Evidence-Based Word Alignment

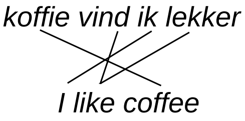
Jörg Tiedemann
Alpha-Informatica, Rijksuniversiteit Groningen,
Groningen, The Netherland,
Department of Linguistics and Philology
Uppsala University, Uppsala/Sweden
j.tiedemann@rug.nl

Abstract

In this paper we describe word alignment experiments using an approach based on a disjunctive combination of alignment evidence. A wide range of statistical, orthographic and positional clues can be combined in this way. Their weights can easily be learned from small amounts of hand-aligned training data. We can show that this "evidence-based" approach can be used to improve the baseline of statistical alignment and also outperforms a discriminative approach based on a maximum entropy classifier.

1 Introduction

Automatic word alignment has received a lot of attention mainly due to the intensive research on statistical machine translation. However, parallel corpora and word alignment are not only useful in that field but may be applied to various tasks such as computer aided language learning (see for example [15]) and bilingual terminology extraction (for example [8, 10]). The automatic alignment of corresponding words in translated sentences is a challenging task even for small translation units as the following Dutch-English example tries to illustrate.



Word alignment approaches have to consider crossing links and multiple links per word in both directions. Discontinuous units may also be aligned to corresponding parts in the other language as shown in the example above (*vind...lekker - like*). Various other issues due to translation divergency make word alignment a much more challenging task than, for instance, sentence alignment. Generative statistical models for word alignment usually have problems with non-monotonic alignments and many-to-many links. In the literature several attempts are described in which additional features are integrated besides the distribution of surface words to overcome these difficulties. In recent years various discriminative approaches have been proposed for this task [18, 9, 13, 14, 11, 1]. They require word-aligned training data for estimating model parameters in contrast to the traditional IBM alignment models that work on raw parallel (sentence aligned) corpora [2, 16]. However, previous studies have shown that only a small number of training examples (around 100 word-aligned sentence pairs) are sufficient to train discriminative models that outperform the traditional generative models.

In this paper we present another supervised alignment approach based on association clues trained on small amounts of word-aligned data. This approach differs from previous discriminative ones in the way the evidence for alignment is combined as we will explain in the following section.

2 Evidence-based alignment

The evidence-based alignment approach is based on the techniques proposed by [19]. This approach applies the notion of link evidence derived from word alignment clues. An *alignment clue* $C(r_k|s_i,t_j)$ is used as a probabilistic score indicating a (positive) relation r_k between two items s_i, t_j in their contexts. *Link evidence* $E(a,r_k|s_i,t_j)$ is then defined as the product of this score and the likelihood of establishing a link given the relation indicated by that clue:

$$E(a,r_k|s_i,t_j) = C(r_k|s_i,t_j)P(a|r_k)$$

Various types of alignment clues can be found in parallel data. Association scores and similarity measures can be used to assign their values. For example, the relation of "cognateness" may be indicated by string similarity measures. Translational equivalence relations can be indicated by co-occurrence measures. For the estimation of these scores, no word-aligned training data is required. However, for the estimation of the likelihood values we need training data as we will explain below. They can be seen as weights that correspond to the quality of clues in predicting links properly. Note that we can also use binary clues. Their influence on alignment decisions is determined by the alignment likelihood values only.

So far, this model is not so much different from previous discriminative alignment approaches in which weighted features are used in a classification approach (see, for example, [18], [13]). However, we use our weighted features as individual pieces of evidence that are combined in a disjunctive way, i.e. the overall alignment evidence for two given items is defined as the union of individual evidence scores:

NLP Methods and Corpora in Translation, Lexicography, and Language Learning 2009 – Borovets, Bulgaria, pages 28–32

$$E(a|s_i, t_j) = E(a, r_1 \vee r_2 \vee .. \vee r_k|s_i, t_j)$$

Note that alignment clues are not mutually exclusive and, therefore, we need to subtract the overlapping parts when computing the union. Using the addition rule of probabilities we obtain, for example, for two clues:

$$\begin{aligned} E(a, r_1 \vee r_2|s_i, t_j) &= E(a, r_1|s_i, t_j) + E(a, r_2|s_i, t_j) - \\ &\quad E(a, r_1 \wedge r_2|s_i, t_j) \end{aligned}$$

Hence we combine individual pieces of evidence in a non-linear way. Figure 1 tries to illustrate such a combination for two given cases.

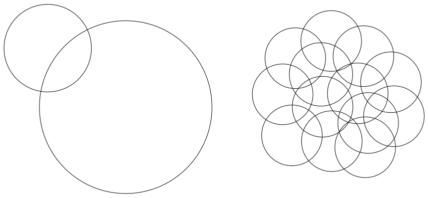


Fig. 1: *Combining alignment evidence. The size of the circles refers to the strength of the evidence given.*

The intuition behind this way of combining features is to give stronger pieces of evidence a larger influence on alignment decisions. As illustrated in figure 1 strong evidence is hard to overrule even by many other weaker clues. A few solid clues are sufficient just like a reliable witness in a murder case overrules all kinds of other weaker pieces of evidence indicating a different scenario. A consequence of our model is that alignment evidence with a value of 1.0 can not be outranked by any other combination of evidence. However, this is not as strange as it sounds if we consider that evidence giving 100% certainty should always be trusted. These cases should be very exceptional, though.

One difficulty arises in our approach: We need to estimate the overlapping parts of our collected evidence. For simplicity, we assume that all relation types are independent of each other (but not mutually exclusive) and, therefore, we can define the joint probability score of the overlapping part as $E(a, r_1 \wedge r_2|s_i, t_j) = E(a, r_1|s_i, t_j)E(a, r_2|s_i, t_j)$. The combination of independent evidence is illustrated in figure 2.

Altogether this model is similar to noisy OR-gates frequently used in belief networks in which causes are modeled to act independently of others to produce a determined effect [17]. Certainly, the independence assumption is violated in most cases. However, we will see in our experiments that this simplification still works well for alignment purposes. Note, that complex features can easily be constructed in order to reduce the impact of this violation on alignment performance.

2.1 Parameter estimation

As we have said earlier, the only parameters that need to be estimated from word-aligned training data are the alignment likelihoods used as weights for individual clues. Due to our independence assumption, we can do this by evaluating each individual clue on the training data. For this, we need to find out to what extent the indicated relations can be used to establish links in our data. Hence, we use each observed clue as a binary classifier and simply count the number of correctly predicted links using that clue (as usual a value above 0.5 is used to predict a positive example). This means that we use the precision of each individual clue on some training data to estimate alignment likelihoods. Intuitively, this seems to fit our approach in which we prefer high precision features as described earlier.

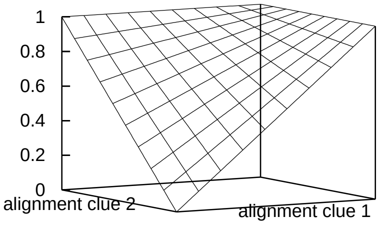


Fig. 2: *The combination of two independent alignment clues.*

Thus, training is extremely simple. The most expensive computation is actually the extraction of features used as alignment clues (see section 3.1 for details). The overhead of training is tiny and can be done in linear time. Note that this model only covers the classification of individual items. For the actual word alignment we need to apply a search algorithm that optimizes the alignment of all words according to the evidence found for individual pairs. This will briefly be discussed in the following section.

2.2 Link dependencies & alignment search

The problem of alignment search has been discussed in related studies on discriminative word alignment. The problem is that the dependency between links has to be considered when creating word alignments. Several approaches have been proposed that either include link dependencies directly in the underlying model [14, 1] or that include contextual features that implicitly add these dependencies [18]. Depending on the model optimal alignments can be found [18, 9, 1] or greedy search heuristics are applied [11, 14].

We will use the second approach and model link dependencies in terms of contextual features. We believe that this gives us more flexibility when defining contextual dependencies and also keeps the model very simple with regards to training. For the alignment search problem we could still apply a model that allows optimal decoding, for example, the approach proposed in [18]. However, we will stick to a simple greedy search heuristics, similar to the "refined" heuristics defined in [16], that is known to produce good results for example for the symmetrization of directional statistical word alignment. The advantages of this approach is that it is fast and easy to apply, it allows n:m alignments, and it makes our results comparable to the statistical alignments that include symmetrization.

3 Experiments

For our experiments we will use well-known data sets that have been used before for word alignment experiments. Most related work on supervised alignment models reports results on the French-English data set from the shared task at WPT03 [12] derived from the parallel Canadian Hansards corpus. This data set caused a lot of discussion especially because of the flaws in evaluation measures used for word alignment experiments [5]. Therefore, we will apply this set for training purposes only (447 aligned sentences with 4,038 sure (S) links and 13,400 (P) possible links) and stick to another set for evaluation [4]. This set includes English-French word alignment data for 100 sentences from the Europarl corpus [6] with a much smaller number of possible links (437 compared to 1,009 sure links) which hopefully leads to more reliable results.

Some of the alignment clues require large parallel corpora for estimating reliable feature values (for example co-occurrence measures). For training we use the Canadian Hansards as provided for the WPT03 workshop and for evaluation these values are taken from the Europarl corpus.

For evaluation we use the standard measures used in related research:

$$
\begin{aligned}
Prec(A,P) &= \frac{|P \cap A|}{|A|} \\
Rec(A,S) &= \frac{|S \cap A|}{|S|} \\
AER(A,P,S) &= 1 - \frac{|P \cap A| + |S \cap A|}{|S| + |A|} \\
F(A,P,S,\alpha) &= 1/\left(\frac{\alpha}{Prec(A,P)} + \frac{(1-\alpha)}{Rec(A,S)}\right)
\end{aligned}
$$

For the F-measure we give balanced values and also unbalanced F-values with $\alpha = 0.4$. The latter is supposed to show a better correlation with BLEU scores. However, we did not perform any tests with statistical MT using our alignment techniques to verify this for the data we have used.

For comparison we use the IBM model 4 alignments and the intersection and grow-diag-final-and symmetrizaton heuristics as implemented in the Moses toolkit [7]. We also compare our results with a discriminative alignment approach using the same alignment search algorithm, the same features and a global maximum entropy classifier [3] trained on the same training data (using default settings of the megam toolkit).

3.1 Alignment features

A wide variety of features can be used to collect alignment evidence. We use, among others, similar features as described in [18]. In particular, we use the Dice coefficient for measuring co-occurrence, the longest common subsequence ratio (LCSR) for string similarity, and other orthographic features such as identical string matching, prefix matching and suffix matching. We use the positional distance measures as described in [18] but turn them into similarity measures. We also model contextual dependencies by including Dice values for the next and the previous words. We use rank similarity derived from word type frequency tables and we use POS labels for the current words and their contexts. Furthermore, we also use evidence derived from the IBM models for statistical alignment. We use lexical probabilities, the directional alignment predictions of Model 4 and the links from the intersection heuristics of Model 4 alignments (produced by Moses/GIZA++; henceforth referred to as *Moses features*). As expected, these features are very powerful as we will see in our experimental results. A small sample from a feature file extracted from a sentence aligned parallel corpus is shown in figure 3.

```
possim 1 mosessrc2trg 1 mosestrg2src 1 pos_NN_VER:pper 1
possim 0.75 pos_NN_PRP 1 lcsr 0.05
possim 0.5 pos_NN_DET:ART 1
possim 0.25 pos_NN_NOM 1 lcsr 0.0714285714285714
possim 0.75 pos_IN_VER:pper 1
possim 1 mosessrc2trg 1 mosestrg2src 1 pos_IN_PRP 1
possim 0.75 pos_IN_DET:ART 1
possim 0.5 lcsr 0.142857142857143 pos_IN_NOM 1
possim 0.5 pos_DT_VER:pper 1 lcsr 0.142857142857143
....
```

Fig. 3: *A short example of link features extracted for each possible word combination in aligned sentences.* possim = *relative position similarity,* lcsr = *string similarity measure,* pos_* = *POS label pairs*

As we can see, some features are in fact binary (as discussed earlier) even though we use them in the same way as the real-valued features. For example, statistical alignment features derived from GIZA++/Moses (mosessrc2trg, mosestrg2src) are set to 1 if the corresponding word pair has been linked in the statistical Viterbi alignment. Other feature types are used as templates and will be instantiated by various values. For example, the POS label feature template adds a feature to each word pair made out of the labels attached to the corresponding words. Again, these features are used as binary flags as we can see in the example in figure 3.

Note that complex features can easily be created. We consider several combinations, for example the product of Dice scores and positional similarity scores. Contextual features can also be combined with any other feature. Complex features are especially useful in cases where the independence assumption is heavily violated. They are also useful to improve linear classification in cases where the correlation between certain features is non-linear.

3.2 Results

Our results are summarized in table 1.
As we can see, we cannot outperform the strong baselines without the features derived from statistical word

baselines	Rec	Prec	$F_{0.5}$	$F_{0.4}$	AER
intersection	72.1	**95.2**	82.0	79.8	17.5
grow-diag-final	**84.5**	78.7	81.5	82.1	18.8
best setting without Moses features					
MaxEnt	71.5	73.0	72.2	72.1	27.7
Clues	68.9	70.1	69.5	69.3	30.5
best setting with all features					
MaxEnt	82.3	84.4	83.3	83.1	16.6
Clues	82.6	85.4	**84.0**	**83.7**	**15.9**

Table 1: *Overview of results: Statistical word alignment derived from GIZA++/Moses (intersection/grow-diag-final), discriminative word alignment using a maximum entropy classifier (MaxEnt), and the evidence-based alignment (Clues).*

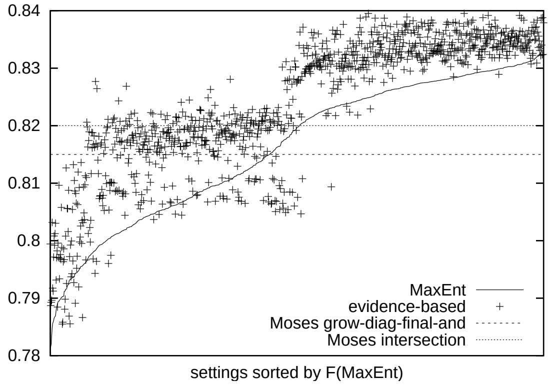


Fig. 4: *A comparison of $F_{0.5}$ scores obtained with settings that include statistical word alignment features.*

alignment. However, adding these features makes it possible to improve alignment results according to AER and F-scores. We can also observe that the MaxEnt classifier is better in handling the dependencies between non-Moses features. The scores are in general slightly above the corresponding clue-based scores. However, including the strong Moses features, our approach outperforms the maximum entropy classifier and yields the overall best result. As expected, our approach seems to handle the combination of strong evidence and weak clues well. It learns to trust these strong clues and still includes additional evidence from other alignment clues. Figure 4 illustrates this by plotting the results ($F_{0.5}$ scores) for settings that include Moses features for both, the MaxEnt classifier approach and the evidence-based approach. The settings are sorted by the F-scores obtained by the MaxEnt classifier approach (solid line) along the x-axis. Corresponding F-scores obtained by the evidence-based approach using the same feature set and alignment search algorithm are plotted as points in the graph. As we can see in most cases, our simple evidence-based approach yields similar or better results than the MaxEnt approach. We can also see that both discriminative approaches improve the baseline scores obtained by the generative statistical word alignment after symmetrization (dashed and dotted lines in the graph).
The best result is obtained with the following features: Dice for the current word pair and the previous one, positional similarity, POS labels, rank similarity, lexical probabilities and link predictions of the two IBM 4 Viterbi alignments. Surprisingly, the orthographic features (LCSR etc) do not perform well at all. Some example weights learned from the training data using the alignment prediction precision are shown in table 2.

feature	**prediction precision**
dice	0.8120830711139080
prevdice	0.8228682170542640
possim	0.2656349270994540
ranksim	0.4259383603034980
lexe2f	0.9634980007738940
lexf2e	0.9348459880846750
lexe2f*lexf2e	0.9900965585540980
mosessrc2trg	0.9601313748745550
mosestrg2src	0.9514683153013910
pos_VBZ_VER:pres	0.7395577395577400
pos_NNS_NOM	0.5319049836981840
pos_)_PUN	0.7142857142857140
pos_VV_ADJ	0.0393013100436681
pos_NNS_VER:pper	0.0593607305936073

Table 2: *Examples of weights learned from prediction precision of individual clues.*

We can see that features derived from statistical word alignment have a high precision and, therefore, the evidence-based alignment approach trusts them a lot. This includes the lexical probabilities taken from the translation model as estimated by Moses. Especially their product is very accurate which is maybe not so surprising considering that this score will be very low for most word pairs and, therefore, only a few links will be predicted by this feature. Co-occurrence measures score also very high. Note that the Dice score of the previous words (*prevdice*) also seems to be very useful for alignment prediction. On the other hand, positional similarity (*possim*) is a rather weak clue according to the precision computed. However, it is still very useful to make alignment decisions in cases where other evidence is missing or not discriminative enough. Frequency rank similarity (*ranksim*) is also surprisingly strong. This is probably due to the similarity between English and French especially in terms of inflectional complexity. Finally, we can see examples of the weights estimated for binary features such as POS label pairs. Here, we use a threshold of a minimum of five occurrences to obtain reliable estimates. We can see that some of them are very useful in predicting links whereas others are very low. Probably, negative clues could be useful as well, for example, using POS labels that indicate a preference for not linking the corresponding items. However, for this the alignment model has to be adjusted to account for such clues as well.

Finally, we also include the plot of alignment error rates for settings that include Moses features (see figure 5).

We can see that the curve follows the same trend as we have seen for the F-scores in figure 4. Most of the evidence-based alignment results are below the corresponding runs with a linear classifier. Again, we also outperform the generative alignment approach,

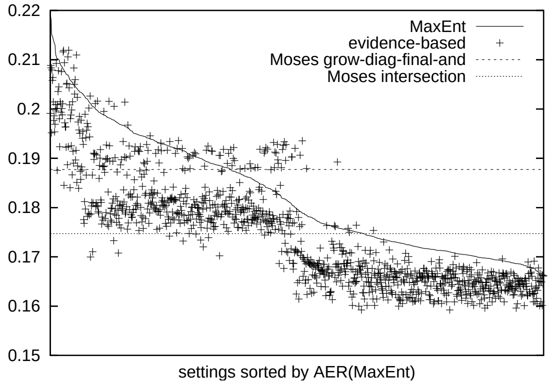


Fig. 5: *A comparison of AER scores obtained with settings that include statistical word alignment features.*

however, only when using features derived from these alignments.

4 Conclusions

In this paper we describe our experiments with evidence-based word alignment. Features (alignment clues) in this approach are combined in a non-linear way in contrast to related discriminative word alignment approaches that usually apply linear classification techniques in the underlying model. We have shown that this kind of combination can be beneficial when comparing to a straightforward linear classification approach especially when high precision features are applied. Another advantage is the simplicity of training feature weights using individual link prediction precision. However, this requires the assumption that each feature can be used as an independent base classifier. This assumption is often violated which can be seen in the degrading performance of the evidence-based approach when applying it in connection with weaker clues. However, the approach seems to work well in terms of picking up strong clues and learns to trust them appropriately. It remains to be investigated to what extend this approach can be used to improve subsequent applications such as machine translation or bilingual terminology extraction. Furthermore, it should be embedded in a proper structural prediction framework in which output space dependencies (between predicted links in a sentence pair) are modeled explicitly. This will boost the performance even further as it has been shown for other discriminative word alignment approaches.

References

[1] P. Blunsom and T. Cohn. Discriminative word alignment with conditional random fields. In *Proceedings of ACL*, Sydney, Australia, 2006.

[2] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The Mathematics of Statistcal Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June 1993.

[3] H. Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, 2004.

[4] J. de Almeida Varelas Graa, J. P. Pardal, L. Coheur, and D. A. Caseiro. Building a golden collection of parallel multi-language word alignment. In *Proceedings of LREC*, 2008.

[5] A. Fraser and D. Marcu. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3):293–303, 2007.

[6] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*, 2005.

[7] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, Prague, Czech Republic, 2007.

[8] J. Kupiec. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 17–22, Morristown, NJ, USA, 1993. Association for Computational Linguistics.

[9] S. Lacoste-Julien, B. Taskar, D. Klein, and M. Jordan. Word alignment via quadratic assignment. In *Proceedings of HLT-NAACL*, New York, 2006.

[10] E. Lefever, L. Macken, and V. Hoste. Language-independent bilingual terminology extraction from a multilingual parallel corpus. In *EACL*, pages 496–504. The Association for Computer Linguistics, 2009.

[11] Y. Liu, Q. Liu, and S. Lin. Log-linear models for word alignment. In *Proceedings of ACL*, Ann Arbor, Michigan, 2005.

[12] R. Mihalcea and T. Pedersen. An evaluation exercise for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, 2003.

[13] R. C. Moore. A discriminative framework for bilingual word alignment. In *Proceedings of HLT/EMNLP*, Vancouver, Canada, 2005.

[14] R. C. Moore, W. tau Yih, and A. Bode. Improved discriminative bilingual word alignment. In *Proceedings of ACL*, 2006.

[15] J. Nerbonne. Parallel texts in computer-assisted language learning. In J. Veronis, editor, *Parallel Text Processing*, pages 354–369. Kluwer, Dordrecht and Boston, 2000.

[16] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[17] J. Pearl. Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288, 1986.

[18] B. Taskar, S. Lacoste-Julien, and D. Klein. A discriminative matching approach to word alignment. In *Proceedings of HLT/EMNLP*, Vancouver, Canada, 2005.

[19] J. Tiedemann. Combining clues for word alignment. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 339–346, Budapest, Hungary, April 2003.