Alignment of Parallel Corpora Exploiting Asymmetrically Aligned Phrases

Patrik Lambert and Núria Castell

TALP Research Center
Universitat Politècnica de Catalunya
Jordi Girona Salgado, 1-3, 08034 Barcelona, Spain
{lambert,castell}@talp.upc.es

Abstract

This paper presents a simple way of producing symmetric, phrase-based alignments, combining two single-word based alignments. Our algorithm exploits the asymmetries in the superposition of the two word alignments to detect the phrases that must be aligned as a whole. It was run with baseline word alignments produced by the Giza++ software and improved these alignments. The ability to treat some groups of words as a whole is essential in applications like machine translation. The paper also addresses the difficulty of the alignment evaluation task.

1. Introduction

A parallel corpus aligned at a word level is a resource directly usable for the building of bilingual lexica and terminology. It is also a valuable resource for several natural language processing applications such as machine translation and word sense disambiguation. A publicly available, widely used software to produce baseline single word based alignments is Giza++ (Och, 2000; Och and Ney, 2003). It implements various translation models: the so-called IBM models 1 to 5, introduced by (Brown et al., 1993), and the HMM model, introduced by (Vogel et al., 1996). The models are trained on a bilingual corpus with the EM algorithm (Baum, 1972), "bootstrapping" from a simpler model to a more complex model. The final alignment (Viterbi alignment) is the best one according to a Viterbi search.

The models implemented by the Giza++ software have limitations. The first one is a consequence of the mapping used, which only allows to link one source word to each target word. The second one is inherent to single-word based alignments: alignment of multiple word or phrases which do not decompose easily in word-into-word translations are not possible.

As pointed out in (Och and Ney, 2003), the first problem can be solved if the Viterbi alignment is calculated in both source-target and target-source directions. If the alignment in one direction is not complete, the alignment in the other direction completes it. The combination of source-target and target-source alignments is also a useful resource to detect the second problem. This is because the phrases that cannot be aligned word-to-word (like idiomatic expressions) are not well aligned by Giza++, so that the source-target and target-source alignments are typically not symmetric.

In section 2., we present an algorithm that detects these asymmetries in the superposition of source-target and target-source alignments, and replaces them by appropriate symmetric alignments. Section 3. discusses the alignment evaluation task. Section 4. describes the experiments. Some conclusions are given in section 5..

2. Symmetrisation Algorithm

The central idea is that if the asymmetry is caused by a language feature such as an idiomatic expression, it will be repeated various times in the corpus, otherwise it will occur only once. Our symmetrisation process has the following two stages:

Building of asymmetries memory. Detect all the asymmetries present in the corpus and store them with their number of occurrences. A word does not belong to an asymmetry if it is linked to exactly one word, which in turn has exactly one link to it.

Alignment correction. Detect again asymmetric zones and for each asymmetry, try to correct the alignment:

1. Look if the limitation associated to the mapping can be solved: if the asymmetry contains various words linked to a word x, itself aligned to only one of them, links are added so that x be aligned to the other words.
2. Look if the asymmetry contains phrases qualified to be aligned as a group: it should include at least one source and one target word. Two parts of a non-contiguous phrase can't be more than three words away from each other. If the asymmetry is suitable for group alignment, follow steps 3 and 4. Otherwise, the asymmetry has generally no linguistic basis and it is advisable to take the intersection of source-target and target-source alignments.
3. Split the source and target strings in fragments, combine each source fragment with each target fragment and see how many times the combination has occurred in an asymmetry. Select the combination that has occurred more times in the corpus. If it is above a predefined threshold, add links so that both fragments be aligned as a group (many-to-many alignment). Continue with the other fragments until all words have been grouped or until no remaining combination has more than the threshold number of occurrences in the corpus.
4. If no combination had occurred more than the threshold, apply a combination of source-target and target-source alignments, like their intersection or union.

3. Alignment Evaluation

A consensus on word alignment evaluation methods has started to appear. These methods are described in (Mihalcea and Pedersen, 2003). Submitted alignments are compared to a manually aligned reference corpus (gold standard) and scored with respect to precision, recall, F-measure and Alignment Error Rate (AER). An inherent problem of the evaluation is the ambiguity of the manual alignment task. The annotation criteria depend on each annotator. Therefore, (Och and Ney, 2003) introduced a reference corpus with explicit ambiguous (called P or Possible) links and unambiguous (called S or Sure) links. Given an alignment $\mathcal{A}$, and a gold standard alignment $\mathcal{G}$, we can define sets $\mathcal{A_S}$, $\mathcal{A_P}$ and $\mathcal{G_S}$, $\mathcal{G_P}$, corresponding to the sets of Sure and Possible links of each alignment. The set of Possible links is also the union of S and P links, or equivalently $\mathcal{A_S} \subseteq \mathcal{A_P}$ and $\mathcal{G_S} \subseteq \mathcal{G_P}$. The following measures are defined (where T is the alignment type, and can be set to either S or P):

$$P_T = \frac{|A_T \cap G_T|}{|A_T|}, \quad R_T = \frac{|A_T \cap G_T|}{|G_T|}, \quad F_T = \frac{2P_T R_T}{P_T + R_T}$$

$$AER = 1 - \frac{|A_P \cap G_S| + |A_P \cap G_P|}{|A_P| + |G_S|}$$

Note that $|G_P|$ is not taken into account in the AER. Therefore, including more P links in the reference alignment can only lower the error rate. The definition also implies that if $G_S \subseteq A_P \subseteq G_P$, the AER is equal to zero.

The next step in the evaluation is to be able to compare the values obtained. However, it is a delicate task because they are very dependent on the exact method used as well as on the reference corpus.

3.1. Influence of the Evaluation Method

The scores are greatly affected by the representation of NULL links (between a word and no other word: whether they are assigned an explicit link to NULL or removed from the alignments). Explicit NULL links contribute to a higher error rate because in this case the errors are penalised twice: for the incorrect link to NULL and for the missing link to the correct word.

Another influent factor is the way of weighting each link: n words linked as a group represent n^2 links instead of n links. To correct this effect, (Melamed, 1998) proposed to attach a weight to each link. The weight $w(\mathrm{x}, \mathrm{y})$ of a link between two words x and y would be inversely proportional to the number of links in which x and y are involved.

In conclusion, experiments are not comparable unless they are evaluated with exactly the same method.

3.2. Influence of the Reference Corpus

Apart from their dependence in the annotator's criteria (the decision of what is translation of what), the results vary in function of the proportion of ambiguous and unambiguous links. If the reference corpus contains a small number of very sure S links and many P links, adding more links to the submitted alignment will only slightly modify the value of $|A_P \cap G_S|$ and $|A_P \cap G_P|$ since they tend easily to $|G_S|$ and $|A_P|$, respectively. However the increase of $|A_P|$ will lower the AER. So this reference corpus will favour high precision alignments. On the contrary, if the reference corpus only contains S links, more submitted links will be needed to increase $|A_P \cap G_S|$ and high recall alignments will be more rewarded than in the previous case.

A related issue is that a reference corpus with many ambiguous links allows many different submitted alignments to have the same AER, while some of them are obviously poorer. Consider for instance the sentence pair 76 of the reference corpus of (Och and Ney, 2000), displayed in figure 1.

nous souhaitons parvenir à une décision cette semaine .
it is our hope to make a decision this week .

	NULL	it	is	our	hope	to	make	a	decision	this	week	.
.	.	.	.	.	.	.	.	.	.	.	.	S
semaine	.	.	.	.	.	.	.	.	.	.	S	.
cette	.	.	.	.	.	.	.	.	.	S	.	.
décision	.	.	.	.	.	P	P	P	S	.	.	.
une	.	.	.	.	.	P	P	S	P	.	.	.
à	.	.	.	.	.	P	P	P	P	.	.	.
parvenir	.	.	.	.	.	P	P	P	P	.	.	.
souhaitons	.	P	P	P	P	P	.	.	.	.	.	.
nous	.	P	P	P	P	P	.	.	.	.	.	.
NULL	.	.	.	.	.	.	.	.	.	.	.	.

Figure 1: Example alignment with few Sure links and many ambiguous links

With such a reference, both alignments of figure 2 would get the same score of zero error rate (as well as all the alignments for which $G_S \subseteq A_P \subseteq G_P$), although the lower one is much poorer.

Therefore, if the gold standard contains ambiguous links, they should only allow alignment combinations that are considered equally correct.

4. Alignment Symmetrisation Experiments

We present results on two corpora. First we give their characteristics. Next, we detail the evaluation of the Giza++ alignments and their symmetrisation.

In all the experiments the NULL links were removed. Here we only show results in which each link has the same weight. The first 200 sentence pairs of each test corpus were used to optimise some parameters of the symmetrisation application (this doesn't require training). The whole test corpus, including these 200 sentence pairs, was used for the evaluation.

4.1. Training and Test Data

4.1.1. Verbmobil Corpus

These data come from a selection of spontaneous speech databases available from the Verbmobil project[1].

[1] http://verbmobil.dfki.de/verbmobil

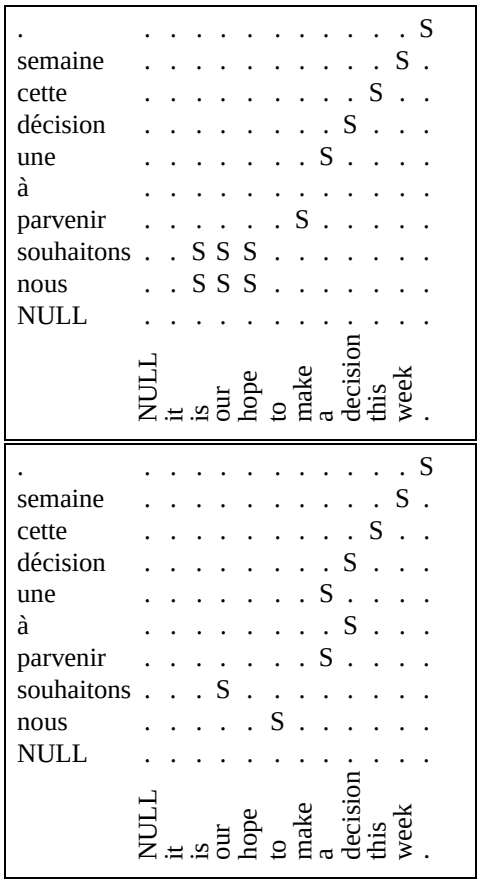

Figure 2: Two possible submission alignments with AER=0. Only the upper one is acceptable.

The databases have been selected to contain only recordings in US-English and to focus on the appointment scheduling domain. Then their counterparts in Catalan and Spanish have been generated by means of human translation (Arranz et al., 2003)[2]. Dates and times were categorised automatically (and revised manually). The test corpus consists of four hundred sentence pairs manually aligned by a single annotator. See the characteristics of the data in table 1.

		Spanish	English
Training	Sentences	28000 ≈ 28K	
	Words	201893	209653
	Vocabulary	4894	3167
	Singletons	2139	1251
Test	Sentences	400	
	Words	3124	3188

Table 1: Characteristics of Verbmobil corpus

4.1.2. Hansards Corpus

The corpus consists of the debates in the 36th Canadian parliament. We used a version of the Hansards aligned by Ullrich Germann at the level of sentences or smaller fragments (Germann, 2001). From the over 1.3 million of parallel text chunks, we selected those of 40 words or less. The size of this corpus is much larger than that of Verbmobil and the domain much more open so that the vocabulary is very large (see table 2). The test data were created by Franz Och and Hermann Ney (Och and Ney, 2000). They contain a restricted set of sure links and a large set of possible links.

		French	English
Training	Sentences	1008K	
	Words	16,95M	14,60M
	Vocabulary	76130	59534
	Singletons	32644	24370
Test	Sentences	484	
	Words	8482	7681

Table 2: Characteristics of Hansards corpus

4.2. Giza++ Baseline

The first decision to take in the symmetrisation process is the default starting point, which is systematically selected when our algorithm can't find an adequate group (step 4 of the algorithm). Combining the source-target and target-source information of the Giza++ alignments, we can obtain a high precision with low recall alignment (taking the intersection), a low precision with high recall alignment (taking the union), or intermediate combinations. The evaluation of different possible sets are presented in table 3.

As outlined in section 3.2., the best combination depends on the reference corpus. Both reference corpora contain more links than the Giza++ alignments because they have many-to-many alignments whereas Giza++ only produces one-to-one alignments. For Verbmobil, the reference corpus contains only S links. The recall plays an important role and the union is the best combination. The reference corpus for the Hansards task contains few S links and many P links. The intersection is the best combination because it keeps fewer, more precise links.

Results with weighted links, as described in section 3.1., are presented in a research report (Lambert and Castell, 2004). In most cases the effect of the weighting of the links is simply to move up the scores. However for the Hansards corpus it produces a qualitative change: the intersection gets a score worse than the union.

4.3. Symmetrisation Evaluation

From the results of the previous section and further experiments, the default starting point of the symmetrisation was set to be the union (of source-target and target-source Giza++ alignments) for the Verbmobil corpus and their intersection for the Hansards corpus.

Table 4 presents the evaluation of the symmetrisation process in these two cases. The symmetrisation increases the recall but introduces also some noise, so the precision is lower. However the outcome is a decrease of the error rate from 18.6 to 17.7 in the case of Verbmobil, and from 9.1 to 7.4 in the case of Hansards. The larger effect in the case of the Hansards could be due to the much greater size of the asymmetries repository. This allows a higher coverage but also permits to increase the threshold number of occurrences of an asymmetry, which implies a gain in precision. This threshold number was 3 for the Hansards, and 2 for Verbmobil.

[2]It is referred to as "subset-1" in the paper

Verbmobil corpus

Experiment	P_S (%)	R_S (%)	F_S (%)	P_P (%)	R_P (%)	F_P (%)	AER (%)
English to Spanish	92.82	64.18	75.89	92.82	64.18	75.89	24.11
Spanish to English	93.95	67.51	78.57	93.95	67.51	78.57	21.43
Intersection	**97.62**	57.59	72.44	**97.62**	57.59	72.44	27.56
Union	90.37	**74.11**	**81.43**	90.37	**74.11**	**81.43**	**18.57**

Hansards corpus

Experiment	P_S (%)	R_S (%)	F_S (%)	P_P (%)	R_P (%)	F_P (%)	AER (%)
English to French	60.89	91.04	72.97	90.29	30.74	45.86	9.41
French to English	62.08	85.81	72.04	90.58	28.50	43.36	11.42
Intersection	**74.06**	82.79	**78.18**	**98.10**	24.97	39.80	**9.13**
Union	53.45	**94.06**	68.17	85.56	**34.27**	**48.94**	11.36

Table 3: Giza++ evaluation

Verbmobil corpus

Experiment	P_S (%)	R_S (%)	F_S (%)	P_P (%)	R_P (%)	F_P (%)	AER (%)
Giza++ Union	90.37	74.11	81.43	90.37	74.11	81.43	18.57
Symmetrisation	88.68	76.75	82.28	88.68	76.75	82.28	17.72

Hansards corpus

Experiment	P_S (%)	R_S (%)	F_S (%)	P_P (%)	R_P (%)	F_P (%)	AER (%)
Giza++ Intersection	74.06	82.79	78.18	98.10	24.97	39.80	9.13
Symmetrisation	65.05	89.49	75.34	94.92	29.73	45.27	7.37

Table 4: Evaluation of the symmetrisation process

5. Conclusions

We used the Giza++ application to produce symmetric, phrase-based alignments with lower alignment error rate. In fact, our symmetrisation process could be applied to any two alignments of the same sentence pairs. The resulting alignments can in turn improve those applications where aligned corpora are a valuable resource. For instance, the obtained alignments could be used as phrase tuples in transducer machine translation. Thus our algorithm may be a simple way of improving machine translation results.

In this paper we also pointed out some critical issues concerning the evaluation methods. All of them stress the care with which evaluation results must be compared.

6. Acknowledgements

This work has been granted by the Spanish Government under grant TIC2002-04447-C02.

7. References

Arranz, Victoria, Núria Castell, and Jesús Giménez, 2003. Development of language resources for speech-to-speech translation. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*. Borovets, Bulgary.

Baum, L. E., 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3:1–8.

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

Germann, Ullrich, 2001. Aligned hansards of the 36th parliament of canada. release 2001-1a. http://www.isi.edu/natural-language/download/hansard/index.html.

Lambert, Patrik and Núria Castell, 2004. Evaluation and symmetrisation of alignments obtained with the giza++ software. Technical Report LSI–04–15–R, Technical University of Catalonia. http://www.lsi.upc.es/dept/techreps/techreps.html.

Melamed, I. Dan, 1998. Manual annotation of translational equivalence. Technical Report 98-07, IRCS.

Mihalcea, Rada and Ted Pedersen, 2003. An evaluation exercise for word alignment. In Rada Mihalcea and Ted Pedersen (eds.), *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*. Edmonton, Alberta, Canada: Association for Computational Linguistics.

Och, Franz Josef, 2000. Giza++: Training of statistical translation models. http://www.isi.edu/~och/GIZA++.html.

Och, Franz Josef and Hermann Ney, 2000. Improved statistical alignment models. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*. Hongkong, China.

Och, Franz Josef and Hermann Ney, 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann, 1996. HMM-based word alignment in statistical translation. In *COLING'96: The 16thInt. Conf. on Computational Linguistics*. Copenhagen, Denmark.