Discriminative Word Alignment with Conditional Random Fields

Phil Blunsom and **Trevor Cohn**
Department of Software Engineering and Computer Science
University of Melbourne
{pcbl,tacohn}@csse.unimelb.edu.au

Abstract

In this paper we present a novel approach for inducing word alignments from sentence aligned data. We use a Conditional Random Field (CRF), a discriminative model, which is estimated on a small supervised training set. The CRF is conditioned on both the source and target texts, and thus allows for the use of arbitrary and overlapping features over these data. Moreover, the CRF has efficient training and decoding processes which both find globally optimal solutions.

We apply this alignment model to both French-English and Romanian-English language pairs. We show how a large number of highly predictive features can be easily incorporated into the CRF, and demonstrate that even with only a few hundred word-aligned training sentences, our model improves over the current state-of-the-art with alignment error rates of 5.29 and 25.8 for the two tasks respectively.

1 Introduction

Modern phrase based statistical machine translation (SMT) systems usually break the translation task into two phases. The first phase induces word alignments over a sentence-aligned bilingual corpus, and the second phase uses statistics over these predicted word alignments to decode (translate) novel sentences. This paper deals with the first of these tasks: word alignment.

Most current SMT systems (Och and Ney, 2004; Koehn et al., 2003) use a generative model for word alignment such as the freely available GIZA++ (Och and Ney, 2003), an implementation of the IBM alignment models (Brown et al., 1993). These models treat word alignment as a hidden process, and maximise the probability of the observed $(\mathbf{e}, \mathbf{f})$ sentence pairs[1] using the expectation maximisation (EM) algorithm. After the maximisation process is complete, the word alignments are set to maximum posterior predictions of the model.

While GIZA++ gives good results when trained on large sentence aligned corpora, its generative models have a number of limitations. Firstly, they impose strong independence assumptions between features, making it very difficult to incorporate non-independent features over the sentence pairs. For instance, as well as detecting that a source word is aligned to a given target word, we would also like to encode syntactic and lexical features of the word pair, such as their parts-of-speech, affixes, lemmas, etc. Features such as these would allow for more effective use of sparse data and result in a model which is more robust in the presence of unseen words. Adding these non-independent features to a generative model requires that the features' inter-dependence be modelled explicitly, which often complicates the model (eg. Toutanova et al. (2002)). Secondly, the later IBM models, such as Model 4, have to resort to heuristic search techniques to approximate forward-backward and Viterbi inference, which sacrifice optimality for tractability.

This paper presents an alternative discriminative method for word alignment. We use a conditional random field (CRF) sequence model, which allows for globally optimal training and decoding (Lafferty et al., 2001). The inference algo-

[1] We adopt the standard notation of $\mathbf{e}$ and $\mathbf{f}$ to denote the target (English) and source (foreign) sentences, respectively.

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 65–72,
Sydney, July 2006.

rithms are tractable and efficient, thereby avoiding the need for heuristics. The CRF is conditioned on both the source and target sentences, and therefore supports large sets of diverse and overlapping features. Furthermore, the model allows regularisation using a prior over the parameters, a very effective and simple method for limiting over-fitting. We use a similar graphical structure to the directed hidden Markov model (HMM) from GIZA++ (Och and Ney, 2003). This models one-to-many alignments, where each target word is aligned with zero or more source words. Many-to-many alignments are recoverable using the standard techniques for superimposing predicted alignments in both translation directions.

The paper is structured as follows. Section 2 presents CRFs for word alignment, describing their form and their inference techniques. The features of our model are presented in Section 3, and experimental results for word aligning both French-English and Romanian-English sentences are given in Section 4. Section 5 presents related work, and we describe future work in Section 6. Finally, we conclude in Section 7.

2 Conditional random fields

CRFs are undirected graphical models which define a conditional distribution over a label sequence given an observation sequence. We use a CRF to model many-to-one word alignments, where each source word is aligned with zero or one target words, and therefore each target word can be aligned with many source words. Each source word is labelled with the index of its aligned target, or the special value *null*, denoting no alignment. An example word alignment is shown in Figure 1, where the hollow squares and circles indicate the correct alignments. In this example the French words *une* and *autre* would both be assigned the index 24 – for the English word *another* – when French is the source language. When the source language is English, *another* could be assigned either index 25 or 26; in these ambiguous situations we take the first index.

The joint probability density of the alignment, $\mathbf{a}$ (a vector of target indices), conditioned on the source and target sentences, $\mathbf{e}$ and $\mathbf{f}$, is given by:

$$p_\Lambda(\mathbf{a}|\mathbf{e},\mathbf{f}) = \frac{\exp \sum_t \sum_k \lambda_k h_k(t, a_{t-1}, a_t, \mathbf{e}, \mathbf{f})}{Z_\Lambda(\mathbf{e},\mathbf{f})} \quad (1)$$

where we make a first order Markov assumption

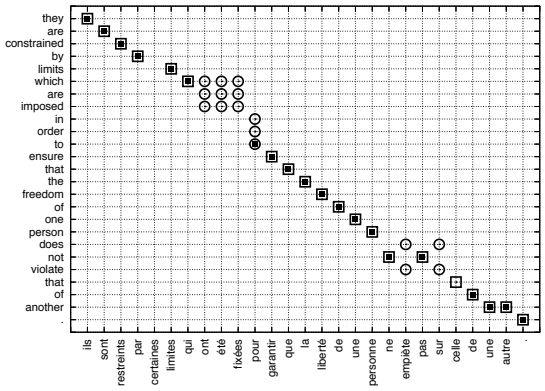


Figure 1. A word-aligned example from the Canadian Hansards test set. Hollow squares represent gold standard sure alignments, circles are gold possible alignments, and filled squares are predicted alignments.

over the alignment sequence. Here t ranges over the indices of the source sentence ($\mathbf{f}$), k ranges over the model's features, and $\Lambda = \{\lambda_k\}$ are the model parameters (weights for their corresponding features). The feature functions h_k are predefined real-valued functions over the source and target sentences coupled with the alignment labels over adjacent times (source sentence locations), t. These feature functions are unconstrained, and may represent overlapping and non-independent features of the data. The distribution is globally normalised by the partition function, $Z_\Lambda(\mathbf{e},\mathbf{f})$, which sums out the numerator in (1) for every possible alignment:

$$Z_\Lambda(\mathbf{e},\mathbf{f}) = \sum_{\mathbf{a}} \exp \sum_t \sum_k \lambda_k h_k(t, a_{t-1}, a_t, \mathbf{e}, \mathbf{f})$$

We use a linear chain CRF, which is encoded in the feature functions of (1).

The parameters of the CRF are usually estimated from a fully observed training sample (word aligned), by maximising the likelihood of these data. I.e. $\Lambda^{ML} = \arg\max_\Lambda p_\Lambda(\mathcal{D})$, where $\mathcal{D} = \{(\mathbf{a},\mathbf{e},\mathbf{f})\}$ are the training data. Because maximum likelihood estimators for log-linear models have a tendency to overfit the training sample (Chen and Rosenfeld, 1999), we define a prior distribution over the model parameters and derive a maximum *a posteriori* (MAP) estimate, $\Lambda^{MAP} = \arg\max_\Lambda p_\Lambda(\mathcal{D})p(\Lambda)$. We use a zero-mean Gaussian prior, with the probability density function $p_0(\lambda_k) \propto \exp\left(-\frac{\lambda_k^2}{2\sigma_k^2}\right)$. This yields a log-likelihood objective function of:

$$\mathcal{L} = \sum_{(\mathbf{a},\mathbf{e},\mathbf{f})\in\mathcal{D}} \log p_\Lambda(\mathbf{a}|\mathbf{e},\mathbf{f}) + \sum_k \log p_0(\lambda_k)$$

$$= \sum_{(\mathbf{a},\mathbf{e},\mathbf{f})\in\mathcal{D}} \sum_t \sum_k \lambda_k h_k(t, a_{t-1}, a_t, \mathbf{e}, \mathbf{f})$$
$$- \log Z_\Lambda(\mathbf{e}, \mathbf{f}) - \sum_k \frac{\lambda_k^2}{2\sigma_k^2} + \textit{const.} \quad (2)$$

In order to train the model, we maximize (2). While the log-likelihood cannot be maximised for the parameters, Λ, in closed form, it is a convex function, and thus we resort to numerical optimisation to find the globally optimal parameters. We use L-BFGS, an iterative quasi-Newton optimisation method, which performs well for training log-linear models (Malouf, 2002; Sha and Pereira, 2003). Each L-BFGS iteration requires the objective value and its gradient with respect to the model parameters. These are calculated using forward-backward inference, which yields the partition function, $Z_\Lambda(\mathbf{e}, \mathbf{f})$, required for the log-likelihood, and the pair-wise marginals, $p_\Lambda(a_{t-1}, a_t|\mathbf{e}, \mathbf{f})$, required for its derivatives.

The Viterbi algorithm is used to find the maximum posterior probability alignment for test sentences, $\mathbf{a}^* = \arg\max_{\mathbf{a}} p_\Lambda(\mathbf{a}|\mathbf{e}, \mathbf{f})$. Both the forward-backward and Viterbi algorithm are dynamic programs which make use of the Markov assumption to calculate efficiently the exact marginal distributions.

3 The alignment model

Before we can apply our CRF alignment model, we must first specify the feature set – the functions h_k in (1). Typically CRFs use binary indicator functions as features; these functions are only active when the observations meet some criteria and the label a_t (or label pair, (a_{t-1}, a_t)) matches a pre-specified label (pair). However, in our model the labellings are word indices in the target sentence and cannot be compared readily to labellings at other sites in the same sentence, or in other sentences with a different length. Such naive features would only be active for one labelling, therefore this model would suffer from serious sparse data problems.

We instead define features which are functions of the source-target word match implied by a labelling, rather than the labelling itself. For example, from the sentence in Figure 1 for the labelling of $f_{24} = de$ with $a_{24} = 16$ (for $e_{16} = of$) we might detect the following feature:

$$h(t, a_{t-1}, a_t, \mathbf{f}, \mathbf{e}) = \begin{cases} 1, & \text{if } e_{a_t} = \textit{`of'} \wedge f_t = \textit{`de'} \\ 0, & \text{otherwise} \end{cases}$$

Note that it is the target word indexed by a_t, rather than the index itself, which determines whether the feature is active, and thus the sparsity of the index label set is not an issue.

3.1 Features

One of the main advantages of using a conditional model is the ability to explore a diverse range of features engineered for a specific task. In our CRF model we employ two main types of features: those defined on a candidate aligned pair of words; and Markov features defined on the alignment sequence predicted by the model.

Dice and Model 1 As we have access to only a small amount of word aligned data we wish to be able to incorporate information about word association from any sentence aligned data available. A common measure of word association is the Dice coefficient (Dice, 1945):

$$Dice(e, f) = \frac{2 \times C_{EF}(e, f)}{C_E(e) + C_F(e)}$$

where C_E and C_F are counts of the occurrences of the words e and f in the corpus, while C_{EF} is their co-occurrence count. We treat these Dice values as *translation scores*: a high (low) value incidates that the word pair is a good (poor) candidate translation.

However, the Dice score often over-estimates the association between common words. For instance, the words *the* and *of* both score highly when combined with either *le* or *de*, simply because these common words frequently co-occur. The GIZA++ models can be used to provide better translation scores, as they enforce competition for alignment beween the words. For this reason, we used the translation probability distribution from Model 1 in addition to the DICE scores. Model 1 is a simple position independent model which can be trained quickly and is often used to bootstrap parameters for more complex models. It models the conditional probability distribution:

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) = \frac{p(|\mathbf{f}||\mathbf{e}|)}{(|\mathbf{e}|+1)^{|\mathbf{f}|}} \times \prod_{t=1}^{|\mathbf{f}|} p(f_t|e_{a_t})$$

where $p(f|e)$ are the word translation probabilities.

We use both the Dice value and the Model 1 translation probability as real-valued features for each candidate pair, as well as a normalised score

over all possible candidate alignments for each target word. We derive a feature from both the Dice and Model 1 translation scores to allow competition between sources words for a particular target alignment. This feature indicates whether a given alignment has the highest translation score of all the candidate alignments for a given *target* word. For the example in Figure 1, the words *la, de* and *une* all receive a high translation score when paired with *the*. To discourage all of these French words from aligning with *the*, the best of these (*la*) is flagged as the best candidate. This allows for competition between source words which would otherwise not occur.

Orthographic features Features based on string overlap allow our model to recognise cognates and orthographically similar translation pairs, which are particularly common between European languages. Here we employ a number of string matching features inspired by similar features in Taskar et al. (2005). We use an indicator feature for every possible source-target word pair in the training data. In addition, we include indicator features for an exact string match, both with and without vowels, and the edit-distance between the source and target words as a real-valued feature. We also used indicator features to test for matching prefixes and suffixes of length three. As stated earlier, the Dice translation score often erroneously rewards alignments with common words. In order to address this problem, we include the absolute difference in word length as a real-valued feature and an indicator feature testing whether both words are shorter than 4 characters. Together these features allow the model to disprefer alignments between words with very different lengths – i.e. aligning rare (long) words with frequent (short) determiners, verbs etc.

POS tags Part-of-speech tags are an effective method for addressing the sparsity of the lexical features. Observe in Figure 2 that the noun-adjective pair *Canadian experts* aligns with the adjective-noun pair *spécialistes canadiens*: the alignment exactly matches the parts-of-speech. Access to the words' POS tags will allow simple modelling of such effects. POS can also be useful for less closely related language pairs, such as English and Japanese where English determiners are never aligned; nor are Japanese case markers.

For our French-English language pair we POS tagged the source and target sentences with TreeTagger.[2] We created indicator features over the POS tags of each candidate source and target word pair, as well as over the source word and target POS (and vice-versa). As we didn't have access to a Romanian POS tagger, these features were not used for the Romanian-English language pair.

Bilingual dictionary Dictionaries are another source of information for word alignment. We use a single indicator feature which detects when the source and target words appear in an entry of the dictionary. For the English-French dictionary we used FreeDict,[3] which contains 8,799 English words. For Romanian-English we used a dictionary compiled by Rada Mihalcea,[4] which contains approximately 38,000 entries.

Markov features Features defined over adjacent aligment labels allow our model to reflect the tendency for monotonic alignments between European languages. We define a real-valued alignment index jump width feature:

$$jump\_width(t-1,t) = abs(a_t - a_{t-1} - 1)$$

this feature has a value of 0 if the alignment labels follow the downward sloping diagonal, and is positive otherwise. This differs from the GIZA++ hidden Markov model which has individual parameters for each different jump width (Och and Ney, 2003; Vogel et al., 1996): we found a single feature (and thus parameter) to be more effective.

We also defined three indicator features over **null** transitions to allow the modelling of the probability of transition between, to and from **null** labels.

Relative sentence postion A feature for the absolute difference in relative sentence position ($abs(\frac{a_t}{|\mathbf{e}|} - \frac{t}{|\mathbf{f}|})$) allows the model to learn a preference for aligning words close to the alignment matrix diagonal. We also included two conjunction features for the relative sentence position multiplied by the Dice and Model 1 translation scores.

Null We use a number of variants on the above features for alignments between a source word and the *null* target. The maximum translation score between the source and one of the target words

[2] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger
[3] http://www.freedict.de
[4] http://lit.csci.unt.edu/~rada/downloads/RoNLP/R.E.tralex

model	precision	recall	f-score	AER
Model 4 refined	87.4	95.1	91.1	9.81
Model 4 intersection	97.9	86.0	91.6	7.42
French → English	96.7	85.0	90.5	9.21
English → French	97.3	83.0	89.6	10.01
intersection	98.7	78.6	87.5	12.02
refined	95.7	89.2	92.3	7.37

Table 1. Results on the Hansard data using all features

model	precision	recall	f-score	AER
Model 4 refined	80.49	64.10	71,37	28.63
Model 4 intersected	95.94	53.56	68.74	31.26
Romanian → English	82.9	61.3	70.5	29.53
English → Romanian	82.8	60.6	70.0	29.98
intersection	94.4	52.5	67.5	32.45
refined	77.1	68.5	72.6	27.41

Table 2. Results on the Romanian data using all features

is used as a feature to represent whether there is a strong alignment candidate. The sum of these scores is also used as a feature. Each source word and POS tag pair are used as indicator features which allow the model to learn particular words of tags which tend to commonly (or rarely) align.

3.2 Symmetrisation

In order to produce many-to-many alignments we combine the outputs of two models, one for each translation direction. We use the refined method from Och and Ney (2003) which starts from the intersection of the two models' predictions and 'grows' the predicted alignments to neighbouring alignments which only appear in the output of one of the models.

4 Experiments

We have applied our model to two publicly available word aligned corpora. The first is the English-French Hansards corpus, which consists of 1.1 million aligned sentences and 484 word-aligned sentences. This data set was used for the 2003 NAACL shared task (Mihalcea and Pedersen, 2003), where the word-aligned sentences were split into a 37 sentence trial set and a 447 sentence testing set. Unlike the unsupervised entrants in the 2003 task, we require word-aligned training data, and therefore must cannibalise the test set for this purpose. We follow Taskar et al. (2005) by using the first 100 test sentences for training and the remaining 347 for testing. This means that our results should not be directly compared to those entrants, other than in an approximate manner. We used the original 37 sentence trial set for feature engineering and for fitting a Gaussian prior.

The word aligned data are annotated with both sure (S) and possible (P) alignments ($S \subseteq P$; Och and Ney (2003)), where the possible alignments indicate ambiguous or idiomatic alignments. We measure the performance of our model using *alignment error rate* (AER), which is defined as:

$$AER(A, S, P) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}$$

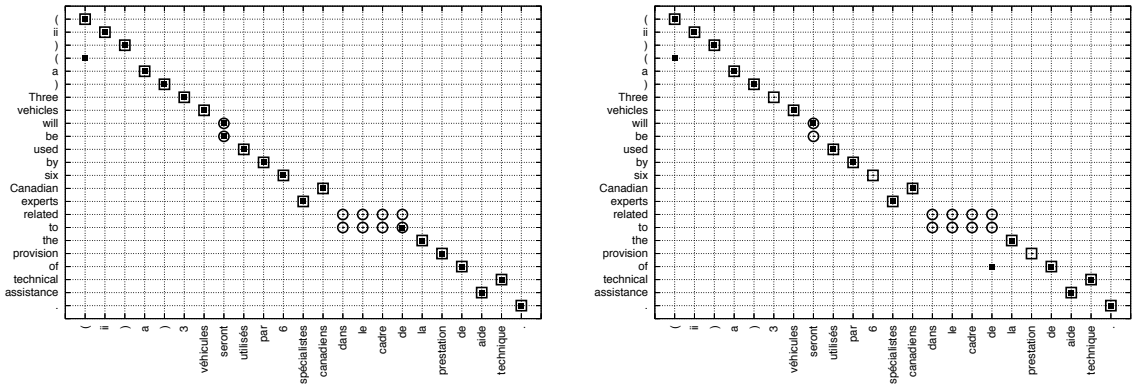
where A is the set of predicted alignments.

The second data set is the Romanian-English parallel corpus from the 2005 ACL shared task (Martin et al., 2005). This consists of approximately 50,000 aligned sentences and 448 word-aligned sentences, which are split into a 248 sentence trial set and a 200 sentence test set. We used these as our training and test sets, respectively. For parameter tuning, we used the 17 sentence trial set from the Romanian-English corpus in the 2003 NAACL task (Mihalcea and Pedersen, 2003). For this task we have used the same test data as the competition entrants, and therefore can directly compare our results. The word alignments in this corpus were only annotated with sure (S) alignments, and therefore the AER is equivalent to the F_1 score. In the shared task it was found that models which were trained on only the first four letters of each word obtained superior results to those using the full words (Martin et al., 2005). We observed the same result with our model on the trial set and thus have only used the first four letters when training the Dice and Model 1 translation probabilities.

Tables 1 and 2 show the results when all feature types are employed on both language pairs. We report the results for both translation directions and when combined using the refined and intersection methods. The Model 4 results are from GIZA++ with the default parameters and the training data lowercased. For Romanian, Model 4 was trained using the first four letters of each word.

The Romanian results are close to the best reported result of 26.10 from the ACL shared task (Martin et al., 2005). This result was from a system based on Model 4 plus additional parameters such as a dictionary. The standard Model 4 implementation in the shared task achieved a result of 31.65, while when only the first 4 letters of each word were used it achieved 28.80.[5]

[5]These results differ slightly our Model 4 results reported in Table 2.



(a) With Markov features

(b) Without Markov features

Figure 2. An example from the Hansard test set, showing the effect of the Markov features.

Table 3 shows the effect of removing each of the feature types in turn from the full model. The most useful features are the Dice and Model 1 values which allow the model to incorporate translation probabilities from the large sentence aligned corpora. This is to be expected as the amount of word aligned data are extremely small, and therefore the model can only estimate translation probabilities for only a fraction of the lexicon. We would expect the dependence on sentence aligned data to decrease as more word aligned data becomes available.

The effect of removing the Markov features can be seen from comparing Figures 2 (a) and (b). The model has learnt to prefer alignments that follow the diagonal, thus alignments such as *3* $\leftrightarrow$ *three* and *prestation* $\leftrightarrow$ *provision* are found, and missalignments such as *de* $\leftrightarrow$ *of*, which lie well off the diagonal, are avoided.

The differing utility of the alignment word pair feature between the two tasks is probably a result of the different proportions of word- to sentence-aligned data. For the French data, where a very large lexicon can be estimated from the million sentence alignments, the sparse word pairs learnt on the word aligned sentences appear to lead to overfitting. In contrast, for Romanian, where more word alignments are used to learn the translation pair features and much less sentence aligned data are available, these features have a significant impact on the model. Suprisingly the orthographic features actually worsen the performance in the tasks (incidentally, these features help the trial set). Our explanation is that the other features (eg. Model 1) already adequately model these correspondences, and therefore the orthographic features do not add much additional modelling power. We expect that with further careful feature engineering, and a larger trial set, these orthographic features could be much improved.

feature group	**Rom $\leftrightarrow$ Eng**	**Fre $\leftrightarrow$ Eng**
ALL	27.41	7.37
–orthographic	27.30	7.25
–Dice	27.68	7.73
–dictionary	27.72	7.21
–sentence position	28.30	8.01
–POS	–	8.19
–Model 1	28.62	8.45
–alignment word pair	32.41	7.20
–Markov	32.75	12.44
–Dice & –Model 1	35.43	14.10

Table 3. The resulting AERs after removing individual groups of features from the full model.

The Romanian-English language pair appears to offer a more difficult modelling problem than the French-English pair. With both the translation score features (Dice and Model 1) removed – the sentence aligned data are not used – the AER of the Romanian is more than twice that of the French, despite employing more word aligned data. This could be caused by the lack of possible (P) alignment markup in the Romanian data, which provide a boost in AER on the French data set, rewarding what would otherwise be considered errors. Interestingly, without any features derived from the sentence aligned corpus, our model achieves performance equivalent to Model 3 trained on the full corpus (Och and Ney, 2003). This is a particularly strong result, indicating that this method is ideal for data-impoverished alignment tasks.

4.1 Training with *possible* alignments

Up to this point our Hansards model has been trained using only the sure (S) alignments. As the data set contains many possible (P) alignments, we would like to use these to improve our model. Most of the possible alignments flag blocks of ambiguous or idiomatic (or just difficult) phrase level alignments. These many-to-many alignments cannot be modelled with our many-to-one setup. However, a number of possibles flag one-to-one or many-to-one aligments: for this experiment we used these possibles in training to investigate their effect on recall. Using these additional alignments our refined precision decreased from 95.7 to 93.5, while recall increased from 89.2 to 92.4. This resulted in an overall decrease in AER to 6.99. We found no benefit from using many-to-many possible alignments as they added a significant amount of noise to the data.

4.2 Model 4 as a feature

Previous work (Taskar et al., 2005) has demonstrated that by including the output of Model 4 as a feature, it is possible to achieve a significant decrease in AER. We trained Model 4 in both directions on the two language pairs. We added two indicator features (one for each direction) to our CRF which were active if a given word pair were aligned in the Model 4 output. Table 4 displays the results on both language pairs when these additional features are used with the refined model. This produces a large increase in performance, and when including the possibles, produces AERs of 5.29 and 25.8, both well below that of Model 4 alone (shown in Tables 1 and 2).

4.3 Cross-validation

Using 10-fold cross-validation we are able to generate results on the whole of the Hansards test data which are comparable to previously published results. As the sentences in the test set were randomly chosen from the training corpus we can expect cross-validation to give an unbiased estimate of generalisation performance. These results are displayed in Table 5, using the possible (P) alignments for training. As the training set for each fold is roughly four times as big previous training set, we see a small improvement in AER.

The final results of 6.47 and 5.19 with and without Model 4 features both exceed the performance of Model 4 alone. However the unsupervised Model 4 did not have access to the word-alignments in our training set. Callison-Burch et al. (2004) demonstrated that the GIZA++ models could be trained in a semi-supervised manner, leading to a slight decrease in error. To our knowledge, our AER of 5.19 is the best reported result, generative or discriminative, on this data set.

model	precision	recall	f-score	AER
Rom ↔ Eng	79.0	70.0	74.2	25.8
Fre ↔ Eng	97.9	90.8	94.2	5.49
Fre ↔ Eng (P)	95.5	93.7	94.6	5.29

Table 4. Results using features from Model 4 bi-directional alignments, training with and without the possible (P) alignments.

model	precision	recall	f-score	AER
Fre ↔ Eng	94.6	92.2	93.4	6.47
Fre ↔ Eng (Model 4)	96.1	93.3	94.7	5.19

Table 5. 10-fold cross-validation results, with and without Model 4 features.

5 Related work

Recently, a number of discriminative word alignment models have been proposed, however these early models are typically very complicated with many proposing intractable problems which require heuristics for approximate inference (Liu et al., 2005; Moore, 2005).

An exception is Taskar et al. (2005) who presented a word matching model for discriminative alignment which they they were able to solve optimally. However, their model is limited to only providing one-to-one alignments. Also, no features were defined on label sequences, which reduced the model's ability to capture the strong monotonic relationships present between European language pairs. On the French-English Hansards task, using the same training/testing setup as our work, they achieve an AER of 5.4 with Model 4 features, and 10.7 without (compared to 5.29 and 6.99 for our CRF). One of the strengths of the CRF MAP estimation is the powerful smoothing offered by the prior, which allows us to avoid heuristics such as early stopping and hand weighted loss-functions that were needed for the maximum-margin model.

Liu et al. (2005) used a conditional log-linear model with similar features to those we have employed. They formulated a global model, without making a Markovian assumption, leading to the need for a sub-optimal heuristic search strategies.

Ittycheriah and Roukos (2005) trained a dis-

criminative model on a corpus of ten thousand word aligned Arabic-English sentence pairs that outperformed a GIZA++ baseline. As with other approaches, they proposed a model which didn't allow a tractably optimal solution and thus had to resort to a heuristic beam search. They employed a log-linear model to learn the observation probabilities, while using a fixed transition distribution. Our CRF model allows both the observation and transition components of the model to be jointly optimised from the corpus.

6 Further work

The results presented in this paper were evaluated in terms of AER. While a low AER can be expected to improve end-to-end translation quality, this is may not necessarily be the case. Therefore, we plan to assess how the recall and precision characteristics of our model affect translation quality. The tradeoff between recall and precision may affect the quality and number of phrases extracted for a phrase translation table.

7 Conclusion

We have presented a novel approach for inducing word alignments from sentence aligned data. We showed how conditional random fields could be used for word alignment. These models allow for the use of arbitrary and overlapping features over the source and target sentences, making the most of small supervised training sets. Moreover, we showed how the CRF's inference and estimation methods allowed for efficient processing without sacrificing optimality, improving on previous heuristic based approaches.

On both French-English and Romanian-English we showed that many highly predictive features can be easily incorporated into the CRF, and demonstrated that with only a few hundred word-aligned training sentences, our model outperforms the generative Model 4 baseline. When no features are extracted from the sentence aligned corpus our model still achieves a low error rate. Furthermore, when we employ features derived from Model 4 alignments our CRF model achieves the highest reported results on both data sets.

Acknowledgements

Special thanks to Miles Osborne, Steven Bird, Timothy Baldwin and the anonymous reviewers for their feedback and insightful comments.

References

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

C. Callison-Burch, D. Talbot, and M. Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of ACL*, pages 175–182, Barcelona, Spain, July.

S. Chen and R. Rosenfeld. 1999. A survey of smoothing techniques for maximum entropy models. *IEEE Transactions on Speech and Audio Processing*, 8(1):37–50.

L. R. Dice. 1945. Measures of the amount of ecologic association between species. *Journal of Ecology*, 26:297–302.

A. Ittycheriah and S. Roukos. 2005. A maximum entropy word aligner for Arabic-English machine translation. In *Proceedings of HLT-EMNLP*, pages 89–96, Vancouver, British Columbia, Canada, October.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*, pages 81–88, Edmonton, Alberta.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In *Proceedings of ICML*, pages 282–289.

Y. Liu, Q. Liu, and S. Lin. 2005. Log-linear models for word alignment. In *Proceedings of ACL*, pages 459–466, Ann Arbor.

R. Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of CoNLL*, pages 49–55.

J. Martin, R. Mihalcea, and T. Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74, Ann Arbor, Michigan, June.

R. Mihalcea and T. Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of HLT-NAACL 2003 Workshop, Building and Using Parrallel Texts: Data Driven Machine Translation and Beyond*, pages 1–6, Edmonton, Alberta.

R. C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of HLT-EMNLP*, pages 81–88, Vancouver, Canada.

F. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

F. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*, pages 213–220.

B. Taskar, S. Lacoste-Julien, and D. Klein. 2005. A discriminative matching approach to word alignment. In *Proceedings of HLT-EMNLP*, pages 73–80, Vancouver, British Columbia, Canada, October.

K. Toutanova, H. Tolga Ilhan, and C Manning. 2002. Extentions to HMM-based statistical word alignment models. In *Proceedings of EMNLP*, pages 87–94, Philadelphia, July.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of 16th Int. Conf. on Computational Linguistics*, pages 836–841.