Natural Language Engineering **11** (3): 263–277. © 2005 Cambridge University Press
doi:10.1017/S1351324905003852 Printed in the United Kingdom

Constrained EM for parallel text alignment

DAVID TALBOT
School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK
e-mail: d.r.talbot@sms.ed.ac.uk

(*Received May 1 2004; revised November 30 2004*)

Abstract

Standard parameter estimation schemes for statistical translation models can struggle to find reasonable settings on some parallel corpora. We show how auxiliary information can be used to constrain the procedure directly by restricting the set of alignments explored during parameter estimation. This enables the integration of bilingual and monolingual knowledge sources while retaining the flexibility of the underlying models. We demonstrate the effectiveness of this approach for incorporating linguistic and domain-specific constraints on various parallel corpora, and consider the importance of using the context of the parallel text to guide the application of such constraints.

1 Introduction

The automatic estimation of translation models from parallel corpora was proposed in Brown, Della Pietra, Della Pietra and Mercer (1993). These models (below the IBM models) and more recent phrase-based models (Marcu and Wong 2002) are fundamentally language and domain independent. Their parameters can be estimated from parallel sentence-aligned text with minimal preprocessing.

Lack of sufficient parallel training corpora has prevented these models being applied to many new domains and language pairs. Domain specific phenomena can also exacerbate the parameter estimation problem. In particular, we observe that these models struggle to find plausible bilingual word alignments when a corpus has a high vocabulary growth rate and average sentence length. To counter this, we show that auxiliary information such as linguistic or domain-specific knowledge sources can be used to constrain the parameter estimation procedure.

Many approaches for incorporating auxiliary information during parameter estimation have been proposed. Here, we present a framework that enables the integration of a wide range of constraints directly within the Expectation Maximisation parameter estimation procedure (EM; Dempster, Laird and Rubin 1977). This facilitates the interaction of multiple constraints and enables their application to be conditioned more closely on the context of the parallel text.

After a brief review of the standard EM parameter estimation scheme, here we:

- Discuss the relationship between word-level alignment error rate (AER), vocabulary growth rate and sentence length;

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

- Describe a modified EM scheme to constrain parameter estimation;
- Show significantly lower AER incorporating monolingual and bilingual constraints from linguistic and domain-specific knowledge sources in this way;
- Examine the effect improved alignments have on translation accuracy.

2 Related work

Many methods for incorporating external knowledge sources in the estimation of statistical translation models have been proposed. Brown, Della Pietra, Della Pietra, Goldsmith, Hajic, Mercer and Mohanty (1993a) modelled a bilingual dictionary alongside the corpus, treating it as additional training data and maximising the likelihood of both simultaneously. Och and Ney (2003) proposed giving a higher weighting for dictionary entries that co-occur in the parallel corpus. A similar approach was taken in Kondrak, Marcu and Knight (2003) with cognate word pairs from the parallel corpus appended multiple times to the training set.

In previous work (Callison-Burch, Talbot and Osborne (2004), we demonstrated that the inclusion of sentence pairs that are explicitly aligned at the word-level can improve the alignments obtained by EM. Recognising this parameter estimation problem as a semi-supervised learning task (e.g. Nigam, McCallum, Thrun and Mitchell 2000), we proposed a mixed likelihood function that amplifies the contribution of the word-aligned data during parameter estimation, reducing the amount of manually annotated data required.

Linguistic structure can also be modelled directly within an alignment model. For example, Toutanova, Ilhan and Manning (2002) modified a hidden Markov alignment model proposed in Vogel, Ney and Tillman (1996) to condition transitions between alignment positions on the part of speech tags of the aligned source and target words. By extending the underlying model to include parameters relating alignment events and linguistic structure, such correlations can be captured in a data-driven manner. However, the resulting model requires annotation resources that may not be available.

The underlying method we use here has been applied to other tasks in computational linguistics, notably the learning of context-free grammars from partially bracketed corpora (Pereira and Schabes 1992).

3 Parameter estimation for machine translation

Our approach for incorporating constraints in the estimation of word alignments models is applicable to any parameter estimation scheme that includes an *expectation* step during which permissible word-level alignments are proposed for each parallel sentence pair in the corpus and then scored by the model. We describe the original and revised parameter estimation schemes with reference to the hidden Markov alignment model.

In the source-channel paradigm (Brown, Cocke, Della Pietra, Della Pietra, Jelinek, Mercer and Poossin (1988), the task of translation is decomposed into language

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

modelling, $p(\mathbf{e})$, and translation modelling, $p(\mathbf{f}|\mathbf{e})$, using Bayes' rule,

$$\mathbf{e}^* = \arg\max_{\mathbf{e}} p(\mathbf{e})\mathrm{p}(\mathbf{f}|\mathbf{e}). \tag{1}$$

Brown *et al.* (1993a) make the assumption that the translation of a sentence can be decomposed into a series of word-level mappings; in addition no *a priori* knowledge of either monolingual or bilingual relationships that may hold between words in the corpus is assumed. The first assumption is essential to keep the number of parameters manageable while the second helps avoid making potentially language-specific assumptions.

That their parameters can be estimated solely from a parallel corpus, has been a clear advantage of the IBM models. However, when the available parallel corpus is insufficient to obtain robust parameter estimates, as with one of the corpora we consider here, it is not immediately clear how external linguistic or domain-specific knowledge sources should be incorporated.

We use the following notation: $\mathbf{e}$ and $\mathbf{f}$ represent aligned source and target sentences respectively; sentences consist of tokens $\{e_i : i = 1, 2, \ldots, I\}$, and $\{f_j : j = 1, 2, \ldots, J\}$ where I and J are the respective sentence lengths. Word level alignments, $\{a_j : j = 1, 2, \ldots J\}$, associated with each target word, f_j, specify which source token the target word in position j is aligned to. In the HMM, word alignments can take values between zero and I (the length of the source sentence). A value of zero implies that the target word is not aligned to any source word.

As with the IBM models, the HMM is parameterised by word-level *translation* parameters, $t(f|e)$, describing the probability that a given source type, e, will be translated as the target type, f. In addition, the HMM contains *alignment* parameters, $a(a_j|a_{j-1}, I)$, that describe the probability that a target word in position j is aligned with a particular source word, given the position in the source sentence with which the previous target word was aligned and the length of the source sentence (these are the state transition probabilities of the HMM). By conditioning each alignment on the alignment position of the preceding target token, the HMM is able to capture certain localised alignment effects (Vogel *et al.* 1996).

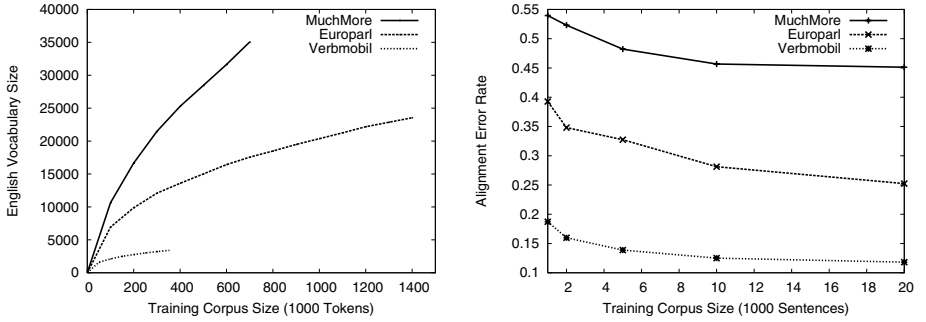
Since explicit word-level alignments are not present in standard parallel corpora, the alignment position for each target word is treated as a hidden variable. The Expectation Maximisation (EM) framework for maximum likelihood estimation from incomplete data (Dempster *et al.* 1977) is used to estimate the expected values of these alignments (the E-step) and the model parameters are then updated treating these expected values as sufficient statistics (the M-step).[1] The probability of a sentence pair is given by summing over alignments permitted by the model,

$$p(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \tag{2}$$

$$= \sum_{a_1, a_2, \ldots, a_J} \prod_{j=1}^{J} a(a_j|a_{j-1}, I)t(f_j|e_{a_j}). \tag{3}$$

[1] We effectively collect fractional counts of possible alignments.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press



(a) Vocabulary growth (b) AER with corpus size

Fig. 1. Relationship between AER and vocabulary size for three parallel texts.

For the HMM this summation can be performed exhaustively using the *forward-backward* procedure (Baum 1972) which exploits the model's conditional independence assumptions; for other models such as IBM model 4 and phrase-based models a subset of *promising* alignments is all that can be searched.

Since *a priori* all positions in the source sentence including the null position are permissible alignments for each target word, all pairwise combinations of source and target tokens will be considered during HMM parameter estimation. Only a fraction of these are linguistically plausible. Given sufficient parallel data, the parameters will converge to plausible settings since pairs of words that are mutual translations co-occur more frequently than non-translations (Brown *et al.* 1993a).

4 Factors affecting alignment error rate

Och and Ney (2003) provides a breakdown of the *alignment error rate* (AER; see section 6.1) on two parallel corpora for various IBM models, including a comparison of specific model features and training regimes as they effect alignment accuracy. The relationship between corpus size and AER for the IBM models may also depend on factors such as the rate of vocabulary growth and average length of sentences in the corpus.

The number of parameters in these models is directly related to the size of the source and target language vocabularies, so we anticipate that AER will fall more slowly when the vocabulary grows at a faster rate. This may be seen in the differences between AER for the Verbmobil and Canadian Hansard corpora for various sample sizes reported in Och and Ney (2003). One of the corpora we consider here suffers more seriously from this effect.

Figure 1(a) shows the source vocabulary growth for three English-German parallel corpora: the Proceedings of the European Parliament, the Verbmobil corpus of travel dialogues and the MuchMore German-English corpus of medical abstracts. AER falls much more slowly for the MuchMore corpus (Figure 1(b)).

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

Average sentence length in the training corpus may also contribute to AER. Together with vocabulary size, sentence length determines the average number of distinct source types with which each target type co-occurs in the corpus. Within the unconstrained model, all source types co-occuring with a given target word will compete to determine the lexical translation parameters, $t(f|e)$. Alignment may also be intrinsically more difficult for longer sentences given the larger choice of source candidates with which each target word can be aligned in such sentences, and also the greater opportunity there is for alignment errors to be compounded.

These observations motivate our constrained EM framework.

5 Constraining EM

We use a range of different knowledge sources to provide constraints on parameter estimation in our experiments (see section 6). For the time being, we take for granted the availability of an external component that can, on occasion, propose accurate judgments regarding the suitability of particular subsets of alignments. We assume that the external component formulates its alignment judgments as restrictions on the set of permissible alignments for a given sentence pair. In this way we can handle a variety of constraints within a single framework.

As described above, the EM framework involves iteratively enumerating as many permissible word-level alignments as possible in each E-step. In standard EM parameter estimation, given no explicit word alignments, the posterior probabilities of all permissible alignments within a sentence pair are coupled. Information regarding any single alignment may therefore change the model's distribution over all alignments in a given sentence pair.

To exclude alignments that pass through a source-target pairing, $< e_i, f_j >$, we set the alignment probability of that pairing to zero. In the HMM this can be implemented by modifying the alignment parameters, $a(a_j|a_{j-1}, I)$, for constrained positions.

The constrained alignment parameters, $a^c(a_j|a_{j-1}, I)$, are calculated by setting all disallowed positions to zero and then renormalising. These are then used in the standard forward-backward recursions to directly restrict the posterior distribution inferred in the E-step,

$$\alpha_j(i) = t(f_j|e_{a_j}) \sum_{a_{j-1}} \alpha_{j-1} a^c(a_j|a_{j-1}, I) \tag{4}$$

$$\beta_{j-1}(i) = \sum_{a_j} t(f_j|e_{a_j}) \beta_j a^c(a_j|a_{j-1}, I). \tag{5}$$

In this way, information from the external knowledge source is incorporated directly within the original inference routine.

In our experiments we consider both exclusive and non-exclusive constraints. Exclusive constraints, which we term *anchor* constraints, force a given target token to align with a specific source token by setting the alignment probabilities for all other source tokens to zero at that position.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

Table 1. *Parallel corpora used in the experiments (number of sentence pairs)*

Verbmobil	English	German	MuchMore	English	German
Vocab. (types)	3505	5936	Vocab. (types)	37102	76012
Singletons	1305	2600	Singletons	18185	41610
Sentences	34K	34K	Sentences	33K	33K
Sentence Len.	10.3	9.74	Sentence Len.	25.6	24.8

Europarl	English	German
Vocab. (types)	34386	77188
Singletons	12811	37939
Sentences	100K	100K
Sentence Len.	28.2	27.34

The impact constraints have on parameter estimation will depend on the structure of the alignment model used as well as the type of constraint applied. In the HMM, for instance, the conditional independence assumptions of the model will render the alignments of all target tokens that precede an anchor constraint, independent of the alignments of all subsequent target tokens in the sentence, effectively decomposing the alignment problem. Non-exclusive constraints may also be formulated that decouple regions of the alignment space. We therefore anticipate that by accurately constraining the alignments at certain positions, AER may be significantly reduced throughout a sentence pair.

6 Experimental setup

We performed a number of word alignment experiments to evaluate the impact of the constrained parameter estimation procedure described above on three different German-English parallel texts. We also conducted some simple machine translation experiments using the word alignments thus obtained.

The Verbmobil corpus (Wahlster 2000) is a speech translation task consisting of transcribed conversations in the travel and tourism domain. The MuchMore corpus (Buitelaar, Sacaleanu, Vintar, Steffen, Volk, Dejean, Gaussier, Widdows, Weiser and Frederking (2003) is made up of around seven thousand pairs of abstracts in German and English from a range of specialist medical journals. The Europarl corpus (Koehn 2000) is taken from the proceedings of the European Parliament. Table 1 gives details of these corpora. All three were initially aligned at the sentence level using the method described in Gale and Church (1991).

6.1 Word alignment experiments

Bilingual word alignments induced on a held out set of parallel data can be evaluated by comparing them to a set of manually annotated reference alignments. We used non-null alignment error rate (AER; Och and Ney 2003) to measure the quality of alignments in these experiments.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

Table 2. *Reference word alignment sets used in the experiments*

Corpora	Sure alignments	Probable alignments
MuchMore	843	997
Europarl	790	1086
Verbmobil	1584	3447

AER is an F-score-type metric that takes into account alignment precision and recall. It is calculated with reference to a set of *sure* and *probable* gold-standard word alignments. In Och and Ney (2003), these categories correspond respectively to the intersection and union of alignments proposed by multiple annotators on the same parallel data. Two of our sets of reference word alignments were created by a single bilingual annotator who was instructed to mark definite word-to-word translations as sure, and more elaborate phrasal correspondences as probable. In both cases, sure alignments are counted as a subset of probable alignments. Details of the reference word alignment sets used in the experiments are given in Table 2.

Formally, AER is defined as,

$$AER(S,P;A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{6}$$

where A are the alignments induced by the model under evaluation, and S, and P, are the sure and probable reference alignments respectively.

To alleviate differences in alignment accuracy due to the translation direction (models trained with German as the source language tend to have significantly lower AER than those with English) and for consistency with the machine translation experiments below, we estimated two models for each training set with English and then German as the source language. We report AER on a set that combines the Viterbi alignments from both these models using the heuristic given in Och and Ney (2000). This procedure starts with the intersection of the alignments induced in both directions and iteratively adds adjacent alignments from their union (see Och and Ney (2000) for details). Such heuristics have been shown to result in higher translation accuracy for phrase-based models (Koehn *et al.* 2003).

For each corpus we conducted experiments with and without constraints for a range of corpus sizes. In each case, constraints were included for the training set only. The constrained EM procedure was implemented by modifying the GIZA++ word alignment training tool (Och and Ney 2003) which was also used unmodified for the unconstrained experiments. Specifically for IBM model 1 and the HMM we modified the alignment probabilities as described above, while for IBM model 4 we used the constraints to restrict the exploration of alignments neighbouring the pseudo-Viterbi alignment (Och and Ney 2003).[2] We estimated parameters using five iterations of model 1, the HMM and model 4.

[2] Model 4 modifications had a negligible impact on AER and are not focused on here.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

Table 3. *Average per sentence frequency of constraints by type and corpus*

Corpus	Dictionary entries	Cognates	Numeric	UMLS	Total unique
Verbmobil	5.14	2.07	N/A	N/A	5.21
Europarl	6.78	2.09	0.17	N/A	7.24
MuchMore	4.11	3.03	2.9	1.2	6.77

As well as comparing our method with unconstrained EM parameter estimation, we also ran experiments to compare this method for including constraints from a bilingual dictionary to that proposed in Och and Ney (2003) where co-occurring bilingual dictionary entries in the corpus are given additional counts during the M-step for model 1. We used the built-in dictionary function of the GIZA++ alignment tool to perform these experiments using the default settings. In addition, we ran experiments using subsets of the constraints on the MuchMore corpus, to compare the effectiveness of constraints based on cognate relations and a general bilingual dictionary on this text from the medical domain.

6.2 Constraint types

We used two types of constraint, *anchor* and *cohesion constraints* (Fox 2002) in the experiments. As mentioned above, the former are exclusive constraints where a target token is forced to align with a specific source token. The latter is a particular type of non-exclusive constraint where a contiguous subset of target tokens are prevented from aligning with a contiguous subset of source tokens. Both types of constraint are shown in Figure 2 below. Due to the different information embodied in these two types of constraint, we treat them slightly differently.

6.2.1 Anchor constraints

The anchor points can be regarded as *labelled* data and estimation with a mixture of constrained and unconstrained tokens treated as a semi-supervised learning problem (Nigam *et al.* 2000; Callison-Burch *et al.* 2004).

Since the ratio of anchor points to unconstrained positions (see Table 3) is relatively small, standard maximum likelihood estimation, in which all tokens in the training corpus are weighted equally, will not necessarily make the most efficient use of the constraint information. Assuming knowledge sources propose accurate alignments, we would like the global parameter estimates to reflect this information more readily than that simply inferred during the E-step. To amplify the contribution of anchor points, we therefore weight the statistics collected in the M-step differently for exclusively constrained target positions. Following Callison-Burch *et al.* (2004), we use a single parameter λ that is optimised on a held out set of reference word alignments. The posterior alignment probabilities, used as counts in the M-step, are multiplied by λ when the target position is exclusively constrained and by $(1 - \lambda)$ otherwise. In these experiments λ was set to 0.95.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

We used the following knowledge sources to generate anchor constraints:

- cognate relations;
- a bilingual dictionary;
- domain specific bilingual semantic annotation (MuchMore only).
- numeric pattern matching (MuchMore only).

These knowledge sources are described below. Table 3 shows the average per sentence frequency of each type of constraint for the three training corpora.

Cognate Pairs English and German share a large number of cognates.[3] This is especially true in the medical domain. For the MuchMore corpus, in particular, it appears safe to assume that cognate pairs are translations of one another when they co-occur in an aligned sentence pair. Since we did not have access to a list of cognate pairs, we had to acquire these for each corpus. In practice, it was relatively simple to learn the parameters of a probabilistic string edit distance to model the cognate relationship with very high precision and relatively high recall.

We started with a set of candidate word pairs that co-occurred in the corpus and had at least four letters in common. Parameters for a probabilistic string edit distance were then estimated for deletions, insertions and substitutions at the character level, taking into account the total number of co-occurrences of each candidate pair in the corpus. This allowed us to construct a model that could distinguish regular orthographic differences that characterise the cognate relationship between English and German types. For instance, pairs such as *localized, lokalisieren* were picked up while unrelated but similarly spelt words, such as *disease, diese* were rejected. In this case, the parallel corpus provided information, in the form of co-occurrence frequencies, that helped improve the accuracy of the external knowledge source. After optimising a threshold parameter for this cognate model on a small held-out set of sentence pairs, we used it to identify cognates in each corpus.

Selecting Dictionary Entries We had access to a general purpose German-English dictionary, containing around 25,000 entries. We created a constraint for each pair of English and German tokens in an aligned sentence pair that appeared together in the dictionary.

Domain-specific Semantic Annotation Concepts covered by the Unified Medical Language System (UMLS) had been annotated on both halves of the MuchMore corpus. Although there were many discrepancies, concepts which were linked via a common bilingually indexed UMLS code, consistently corresponded to accurate translations.

Numeric Anchor Pairs The MuchMore corpus contains a large number of tokens containing numeric characters. These range from chemical compounds to percentages.

[3] We use the term for any pair of words that are show strong orthographic similarities.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

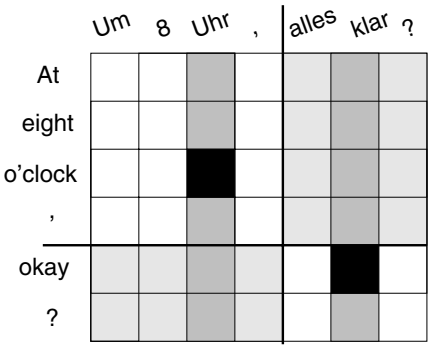


Fig. 2. Anchor constraints (black) result in all other source (English) words having zero alignment probability (dark grey). Cohesion constraints additionally set all alignment positions between unanchored chunks to zero (light grey). Chunks marked by bold lines.

Table 4. *Chunks selected by different criteria (MuchMore)*

	Per Sentence Freq.		Ave. Length	
Chunk criteria	English	German	English	German
All Constituents	6.06	9.09	4.09	2.64
All Clauses	1.33	1.36	17.64	17.23
Filtered Clauses	1.23	1.23	22.58	21.04

We extracted source target pairs which co-occurred and predominantly consisted of the same numeric characters.

6.2.2 Cohesion constraints

We infer cohesion constraints by combining anchor points with monolingual *chunk* information; these constraints rely on the assumption that translations display a degree of compositionality (Simard and Langlais 2003). Given anchor points aligning tokens in a source and target chunk, we constrain EM to prevent any tokens within each of these anchored chunks from aligning with any other chunks (see Figure 2). We used the Treetagger chunk tagger (Schmid 1994) to identify monolingual English and German chunk boundaries.

As noted in Simard and Langlais (2003), although compositionality is common in translation, it is by no means a hard rule. From a translator's perspective, breaking the source sentence into shorter segments that can be processed independently, may well simplify the task of translation; however, the smaller such segments are, the less natural the resulting translation will tend to be.

We experimented with cohesion constraints using chunks selected according to various criteria. Firstly, we took chunks corresponding to all constituents proposed by the chunk tagger; secondly, that subset of chunks corresponding to clause boundaries; and thirdly those clause-level segments for which the tagger identified the same number of clauses in both languages of an aligned sentence pair. Table 4

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

gives details of the average frequency and length of chunks selected according to these criteria for the MuchMore corpus. For the main word alignment and machine translation experiments, we used the third criteria.

6.3 Machine translation experiments

The constrained parameter estimation scheme presented here may help any task that requires accurate bilingual word alignments. Machine translation is one obvious application. We conducted experiments using some of the word alignments created in the above experiments to estimate phrase-based translation models (Koehn *et al.* 2003).

The translation models were estimated using the relative frequencies of aligned phrase pairs extracted from the merged bidirectional word alignments. Here, phrases were restricted to contiguous sequences of source and target tokens that were exclusively aligned to one another. Lexical weightings were calculated for each phrase pair (Koehn *et al.* 2003). A trigram language model was estimated from the target side of each parallel corpus using the SRI Language Modelling Toolkit (Stolcke 2002). Translations were generated using the Pharaoh Beam-Search Decoder (Koehn 2004) on a held out sample of 500 sentence pairs from each corpus. Translation quality was evaluated using the BLEU metric (Papineni, Roukos and Ward (1998), an exponential loss function based on the *n*-gram precision and length of a candidate translation against a reference translation.

7 Results

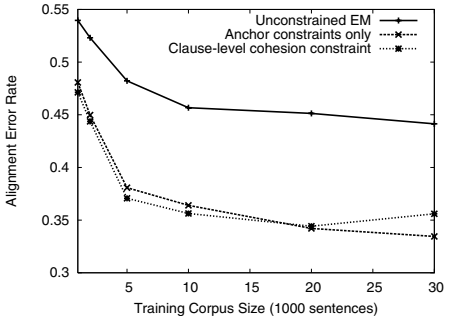
AER for different sized training sets using standard EM parameter estimation is shown in Figure 1(b) (section 4) for all three corpora. While unconstrained EM reaches an impressive AER of 0.15 with only slightly more than 2000 sentences on the Verbmobil task, it appears to struggle on the MuchMore corpus.

Figures 3(a), 3(b) and 3(c) show significant reductions in AER on all three corpora incorporating constraints within EM using the method we have described. These reductions are particularly striking for smaller training sets, although they are maintained for all sample sizes considered here (in the case of Europarl up to 100 thousand sentence pairs). The addition of the cohesion constraint consistently appears to provide some further reduction in AER for smaller training sets.

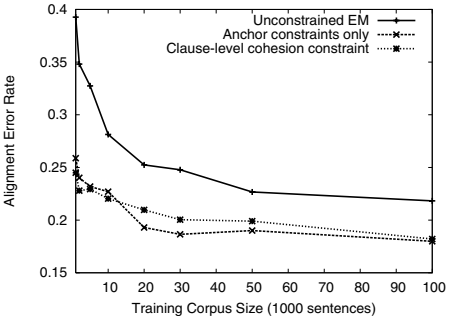
Figure 3(c) also shows that the proposed method is more effective at using the same knowledge source (here a bilingual dictionary) to constrain parameter estimates than a common approach whereby counts are added in model 1.

Figure 4(a) suggests that for the MuchMore corpus cognate information was more useful than a general bilingual dictionary in constraining the parameter estimation. While there were fewer cognate constraints than dictionary matches (see Table 3), the cognate pairs tended to be much less frequent items, in general medical terminology and compounds, whose parameters may otherwise have been particularly badly estimated by the statistical model.

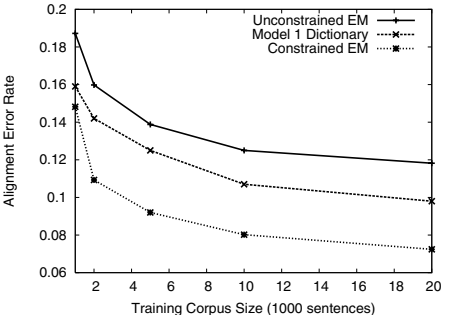
https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press



(a) Constraining EM for MuchMore

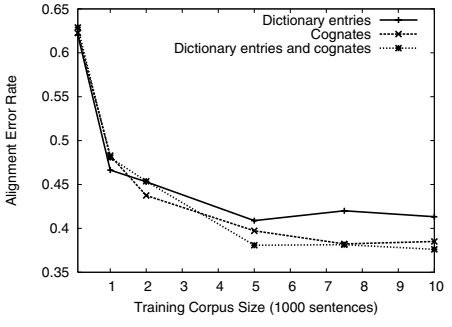


(b) Constraining EM for Europarl

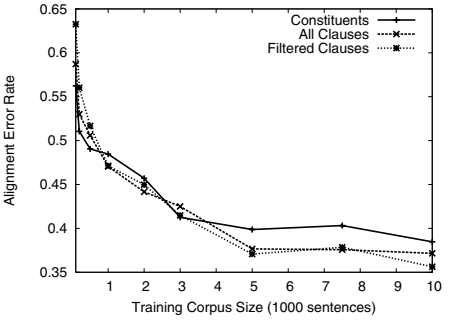


(c) Constraining EM for Verbmobil

Fig. 3. Constrained and Standard EM Parameter Estimation



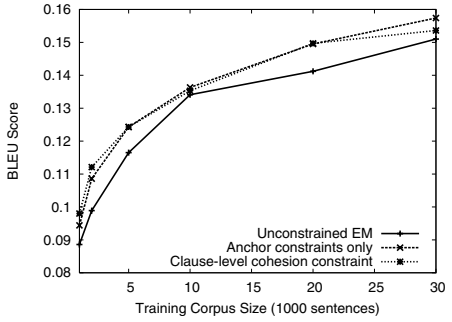
(a) Dictionary entries vs. cognates



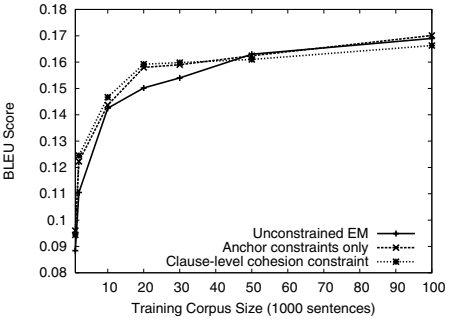
(b) Cohesion constraints

Fig. 4. Comparison of various constraints (MuchMore).

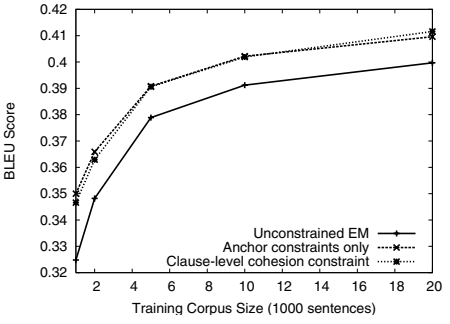
https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press



(a) MuchMore Translation Accuracy



(b) Europarl Translation Accuracy



(c) Verbmobil Translation Accuracy

Fig. 5. Effect of Constrained EM on Translation Quality (BLEU).

Figure 4(b) suggests that larger chunks, defined at the clause-level, enable more accurate inference regarding the compositionality of translations in this parallel corpus. Although for very small sample sizes (below 1000 sentence pairs), the constituent-level cohesion constraint gave the greatest reductions in AER, clause-level constraints were more accurate for all larger samples. From this we might conclude that constituent-level compositionality is less common than clause-level compositionality in translations in this corpus.

Figures 5(a), 5(b) and 5(c) show that the reductions in AER are, in general, passed on as significant improvements in translation accuracy as measured by BLEU. As with the word alignment experiments, the constraints provide the largest relative improvements for smaller sample sizes.

8 Discussion

Statistical word alignment models can struggle to align parallel texts in certain domains. The method proposed here for constraining parameter estimation in such

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

models has been shown to enable the combination of a variety of constraints in a highly effective manner.

By allowing constraints to apply conditionally to specific sentences and source-target pairings within them, the method we have proposed may offer greater precision. Previous approaches which add information such as dictionary entries or cognate pairs to the training corpus as a whole (literally or as additional counts) rely to some extent on the information being globally valid for the corpus. In the method proposed here a constraint need not be applied uniformly. On the other hand, by directly restricting specific alignments, our approach clearly has more potential to do damage when the constraints are incorrect.

9 Future Work

There are a number of ways in which this framework could be developed. For computationally more demanding alignment models such as the joint phrase-based models (Marcu and Wong 2002) constraints may help guide a partial search of the alignment space. Soft constraints, where alignment probabilities are not necessarily set to zero in the E-step, but rather given a weighting by the external model could be considered. This would result in a form of product model potentially enabling the integration of other data-driven constraints into the word alignment process. Finally, it would be interesting to consider ways in which this framework could incorporate constraints learned from the projection of linguistic structure across parallel corpora (Hwa *et al.* 2005, this volume) and conversely how this framework could assist in such tasks.

Acknowledgements

This work was supported by an MRC Priority Area Studentship to the School of Informatics, University of Edinburgh. The author would like to thank Chris Callison-Burch for useful feedback on this work and providing a manual word alignment tool used in the experiments.

References

Baum, L. E. (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**: 1–8.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R. and Poossin, P. (1988) A statistical approach to language translation. *12th International Conference on Computational Linguistics.*

Brown, P., Della Pietra, S., Della Pietra, V. and Mercer, R. (1993a) The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Goldsmith, M. J., Hajic, J., Mercer, R. L. and Mohanty, S. (1993b) But dictionaries are data too. *Human Language Technology 1993*, pp. 202–205.

Buitelaar, P., Sacaleanu, B., Vintar, S., Steffen, D., Volk, M., Dejean, H., Gaussier, E., Widdows, D., Weiser, O. and Frederking, R. (2003) Final Report: MuchMore (Multilingual Concept Hierarchies for Medical Information Organization and Retrieval). Muchmore final, DFKI.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press

Callison-Burch, C., Talbot, D. and Osborne, M. (2004) Statistical Machine Translation with Word- and Sentence-Aligned Parallel Corpora. *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics.*

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**(1):1–38.

Fox, H. J. (2002) Phrasal cohesion and statistical machine translation. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing(EMNLP-2002)*, pp. 95–100.

Gale, W. A. and Church, K. W. (1991) A program for aligning sentences in bilingual corpora. *Meeting of the Association for Computational Linguistics*, pp. 177–184.

Hwa, R., Resnik, P. and Weinberg, A. (2005) Bootstrapping parsers via syntactic projection across parallel texts. *Journal of Natural Language Engineering*, Special Issue on Parallel Texts, **11**(3).

Koehn, P., Och, F. J. and Marcu, D. (2003) Statistical phrase-based translation. *Proceedings of the HLT/NAACL.*

Koehn, P. (2000) Europarl: A Multilingual Corpus for Evaluation of Machine Translation. Unpublished draft.

Koehn, P. (2004) PHARAOH, a beam search decoder for phrase-based statistical machine translation models. *AMTA-2004 Proc. of the 6th Conference of the Association for Machine Translation in the Americas*, Washington, DC.

Kondrak, G., Marcu, D. and Knight, K. (2003) Cognates Can Improve Statistical Translation Models. *Human Language Technology Conference of the North American Chapter of the ACL (HLT-NAACL 2003)*, pp. 46–48.

Marcu, D. and Wong, W. (2002) A phrase-based, joint probability model for statistical machine translation. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002).*

Nigam, K., McCallum, A. K., Thrun, S. and Mitchell, T. M. (2000) Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning* **39**(2/3): 103–134.

Och, F. J. and Ney, H. (2000) Improved statistical alignment models. *ACL00, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440–447. Hongkong, China.

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

Papineni, K., Roukos, S. and Ward, R. T. (1998) Maximimum likelihood and discriminative training of direct translation models. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 189–192.

Pereira, F. and Schabes, Y. (1992) Inside-outside reestimation from partially bracketed corpora. *Proc. of the 20th Annual Meeting of the Association for Computational Linguistics*, Neward, Delaware.

Schmid, H. (1994) Probabilistic part-of-speech tagging using decision trees. *Proc. Intl. Conf. on New Methods in Language Processing.*

Simard, M. and Langlais, P. (2003) Statistical translation alignment with compositionality constraints. *NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Canada.

Stolcke, A. (2002) SRILM – An extensible language modeling toolkit. *Proc. Intl. Conf. on Spoken Language Processing*, volume 2, pp. 901–904. Denver, CO.

Toutanova, K., Ilhan, H. T. and Manning, C. D. (2002) Extensions to HMM-based Statistical Word Alignment Models. *Proc. of the 2002 Conf. on Empirical Methods in Natural Language Processing*, pp. 87–94.

Vogel, S., Ney, H. and Tillmann, C. (1996) HMM-based word alignment in statistical translation. *The 16th Int. Conf. on Computational Ling.*, pp. 836–841.

Wahlster, W. (editor) (2000) *Verbmobil: Foundations of Speech-to-Speech Translation.* Springer.

https://doi.org/10.1017/S1351324905003852 Published online by Cambridge University Press