are from the gold standard.

We evaluate our coverage-based aligner under exactly the same conditions as Bleualign was originally evaluated (Sennrich and Volk, 2010): we use the same gold-standard alignments and evaluation scripts.[1]. Furthermore, our phrase table was obtained from the Europarl corpus, which they also used to train their Moses system.

The Text+Berg corpus is a small[2] German-French corpus available together with the source code of Bleualign[3]. The corpus is composed of yearbooks from Swiss Alpine Clube and contains reports on mountain expeditions as well as some scientific articles. Because the domain of this evaluation corpus is quite different from the domain of the corpus from where we obtained the phrase table (parliamentary debates) we believe that the results obtained are not domain-specific.

Aligner	Strict			Lax		
	Prec	Rec	F_1	Prec	Rec	F_1
Length	0.67	0.68	0.68	0.79	0.80	0.80
Moore	**0.86**	0.71	0.78	0.96	0.80	0.87
Bleualign	0.83	0.78	0.81	0.98	0.92	0.95
Coverage	0.85	**0.84**	**0.85**	**0.99**	**0.96**	**0.98**

Table 1: Precision, recall and combined F_1 score for alignments produced by four different aligners on the same gold standard corpus.

Table 1 summarizes the evaluation results for our coverage-based aligner, Bleualign, the Microsoft Bilingual Aligner and the length-based aligner by Gale and Church (1993)[4].

Our aligner has the best overall performance with an F_1 score (in strict evaluation mode) that is 4 points higher than Bleualign, 7 points higher than the Microsoft Bilingual Aligner and 17 points higher than the length based-aligner. The Microsoft Bilingual Aligner seems to favor precision at expense of recall, whereas all other aligners have more balanced precision and recall scores.

5. Conclusions and Future Work

We presented a new coverage-based scoring function that improves both precision and recall of Bleualign. Furthermore, we avoid the need to translate the texts to be aligned, which is time-consuming, requires access to an external MT system and inevitably introduces errors (even if the MT system produced "perfect" translations they would probably contain some lexical choices that are different than the ones in texts to be aligned).

We have demonstrated that even with a simple scoring function the coverage-based approach to sentence alignment yields considerable gains over state-of-the-art aligners. More than the particular scoring function presented here, we believe that the greatest contribution of this paper is the general idea of phrase table coverage-based alignment, which effectively takes advantage of previously acquired translation knowledge instead of trying to infer a weaker word-based lexical model from the texts being aligned, as most other aligners do.

In future experiments, we will investigate other coverage-based scoring functions, taking into account phrase translation probabilities and lexical weigths. We also intend to extend evaluation to other language pairs and domains, though gold standard alignments are scarce and costly to produce by hand.

We are currently investigating how to improve the overall alignment algorithm computational efficiency, to overcome the need for heavy phrase table filtering and the need for

[1] we thank the authors of Bleualign for making their experimental conditions easy to replicate

[2] the Text+Berg corpus distributed with bleualign seems to be only a small part of the full Text+Berg corpus, but we refer to it as Text+Berg in this paper

[3] `https://github.com/rsennrich/bleualign`

[4] the results for Microsoft Bilingual Aligner and the lenght-based aligner are reproduced from the Bleualign paper (Sennrich and Volk, 2010)

hard delimiters (inherited from Bleualign) to keep the hypothesis space within amenable bounds, as time and memory requirements are asymptotically quadratic with respect to the number of sentences between consecutive hard delimiters.

Acknowledgements

This work was supported by the Portuguese Foundation for Science and Technology (FCT/MCTES) through individual PhD grant SFRH/BD/65059/2009 (LG), funded research project ISTRION (ref. PTDC/EIA-EIA/114521/2009) and NOVA LINCS (ref. UID/CEC/04516/2013).

6. Bibliographical References

Abdul-Rauf, S., Fishel, M., Lambert, P., Noubours, S., and Sennrich, R. (2012). Extrinsic evaluation of sentence alignment systems. In *Proceedings of LREC Workshop on Creating Cross-language Resources for Disconnected Languages and Styles*.

Braune, F. and Fraser, A. (2010). Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 81–89. Association for Computational Linguistics.

Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 169–176. Association for Computational Linguistics.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Comput. Linguist.*, 19(1):75–102, March.

Johnson, J. H., Martin, J., Foster, G., and Kuhn, R. (2007). Improving translation quality by discarding most of the phrasetable. *EMNLP-CoNLL 2007*, page 967.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180. Association for Computational Linguistics.

Li, P., Sun, M., and Xue, P. (2010). Fast-champollion: A fast and robust sentence alignment algorithm. In *Coling 2010: Posters*, pages 710–718, Beijing, China, August. Coling 2010 Organizing Committee.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA 2002 Tiburon, CA, USA, October 6-12, 2002, Proceedings*, pages 135–144.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Sennrich, R. and Volk, M. (2010). MT-based sentence alignment for OCR-generated parallel texts. In *The Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010), Denver, Colorado*.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)*, pages 590–596.

Zens, R., Stanton, D., and Xu, P. (2012). A systematic comparison of phrase table pruning techniques. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 972–983. Association for Computational Linguistics.