Context-Dependent Alignment Models for Statistical Machine Translation

Jamie Brunning, Adrià de Gispert and William Byrne
Machine Intelligence Laboratory
Department of Engineering, Cambridge University
Trumpington Street, Cambridge, CB2 1PZ, U.K.
{jjjb2,ad465,wjb31}@eng.cam.ac.uk

To appear in proceedings of NAACL-HLT 2009

Abstract

We introduce alignment models for Machine Translation that take into account the context of a source word when determining its translation. Since the use of these contexts alone causes data sparsity problems, we develop a decision tree algorithm for clustering the contexts based on optimisation of the EM auxiliary function. We show that our context-dependent models lead to an improvement in alignment quality, and an increase in translation quality when the alignments are used in Arabic-English and Chinese-English translation.

1 Introduction

Alignment modelling for Statistical Machine Translation (SMT) is the task of determining translational correspondences between the words in pairs of sentences in parallel text. Given a source language word sequence f_1^J and a target language word sequence e_1^I, we model the translation probability as $\mathrm{P}(e_1^I|f_1^J)$ and introduce a hidden variable a_1^I representing a mapping from the target word positions to source word positions such that e_i is aligned to f_{a_i}. Then $\mathrm{P}(e_1^I|f_1^j) = \sum_{a_1^I} \mathrm{P}(e_1^I, a_1^I|f_1^j)$ (Brown et al., 1993).

Previous work on statistical alignment modelling has not taken into account the source word context when determining translations of that word. It is intuitive that a word in one context, with a particular part-of-speech and particular words surrounding it, may translate differently when in a different context. We aim to take advantage of this information to provide a better estimate of the word's translation. The challenge of incorporating context information is maintaining computational tractability of estimation and alignment, and we develop algorithms to overcome this.

The development of efficient estimation procedures for context-dependent acoustic models revolutionised the field of Automatic Speech Recognition (ASR) (Young et al., 1994). Clustering is used extensively for improving parameter estimation of triphone (and higher order) acoustic models, enabling robust estimation of parameters and reducing the computation required for recognition. Kannan et al. (1994) introduce a binary tree-growing procedure for clustering Gaussian models for triphone contexts based on the value of a likelihood ratio. We adopt a similar approach to estimate context-dependent translation probabilities.

We focus on alignment with IBM Model 1 and HMMs. HMMs are commonly used to generate alignments from which state of the art SMT systems are built. Model 1 is used as an intermediate step in the creation of more powerful alignment models, such as HMMs and further IBM models. In addition, it is used in SMT as a feature in Minimum Error Training (Och et al., 2004) and for rescoring lattices of translation hypotheses (Blackwood et al., 2008). It is also used for lexically-weighted phrase extraction (Costa-jussà and Fonollosa, 2005) and sentence segmentation of parallel text (Deng et al., 2007) prior to machine translation.

1.1 Overview

We first develop an extension to Model 1 that allows the use of arbitrary context information about a source word to estimate context-dependent word-to-word translation probabilities. Since there is insufficient training data to accurately estimate translation probabilities for less frequently occurring contexts, we develop a decision tree clustering algorithm to form context classes. We go on to develop a context-dependent HMM model for alignment.

In Section 3, we evaluate our context-dependent models on Arabic-English parallel text, comparing them to our baseline context-independent models. We perform morphological decomposition of the Arabic text using MADA, and use part-of-speech taggers on both languages. Alignment quality is measured using Alignment Error Rate (AER) measured against a manually-aligned parallel text. Section 4 uses alignments produced by

our improved alignment models to initialise a statistical machine translation system and evaluate the quality of translation on several data sets. We also apply part-of-speech tagging and decision tree clustering of contexts to Chinese-English parallel text; translation results for these languages are presented in Section 4.2.

1.2 Previous and related work

Brown et al. (1993) introduce IBM Models 1-5 for alignment modelling; Vogel et al. (1996) propose a Hidden Markov Model (HMM) model for word-to-word alignment, where the words of the source sentence are viewed as states of an HMM and emit target sentence words; Deng and Byrne (2005a) extend this to an HMM word-to-phrase model which allows many-to-one alignments and can capture dependencies within target phrases.

Habash and Sadat (2006) perform morphological decomposition of Arabic words, such as splitting of prefixes and suffixes. This leads to gains in machine translation quality when systems are trained on parallel text containing the modified Arabic and processing of Arabic text is carried out prior to translation. Nießen and Ney (2001a) perform pre-processing of German and English text before translation; Nießen and Ney (2001b) use morphological information of the current word to estimate hierarchical translation probabilities.

Berger et al. (1996) introduce maximum entropy models for machine translation, and use a window either side of the target word as context information. Varea et al. (2002) test for the presence of specific words within a window of the current source word to form features for use inside a maximum entropy model of alignment.

Toutanova et al. (2002) use part-of-speech information in both the source and target languages to estimate alignment probabilities, but this information is not incorporated into translation probabilities. Popović and Ney (2004) use the base form of a word and its part-of-speech tag during the estimation of word-to-word translation probabilities for IBM models and HMMs, but do not defined context-dependent estimates of translation probabilities.

Stroppa et al. (2007) consider context-informed features of phrases as components of the log-linear model during phrase-based translation, but do not address alignment.

2 Use of source language context in alignment modelling

Consider the alignment of the target sentence $\mathbf{e} = e_1^I$ with the source sentence $\mathbf{f} = f_1^J$. Let $\mathbf{a} = a_1^I$ be the alignments of the target words to the source words. Let c_j be the context information of f_j for $j = 1, \ldots, J$. This context information can be any information about the word, e.g. part-of-speech, previous and next words, part-of-speech of previous and next words, or longer range context information.

We follow Brown et al. (1993), but extend their modelling framework to include information about the source word from which a target word is emitted. We model the alignment process as:

$$\begin{aligned}\mathrm{P}(e_1^I, a_1^I, I | f_1^J, c_1^J) = \\ \mathrm{P}(I|f_1^J, c_1^J) \prod_{i=1}^{I} \big[\mathrm{P}(e_i | a_1^i, e_1^{i-1}, f_1^J, c_1^J, I) \\ \times \mathrm{P}(a_i | e_1^{i-1}, a_1^{i-1}, f_1^J, c_1^J, I) \big] \quad (1)\end{aligned}$$

We introduce word-to-word translation tables that depend on the source language context for each word, i.e. the probability that f translates to e given f has context c is $t(e|f,c)$. We assume that the context sequence is given for a source word sequence. This assumption can be relaxed to allow for multiple tag sequences as hidden processes, but we assume here that a tagger generates a single context sequence c_1^J for a word sequence f_1^J. This corresponds to the assumption that, for a context sequence $\tilde{c}_1^J$, $\mathrm{P}(\tilde{c}_1^J | f_1^J) = \delta_{c_1^J}(\tilde{c}_1^J)$; hence

$$\mathrm{P}(e_1^I, a_1^I | f_1^J) = \sum_{\tilde{c}_1^J} \mathrm{P}(e_1^I, a_1^I, \tilde{c}_1^J | f_1^J) = \mathrm{P}(e_1^I, a_1^I | c_1^J, f_1^J)$$

For Model 1, ignoring the sentence length distribution,

$$\mathrm{P}_{\mathrm{M1}}(e_1^I, a_1^I | f_1^J, c_1^J) = \frac{1}{(J+1)^I} \prod_{i=1}^{I} t(e_i | f_{a_i}, c_{a_i}). \quad (2)$$

Estimating translation probabilities separately for every possible context of a source word individually leads to problems with data sparsity and rapid growth of the translation table. We therefore wish to cluster source contexts which lead to similar probability distributions. Let C_f denote the set of all observed contexts of source word f. A particular clustering is denoted

$$\mathcal{K}_f = \{K_{f,1}, \ldots, K_{f,N_f}\},$$

where $\mathcal{K}_f$ is a partition of C_f. We define a class membership function μ_f such that for any context c, $\mu_f(c)$ is the cluster containing c. We assume that all contexts in a cluster give rise to the same translation probability distribution for that source word, i.e. for a cluster K, $t(e|f,c) = t(e|f,c')$ for all contexts $c, c' \in K$ and all target words e; we write this shared translation probability as $t(e|f,K)$.

The Model 1 sentence translation probability for a given alignment (Equation 2) becomes

$$\mathrm{P}_{\mathrm{M1}}(e_1^I, a_1^I | f_1^J, c_1^J) = \frac{1}{(J+1)^I} \prod_{i=1}^{I} t(e_i | f_{a_i}, \mu_f(c_{a_i})). \quad (3)$$

For HMM alignment, we assume that the transition probabilities $a(a_i|a_{i-1})$ are independent of the word contexts and the sentence translation probability is

$$\mathrm{P_H}(e_1^I, a_1^I|f_1^J, c_1^J) = \prod_{i=1}^{I} a(a_i|a_{i-1}, J)t(e_i|f_{a_i}, \mu_f(c_{a_i})). \quad (4)$$

Section 2.1.1 describes how the context classes are determined by optimisation of the EM auxiliary function. Although the translation model is significantly more complex than that of context-independent models, once class membership is fixed, alignment and parameter estimation use the standard algorithms.

2.1 EM parameter estimation

We train using Expectation Maximisation (EM), optimising the log probability of the training set $\{\mathbf{e}^{(s)}, \mathbf{f}^{(s)}\}_{s=1}^{S}$ (Brown et al., 1993). Given model parameters θ', we estimate new parameters θ by maximisation of the EM auxiliary function

$$\sum_{s,\mathbf{a}} \mathrm{P}_{\theta'}(\mathbf{a}|\mathbf{f}^{(s)}, \mathbf{c}^{(s)}, \mathbf{e}^{(s)}) \log \mathrm{P}_{\theta}(\mathbf{e}^{(s)}, \mathbf{a}, I^{(s)}|\mathbf{f}^{(s)}, \mathbf{c}^{(s)}).$$

We assume the sentence length distribution and alignment probabilities do not depend on the contexts of the source words; hence the relevant part of the auxiliary function is

$$\sum_e \sum_f \sum_{c \in C_f} \gamma'(e|f,c) \log t(e|f,c), \quad (5)$$

where

$$\gamma'(e|f,c) = \sum_s \sum_{i=1}^{I^{(s)}} \sum_{j=1}^{J^{(s)}} \Big[\delta_c(c_j^{(s)})\delta_e(e_i^{(s)})\delta_f(f_j^{(s)}) \times \mathrm{P}_{\theta'}(a_i = j|\mathbf{e}^{(s)}, \mathbf{f}^{(s)}, \mathbf{c}^{(s)})\Big]$$

Here γ' can be computed under Model 1 or the HMM, and is calculated using the forward-backward algorithm for the HMM.

2.1.1 Parameter estimation with clustered contexts

We can re-write the EM auxiliary function (Equation 5) in terms of the cluster-specific translation probabilities:

$$\begin{aligned} &\sum_e \sum_f \sum_{l=1}^{|\mathcal{K}_f|} \sum_{c \in K_{f,l}} \gamma'(e|f,c) \log t(e|f,c) \\ &= \sum_e \sum_f \sum_{K \in \mathcal{K}_f} \gamma'(e|f,K) \log t(e|f,K) \quad (6) \\ &\text{where } \gamma'(e|f,K) = \sum_{c \in K} \gamma'(e|f,c) \end{aligned}$$

Following the usual derivation, the EM update for the class-specific translation probabilities becomes

$$\hat{t}(e|f,K) = \frac{\gamma'(e|f,K)}{\sum_{e'} \gamma'(e'|f,K)}. \quad (7)$$

Standard EM training can be viewed a special case of this, with every context of a source word grouped into a single cluster. Another way to view these clustered context-dependent models is that contexts belonging to the same cluster are tied and share a common translation probability distribution, which is estimated from all training examples in which any of the contexts occur.

2.2 Decision trees for context clustering

The objective for each source word is to split the contexts into classes to maximise the likelihood of the training data. Since it is not feasible to maximise the likelihood of the observations directly, we maximise the expected log likelihood by considering the EM auxiliary function, in a similar manner to that used for modelling contextual variations of phones for ASR (Young et al., 1994; Singer and Ostendorf, 1996). We perform divisive clustering independently for each source word f, by building a binary decision tree which forms classes of contexts which maximise the EM auxiliary function. Questions for the tree are drawn from a set of questions $\mathcal{Q} = \{q_1, \ldots, q_{|\mathcal{Q}|}\}$ concerning the context information of f.

Let K be any set of contexts of f, and define

$$\begin{aligned} L(K) &= \sum_e \sum_{c \in K} \gamma'(e|f,c) \log t(e|f,c) \\ &= \sum_e \sum_{c \in K} \gamma'(e|f,c) \log \frac{\sum_{c \in K} \gamma'(e|f,c)}{\sum_{e'} \sum_{c \in K} \gamma'(e'|f,c)}. \end{aligned}$$

This is the contribution to the EM auxiliary function of source word f occurring in the contexts of K. Let q be a binary question about the context of f, and consider the effect on the partial auxiliary function (Equation 6) of splitting K into two clusters using question q. Define K_q be the set of contexts in K which answer 'yes' to q and $K_{\bar{q}}$ be the contexts which answer 'no'. Define the objective function

$$\begin{aligned} Q_{f,q}(K) &= \sum_e \sum_{c \in K_q} \gamma'(e|f,c) \log t(e|f,c) \\ &\quad + \sum_e \sum_{c \in K_{\bar{q}}} \gamma'(e|f,c) \log t(e|f,c) \\ &= L(K_q) + L(K_{\bar{q}}) \end{aligned}$$

When the node is split using question q, the increase in objective function is given by

$$Q_{f,q}(K) - L(K) = L(K_{\bar{q}}) + L(K_q) - L(K).$$

We choose q to maximise this.

In order to build the decision tree for f, we take the set of all contexts C_f as the initial cluster at the root node. We then find the question $\hat{q}$ such that $Q_{f,q}(C_f)$ is maximal, i.e.

$$\hat{q} = \arg\max_{q \in \mathcal{Q}} Q_{f,q}(C_f)$$

This splits C_f, so our decision tree now has two nodes. We iterate this process, at each iteration splitting (into two further nodes) the leaf node that leads to the greatest increase in objective function. This leads to a greedy search to optimise the log likelihood over possible state clusterings.

In order to control the growth of the tree, we put in place two thresholds:

- T_{imp} is the minimum improvement in objective function required for a node to be split; without it, we would continue splitting nodes until each contained only one context, even though doing so would cause data sparsity problems.
- T_{occ} is the minimum occupancy of a node, based on how often the contexts at that node occur in the training data; we want to ensure that there are enough examples of a context in the training data to estimate accurately the translation probability distribution for that cluster.

For each leaf node l and set of contexts K_l at that node, we find the question q_l that, when used to split K_l, produces the largest gain in objective function:

$$\begin{aligned} q_l &= \arg\max_{q \in \mathcal{Q}} [L(K_{l,q}) + L(K_{l,\bar{q}}) - L(K_l)] \\ &= \arg\max_{q \in \mathcal{Q}} [L(K_{l,q}) + L(K_{l,\bar{q}})] \end{aligned}$$

We then find the leaf node for which splitting gives the largest improvement:

$$\hat{l} = \arg\max_{l} [L(K_{l,q_l}) + L(K_{l,\bar{q}_l}) - L(K_l)]$$

If the following criteria are both satisfied at that node, we split the node into two parts, creating two leaf nodes in its place:

- The objective function increases sufficiently

$$L(K_{l,q_l}) + L(K_{l,\bar{q}_l}) - L(K_l) > T_{\text{imp}}$$

- The occupancy threshold is exceeded for both child nodes:

$$\sum_e \sum_{c \in K_{l,x}} \gamma'(e|f,c) > T_{\text{occ}} \text{ for } x = q, \bar{q}$$

We perform such clustering for every source word in the parallel text.

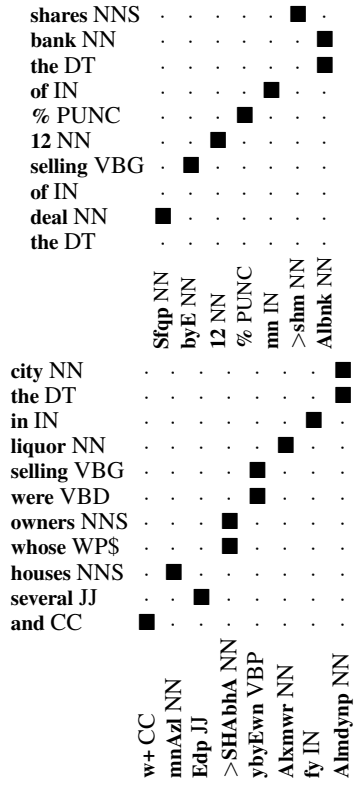

Figure 1: Alignment of the English *selling* in different contexts. In the first, it is preceded by *of* and links to the infinitive of the Arabic verb *byE*; in the second, it is preceded by *were* and links to an inflected form of the same Arabic verb, *ybyEwn*.

3 Evaluation of alignment quality

Our models were built using the MTTK toolkit (Deng and Byrne, 2005b). Decision tree clustering was implemented and the process parallelised to enable thousands of decision trees to be built. Our context-dependent (CD) Model 1 models trained on context-annotated data were compared to the baseline context-independent (CI) models trained on untagged data.

The models were trained using data allowed for the NIST 08 Arabic-English evaluation[1], excluding the UN collections, comprising 300k parallel sentence pairs, a total of 8.4M words of Arabic and 9.5M words of English.

The Arabic language incorporates into its words several prefixes and suffixes which determine grammatical features such as gender, number, person and voice. The MADA toolkit (Habash and Sadat, 2006) was used to perform Arabic morphological word decomposition and part-of-speech tagging. It determines the best analysis for each word in a sentence and splits word prefixes and suffixes, based on the alternative analyses provided by BAMA (Buckwalter, 2002). We use tokenisation scheme

[1] http://nist.gov/speech/tests/mt/2008

'D2', which splits certain prefixes and has been reported to improve machine translation performance (Habash and Sadat, 2006). The alignment models are trained on this processed data, and the prefixes and suffixes are treated as words in their own right; in particular their contexts are examined and clustered.

The TnT tagger (Brants, 2000), used as distributed with its model trained on the Wall Street Journal portion of the Penn treebank, was used to obtain part-of-speech tags for the English side of the parallel text. Marcus et al. (1993) gives a complete list of part-of-speech tags produced. No morphological analysis is performed for English.

Automatic word alignments were compared to a manually-aligned corpus made up of the IBM Arabic-English Word Alignment Corpus (Ittycheriah et al., 2006) and the word alignment corpora LDC2006E86 and LDC2006E93. This contains 28k parallel text sentences pairs: 724k words of Arabic and 847k words of English. The alignment links were modified to reflect the MADA tokenisation; after modification, there are 946k word-to-word alignment links.

Alignment quality was evaluated by computing Alignment Error Rate (AER) (Och and Ney, 2000) relative to the manual alignments. Since the links supplied contain only 'sure' links and no 'possible' links, we use the following formula for computing AER given reference alignment links S and hypothesised alignment links A: $\text{AER} = 1 - \frac{2|S \cap A|}{|S|+|A|}$.

3.1 Questions about contexts

The algorithm presented in Section 2 allows for any information about the context of the source word to be considered. We could consider general questions of the form *'Is the previous word x?'* and *'Does word y occur within n words of this one?'*. To maintain computational tractability, we restrict the questions to those concerning the part-of-speech tag assigned to the current, previous and next words. We do not ask questions about the identities of the words themselves. For each part-of-speech tag T, we ask the question *'Does w have tag T?'*. In addition, we group part-of-speech tags to ask more general questions: e.g. the set of contexts which satisfies *'Is w a noun?'* contains those that satisfy *'Is w a proper noun?'* and *'Is w a singular or mass noun?'*. We also ask the same questions of the previous and next words in the source sentence. In English, this gives a total of 152 distinct questions, each of which is considered when splitting a leaf node. The MADA part-of-speech tagger uses a reduced tag set, which produces a total of 68 distinct questions.

Figure 1 shows the links of the English source word *selling* in two different contexts where it links to different words in Arabic, which are both forms of the same verb. The part-of-speech of the previous word is useful for discriminating between the two cases, whereas a context-independent model would assign the same probability to both Arabic words.

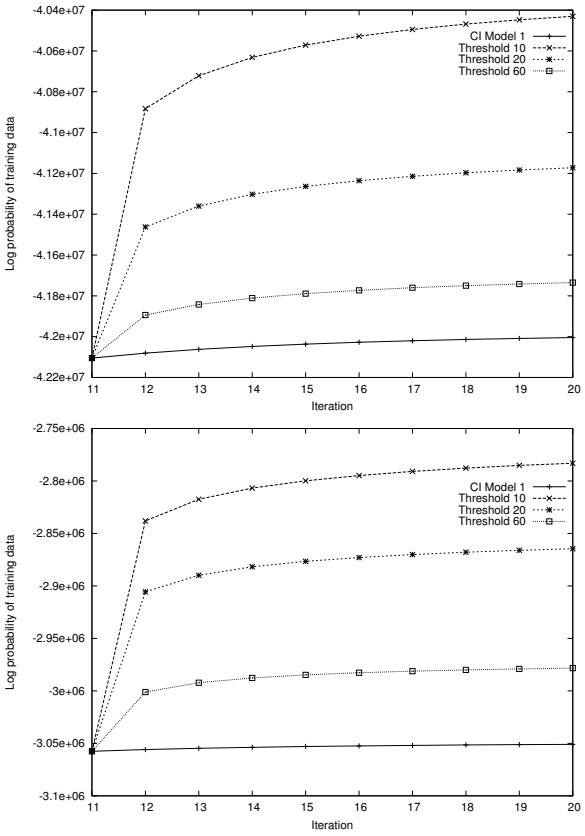

Figure 2: Increase in log probability of training data during training for varying T_{imp}, with Model 1, for Arabic to English (top) and English to Arabic (bottom)

3.2 Training Model 1

Training is carried out in both translation directions. For Arabic to English, the Arabic side of the parallel text is tagged and the English side remains untagged; we view the English words as being generated from the Arabic words and questions are asked about the context of the Arabic words to determine clusters for the translation table. For English to Arabic, the situation is reversed: we used tagged English text as the source language and untagged Arabic text, with morphological decomposition, as the target language.

Standard CI Model 1 training, initialised with a uniform translation table so that $t(e|f)$ is constant for all source/target word pairs (f, e), was run on untagged data for 10 iterations in each direction (Brown et al., 1993; Deng and Byrne, 2005b). A decision tree was built to cluster the contexts and a further 10 iterations of training were carried out using the tagged words-with-context to produce context-dependent models (CD Model 1). The

English question	Frequency
Is_Next_Preposition	1523
Is_Prev_Determiner	1444
Is_Prev_Preposition	1209
Is_Prev_Adjective	864
Is_Next_Noun_Singular_Mass	772
Is_Prev_Noun_Singular_Mass	690
Is_Next_Noun_Plural	597
Is_Next_Noun	549
Arabic question	**Frequency**
Is_Prev_Preposition	1110
Is_Next_Preposition	993
Is_Prev_Noun	981
Is_Next_Noun	912
Is_Prev_Coordinating_Conjunction	627
Is_Prev_Noun_SingularMass	607
Is_Next_Punctuation	603
Is_Next_Adjective_Adverb	559

Table 1: Most frequent root node context questions

models were then evaluated using AER at each training iteration. A number of improvement thresholds T_{imp} were tested, and performance compared to that of models found after further iterations of CI Model 1 training on the untagged data. In both alignment directions, the log probability of the training data increases during training (see Figure 2). As expected, the training set likelihood increases as the threshold T_{imp} is reduced, allowing more clusters and closer fitting to the data.

3.2.1 Analysis of frequently used questions

Table 1 shows the questions used most frequently at the root node of the decision tree when clustering contexts in English and Arabic. Because they are used first, these are the questions that individually give the greatest ability to discriminate between the different contexts of a word. The list shows the importance of the left and right contexts of the word in predicting its translation: of the most common 50 questions, 25 concern the previous word, 19 concern the next, and only 6 concern the part-of-speech of the current word. For Arabic, of the most frequent 50 questions, 21 concern the previous word, 20 concern the next and 9 the current word.

3.2.2 Alignment Error Rate

Since MT systems are usually built on the union of the two sets of alignments (Koehn et al., 2003), we consider the union of alignments in the two directions as well as those in each direction. Figure 3 shows the change in AER of the alignments in each direction, as well as the alignment formed by taking their union at corresponding thresholds and training iterations.

T_{imp}	Arabic-English (%)	English-Arabic (%)
10	30601 (25.33)	26011 (39.87)
20	11193 (9.27)	18365 (28.15)
40	1874 (1.55)	9104 (13.96)
100	307 (0.25)	1128 (1.73)

Table 2: Words [number (percentage)] with context-dependent translation for varying T_{imp}

3.2.3 Variation of improvement threshold T_{imp}

There is a trade-off between modelling the data accurately, which requires more clusters, and eliminating data sparsity problems, which requires each cluster to contain contexts that occur frequently enough in the training data to estimate the translation probabilities accurately. Use of a smaller threshold T_{imp} leads to more clusters per word and an improved ability to fit to the data, but this can lead to reduced alignment quality if there is insufficient data to estimate the translation probability distribution accurately for each cluster. For lower thresholds, we observe over-fitting and the AER rises after the second iteration of CD training, similar to the behaviour seen in Och (2002). Setting $T_{\text{imp}} = 0$ results in each context of a word having its own cluster, which leads to data sparsity problems.

Table 2 shows the percentage of words for which the contexts are split into multiple clusters for CD Model 1 with varying improvement thresholds. This occurs when there are enough training data examples and sufficient variability between the contexts of a word that splitting the contexts into more than one cluster increases the EM auxiliary function. For words where the contexts are not split, all the contexts remain in the same cluster and parameter estimation is exactly the same as for the unclustered context-independent models.

3.3 Training HMMs

Adding source word context to translation has so far led to improvements in AER for Model 1, but the performance does not match that of HMMs trained on untagged data; we therefore train HMMs on tagged data.

We proceed with Model 1 and Model 2 trained in the usual way, and context-independent (CI) HMMs were trained for 5 iterations on the untagged data. Statistics were then gathered for clustering at various thresholds, after which 5 further EM iterations were performed with tagged data to produce context-dependent (CD) HMMs. The HMMs were trained in both the Arabic to English and the English to Arabic directions. The log likelihood of the training set varies with T_{imp} in much the same way as for Model 1, increasing at each iteration, with greater likelihood at lower thresholds. Figure 4 shows how the AER of the union alignment varies with T_{imp} during training. As with Model 1, the clustered HMM

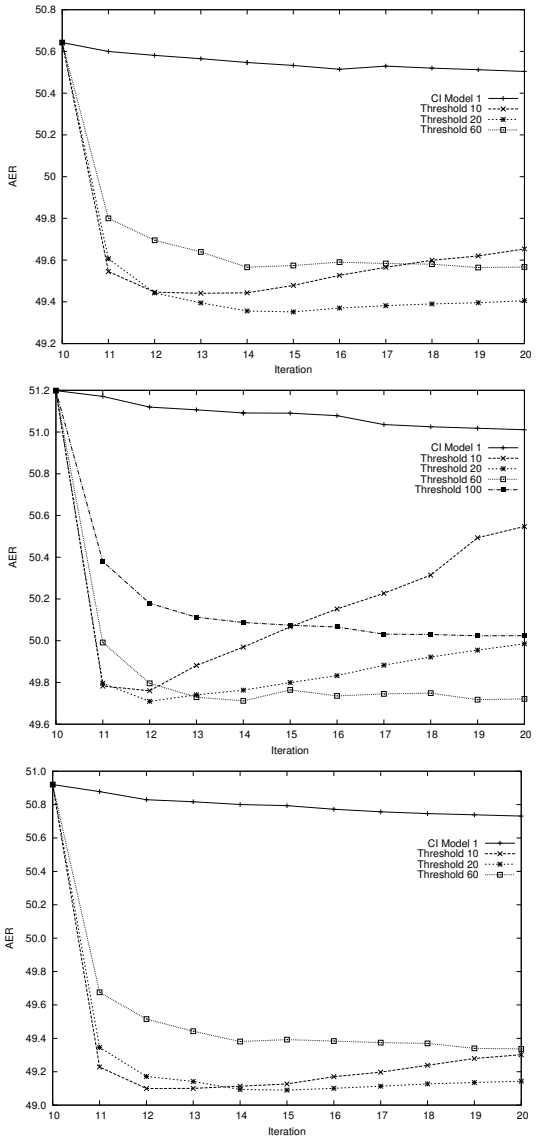

Figure 3: Variation of AER during Model 1 training for varying T_{imp}, for Arabic to English (top), English to Arabic (middle) and their union (bottom)

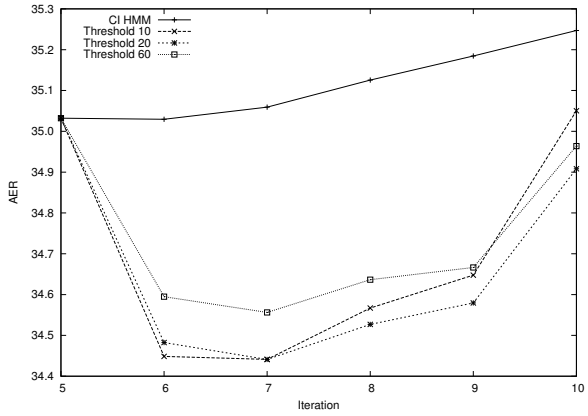

Figure 4: AER of the union alignment for varying T_{imp} with the HMM model

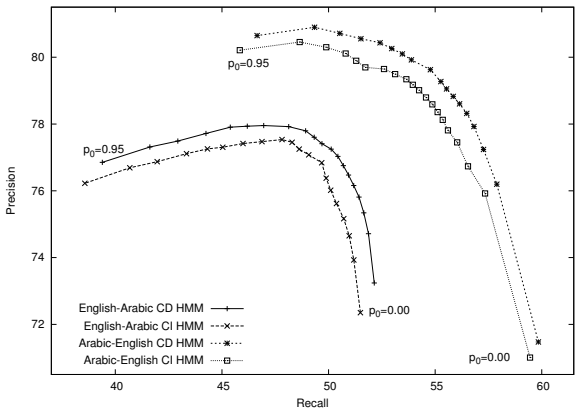

Figure 5: Precision/recall curves for the context-dependent HMM and the baseline context-independent HMM, for Arabic to English and English to Arabic. p_0 varies from 0.00 to 0.95 in steps of 0.05.

models produce alignments with a lower AER than the baseline model, and there is evidence of over-fitting to the training data.

3.3.1 Alignment precision and recall

The HMM models include a *null transition probability*, p_0, which can be modified to adjust the number of alignments to the null token (Deng and Byrne, 2005a). Where a target word is emitted from null, it is not included in the alignment links, so this target word is viewed as not being aligned to any source word; this affects the precision and recall. The results reported above use $p_0 = 0.2$ for English-Arabic and $p_0 = 0.4$ for Arabic-English; we can tune these values to produce alignments with the lowest AER. Figure 5 shows precision-recall curves for the CD HMMs compared to the CI HMMs for both translation directions. For a given value of precision, the CD HMM has higher recall; for a given value of recall, the CD HMM has higher precision.

We do not report F-score (Fraser and Marcu, 2006) since in our experiments we have not found strong correlation with translation performance, but we note that these results for precision and recall should lead to improved F-scores as well.

4 Evaluation of translation quality

We have shown that the context-dependent models produce a decrease in AER measured on manually-aligned data; we wish to show this improved model performance leads to an increase in translation quality, measured by BLEU score (Papineni et al., 2001). In addition to the Arabic systems already evaluated by AER, we also report results for a Chinese-English translation system.

Alignment models were evaluated by aligning the training data using the models in each translation direc-

tion. HiFST, a WFST-based hierarchical translation system described in (Iglesias et al., 2009), was trained on the union of these alignments. MET (Och, 2003) was carried out using a development set, and the BLEU score evaluated on two test sets. Decoding used a 4-gram language model estimated from the English side of the entire MT08 parallel text, and a 965M word subset of monolingual data from the English Gigaword Third Edition.

For both Arabic and English, the CD HMM models were evaluated as follows. Iteration 5 of the CI HMM was used to produce alignments for the parallel text training data: these were used to train the baseline system. The same data is aligned using CD HMMs after two further iterations of training and a second WFST-based translation system built from these alignments. The models are evaluated by comparing BLEU scores with those of the baseline model.

4.1 Arabic to English translation

Alignment models were trained on the NIST MT08 Arabic-English parallel text, excluding the UN portion. The null alignment probability was chosen based on the AER, resulting in values of $p_0 = 0.05$ for Arabic to English and $p_0 = 0.10$ for English to Arabic. We perform experiments on the NIST Arabic-English translation task. The *mt02_05_tune* and *mt02_05_test* data sets are formed from the odd and even numbered sentences of the NIST MT02 to MT05 evaluation sets respectively; each contains 2k sentences and 60k words. We use *mt02_05_tune* as a development set and evaluate the system on *mt02_05_test* and the newswire portion of the MT08 set, *MT08-nw*. Table 3 shows a comparison of the system trained using CD HMMs with the baseline system, which was trained using CI HMM models on untagged data. The context-dependent models result in a gain in BLEU score of 0.3 for *mt02_05_test* and 0.6 for *MT08-nw*.

4.2 Chinese to English translation

The Chinese training set was 600k random parallel text sentences of the newswire LDC collection allowed for NIST MT08, a total of 15.2M words of Chinese and 16.6M words of English. The Chinese text was tagged using the MXPOST maximum-entropy part of speech tagging tool (Ratnaparkhi, 1996) trained on the Penn Chinese Treebank 5.1; the English text was tagged using the TnT part of speech tagger (Brants, 2000) trained on the Wall Street Journal portion of the English Penn treebank.

The development set *tune-nw* and validation set *test-nw* contain a mix of the newswire portions of MT02 through MT05 and additional developments sets created by translation within the GALE program. We also report results on the newswire portion of the MT08 set. Again we see an increase in BLEU score for both test sets: 0.5 for *test-nw* and 0.8 for *MT08-nw*.

Arabic-English			
Alignments	tune	mt02_05_test	MT08-nw
CI HMM	50.0	49.4	46.3
CD HMM	50.0	49.7	46.9
Chinese-English			
Alignments	tune	test-nw	MT08-nw
CI HMM	28.1	28.5	26.9
CD HMM	28.5	29.0	27.7

Table 3: Comparison, using BLEU score, of the CD HMM with the baseline CI HMM

5 Conclusions and future work

We have introduced context-dependent Model 1 and HMM alignment models, which use context information in the source language to improve estimates of word-to-word translation probabilities. Estimation of parameters using these contexts without smoothing leads to data sparsity problems; therefore we have developed decision tree clustering algorithms to cluster source word contexts based on optimisation of the EM auxiliary function. Context information is incorporated by the use of part-of-speech tags in both languages of the parallel text, and the EM algorithm is used for parameter estimation.

We have shown that these improvements to the model lead to decreased AER compared to context-independent models. Finally, we compare machine translation systems built using our context-dependent alignments. For both Arabic- and Chinese-to-English translation, we report an increase in translation quality measured by BLEU score compared to a system built using context-independent alignments.

This paper describes an initial investigation into context-sensitive alignment models, and there are many possible directions for future research. Clustering the probability distributions of infrequently occurring may produce improvements in alignment quality, different model training schemes and extensions of the context-dependence to more sophisticated alignment models will be investigated. Further translation experiments will be carried out.

Acknowledgements

This work was supported in part by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. J. Brunning is supported by a Schiff Foundation graduate studentship. Thanks to Yanjun Ma, Dublin City University, for training the Chinese part of speech tagger.

References

A. L. Berger, S. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Graeme Blackwood, Adrià de Gispert, Jamie Brunning, and William Byrne. 2008. European language translation with weighted finite state transducers: The CUED MT system for the 2008 ACL workshop on SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 131–134, Columbus, Ohio, June. Association for Computational Linguistics.

Thorsten Brants. 2000. TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied Natural Language Processing Conference: ANLP-2000*, Seattle, USA.

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

T. Buckwalter. 2002. Buckwalter Arabic morphological analyzer.

Marta Ruiz Costa-jussà and Jos´e A. R. Fonollosa. 2005. Improving phrase-based statistical translation by modifying phrase extraction and including several features. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 149–154, June.

Yonggang Deng and William Byrne. 2005a. HMM word and phrase alignment for statistical machine translation. In *Proc. of HLT-EMNLP*.

Yonggang Deng and William Byrne. 2005b. JHU-Cambridge statistical machine translation toolkit (MTTK) user manual.

Yonggang Deng, Shankhar Kumar, and William Byrne. 2007. Segmentation and alignment of parallel text for statistical machine translation. *Journal of Natural Language Engineering*, 13:3:235–260.

Alexander Fraser and Daniel Marcu. 2006. Measuring word alignment quality for statistical machine translation. Technical Report ISI-TR-616, ISI/University of Southern California, May.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *HLT-NAACL*.

G. Iglesias, A. de Gispert, E. R. Banga, and W. Byrne. 2009. Hierarchical phrase-based translation with weighted finite state transducers. In *Procedings of NAACL-HLT, 2009*, Boulder, Colorado.

Abraham Ittycheriah, Yaser Al-Onaizan, and Salim Roukos. 2006. The IBM Arabic-English word alignment corpus, August.

A. Kannan, M. Ostendorf, and J. R. Rohlicek. 1994. Maximum likelihood clustering of Gaussians for speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 2(3):453–455, July.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Sonja Nießen and Hermann Ney. 2001a. Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of MT Summit VIII*, pages 247–252, September.

Sonja Nießen and Hermann Ney. 2001b. Toward hierarchical models for statistical machine translation of inflected languages. In *Proceedings of the workshop on Data-driven methods in machine translation*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2000. A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th conference on Computational Linguistics*, pages 1086–1090.

F. Och, D. Gildea, S. Khudanpur, A. Sarkar, K. Yamada, A. Fraser, S. Kumar, L. Shen, D. Smith, K. Eng, V. Jain, Z. Jin, and D. Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proceedings of NAACL*.

Franz Josef Och. 2002. *Statistical Machine Translation: From Single Word Models to Alignment Templates*. Ph.D. thesis, Franz Josef Och.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Maja Popović and Hermann Ney. 2004. Improving word alignment quality using morpho-syntactic information. In *In Proceedings of COLING*, page 310.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

H. Singer and M. Ostendorf. 1996. Maximum likelihood successive state splitting. *Proceedings of ICASSP*, 2:601–604.

Nicolas Stroppa, Antal van den Bosch, and Andy Way. 2007. Exploiting source similarity for SMT using context-informed features. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, pages 231 – 240.

Kristina Toutanova, H. Tolga Ilhan, and Christopher D. Manning. 2002. Extensions to HMM-based statistical word alignment models. In *Proceedings of EMNLP*, pages 87–94.

Ismael García Varea, Franz J. Och, Hermann Ney, and Francisco Casacuberta. 2002. Improving alignment quality in statistical machine translation using context-dependent maximum entropy models. In *Proceedings of COLING*, pages 1–7.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING*, pages 836–841.

S. J. Young, J. J. Odell, and P. C. Woodland. 1994. Tree-based state tying for high accuracy acoustic modelling. In *HLT '94: Proceedings of the workshop on Human Language Technology*, pages 307–312.