Domain Adaptation in Statistical Machine Translation with Mixture Modelling *

Jorge Civera and **Alfons Juan**
Universidad Politécnica de Valencia
Camino de Vera s/n
46022 Valencia, Spain
{jorcisai,ajuan}@iti.upv.es

Abstract

Mixture modelling is a standard technique for density estimation, but its use in statistical machine translation (SMT) has just started to be explored. One of the main advantages of this technique is its capability to learn specific probability distributions that better fit subsets of the training dataset. This feature is even more important in SMT given the difficulties to translate polysemic terms whose semantic depends on the context in which that term appears. In this paper, we describe a mixture extension of the HMM alignment model and the derivation of Viterbi alignments to feed a state-of-the-art phrase-based system. Experiments carried out on the Europarl and News Commentary corpora show the potential interest and limitations of mixture modelling.

1 Introduction

Mixture modelling is a popular approach for density estimation in many scientific areas (G. J. McLachlan and D. Peel, 2000). One of the most interesting properties of mixture modelling is its capability to model multimodal datasets by defining soft partitions on these datasets, and learning specific probability distributions for each partition, that better explains the general data generation process.

In Machine Translation (MT), it is common to encounter large parallel corpora devoted to heterogeneous topics. These topics usually define sets of topic-specific lexicons that need to be translated taking into the semantic context in which they are found. This semantic dependency problem could be overcome by learning topic-dependent translation models that capture together the semantic context and the translation process.

However, there have not been until very recently that the application of mixture modelling in SMT has received increasing attention. In (Zhao and Xing, 2006), three fairly sophisticated bayesian topical translation models, taking IBM Model 1 as a baseline model, were presented under the bilingual topic admixture model formalism. These models capture latent topics at the document level in order to reduce semantic ambiguity and improve translation coherence. The models proposed provide in some cases better word alignment and translation quality than HMM and IBM models on an English-Chinese task. In (Civera and Juan, 2006), a mixture extension of IBM model 2 along with a specific dynamic-programming decoding algorithm were proposed. This IBM-2 mixture model offers a significant gain in translation quality over the conventional IBM model 2 on a semi-synthetic task.

In this work, we present a mixture extension of the well-known HMM alignment model first proposed in (Vogel and others, 1996) and refined in (Och and Ney, 2003). This model possesses appealing properties among which are worth mentioning, the simplicity of the first-order word alignment distribution that can be made independent of absolute positions while

* Work supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01, the *Consellería d'Empresa, Universitat i Ciència - Generalitat Valenciana* under contract GV06/252, the *Universidad Politécnica de Valencia* with ILETA project and Ministerio de Educación y Ciencia.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 177–180,
Prague, June 2007.

taking advantage of the localization phenomenon of word alignment in European languages, and the efficient and exact computation of the E-step and Viterbi alignment by using a dynamic-programming approach. These properties have made this model suitable for extensions (Toutanova et al., 2002) and integration in a phrase-based model (Deng and Byrne, 2005) in the past.

2 HMM alignment model

Given a bilingual pair (x, y), where x and y are mutual translation, we incorporate the hidden variable $a = a_1 a_2 \cdots a_{|x|}$ to reveal, for each source word position j, the target word position $a_j \in \{0, 1, \ldots, |y|\}$ to which it is connected. Thus,

$$p(x \mid y) = \sum_{a \in \mathcal{A}(x,y)} p(x, a \mid y) \tag{1}$$

where $\mathcal{A}(x, y)$ denotes the set of all possible alignments between x and y. The *alignment-completed* probability $P(x, a \mid y)$ can be decomposed in terms of source position-dependent probabilities as:

$$p(x, a \mid y) = \prod_{j=1}^{|x|} p(a_j \mid a_1^{j-1}, x_1^{j-1}, y)\, p(x_j \mid a_1^j, x_1^{j-1}, y) \tag{2}$$

The original formulation of the HMM alignment model assumes that each source word is *connected to exactly one* target word. This connection depends on the target position to which was aligned the previous source word and the length of the target sentence. Here, we drop both dependencies in order to simplify to a jump width alignment probability distribution:

$$p(a_j \mid a_1^{j-1}, x_1^{j-1}, y) \approx \begin{cases} p(a_j) & j = 1 \\ p(a_j - a_{j-1}) & j > 1 \end{cases} \tag{3}$$

$$p(x_j \mid a_1^j, x_1^{j-1}, y) \approx p(x_j \mid y_{a_j}) \tag{4}$$

Furthermore, the treatment of the NULL word is the same as that presented in (Och and Ney, 2003).

Finally, the HMM alignment model is defined as:

$$p(x \mid y) = \sum_{a \in \mathcal{A}(x,y)} p(a_1) \prod_{j=2}^{|x|} p(a_j - a_{j-1}) \prod_{j=1}^{|x|} p(x_j | y_{a_j}) \tag{5}$$

3 Mixture of HMM alignment models

Let us suppose that $p(x \mid y)$ has been generated using a T-component mixture of HMM alignment models:

$$\begin{aligned} p(x \mid y) &= \sum_{t=1}^{T} p(t \mid y)\, p(x \mid y, t) \\ &= \sum_{t=1}^{T} p(t \mid y) \sum_{a \in \mathcal{A}(x,y)} p(x, a \mid y, t) \end{aligned} \tag{6}$$

In Eq. 6, we introduce mixture coefficients $p(t \mid y)$ to weight the contribution of each HMM alignment model in the mixture. While the term $p(x, a \mid y, t)$ is decomposed as in the original HMM model.

The assumptions of the constituent HMM models are the same than those of the previous section, but we obtain topic-dependent statistical dictionaries and word alignments. Apropos of the mixture coefficients, we simplify these terms dropping its dependency on y, leaving as future work its inclusion in the model. Formally, the assumptions are:

$$p(t \mid y) \approx p(t) \tag{7}$$

$$p(a_j \mid a_1^{j-1}, x_1^{j-1}, y, t) \approx \begin{cases} p(a_j \mid t) & j = 1 \\ p(a_j - a_{j-1} \mid t) & j > 1 \end{cases} \tag{8}$$

$$p(x_j \mid a_1^j, x_1^{j-1}, y, t) \approx p(x_j \mid y_{a_j}, t) \tag{9}$$

Replacing the assumptions in Eq. 6, we obtain the (incomplete) HMM mixture model as follows:

$$\begin{aligned} p(x \mid y) = \sum_{t=1}^{T} p(t) \sum_{a \in \mathcal{A}(x,y)} p(a_1 \mid t) \times \\ \times \prod_{j=2}^{|x|} p(a_j - a_{j-1} \mid t) \prod_{j=1}^{|x|} p(x_j | y_{a_j}, t) \end{aligned} \tag{10}$$

and the set of unknown parameters comprises:

$$\vec{\Theta} = \begin{cases} p(t) & t = 1 \ldots T \\ p(i \mid t) & j = 1 \\ p(i - i' \mid t) & j > 1 \\ p(u \mid v, t) & \forall u \in \mathcal{X} \text{ and } v \in \mathcal{Y} \end{cases} \tag{11}$$

$\mathcal{X}$ and $\mathcal{Y}$, being the source and target vocabularies.

The estimation of the unknown parameters in Eq. 10 is troublesome, since topic and alignment data are missing. Here, we revert to the EM optimisation algoritm to compute these parameters.

In order to do that, we define the complete version of Eq. 10 incorporating the indicator variables z_t and z_a, uncovering, the until now hidden variables. The variable z_t is a T-dimensional bit vector with 1 in the position corresponding to the component generating (x, y) and zeros elsewhere, while the variable $z_a = z_{a_1} \dots z_{a_{|x|}}$ where z_{a_j} is a $|y|$-dimensional bit vector with 1 in the position corresponding to the target position to which position j is aligned and zeros elsewhere. Then, the complete model is:

$$p(x, z_t, z_a \mid y) \approx \prod_{t=1}^{T} p(t)^{z_t} \prod_{i=1}^{|y|} p(i \mid t)^{z_{a_{1i}} z_t} \times \\ \times \prod_{j=1}^{|x|} \prod_{i=1}^{|y|} p(x_j \mid y_i, t)^{z_{a_{ji}} z_t} \prod_{i'=1}^{|y|} p(i - i' \mid t)^{z_{a_{j-1i'}} z_{a_{ji}} z_t} \tag{12}$$

Given the complete model, the EM algorithm works in two basic steps in each iteration: the E(xpectation) step and the M(aximisation) step. At iteration k, the E step computes the expected value of the hidden variables given the observed data (x, y) and the estimate of the parameters $\vec{\Theta}^{(k)}$.

The E step reduces to the computation of the expected value of z_t, $z_{a_{ji}} z_t$ and $z_{a_{j-1i'}} z_{a_{ji}} z_t$ for each sample n:

$$z_t \propto p(t) \sum_{i=1}^{|y|} \alpha_{|x|it} \tag{13}$$

$$z_{a_{ji}} z_t = z_{a_{ji}t}\, z_t \tag{14}$$

$$z_{a_{j-1i'}} z_{a_{ji}} z_t = (z_{a_{j-1i'}} z_{a_{ji}})_t\, z_t \tag{15}$$

where

$$z_{a_{ji}t} \propto \sum_{k=1}^{|y|} \alpha_{jkt} \beta_{jkt}$$

$$(z_{a_{j-1i'}} z_{a_{ji}})_t \propto \alpha_{j-1it}\, p(i - i' \mid t)\, p(x_j | y_i, t)\, \beta_{jit}$$

and the recursive functions α and β defined as:

$$\alpha_{jit} = \begin{cases} p(i \mid t)\, p(x_j \mid y_i, t) & j = 1 \\ \sum_{k=1}^{|y|} \alpha_{j-1kt}\, p(i - k \mid t)\, p(x_j \mid y_i, t) & j > 1 \end{cases}$$

$$\beta_{jit} = \begin{cases} 1 & j = |x| \\ \sum_{k=1}^{|y|} p(k - i \mid t)\, p(x_{j+1} \mid y_k, t) \beta_{j+1kt} & j < |x| \end{cases}$$

The M step finds a new estimate of $\vec{\Theta}$, by maximising Eq. 12, using the expected value of the missing data from Eqs. 13,14 and 15 over all sample n:

$$p(t) = \frac{1}{N} \sum_{n=1}^{N} z_{nt}$$

$$p(i \mid t) \propto \sum_{n=1}^{N} z_{na_{1i}t}$$

$$p(i - i' \mid t) \propto \sum_{n=1}^{N} \sum_{j=1}^{|x_n|} (z_{na_{j-1i'}} z_{na_{ji}})_t$$

$$p(u \mid v, t) \propto \sum_{n=1}^{N} \sum_{j=1}^{|x_n|} \sum_{i=1}^{|y_n|} z_{na_{ji}t}\, \delta(x_{nj}, u) \delta(y_{ni}, v)$$

3.1 Word alignment extraction

The HMM mixture model described in the previous section was used to generate Viterbi alignments on the training dataset. These optimal alignments are the basis for phrase-based systems.

In the original HMM model, the Viterbi alignment can be efficiently computed by a dynamic-programming algorithm with a complexity $O(|x| \cdot |y|^2)$. In the mixture HMM model, we approximate the Viterbi alignment by maximising over the components of the mixture:

$$\hat{a} \approx \arg\max_a \max_t p(t)\, p(x, a \mid y, t)$$

So we have that the complexity of the computation of the Viterbi alignment in a T-component HMM mixture model is $O(T \cdot |x| \cdot |y|^2)$.

4 Experimental results

The data that was employed in the experiments to train the HMM mixture model corresponds to the concatenation of the Spanish-English partitions of the Europarl and the News Commentary corpora. The idea behind this decision was to let the mixture model distinguish which bilingual pairs should contribute to learn a given HMM component in the mixture. Both corpora were preprocessed as suggested for the baseline system by tokenizing, filtering sentences longer than 40 words and lowercasing.

Regarding the components of the translation system, 5-gram language models were trained on the monolingual version of the corpora for English(En) and Spanish(Es), while phrase-based models with lexicalized reordering model were trained using the Moses toolkit (P. Koehn and others, 2007), but replacing the Viterbi alignments, usually provided by GIZA++ (Och and Ney, 2003), by those of the HMM mixture model with training scheme $mix\,1^5 H^5$. This configuration was used to translate both test development sets, Europarl and News Commentary.

Concerning the weights of the different models, we tuned those weights by minimum error rate training and we employed the same weighting scheme for all the experiments in the same language pair. Therefore, the same weighting scheme was used over different number of components.

BLEU scores are reported in Tables 1 and 2 as a function of the number of components in the HMM mixture model on the preprocessed development test sets of the Europarl and News Commentary corpora.

Table 1: BLEU scores on the Europarl development test data

T	1	2	3	4
En-Es	31.27	31.08	31.12	31.11
Es-En	31.74	31.70	31.80	31.71

Table 2: BLEU scores on the News-Commentary development test data

T	1	2	3	4
En-Es	29.62	30.01	30.17	29.95
Es-En	29.15	29.22	29.11	29.02

As observed in Table 1, if we compare the BLEU scores of the conventional single-component HMM model to those of the HMM mixture model, it seems that there is little or no gain from incorporating more topics into the mixture for the Europarl corpus. However, in Table 2, the BLEU scores on the English-Spanish pair significantly increase as the number of components is incremented. We believe that this is due to the fact that the News Commentary corpus seems to have greater influence on the mixture model than on the single-component model, specializing Viterbi alignments to favour this corpus.

5 Conclusions and future work

In this work, a novel mixture version of the HMM alignment model was introduced. This model was employed to generate topic-dependent Viterbi alignments that were input into a state-of-the-art phrase-based system. The preliminary results reported on the English-Spanish partitions of the Europarl and News-Commentary corpora may raise some doubts about the applicability of mixture modelling to SMT, nonetheless in the advent of larger open-domain corpora, the idea behind topic-specific translation models seem to be more than appropriate, necessary. On the other hand, we are fully aware that indirectly assessing the quality of a model through a phrase-based system is a difficult task because of the different factors involved (Ayan and Dorr, 2006).

Finally, the main problem in mixture modelling is the linear growth of the set of parameters as the number of components increases. In the HMM, and also in IBM models, this problem is aggravated because of the use of statistical dictionary entailing a large number of parameters. A possible solution is the implementation of interpolation techniques to smooth sharp distributions estimated on few events (Och and Ney, 2003; Zhao and Xing, 2006).

References

N. F. Ayan and B. J. Dorr. 2006. Going beyond AER: an extensive analysis of word alignments and their impact on MT. In *Proc. of ACL'06*, pages 9–16.

J. Civera and A. Juan. 2006. Mixtures of IBM Model 2. In *Proc. of EAMT'06*, pages 159–167.

Y. Deng and W. Byrne. 2005. HMM word and phrase alignment for statistical machine translation. In *Proc. of HLT-EMNLP'05*, pages 169–176.

G. J. McLachlan and D. Peel. 2000. *Finite Mixture Models*. Wiley.

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

P. Koehn and others. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of ACL'07 Demo Session*, page To be published.

K. Toutanova, H. T. Ilhan, and C. D. Manning. 2002. Extensions to HMM-based statistical word alignment models. In *Proc. of EMNLP '02*, pages 87–94.

S. Vogel et al. 1996. HMM-based word alignment in statistical translation. In *Proc. of CL*, pages 836–841.

B. Zhao and E. P. Xing. 2006. BiTAM: Bilingual Topic AdMixture Models for Word Alignment. In *Proc. of COLING/ACL'06*.