Bayesian Word Alignment for Massively Parallel Texts

Robert Östling
Department of Linguistics
Stockholm University
robert@ling.su.se

Abstract

There has been a great amount of work done in the field of bitext alignment, but the problem of aligning words in massively parallel texts with hundreds or thousands of languages is largely unexplored. While the basic task is similar, there are also important differences in purpose, method and evaluation between the problems. In this work, I present a non-parametric Bayesian model that can be used for simultaneous word alignment in massively parallel corpora. This method is evaluated on a corpus containing 1144 translations of the New Testament.

1 Introduction

Bitext word alignment is the problem of finding links between words given pairs of translated sentences (Tiedemann, 2011). Initially, this was motivated by Statistical Machine Translation (SMT) applications (Brown et al., 1993), but word-aligned texts have also been used to transfer linguistic annotation between languages (Yarowsky et al., 2001; Täckström, 2013), for Word Sense Disambiguation (WSD) (Diab and Resnik, 2002) and lexicon extraction (Wu and Xia, 1994).

Massively parallel texts, in the sense used by Cysouw and Wälchli (2007), are essentially the same as bitexts, only with hundreds or thousands of languages rather than just two. Parallel corpora used in SMT, for instance the Europarl Corpus (Koehn, 2005), tend to contain few (up to tens of) languages, but many (up to billions of) words in each language. Massively parallel corpora, on the other hand, contain many (hundreds of) languages, but usually fewer (less than a million) words in each language.

Additionally, aligned massively parallel corpora have different applications than traditional parallel corpora with pairwise alignments. Whereas the latter tend to be used for the various NLP tasks mentioned above, massively parallel corpora have mostly been used for investigations in linguistic typology (Cysouw and Wälchli, 2007).

There has been surprisingly few studies on multilingual word alignment. Mayer and Cysouw (2012) treat alignment as a clustering problem, where the words in each sentence are clustered according to some measure of co-occurrence. They provide no evaluation, but alignment methods based on co-occurrence statistics have been found to have lower accuracy than even very simple generative models (Och and Ney, 2003), so this might not be a promising direction as far as accuracy is concerned.

A related line of research is due to Lardilleux et al. (2011), who learn sets of multilingual translation equivalent phrases. Although later work (Lardilleux et al., 2013) uses phrase pairs extracted with this method for (bitext) word alignment, their method solves a somewhat different problem from what is considered here.

Some authors have studied how multilingual parallel corpora can be used to improve bitext alignment. Filali and Bilmes (2005) use (bitext) alignments to addditional languages as features in bitext alignment, while Kumar et al. (2007) interpolate alignments through multiple bridge languages to produce a bitext alignment for another language pair. Since the goal of this research is not multilingual alignment, it will not be considered further here.

2 Multilingual Alignment

In bitext alignment, the goal is to find links between individual word tokens in parallel sentence pairs. The IBM models (Brown et al., 1993) formalize this in a directional fashion where each word j in a *source language* is linked to word i in the *target language* through alignment variables $i = a_j$, thus specifying a 1-to-n mapping from

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 123–127,
Gothenburg, Sweden, April 26-30 2014.

source language words to target language words.

An intuitively appealing way to formalize the multilingual alignment problem is through a *common representation* (or *interlingua*) to which each individual language is aligned. If the common representation is isomorphic to one of the languages in the corpus, this is equivalent to using that language as a bridge. However, since all languages (and all translations) have their own idiosyncrasies that make linking to other translations difficult, it seems better to learn a common representation that corresponds to information in a sentence that is present in as many of the translations as possible.

3 Method

Recently, it has been shown that Bayesian methods that use priors to bias towards linguistically more plausible solutions can improve bitext word alignment (Mermer and Saraçlar, 2011; Riley and Gildea, 2012; Gal and Blunsom, 2013). Given these promising results and the fact that massively parallel texts tend to be rather short, which makes the role of realistic priors more important, I have decided to use a Bayesian alignment model for this work.

3.1 Model

The model used in this work uses a common representation of *concepts* generated by a Chinese Restaurant Process (CRP), which is aligned to each of the languages in a corpus using the model of Mermer and Saraçlar (2011).

Table 1 introduces the variables (observed and latent) as well as the hyperparameters of the model. Basically, the model consists of a common representation c (where token i of sentence s is denoted c_{si}), which is aligned to one or more words w_{lsj} (from language l, sentence s, token j) through a set of alignment variables a_{lsj} which contain the index within c_s that w_{lsj} is linked to.

The probability of an assignment c is:

$$\mathrm{CRP}(c;\alpha) = \frac{\Gamma(1+\alpha)}{\Gamma(n+\alpha)} \cdot \alpha^{|E_c|-1} \cdot \prod_{e \in E_c} (n_e - 1)!$$

where n_e is the number of occurrences of concept type e in the assignment c, and $n = \sum_e n_e$ is the (fixed) total number of tokens in the common representation.

For the translation probabilities, I follow Mermer and Saraçlar (2011) in assuming that $p(f_l|e) \sim \mathrm{Dir}(t_l;\theta_l)$, and that the priors θ_l are *symmetric* (i.e. all values in these vectors are equal, $\theta_{lef} = \beta$). By specifying a low value for β (a *sparse* prior), we can encode our prior knowledge that translation probability functions $p(f_l|e)$ tend to have a low entropy, or in other words, that each concept is typically only translated into a very small number of words in each language.

The joint probability of the common representation and the alignments is given by:

$$\begin{aligned} &p(c,a,w,t;\alpha,\theta) = \\ &\quad p(c;\alpha) \cdot p(w|c,a,t) \cdot p(a|c) \cdot p(t;\theta) \end{aligned} \tag{1}$$

where $p(c;\alpha) = \mathrm{CRP}(c;\alpha)$ and the remaining factors are the same as in Mermer and Saraçlar (2011) with the common representation being the "target language", except that there is a product across all languages l. Note that since word order is not modeled, $p(a|c)$ is constant.

3.2 Learning

The model is trained using a collapsed Gibbs sampler. Due to space limitations, the full derivation is omitted, but the sampling distribution turns out to be as follows for the common representation:

$$\begin{aligned} p(c_{si} = e') \propto \frac{1}{n-1+\alpha} \cdot \begin{cases} \alpha & \text{if } n_{e'} = 1 \\ n_{e'} - 1 & \text{if } n_{e'} > 1 \end{cases} \\ \cdot \prod_l \frac{\prod_{f \in A_{lsi}} \prod_{k=1}^{m_{lsif}} \left(n_{le'f} + \theta_{le'f} - k\right)}{\prod_{k=1}^{\sum_f m_{lsif}} \left(\sum_{f \in F_l} n_{le'f} + \theta_{le'f} - k\right)} \end{aligned} \tag{2}$$

where A_{lsi} is the set of word types f in language l which are aligned to c_{si}, and m_{lsif} is the number of times each such f is aligned to c_{si}. In order to speed up calculations, the product in Equation 2 can be approximated by letting l run over a small random subset of languages. The experiments carried out in this work only use this approximation when the full corpus of 1144 translations is used, then a subset of 24 languages is randomly selected when each c_{si} is sampled. An empirical evaluation of the effects of this approximation is left for future work.

The alignment sampling distribution is:

$$p(a_{lsj} = i) \propto \frac{n_{le'f'} + \theta_{le'f'} - 1}{\sum_f \left(n_{le'f} + \theta_{le'f}\right) - 1} \tag{3}$$

where $e' = c_{s a_{lsj}}$ is the concept type aligned to word type $f' = w_{lsj}$.

Rather than sampling directly from the distributions above, one can sample from $\hat{p}(c_{si} = e') \propto$

Table 1: Variables used in the model.

Observed variables	
F_l	the set of word types in language l
$w_{lsj} \in F_l$	word j of sentence s in language l
$I_s \in \mathbb{N}$	length of sentence s in the common representation
$J_{ls} \in \mathbb{N}$	length of sentence s in language l
Latent variables	
E_c	the set of concepts in the assignment c
$c_{si} \in E_c$	concept i of sentence s in the common representation
$a_{lsj} \in \{1..I_s\}$	alignment of w_{lsj} to c_{si}; $i = a_{lsj}$
$t_{lef} \in \mathbb{R}$	translation probability $p(f_l\|e)$, where $f_l \in F_l$ and $e \in E_c$
Hyperparameters	
α	CRP hyperparameter, fixed to 1000 in the experiments
β	symmetric Dirichlet prior for translation distributions $\theta_\mathbf{l}$, fixed to 0.001 in the experiments

$p(c_{si} = e')^\lambda$ and $\hat{p}(a_{lsj} = i) \propto p(a_{lsj} = i)^\lambda$. The temperature parameter λ can be varied during training to change the amount of randomness while sampling.

3.3 Initialization

In order to obtain a reasonable initial state for the Gibbs sampling, one can simply initialize the common representation to be identical to one of the languages in the corpus. For this language one then (trivially) has a perfect alignment, while the remaining languages are initialized randomly and their alignments are learned. Random initialization of the common representation is possible, but turns out to perform poorly.

4 Experiments

The most basic question about the present model is whether sampling the common representation is helpful, compared to simply choosing a language and aligning all other languages to that one.

In order to test this, I initialize the model as described in section 3.3 and sample alignments (but not the common representation) for 200 iterations with λ linearly increasing from 0 to 2, followed by two iterations with $\lambda \to \infty$. This gives a strong baseline, from which one can start learning the joint model.

4.1 Data

I use a corpus containing verse-aligned translations of the New Testament into a great number of languages. After some exclusions due to e.g. non-standard formatting or improperly segmented text, the version used in this work contains 1144 translations in 986 different languages. The mean number of tokens among the translations is 236 000, and the mean number of types is 9 500.

4.2 Evaluation Measures

Previous authors have tended to avoid multilingual evaluation altogether. Mayer and Cysouw (2012) do not evaluate their method, while Lardilleux et al. (2011) only use bilingual evaluation.

Cysouw et al. (2007) use the fact that some translations of the Bible have been annotated with Strong's Numbers, which map most word tokens to the lemma of its translation equivalent in the original language, to perform bilingual evaluation of Bible corpus alignments.

Strong's Numbers can be used in a different way to evaluate the type of multilingual alignment produced by the method in this work. Both the Strong's Numbers and the common representation can be interpreted as clusterings of the word tokens in each language. Ideally one would want these two clusterings to be identical, as they would be if the original language had been perfectly constructed. Standard clustering evaluation measures can be used for this task, and in this work I use normalized mutual information (also reinvented as *V-measure* by Rosenberg and Hirschberg (2007)). The evaluation is limited to words which are assigned exactly one Strong's Number, in an attempt to avoid some of the problems with scope discussed by Cysouw et al. (2007). Note that even a perfect alignment from one language to itself does not achieve the maximum score using this mea-

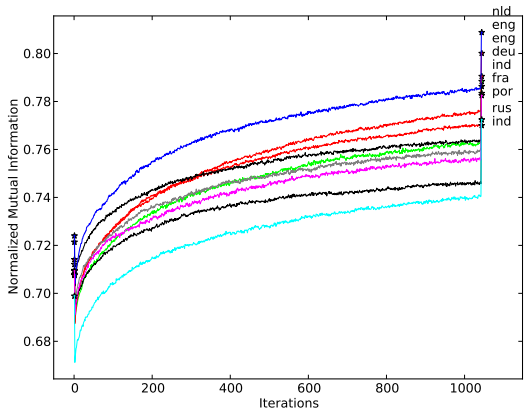

Figure 1: Alignment quality of Mandarin-initialized model.

sure, only a successful reconstruction of the original text (minus inflections) would.

In the Bible corpus used here, nine translations in seven languages contain Strong's Numbers annotations: English and Indonesian (two translations each), as well as German, French, Dutch, Portuguese and Russian (one translation each).

4.3 Results

Figure 1 shows alignment quality during training in a model initialized using a translation in Mandarin, which is not related to any of the languages in the evaluation sample and was chosen to avoid initialization bias. After an initial drop when noise is introduced during the Gibbs sampling process, alignment quality quickly increases as the common representation moves towards the versions in the evaluation sample. The final two iterations (with $\lambda \to \infty$) remove the sampling noise and the model rapidly converges to a local maximum, resulting in a sharp increase in alignment quality at the end. Further iterations only result in minor improvements.

Table 2 contains the baseline and joint model results for models initialized with either English or Mandarin versions. The joint model outperforms the baseline in all cases except when the initialization language is the same as the evaluation language (the two English translations in the left column), which is expected since it is easy to align a text to itself or to a very similar version.

The two models described so far only use the nine-translation evaluation sample to learn the common representation, since using additional languages would unfairly penalize the joint learning model. I have also tested the model on the full corpus of 1144 translations with an English-initialized model and the same training setup as above (initialized from English). In this case, alignment quality decreased somewhat for the languages most similar to English, which is to be expected since the majority of languages in the corpus are unrelated to English and pull the common representation away from the European languages in the evaluation sample. Although it is not possible to directly evaluate alignment quality outside the evaluation sample with Strong's Numbers, the log-probability of the entire data under the model (Equation 1) increases as expected, by about 5%.

	English		Mandarin	
	A	A+J	A	A+J
deu	0.817	**0.824**	0.708	**0.788**
eng	**0.854**	0.851	0.714	**0.800**
eng_2	**0.834**	0.833	0.708	**0.790**
fra	0.807	**0.816**	0.712	**0.783**
ind	0.774	**0.785**	0.710	**0.770**
ind_2	0.791	**0.803**	0.721	**0.786**
nld	0.839	**0.850**	0.724	**0.809**
por	0.807	**0.813**	0.709	**0.782**
rus	0.792	**0.800**	0.699	**0.772**

Table 2: Normalized mutual information with respect to Strong's Numbers, using alignment only (A) or joint alignment + common representation learning (A+J), for models initialized using English or Mandarin.

5 Conclusions and Future Work

As the number of translations in a parallel corpus increases, the problem of aligning them becomes a rather different one from aligning translation *pairs*. I have presented a Bayesian method that jointly learns a common structure along with alignments to each language in the corpus. In an empirical evaluation, the joint method outperforms the baseline where the common structure is one of the languages.

Currently the underlying alignment model is quite simplistic, and preliminary results indicate that including the HMM word order model of Vogel et al. (1996) further improves alignments.

Acknowledgments

Thanks to Jörg Tiedemann, Mats Wirén and the anonymous reviewers for their comments.

References

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

Michael Cysouw and Bernhard Wälchli. 2007. Parallel texts: Using translational equivalents in linguistic typology. *STUF - Language Typology and Universals*, 60(2):95–99.

Michael Cysouw, Chris Biemann, and Matthias Ongyerth. 2007. Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *STUF - Language Typology and Universals*, 60(2):158–171.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Stroudsburg, PA, USA. Association for Computational Linguistics.

Karim Filali and Jeff Bilmes. 2005. Leveraging multiple languages to improve statistical MT word alignments. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 92–97, San Juan, November. IEEE.

Yarin Gal and Phil Blunsom. 2013. A systematic bayesian treatment of the ibm alignment models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit*, Phuket, Thailand.

Shankar Kumar, Franz J. Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech Republic, June. Association for Computational Linguistics.

Adrien Lardilleux, Yves Lepage, and Franois Yvon. 2011. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189–217.

Adrien Lardilleux, Francois Yvon, and Yves Lepage. 2013. Hierarchical sub-sentential alignment with Anymalign. In *Proceedings of the 16th EAMT Conference*, pages 279–286, Trento, Italy, 28-30 May 2012.

Thomas Mayer and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, EACL 2012, pages 54–62, Stroudsburg, PA, USA. Association for Computational Linguistics.

Coşkun Mermer and Murat Saraçlar. 2011. Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 182–187, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Darcey Riley and Daniel Gildea. 2012. Improving the IBM alignment models using variational Bayes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 306–310, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic, June. Association for Computational Linguistics.

Oscar Täckström. 2013. *Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology.

Jörg Tiedemann. 2011. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 836–841, Stroudsburg, PA, USA. Association for Computational Linguistics.

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*, pages 206–213.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, Stroudsburg, PA, USA. Association for Computational Linguistics.