Bilingual Lexicon Induction: Effortless Evaluation of Word Alignment Tools and Production of Resources for Improbable Language Pairs

Adrien Lardilleux, Julien Gosme, Yves Lepage

GREYC, Université de Caen Basse-Normandie, France
firstname.lastname@info.unicaen.fr

Abstract

In this paper, we present a simple protocol to evaluate word aligners on bilingual lexicon induction tasks from parallel corpora. Rather than resorting to gold standards, it relies on a comparison of the outputs of word aligners against a reference bilingual lexicon. The quality of this reference bilingual lexicon does not need to be particularly high, because evaluation quality is ensured by systematically filtering this reference lexicon with the parallel corpus the word aligners are trained on. We perform a comparison of three freely available word aligners on numerous language pairs from the Bible parallel corpus (Resnik et al., 1999): MGIZA++ (Gao and Vogel, 2008), BerkeleyAligner (Liang et al., 2006), and Anymalign (Lardilleux and Lepage, 2009). We then select the most appropriate one to produce bilingual lexicons for all language pairs of this corpus. These involve Cebuano, Chinese, Danish, English, Finnish, French, Greek, Indonesian, Latin, Spanish, Swedish, and Vietnamese. The 66 resulting lexicons are made freely available.

1. Introduction

Bilingual lexicons are a valuable resource for many NLP tasks, such as machine translation or multilingual information retrieval.

High quality bilingual lexicons are widely available for well-resourced language pairs, such as English-French, or English-Chinese. However, the number of such well-resourced language pairs still remains very low today. Even well-resourced languages, such as Chinese or French, do not necessarily constitute a well-resourced language pair when paired together (*e.g.*, Chinese-French). Obviously, any language pair involving a less-resourced language typically falls into the category of less-resourced language pairs. Hand-made bilingual lexicons thus remain a rare resource for most language pairs.

Many researchers have investigated the issue of automatic constitution of bilingual lexicons for this reason. For example, dictionaries for new language pairs can be obtained by combining existing bilingual dictionaries sharing a common language (Tanaka and Umemura, 1994; Bond et al., 2001; Nerima and Wehrli, 2008). Another approach consists in using sub-sentential alignment techniques to induce a bilingual lexicon from a parallel corpus, be it sentence-aligned (Wu and Xia, 1994) or not (Fung and Church, 1994). More recently, extraction of translation equivalents from comparable corpora rather than paralell corpora has drawn an increasing attention (Fung, 1995; Chiao and Zweigenbaum, 2002; Yu and Tsujii, 2009).

We propose to induce bilingual lexicons from sentence-aligned parallel corpora for *improbable* language pairs, *i.e.*, pairs of languages typically neglected because they would be of use for only a small number of users. This work originates from our need for bilingual lexicons involving some less-resourced language pairs. These languages happen to be in the Bible parallel corpus, created several years ago by Resnik et al. (1999). It covers 13 languages. First, we perform an evaluation of freely available word aligners on language pairs for which resources for evaluation are available. We then choose the most promising one and produce resources for all language pairs from this corpus. We eventually make the lexicons freely available.

This paper is organized as follows. Section 2 details the word aligner evaluation process. In Section 3 we conduct an experiment on many language pairs and report the results. Section 4 describes the newly created bilingual lexicons for all language pairs.

2. Evaluation of word aligners

In this section, we focus on defining an effortless evaluation protocol of word alignment tools on the task of bilingual lexicon induction from parallel corpora.

2.1. Lexicon evaluation vs. word aligner evaluation

Typically, bilingual lexicons obtained from parallel corpora are built by

- running a word aligner on a parallel corpus,
- extracting translation pairs and their associated translation probabilities, and
- filtering out translation pairs which have associated probabilities below a certain threshold.

To assess the quality of bilingual lexicons obtained in this way, a representative sample is usually manually evaluated, in order to obtain a measure of precision. Automatic evaluations were also proposed, for example by comparing the bilingual lexicon to a test parallel corpus (Melamed, 1995). Subjective evaluation may be difficult in the process of developing translation lexicons involving exotic languages, even on small representative samples.

We propose to evaluate the *potential* of word aligners to produce good quality lexicons instead of the final lexicons themselves, which highly depend on the filtering criteria used. To this end, we improve the quality of traditional measures by taking into account all raw translation

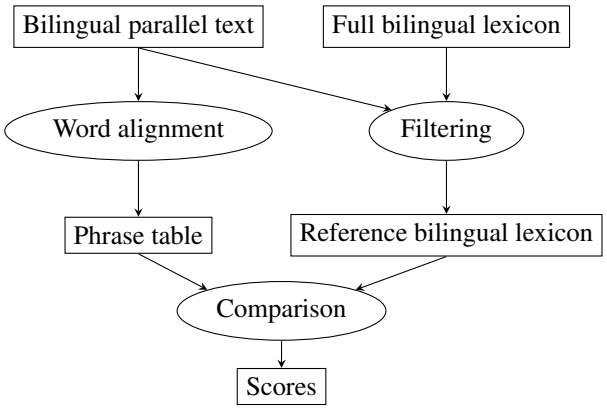

Figure 1: Evaluation protocol overview.

pairs output by a word aligner, weighted by their associated translation probabilities. It will then be up to the practitioner to decide whether some particular set of filtering rules have to be applied, or if the translation probabilities are to be included in the final lexicon.

2.2. Evaluation specifications

In our evaluation, we compare the output of word aligners to an existing reference bilingual lexicon, possibly containing multi-word entries. Given a sentence-aligned parallel corpus in two languages, the steps are the following:

1. run the word aligner on the parallel corpus. The result is a list of translation pairs of n-grams of words, along with their associated translation probabilities (hereafter referred to as a phrase table);

2. filter the reference bilingual lexicon so that the actual reference contains only entries that can actually be extracted from the training parallel corpus. Practically, an entry in the initial bilingual lexicon is kept as a reference if it is a subsequence of a pair of sentences from the training corpus;

3. compute the f-measure, which is the harmonic mean of precision and recall, with:

   precision: sum of the source-to-target translation probabilities for those alignments from a phrase table that match an entry in the (filtered) reference bilingual lexicon, divided by the number of unique source entries covered by these alignments;

   recall: same as precision except that we divide by the number of unique source entries covered by the (filtered) reference bilingual lexicon.

A summary is shown on Figure 1.

2.3. Advantages of the approach

We thus rely on a single resource for evaluation, which is an existing reference bilingual lexicon. This may sound contradictory since our ultimate goal is precisely to produce new lexicons. However, we stress the fact that we do not evaluate lexicons, but word aligners. Evaluation is thus typically performed on well-resourced languages, but new lexicons for any language pair can be produced as long as parallel corpora are available.

It is important to note that the quality of the initial reference bilingual lexicon is not much of an issue. Indeed, it is systematically filtered so that it only contains entries from the training parallel corpus, which is tantamount to take the intersection of these two resources. The quality of the resulting filtered reference bilingual lexicon is thus guaranteed, whatever the origin of the initial reference bilingual lexicon. As a result, one can serenely compile the initial bilingual reference lexicons from various sources, erroneous and out-of-domain entries being naturally filtered out when taking the intersection with the parallel corpus.

This last point is particularly useful as it allows to join existing dictionaries together to produce dictionaries between new language pairs by transitivity (Nerima and Wehrli, 2008) without the need for additional processing. For instance, one can safely join a Chinese-English dictionary with an English-Finish dictionary to produce a Chinese-Finish dictionary to be used as a reference bilingual lexicon. No filtering on the spurious entries resulting from the join operation is required, because it will implicitly be done when taking the intersection with the parallel corpus. Such transitivity thus permits evaluation of word aligners even on less-resourced language pairs.

3. Experiments

For our experiments, we use the Bible parallel corpus (Resnik et al., 1999), which consists of rougly 30,000 aligned sentences in 13 languages: Cebuano, Chinese, Danish, English, Finnish, French, Greek, Indonesian, Latin, Spanish, Swahili,[1] Swedish, and Vietnamese. The sentence length in words for the English part is 29 (avg.) $\pm$13 (std. dev.).

3.1. Constitution of reference dictionaries

We intend to evaluate the word aligners on all language pairs. Unfortunately, obtaining exploitable reference bilingual lexicons can be very difficult for most pairs since they constitute less-resourced language pairs. We therefore decide to build them by transitivity, as previously described. We choose English as “pivot" language since it is the most resourced language amongst all languages of our corpus.

We select bilingual dictionaries which number of entries allows us to perform a significant evaluation. These initial bilingual lexicons are listed in Table 1. The English-Chinese lexicon comes from CC-CEDICT[2] and the others from the Freelang project.[3] All missing language pairs are then obtained by joining these initial bilingual lexicons on their English part. The number of entries in the resulting lexicons is presented in Table 2. Eventually, these lexicons are filtered with the parallel corpus. The final bilingual reference lexicons are presented in Table 3. In average, they represent roughly 10% of the vocabulary of the parallel text,

[1] This part of the parallel corpus is unavailable at the time these experiments are conducted.

[2] http://cc-cedict.org/

[3] http://freelang.org/

Language pair	# entries
eng-dan	16,112
eng-fin	30,188
eng-fra	104,775
eng-spa	27,639
eng-swe	36,070
eng-zho	287,651

Table 1: Number of entries in the initial bilingual lexicons used for the experiment.

	fin	fra	spa	swe	zho
dan	25,519	50,532	12,703	22,408	64,989
fin		97,448	21,359	49,051	140,525
fra			41,409	93,470	242,023
spa				37,534	106,142
swe					127,650

Table 2: Number of entries in the lexicons obtained by joining all pairs of initial lexicons, using English as "pivot".

which is quite low, but we assimilate this fraction to a representative sample of the vocabulary of the text.

3.2. The word aligners

We compare three freely available word aligners:

MGIZA++ (Gao and Vogel, 2008)[4] which corrects some bugs from the original GIZA++ (Och and Ney, 2003). It implements the well-known IBM models (Brown et al., 1993) and the HMM model (Vogel et al., 1996). Alignments are made symmetric and phrase alignments extracted and scored with the Moses toolkit (Koehn et al., 2007);

BerkeleyAligner (Liang et al., 2006)[5] in which two simple asymmetric models are trained jointly. The resulting alignments are symmetric, we thus simply extract and score phrase alignments with Moses;

Anymalign (Lardilleux and Lepage, 2009)[6] based on random sampling and string differences. It directly produces symmetric phrase alignments and their associated scores.

[4] http://geek.kyloo.net/software/doku.php/mgiza:overview/

[5] http://nlp.cs.berkeley.edu/Main.html#WordAligner

[6] http://users.info.unicaen.fr/~alardill/anymalign/

	eng	fin	fra	spa	swe	zho
dan	1,290	972	2,928	1,890	2,346	2,271
eng		1,043	4,648	2,381	3,147	3,869
fin			1,908	1,264	1,494	1,381
fra				4,809	5,407	5,127
spa					3,163	3,492
swe						3,986

Table 3: Number of entries in the final reference lexicons. These are obtained by filtering the lexicons resulting from the join operation.

	dan	eng	fin	fra	spa	swe	zho
dan		46	32	35	37	51	29
eng	39		27	36	42	36	26
fin	40	34		25	28	36	26
fra	33	43	25		45	29	24
spa	39	46	28	46		34	27
swe	48	43	32	31	33		25
zho	19	18	17	15	17	17	

Table 4: F-measures (percentages) obtained by MGIZA++/Moses for all language pairs. The source language is indicated in the first column and the target language in the first line.

	dan	eng	fin	fra	spa	swe	zho
dan		–24	–10	–19	–16	– 7	–21
eng	–27		– 8	– 7	– 8	–21	– 7
fin	– 5	**+ 7**		**+26**	**+25**	**+ 1**	**+ 2**
fra	–19	– 1	**+ 5**		– 8	–13	– 9
spa	–20	– 4	**+10**	– 9		–19	**+15**
swe	– 7	–13	– 3	– 6	–10		–11
zho	–25	–13	**+ 2**	**+ 1**	– 9	–24	

Table 5: Relative gain in f-measure (percentages) when using BerkeleyAligner/Moses instead of MGIZA++/Moses. An average loss of 7% is observed relative to MGIZA++/Moses.

The alignment process is unsupervised and based on surface forms only. Although all three tools are suited for parallel processing, for a fair evaluation we only use them on a single processor. The default set of parameter values are used for the three aligners, as they typically yield good results. We compare the f-measures obtained by the three aligners on all language pairs, MGIZA++ serving as a reference. The behavior of Anymalign is special in that it can be stopped at any time; therefore, we first run MGIZA++/Moses and BerkeleyAligner/Moses and measure the time required to process each language pair, and make Anymalign run for the same amount of time. This is roughly 25 minutes in average.

	dan	eng	fin	fra	spa	swe	zho
dan		**+ 3**	–10	–15	**+ 9**	**+8**	– 4
eng	–15		–10	**+13**	**+ 2**	–6	– 5
fin	**+ 2**	**+36**		**+70**	**+53**	**+7**	**+11**
fra	–15	0	– 2		**+ 1**	–3	**+ 5**
spa	– 9	**+15**	**+ 3**	**+13**		–2	**+15**
swe	– 4	**+ 7**	–18	**+19**	**+ 7**		– 1
zho	–13	**+16**	0	**+58**	**+31**	**+3**	

Table 6: Relative gain in f-measure (percentages) when using Anymalign instead of MGIZA++/Moses. An average gain of 7% is observed relative to MGIZA++/Moses.

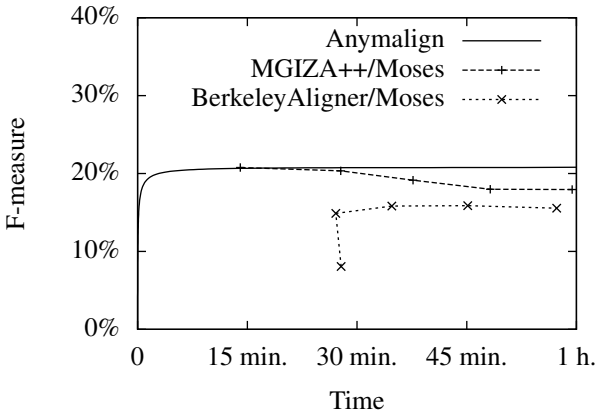

Figure 2: Comparison of the outputs of the three word aligners against a reference bilingual lexicon on the Chinese-English task.

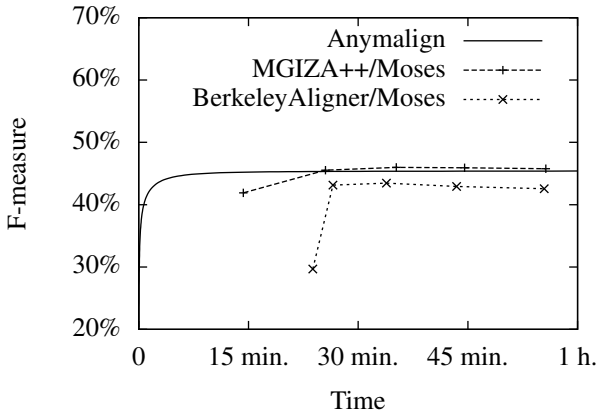

Figure 3: Comparison of the outputs of the three word aligners against a reference bilingual lexicon on the French-Spanish task.

3.3. Results

The f-measures obtained by MGIZA++/Moses are presented in Table 4. As one would expect, better results are obtained between close languages, such as Danish to Swedish (51%), than between distant languages, *e.g.* Chinese to French (15%).

The gain in f-measure relative to MGIZA++/Moses for BerkeleyAligner/Moses and Anymalign are presented in Tables 5 and 6, respectively. The differences in score can be very different from one language to another. For instance, MGIZA++ seems to lead to much better results than the two others when aligning into Danish, while it is clearly outperformed when aligning into Finnish. In average, BerkeleyAligner leads to a loss of 7% in f-measure relatively to MGIZA++/Moses and Anymalign leads to a gain of 7%.

In order to get a better insight of the behaviors of the aligners, we repeat this experiment and we now evaluate the f-measures according to the processing time they require to produce a phrase table. As for MGIZA++ and BerkeleyAligner, we vary the number of iterations (from 1 to 5) of the default models (IBM1, HMM, IBM3, and IBM4 for MGIZA++, IBM1 and HMM for BerkeleyAligner) and measure the elapsed CPU time, including the time required to extract and score phrase alignments. Anymalign can be stopped at any time, so we just perform the same experiment for numerous processing times. The results of the Chinese-English task (distant languages) and of the French-Spanish task (close languages) are visible in Figures 2 and 3, respectively.

Again, all three aligners yield better results when aligning close languages. Generally speaking, the quality of the phrase tables obtained from the three aligners is comparable on these two experiments, with a little advantage to Anymalign and MGIZA++. However Anymalign is much faster: two to five minutes typically suffice to get the most of it in these experiments. For some reason, it also appears that increasing the number of iterations may hinder the quality of the results of MGIZA++ and BerkeleyAligner, which is most visible on the Chinese to English task. The odd position of BerkeleyAligner's first point on the two graphs is due to the fact that this aligner produces spurious links when running for a single iteration, which results in a much larger phrase table, Moses thus needing much more time to extract and score phrase alignments.

Note that Anymalign has already shown to produce better results on unigrams of words than longer n-grams (Lardilleux et al., 2009). It thus appears to be currently more appropriate to produce multilingual lexicons, as we will do in the next section, whereas MGIZA++ and BerkeleyAligner may be more appropriate for other tasks such as statistical machine translation.

4. Constitution of resources for improbable language pairs

We use Anymalign to produce bilingual lexicons for all language pairs of the Bible parallel corpus, as it has shown to be able to produce slightly better results than the two other aligners in our experiments, much faster.

We thus build bilingual lexicons including translation probabilities in two directions (source to target and target to source) for all pairs of languages including: Cebuano, Chinese, Danish, English, Finnish, French, Greek, Indonesian, Latin, Spanish, Swedish, and Vietnamese. Although the original corpus uses specific transliterations for most languages, all our final lexicons are simple and versatile text files encoded in UTF-8. Figure 4 gives an excerpt of the Danish-Vietnamese bilingual lexicon.

These lexicons will be made publicly available at the following address:

`http://users.info.unicaen.fr/~alardill/lexicons/`

Additional bilingual lexicons from various parallel corpora shall be added in the future.

5. Conclusion

In this paper, we have described an effortless protocol to evaluate word aligners on a bilingual lexicon induction task. It relies on the comparison between alignments of words generated from parallel corpora, weighted by their translation probabilities, and a reference bilingual lexicon. The quality of this reference lexicon is not an issue because

```
.           .           0.66  0.68
jesus       yêsu        0.75  0.97
?           ?           0.91  0.96
johannes    yoan        0.99  0.99
:           :           0.72  0.67
,           ,           0.68  0.57
ikke        không       0.68  0.75
paulus      phaolô      0.99  0.99
peter       phêrô       0.99  0.99
"           "           0.76  0.69
jerusalem   yêrusalem   0.98  0.99
kristus     kitô        0.86  0.85
```

Figure 4: First lines of the Danish-Vietnamese bilingual lexicon. Each line consists of four fields separated by tabulations: a source entry, a target entry, a source-to-target translation probability, and a target-to-source translation probability. The lexicons also contain multi-words entries.

it is filtered by the parallel corpus used to train the word aligners. We compared the MGIZA++, BerkeleyAligner, and Anymalign word aligners on many language pairs and showed that, in average the latter was more appropriate for the task of bilingual lexicon induction. We eventually built bilingual lexicons using Anymalign between all language pairs from the Bible parallel corpus, which include mostly less-resourced language pairs, and made them freely available.

6. References

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. 2001. Design and Construction of a machine-tractable Japanese-Malay Dictionary. In *Proceedings of the eight Machine Translation Summit (MT Summit VIII)*, pages 53–58, Santiago de Compostela, Spain, September.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *Proceedings of the 19th International Conference on Computational Linguistics (Coling'02)*, pages 1208–1212, Taipei, Taiwan.

Pascale Fung and Kenneth Ward Church. 1994. K-vec: A New Approach for Aligning Parallel Texts. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling'94)*, volume 2, pages 1096–1102, Kyoto, Japan.

Pascale Fung. 1995. Compiling Bilingual Lexicon Entries From a Non-Parallel English-Chinese Corpus. In *Proceedings of the Third Workshop on Very Large Corpora (VLC'95)*, pages 173–183, Cambridge, USA.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, June.

Adrien Lardilleux and Yves Lepage. 2009. Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, Borovets, Bulgaria, September.

Adrien Lardilleux, Jonathan Chevelu, Yves Lepage, Ghislain Putois, and Julien Gosme. 2009. Lexicons or phrase tables? An investigation in sampling-based multilingual alignment. In *Proceedings of the 3rd Workshop on Example-Based Machine Translation (EBMT3)*, pages 45–52, Dublin, Ireland.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by Agreement. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 104–111, New York City, USA, June. Association for Computational Linguistics.

Dan Melamed. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. In *Proceedings of the Third Workshop on Very Large Corpora (VLC'95)*, pages 184–198, Boston, Massachusetts, USA, June.

Luka Nerima and Eric Wehrli. 2008. Generating Bilingual Dictionaries by Transitivity. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 2584–2587, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Franz Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29:19–51, March.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The Bible as a Parallel Corpus: Annotating the "Book of 2000 Tongues". *Computers and the Humanities*, 23(1-2):129–153.

Kumiko Tanaka and Kyoji Umemura. 1994. Construction of a Bilingual Dictionary Intermediated by a Third Language. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling'94)*, pages 297–303, Kyoto, Japan, August.

Stephan Vogel, Hermann Ney, and Christoph Tillman. 1996. HMM-Based Word Alignment in Statistical Translation. In *Proceedings of the 16th International Conference on Computational Linguistics (Coling'96)*, pages 836–841, Copenhagen, Denmark, August.

Dekai Wu and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA 1994)*, pages 206–213, Columbia, Maryland, USA, October.

Kun Yu and Junichi Tsujii. 2009. Bilingual dictionary extraction from Wikipedia. In *Proceedings of the twelfth Machine Translation Summit (MT Summit XII)*, pages 379–386, Ottawa, Canada, August.