Word Alignment without NULL words

Schulz, P.; Aziz, W.; Sima'an, K.

Word Alignment without NULL Words

Philip Schulz
P.Schulz@uva.nl

Wilker Aziz
W.Aziz@uva.nl

Khalil Sima'an
K.Simaan@uva.nl

ILLC
University of Amsterdam

Abstract

In word alignment certain source words are only needed for fluency reasons and do not have a translation on the target side. Most word alignment models assume a target NULL word from which they generate these untranslatable source words. Hypothesising a target NULL word is not without problems, however. For example, because this NULL word has a position, it interferes with the distribution over alignment jumps. We present a word alignment model that accounts for untranslatable source words by generating them from preceding source words. It thereby removes the need for a target NULL word and only models alignments between word pairs that are actually observed in the data. Translation experiments on English paired with Czech, German, French and Japanese show that the model outperforms its traditional IBM counterparts in terms of BLEU score.

1 Introduction

When the IBM models (Brown et al., 1993) were designed, some way of accounting for words that likely have no translation was needed. The modellers back then decided to introduce a NULL word on the target (generating) side[1]. All words on the source side without a proper target translation would then be generated by that NULL word.

While this solution is technically valid, it neglects that those untranslatable words are required for source fluency. Moreover, the NULL word, although hypothetical in nature, does have a position. It is well-known that this NULL position is problematic for distortion-based alignment models. Alignments to NULL demand a special treatment as they would otherwise induce very long jumps that one does not usually observe in distortion-based alignment models. Examples of this can be found in Vogel et al. (1996), who drop the NULL word entirely and thus force all source words to align lexically, and Och and Ney (2003), who choose a *fixed* NULL probability.

In the present work, we introduce a family of IBM-style alignment models that can express dependencies between translated and untranslated source words. The models do not use NULL words and instead allow untranslatable source words to be generated from translated words in their context. This is achieved by modelling source word collocations. From a technical point of view the model can be seen as a mixture of an alignment and a language model.

2 IBM models 1 and 2

Here, we quickly review the IBM alignment models 1 and 2 (Brown et al., 1993). We assume a random variable E over the English (target) vocabulary[2], a variable F over the French (source) vocabulary and a variable A over alignment links[3]. The IBM models assign probabilities to alignment configurations and source sentences given the target side. Under the assumption that all source words are conditionally independent given the alignment links, these probabilities factorise as

$$P(f_1^m, a_1^m | e_0^l) = P(a_1^m) \prod_{j=1}^{m} P(f_j | e_{a_j}) \quad (1)$$

where x_1^k is a vector of outcomes $x_1, \ldots, x_k$ and e_{a_j} denotes the English word that the French word

---

[1]The target side is often identified with English and the source side is usually taken to be French.

[2]Crucially, this vocabulary includes a NULL word.

[3]We denote realisations of random variables by the corresponding lower case letters.

in the jth position (f_j) is aligned to under a_1^m.

In IBM model 1 $P(a_1^m)$ is uniform. In IBM model 2, all alignment links a_j are assumed to be independent and follow a categorical distribution. Here, we choose to parametrise this categorical based on the distance between the two words to be aligned, as has been done by Vogel et al. (1996) and Liang et al. (2006). Thus, in our IBM model 2

$$P(a_1^m) = \prod_{j=1}^m P(a_j) = \prod_{j=1}^m P\left(i - \left\lfloor \frac{jl}{m} \right\rfloor\right) \quad (2)$$

where i is the position of the English word that a_j links to and the values l and m stand for the target and source sentence lengths. Notice that there is a target position $i = 0$ for the NULL word. Alignment to this NULL position often causes unusually long alignment jumps.

3 Removing the NULL word

3.1 Model description

Our model consists of an alignment model component (which is either IBM model 1 or 2 without NULL words) and a language model component. It also contains a random variable Z that indicates which component to use. If $Z = 0$ we use the alignment model, if $Z = 1$ we instead use the language model. We generate each z_j conditional on f_{j-1}. By making the outcome z_j depend on f_{j-1}, we allow the model to capture the tendency of individual source words to be part of a collocation, i.e. to be followed by a closely related word. A similar strategy has been employed for topic modelling by Griffiths et al. (2007).

When generating the source side, the model does the following for each source word f_j:

1. Depending on the previous source word f_{j-1}, draw z_j.

2. If $z_j = 1$, generate f_j from f_{j-1} and choose a_j according to $P(a_j)$. Otherwise, if $z_j = 0$, generate f_j from the target side and choose a_j according to the probability that it has under the relevant alignment model *without a target NULL word*.

Our model thus induces a joint probability distribution of the form

$$P(f_1^m, a_1^m, z_1^m | e_1^l) \quad (3)$$

$$= P(a_1^m) \prod_{j=1}^m P(z_j | f_{j-1}) P(f_j | e_{a_j}, f_{j-1}, z_j)$$
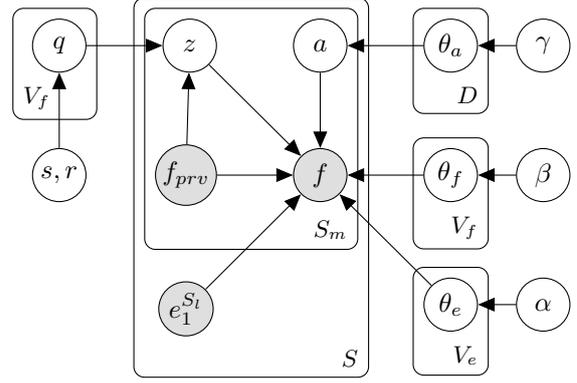

Figure 1: A graphical representation of our model for S sentence pairs. We use $V_{f/e}$ to denote the source/target vocabulary sizes and D to denote the number of possible alignment link configurations. Furthermore, $S_{m/l}$ is the number of source/target words in the current sentence and f_{prv} the source word preceding the one that we currently generate.

where it is crucial to note that there is no E_0 variable, standing for the NULL word, anymore. Therefore, jumps to a NULL position do not need to be modelled. Notice further that the formulation of our model is general enough to be readily extensible to an HMM alignment model (Vogel et al., 1996).

Depending on the value of z_j, F_j is distributed either according to an alignment (4) or a language model[4] (5).

$$P(f_j | e_{a_j}, f_{j-1}, z_j = 0) = P(f_j | e_{a_j}) \quad (4)$$

$$P(f_j | e_{a_j}, f_{j-1}, z_j = 1) = P(f_j | f_{j-1}) \quad (5)$$

3.2 The full model

Our full model is a Bayesian model, meaning that we treat all model parameters as random variables that are drawn from prior distributions. A graphical depiction of the model can be found in Figure 1. We impose Dirichlet priors on the translation (θ_e), language model (θ_f) and distortion parameters (θ_a). This has been done before and improved the standard IBM models.

In order to be able to bias the model against using the language model component (5) too often and instead make it prefer the alignment model component (4), we impose a Beta prior on the Bernoulli distributions over component choices. In effect, the model will only explain a source word with the language model if there is a lot of

---

[4] We use a bigram LM to avoid conditioning Z on longer $(n-1)$-grams.

evidence that this word cannot be translated from the target side. The full model can be summarised as follows:

$$F_j|e, a_j, z_j = 0 \sim Cat(\theta_{e_{aj}}) \qquad \Theta_{e_{aj}} \sim Dir(\alpha)$$
$$F_j|f_{j-1}, z_j = 1 \sim Cat(\theta_{f_{j-1}}) \quad \Theta_{f_{j-1}} \sim Dir(\beta)$$
$$Z_j|f_{j-1} \sim Bernoulli(q) \qquad Q \sim Beta(s, r) .$$

For IBM model 1, A_j is uniformly distributed whereas for model 2 we have

$$A \sim Cat(\theta_a) \qquad \Theta_a \sim Dir(\gamma) .$$

3.3 Inference

We use a Gibbs sampler to perform inference of the alignment and choice variables. Since our priors are conjugate to the model distributions, we integrate over the model parameters, giving us a collapsed sampler[5]. The sampler alternates between sampling alignment links A and component choices Z.

The predictive posterior probabilities for $Z_j = 0$ and $Z_j = 1$ are given in Equations (6) and (7) (up to proportionality). We use $c(\cdot)$ as a (conditional) count function that counts how often an outcome has been observed in a given context. We furthermore use V_f to denote the French (source) vocabulary size. To ease notation, we also introduce the context set $\mathcal{C}_{-X_j}$ which contains the current values of all variables in our model except X_j and the set $\mathcal{H}$ which simply contains all hyperparameters.

$$P\left(Z_j = 0|\mathcal{C}_{-Z_j}, \mathcal{H}\right) \propto \qquad (6)$$
$$(c(z = 0|f_{j-1}) + s) P(a_j) \frac{c(f_j|e_{a_j}, z = 0) + \alpha}{c(e_{a_j}|z = 0) + \alpha V_f}$$

$$P\left(Z_j = 1|\mathcal{C}_{-Z_j}, \mathcal{H}\right) \propto \qquad (7)$$
$$(c(z = 1|f_{j-1}) + r) \frac{c(f_j|f_{j-1}, z = 1) + \beta}{c(z = 1|f_{j-1}) + \beta V_f}$$

When $Z_j = 0$, the predictive probability for alignment link A_j is proportional to Equation (8).

$$P\left(a_j|\mathcal{C}_{-Z_j, -A_j}, Z_j = 0, \mathcal{H}\right) \propto \qquad (8)$$
$$P(a_j) \frac{c(f_j|e_{a_j}, z = 0) + \alpha}{c(e_{a_j}|z = 0) + \alpha V_f}$$

When $Z_j = 1$, it is simply proportional to $P(a_j)$. In the case of IBM model 1, $P(a_j)$ is a constant. For IBM model 2, we use

$$P(a_j) \propto c\left(i - \left\lfloor \frac{jl}{m} \right\rfloor\right) + \gamma .$$

where l and m are the target and source sentence lengths. Notice that target positions start at 1 as we do not use a NULL word.

Notice that a naïve implementation of our sampler is unpractically slow. We therefore augment the sampler with an auxiliary variable (Tanner and Wong, 1987) that uniformly chooses only one possible new assignment per sampled link. The sampling complexity, which would normally be linear in the size of the target sentence, thus becomes constant. In practice this speed up the sampler by several orders of magnitude, making our aligner as fast as Giza++. Unfortunately, this strategy also slightly impairs the mobility of our sampler.

3.4 Decoding

Our samples contain assignments of the A and Z variables. If for a word f_j we have $z_j = 1$, we treat the word as not aligned. We then use maximum marginal decoding (Johnson and Goldwater, 2009) over alignment links to generate final word alignments. This means that we align each source word to the target word it has been aligned to most often in the samples. If the word was unaligned in most samples, we leave it unaligned in the output alignment.

4 Experiments and results

We present translation experiments on English paired with German, French, Czech and Japanese, thereby covering four language families. We compare our model and the Bayesian IBM models 1 and 2 of Mermer et al. (2013) against IBM model 2 as a baseline.

4.1 Experiments

Data We use the news commentary data from the WMT 2014 translation task[6] for German, French and Czech paired with English. We use newstest-2013 as development data and we use the newstest-2014 for testing. We use all available monolingual data from WMT 2014 for language modelling. All data are truecased and sentences

Model	En-De	En-Fr	En-Cs	En-Ja	De-En	Fr-En	Cs-En	Ja-En
Brown et al. (model 2)	14.56	27.16	13.74	25.78	18.12	26.69	18.77	23.29
Mermer et al. (model 1)	-0.09	-0.64	+0.38	-0.13	+0.32	-0.92	+0.66	-0.31
Mermer et al. (model 2)	**+1.07**	-0.17	**+1.76**	+0.39	+1.63	-1.04	**+1.63**	-0.21
This work (model 1)	-0.03	-0.79	-0.42	+0.15	+0.29	-1.49	+0.45	-0.65
This work (model 2)	+0.92	**+1.32**	+1.66	**+1.69**	**+1.73**	**+2.01**	+1.42	**+2.24**
Giza	+0.96	+0.23	+1.58	+2.97	+2.27	+2.26	+1.96	+2.73
fastAlign	+0.88	+0.70	+1.47	+1.97	+2.27	+1.90	+1.86	+2.63

(a) Directional: alignments obtained in target-to-source direction.

Model	En-De	En-Fr	En-Cs	En-Ja	De-En	Fr-En	Cs-En	Ja-En
Brown et al. (model 2)	+0.84	+0.77	+1.14	+3.02	+1.80	+1.77	+1.15	+2.95
Mermer et al. (model 1)	+0.52	+0.80	+1.30	+3.19	+1.51	+1.60	+1.77	+2.44
Mermer et al. (model 2)	+0.63	+0.33	+1.94	+3.00	+2.02	+1.22	**+2.34**	+2.48
This work (model 1)	+0.39	+0.23	+1.31	**+3.33**	+1.61	+0.98	+1.87	+2.56
This work (model 2)	**+1.07**	**+1.47**	**+2.08**	+2.65	**+2.30**	**+2.19**	+2.13	**+3.21**
Giza	+1.59	+0.87	+1.70	+4.24	+2.54	+2.08	+2.36	+3.94
fastAlign	+1.39	+1.23	+1.87	+2.47	+2.44	+2.06	+2.21	+3.58

(b) Symmetrised: alignments obtained in both directions independently and heuristically symmetrised (grow-diag-final-and).

Table 1: Translation results from and into English. Alignments in the top (1a) and bottom (1b) tables were obtained in the target-to-source direction and symmetrised, respectively. Differences are computed with respect to the directional IBM model 2 in its original parameterisation (Brown et al., 1993). The best Bayesian model in each column is boldfaced.

with more than 100 words discarded as is standardly done in SMT.

The Japanese training data consist of 200.000 randomly extracted sentence pairs from the NTCIR-8 Patent Translation Task. The full data are used for language modelling. We use the NTCIR-7 dev sets for tuning and the NTCIR-9 test set for testing.[7]

Training   The maximum likelihood IBM model 2 is initialized with model 1 parameter estimates and trained for 5 EM iterations. Following Mermer and Saraçlar (2011), we initialize the Gibbs samplers of all Bayesian models with the Viterbi alignment from IBM model 1. We run each sampler for 1000 iterations and take a sample after every 25th iteration. We do not use burn-in.[8]

Hyperparameters   All Bayesian models are trained with $\alpha = 0.0001$ and $\beta = 0.0001$ to induce sparse lexical distributions. We also set $s = 1$ and $r = 0.1$ when IBM1 is the alignment component in our model. This has the effect of biasing the model towards using the align-

ment component. For the IBM2 version we even set $r = 0.01$ since IBM2 is a more trustworthy alignment model. For IBM2, we furthermore set $\gamma = 1$ to obtain a flat distortion prior.

Observe that experiments presented here use *the same* fixed hyperparameters for all language pairs. We tried to add another level to our model by imposing Gamma priors on the hyperparameters. The hyperparameters were then inferred using slice sampling after each Gibbs iteration. When run on the German-English and Czech-English data, this strategy increased the posterior probability of the states visited by our sampler but had no effect on BLEU. This may indicate that either the hand-chosen hyperparameters are adequate for the task or that the model generally performs well for a large range of hyperparameters.

Translation   We train Moses systems (Koehn et al., 2007) with 5-gram language models with modified Kneser-Ney-smoothing using KenLM (Heafield et al., 2013) and orientation-based lexicalised reordering. We tune the systems with MERT (Och, 2003) on the dev sets. We report the BLEU score (Papineni et al., 2002) for all models averaged over 5 MERT runs.

4.2   Results

We report the translation results in Tables (1a) and (1b). Results of the full Giza++ pipeline and fastAlign (Dyer et al., 2013) are reported as a com-

---

[7]The Japanese data was provided to us by a colleague with the pre-processing steps already performed, with sentences shortened to at most 40 words. Our algorithm can handle sentences of any length and there is actually no need to restrict the sentence lengths.

[8]Burn-in is simply a heuristic that is not guaranteed to improve the samples in any way. See `http://users.stat.umn.edu/~geyer/mcmc/burn.html` for further details.

parison standard. All symmetrised results were obtained using the `grow-diag-final-and` heuristic.

Using IBM2 as an alignment component, our model mostly outperforms the standard IBM models and their Bayesian variants. Importantly, the improvement that our model 2 achieves over its model 1 variant is much larger than the difference between the corresponding models of Mermer et al. (2013). This indicates that our model makes better use of the distortion distribution that is not altered by NULL alignments. We also observe that our model gains relatively little from symmetrisation, likely because it is a very strong model already. It is interesting that although our model 2 does not use fertility parameters or dependencies between alignment links, it often approaches the performance of Giza which does use these features. Moreover, it also approaches the performance of fastAlign which does not use fertility nor dependencies between alignment links, but has a stronger inductive bias with respect to distortion.

5 Discussion and future work

We have presented an IBM-style word alignment model that does not need to hypothesise a NULL word as it explains untranslatable source words by grouping them with translated words. This also leads to a cleaner handling of distortion probabilities.

In our present work, we have only considered IBM models 1 and 2. As we have mentioned already, our model can easily be extended with the HMM alignment model. We are currently exploring this possibility. Our models also allow symmetrisation (Liang et al., 2006) of *all* translation and distortion parameters where before the NULL distortion parameters had to be fixed. We therefore plan to extend them towards model-based instead of heuristic alignment symmetrisation.

A limitation of our model is that it is only capable of modelling left-to-right linear dependencies in the source language. In languages like German or English, however, where an adjective or determiner is selected by the following noun, this may not be appropriate to model selection biases amongst neighbouring words. An interesting extension to our model is thus to add more structure to it such that it will be able to capture more complex source side dependencies.

Another concern is the inference in our model.

Using the auxiliary variable sampler, inference becomes very fast but may sacrifice performance. This is why we are interested in improving the inference method, e.g. by using a more mobile sampler or by employing a variational Bayes algorithm.

The software used in our experiments can be downloaded from `https://github.com/philschulz/Aligner`.

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*, 19:263–311.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *In Proceedings of NAACL*.

Thomas L. Griffiths, Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological review*, 114(2):211–244.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. NAACL '09, pages 317–325.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 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*, ACL '07, pages 177–180.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. HLT-NAACL '06, pages 104–111. 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*, pages 182–187.

Coşkun Mermer, Murat Saraçlar, and Ruhi Sarikaya. 2013. Improving statistical machine translation using Bayesian word alignment and Gibbs sampling. *IEEE Transactions on Audio, Speech & Language Processing*, 21(5):1090–1101.

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

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. ACL '03, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Martin A. Tanner and Wing Hung Wong. 1987. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–550.

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.