Extended Translation Models in Phrase-based Decoding

Andreas Guta, Joern Wuebker, Miguel Graça, Yunsu Kim, Hermann Ney
Human Language Technology and Pattern Recognition Group
RWTH Aachen University
Aachen, Germany
{surname}@cs.rwth-aachen.de

Abstract

We propose a novel extended translation model (ETM) to counteract some problems in phrase-based translation: The lack of translation context when using single-word phrases and uncaptured dependencies beyond phrase boundaries. The ETM operates on word-level and augments the IBM models by an additional bilingual word pair and a reordering operation. Its implementation in a phrase-based decoder introduces translation and reordering dependencies for single-word phrases and dependencies across phrase boundaries. More, the model incorporates an explicit treatment of multiple and empty alignments. Its integration outperforms competitive systems that include lexical and phrase translation models as well as hierarchical reordering models on 4 language pairs significantly by +0.7% BLEU on average. Although simpler and using fewer dependencies, the ETM proves to be on par with 7-gram operation sequence models (Durrani et al., 2013b).

1 Introduction

The first successful steps in Statistical Machine Translation have been taken by applying word-based models in a source-channel approach (Brown et al., 1990; Brown et al., 1993). Within this framework, the language model (LM) is estimated on monolingual n-grams, whereas the translation models IBM-1 to IBM-5 are trained on bilingual data using word alignments. The disadvantage of word-to-word translation is overcome by phrase-based translation (PBT) (Och et al., 1999; Zens et al., 2002; Koehn et al., 2003) and log-linear model combination (Och and Ney, 2002).

Nevertheless, phrase-based translation models have several drawbacks: (i) Single-word phrases are translated without any context. (ii) Dependencies beyond phrase boundaries are not modelled at all. (iii) Phrase-based translation models have difficulties modelling long-distance dependencies on source words with large gaps inbetween.

The open question is how much actual lexical context is included in decoding. Figure 1 depicts the relative word frequencies plotted against the length of the phrase they were decoded with for the IWSLT dev2010 German→English and English→French corpora.

Figure 1: Relative frequency of words dependent on the length of the phrase they were decoded with for the IWSLT dev2010 German→English and English→French corpora.

Nevertheless, phrase-based translation models have several drawbacks: (i) Single-word phrases are translated without any context. (ii) Dependencies beyond phrase boundaries are not modelled at all. (iii) Phrase-based translation models have difficulties modelling long-distance dependencies on source words with large gaps inbetween.

The open question is how much actual lexical context is included in decoding. Figure 1 depicts the relative word frequencies plotted against the length of the phrase they were translated with for the IWSLT 20141 German→English and English→French tasks. For English→French, more than 40% of the words are translated using single- or two-word phrases, i.e. with a lexical context of at most one word. For the German→English task, more reorderings occur and lead to less monotone alignments. Here, even 60% of all words are translated with a lexical context of at most one single word and over 20% are translated without any lexical context at all.

1http://www.iwslt2014.org
We address this problem by developing two variants of extended translation models (ETM), the direct \( \text{(EdTM)} \) for the Source\( \rightarrow \)Target and the inverse \( \text{(EiTM)} \) for the Target\( \rightarrow \)Source direction. They operate on word-level and augment the IBM models by an additional bilingual word pair and a reordering operation. We introduce them into the log-linear framework of a PBT system. Thus, the decoding of single-word phrases can benefit from lexical and reordering context. Moreover, the ETM allows to capture dependencies across phrase boundaries and long-range source dependencies. It incorporates reordering information for non-monotone and multiple alignments including unaligned words.

As a first step, we implement the ETM as a count model with interpolated Kneser-Ney smoothing \cite{chen1998interpolating} using the Viterbi alignment and apply it in phrase-based decoding. Nevertheless, the long-term goal of this approach is to replace the phrases used in decoding by translation units that predict a single target word, but may depend on several source words, previously translated target words and the reordering context.

2 Previous Work

Various approaches have been taken to compensate the downside of the phrase translation model. Mariño et al. \cite{marcaio2006} introduce a translation model based on \( n \)-grams of bilingual word pairs, i.e. a bilingual language model (BILM), with an \( n \)-gram decoder that requires monotone alignments. In \cite{niehues2011}, this is further advanced by BILMs operating on non-monotone alignments within a PBT framework.

However, this differs from our approach: BILMs treat jointly aligned source words as atomic units, ignore source deletions and do not include reordering context.

The Operation Sequence Model (OSM) introduced in \cite{durrani2011,durrani2013} includes \( n \)-grams of both translation and reordering operations in a consistent framework. It utilizes minimal translation units (MTUs) and is applied in a corresponding OSM decoder. Experiments in \cite{durrani2013b} show that a slightly enhanced version of OSM performs best when integrated into the log-linear framework of a phrase-based decoder. Both the BILM \cite{stewart2014} and the OSM \cite{durrani2014} can be smoothed using word clusters.

In comparison, the ETM is much simpler: Since it predicts probabilities of single words, it has a lower vocabulary size. More, it does not make use of reordering gaps, i.e. it utilizes a simpler reordering approach. The OSM uses one joint model for reorderings and translations. In contrast, the ETM incorporates separate models to estimate the probability of words and the probability of reorderings. Furthermore, the OSM has the drawback that it extracts the MTUs sentence-wise, thus one word can appear in several MTUs extracted from different sentence pairs. Since an MTUs is treated as an atomic unit, this results in a distribution of probability mass on overlapping events. The ETM overcomes this drawback by operating on single words.

Guta et al. \cite{guta2015} propose the conversion of bilingual sentence pairs and word alignments into joint translation and reordering (JTR) sequences. They investigate \( n \)-gram models with modified Kneser-Ney smoothing, feed-forward and recurrent neural networks trained on JTR sequences. In comparison to the OSM, JTR models have smaller vocabulary sizes, as they operate on words, and incorporate simpler reordering structures. Nevertheless, they are shown to perform slightly better than the OSM when included into the log-linear framework of a phrase-based decoder.

Although our approach is similar, there are the following significant differences: On the one hand, the ETM estimates the probability of single words conditioned on an extended lexical and reordering context, whereas the JTR \( n \)-gram model predicts the probability of bilingual word pairs. On the other hand, we do not assume linear sequences of dependencies, but propose and explicit treatment of multiply aligned words.

Deng and Byrne \cite{deng2005} present an HMM approach for word-to-phrase alignments, which performs similar to IBM-4 on the task of bitext alignment and can also be applied for more powerful phrase induction. Feng et al. \cite{feng2013} introduce an reordering model based on sequence labeling techniques by converting the reordering problem into a tagging task. Zhang et al. \cite{zhang2013} explore different Markov chain orderings for an \( n \)-gram model on MTUs. These are not integrated into decoding, but used in \( N \)-best rescoring. Another generative, word-based Markov chain translation model is presented by Feng and Cohn \cite{feng2013}. It exploits
a hierarchical Pitman-Yor process for smoothing, but is only applied to induce word alignments. Their follow-up work (Feng et al., 2014) introduces a Markov-model on MTUs, similar to the OSM described above.

Finally, there has been recent research on applying neural network models for extended context (Le et al., 2012; Auli et al., 2013; Hu et al., 2014; Devlin et al., 2014; Sundermeyer et al., 2014). All of these papers focus on lexical context and ignore the reordering aspect covered in our work.

3 Extended Translation Models

Given a source sentence $f_i^1$ and its translation $e_i^1$, EiTM models the inverse probability $p(f_i^1|e_i^1)$ and EdTM the direct probability $p(e_i^1|f_i^1)$. We allow for source words to be translated to multiple target words and vice versa. The inverted alignment $b_i$ denotes the sequence of source positions $j$ aligned to target position $i$ for $i = 1, \ldots, I$. Its subsequence $b_i^{<j}$ includes all source positions in $b_i$ preceding a given source position $j$:

$$b_i^{<j} = \{ j \in b_i : j < j \}.$$ 

Unaligned target words are aligned to the empty source word $f_0$, unaligned source words to the empty target word $e_0$. $b_0$ denotes the unaligned source positions. We introduce the fertility $\phi_i$ of a target word $e_i$. It determines the number of source words aligned to the target word $e_i$:

$$\phi_i = \begin{cases} 0, & b_i = \{0\} \\ |b_i|, & \text{else} \end{cases}$$

By analogy, we use $\phi_i^{<j}$ to denote the number of source positions in $b_i^{<j}$. Similar to the approach in (Feng and Cohn, 2013), we generalize reorderings to the following jump classes $\Delta_{j,i}^\phi$:

$$\Delta_{j,i}^\phi = \begin{cases} \downarrow \text{ (‘insert’)}, & \phi_i = 0 \\ \cdot & \phi_i > 0, j = j' \\ \rightarrow \text{ (‘forward’)}, & \phi_i > 0, j = j' + 1 \\ \leftarrow \text{ (‘backward’)}, & \phi_i > 0, j = j' - 1 \\ \leftarrow \text{ (‘jump backward’)}, & \phi_i > 0, j < j' - 1. \end{cases}$$

Figure 2 outlines the jump classes for subsequent target positions $i'$ and $i$. As shown in Figure 3, for source positions $j' < j$ which are aligned to the same target position $i$, there are two possible jump classes:

$$\Delta_{j,i} = \begin{cases} \rightarrow \text{ (‘step forward’)}, & j = j' + 1 \\ \leftarrow \text{ (‘jump forward’)}, & j > j' + 1. \end{cases}$$

In the following, we depict the derivations of the EiTM and the EdTM. Although they operate in opposite translation directions, both models incorporate the inverted alignment $b_i$.

3.1 Extended Inverse Translation Model

In order to model the inverse probability $p(f_i^1|e_i^1)$, the unknown inverted alignment $b_i$ is introduced as a hidden variable and approximated by the Viterbi alignment.

$$p(f_i^1|e_i^1) = \max_{b_i} \left\{ p(f_i^1, b_i|e_i^1) \right\} = \max_{b_i} \left\{ p(f_i^1, b_i|e_i^1) \times p(f_i^1, b_i, e_i^1) \right\}$$

The inverse probability has been decomposed into the deletion probability $p(f_i^1, b_i|e_i^1)$ and the joint probability $p(f_i^1, b_i|e_i^1)$. The latter is reformulated using the Markov chain rule:

$$p(f_i^1, b_i|e_i^1) = \prod_{i=1}^{I} p(f_i^1, b_i|e_i^1, f_{i-1}, b_{i-1}).$$

In order to restrict the history, we assume the probability of $(f_i, b_i)$ to be dependent only on the current target word $e_i$, its last aligned predecessor $e_{i-1}$, the corresponding alignment $b_{i-1}$ and the source words $f_{i-1}$:

$$p(f_i^1, b_i|e_i^1, f_{i-1}, b_{i-1}) = \prod_{i=1}^{I} p(f_i^1, b_i|e_i, e_{i-1}, f_{i-1}, b_{i-1}).$$

The conditional joint probability is factorized as

$$p(f_i^1, b_i|e_i, e_{i-1}, f_{i-1}, b_{i-1}) = p(f_i^1|e_i, e_{i-1}, f_{i-1}, b_{i-1}) \times p(b_i|e_i, e_{i-1}, f_{i-1}, b_{i-1}).$$
resulting in the lexicon probability of \( f_{b_1} \) and the alignment probability of \( b_1 \). In a nutshell, we have decomposed the inverse probability into the following three probabilities:

- deletion: \( p(f_{b_0}\mid b_{b_1}', b_1', e_1') \)
- lexicon: \( \prod_{i=1}^{j} p(f_{b_i}\mid e_i', e_i, f_{b_j}, b_j, b_1) \)
- alignment: \( \prod_{i=1}^{j} p(b_i\mid e_i', e_i, f_{b_j}, b_j) \)

Below, we show how to estimate these probabilities using the EiTM deletion, lexicon and alignment models.

### 3.1.1 EiTM: Deletion Model

Due to its artificiality, \( e_0 \) has no preceding target word. We condition the deletion of \( f_{b_0} \) only on \( e_0 \) and assume conditional independence between the unaigned source words \( f_{b_0} \):

\[
p(f_{b_0}\mid b_{b_1}', b_1', e_1') = \prod_{j\notin b_0} p(f_j\mid e_0).
\]

### 3.1.2 EiTM: Lexicon Model

Firstly, we apply the Markov chain rule to obtain the factorized probabilities of single words \( f_j \).

\[
p(f_{b_i}\mid e_i', e_i, f_{b_j}, b_j, b_1) = \prod_{j\in b_i} p(f_j\mid e_i', e_i, f_{b_j}, b_j, b_1)
\]

Each source word \( f_j \) is dependent on all predecessors \( f_{b_i} \) aligned to the same target word \( e_i \) and all previously aligned source words \( f_{b_j} \). If we modelled the probability conditioned on the sets of source words \( f_{b_i} \) and \( f_{b_j} \), this would lead to sparsity problems due to the arbitrary number of source words contained in the sets.

In order to avoid this, we therefore condition the probability on the individual words contained in \( f_{b_i}, f_{b_j} \). Without any additional information, we assume all words \( f_{b_i}, f_{b_j} \) to be equally important for the prediction of \( f_j \). Thus, we average over the probabilities conditioned on:

- all source words \( f_j \) aligned to the preceding target word \( e_i' \),
- all preceding source words \( f_j \) aligned to the current target word \( e_i \).

Moreover, we reduce the alignments \( (b_j, b_i) \) to their corresponding jump classes. As a final result we obtain:

\[
p(f_j\mid e_i', e_i, f_{b_j}, f_{b_i}<, b_j, b_1) =
\frac{1}{\phi_i + \phi_j} \left( \sum_{j\in b_i} p(f_j\mid e_i, f_j, e_i', \Delta_j^b) + \sum_{j\in b_i} p(f_j\mid e_i, f_j, \Delta_j^1) \right).
\]

### 3.1.3 EiTM: Alignment Model

In principle, we follow the same derivation as for the lexicon model above. The probability of a source position \( j \in b_i \) is computed as the average probability of a jump from a previously aligned source position, which either has to be aligned to the target predecessor \( i' \) or is a preceding source position aligned to the same target word \( e_i \).

\[
p(b_i\mid e_i, f_{b_j}, b_j) =
\prod_{j\in b_i} \frac{1}{\phi_i + \phi_j} \left( \sum_{j\in b_j} p(\Delta_j^b\mid e_i, f_j, e_i') + \sum_{j\in b_j} p(\Delta_j^1\mid e_i, f_j) \right).
\]

To emphasize the core idea, Figure 4 demonstrates the application on a German→English translation example. Thin blue arcs denote the probabilities conditioned on distinct target words \( e_i' \) and \( e_i \), the thick red arc denotes the probabilities conditioned on a previous source word \( f_j \) aligned to the current target word. The shape of an arc symbolizes
the jump class, see Figures 2 and 3. The empty words are shown at positions $j, i = 0$. The deletion is indicated by a violet circle. The EiTM probability for the whole sentence pair is computed as follows:

$$p(f_1^b | e_1^a) =$$

$$p(f_1 | e_1, s, s, s, s) \cdot p(\neg | e_1, s, s, s)$$

$$\cdot p(f_2 | e_2, f_1, e_1, \cdot) \cdot p(\neg | e_2, f_1, e_1)$$

$$\cdot p(f_3 | e_3, f_2, e_2, \cdot) + p(f_0 | e_3, f_2, e_2, \cdot)$$

$$\cdot p(\neg | e_2, f_1, e_1) + p(\neg | e_2, f_5)$$

$$\cdot p(f_4 | e_4, f_3, f_2, e_2) + p(\neg | e_3, f_6, e_2)$$

$$\cdot p(f_4 | e_4, f_5, e_3, \cdot) \cdot p(\neg | e_4, f_5, e_3)$$

$$\cdot p(f_0 | e_5, f_7, e_4, \cdot) \cdot p(\neg | e_5, f_7, e_4)$$

$$\cdot p(f_2 | e_6, f_5, e_4, \cdot) \cdot p(\neg | e_6, f_5, e_4)$$

$$\cdot p(f_1 | e_7, f_6, e_5, \cdot) \cdot p(\neg | e_7, f_6, e_5)$$

$$\cdot p(f_3 | e_8, f_4, e_7, \cdot) \cdot p(\neg | e_8, f_4, e_7)$$

$$\cdot p(f_0 | e_9, f_4, e_8, \cdot) \cdot p(\neg | e_9, f_4, e_8)$$

$$\cdot p(\neg | s, s, s, s) \cdot f_0, e_9, \cdot) \cdot p(\neg | s, s, s, s) \cdot f_0, e_9)$$

$$\cdot p(f_3 | e_0).$$

Lines (1), (2), (3) and (7) are dependencies included in the EiTM but not in phrase translation models due to the phrase extraction heuristics. The dependency on multiple preceding word pairs is exemplified in (2) and (3). (4) depicts the insertion of the target word $e_5 = \text{up}$ conditioned on the word pair $(e_4 = \text{us}, f_7 = \text{uns})$. Note that in (5) there is no dependency of $e_6 = \text{is}$ on its predecessor $e_5 = \text{up}$ and the empty word $f_0$, but on its last aligned predecessor $e_4 = \text{us}$ and the corresponding source word $f_7 = \text{uns}$. (6) shows an example of a source word aligned to multiple target words. The deletion probability of the source word $f_3 = \text{es}$ is presented in (8).

### 3.2 Extended Direct Translation Model

So far, we have introduced the EiTM, which models the *inverse* translation probability $p(f_1^b | e_1^a)$. Besides modelling $p(f_1^b | e_1^a)$ using extended translation models, our aim is to employ them to model the *direct* probability $p(e_1^a | f_1^b)$ as well.

For a start, the direct probability $p(e_1^a | f_1^b)$ can be modelled using the EiTM: Simply put, source and target corpora have to be swapped for the training of the EiTM. By doing so, the alignment has to be inverted as well, i.e. one has to use the direct alignment $a_j$ which denotes the sequence of target positions $i$ aligned to source position $j$. As a result, the EiTM models $p(a_0^j, a_1^j | f_1^i)$ when trained with inverted corpora and alignments.

During the decoding process, the partial hypotheses are generated successively. Thus, for each target word $e_i$ that is hypothesized, all its predecessors have already been translated, i.e. its last aligned predecessor $e_f$ and the corresponding alignment $a_f$ and source words $f_a$ are known.

Nevertheless, source words do not have to be translated in monotone order. In general, it cannot be guaranteed that the predecessor $f_{j-1}$ of the first word $f_j$ of a source phrase has been translated yet. Therefore, the last aligned predecessor of $f_j$ and its aligned target words are generally unknown.

As a result, when applying the EiTM within phrase-based decoding for modelling the direct probability $p(e_1^a | f_1^b)$, dependencies beyond phrase boundaries cannot be captured.

Thus, we additionally develop the EdTM which models the direct translation probability $p(e_1^a | f_1^b)$. In comparison to the EiTM trained with swapped corpora and alignments, EdTM incorporates dependencies beyond phrase boundaries by keep-
the inverted alignment \(b'_1\) instead of using \(a'_1\). Analogue to the ETM, the hidden alignment \(b'_1\) is approximated by the Viterbi alignment.

\[
p(e'_1, b'_1 | f_{b_0}, e_0) = \prod_{i=1}^{l} p(e, b_i | f_{b_0}, e_i, f_i, b_i).
\]

We factorize the joint probability to obtain the lexicon probability of \(e_i\) and the alignment probability of \(b_i\).

\[
p(e_i, b_i | f_{b_0}, e_i, f_i, b_i) = p(e_i | f_{b_0}, e_i, f_i, b_i) \cdot p(b_i | f_{b_0}, e_i, f_i, b_i)
\]

The direct probability has been decomposed into the following three probabilities.

- deletion: \(p(e_0 | f_{b_0}^{b_0})\)
- lexicon: \(\prod_{i=1}^{l} p(e_i | f_{b_0}, f_{b_0}, e_i, f_i, b_i)\)
- alignment: \(\prod_{i=1}^{l} p(b_i | f_{b_0}, f_{b_0}, e_i, f_i)\)

Next, we introduce the corresponding EdTM deletion, lexicon and alignment models.

### 3.2.1 EdTM: Deletion Model

The EdTM deletion model approximates the probability of \(e_0\) conditioned on all unaligned source words \(f_{b_0}\) and is obtained by averaging over all unaligned source words:

\[
p(e_0 | f_{b_0}^{b_0}) = \sum_{j \in b_{0}^{b_0}} \frac{p(e_0 | f_j)}{\phi_0}.
\]

### 3.2.2 EdTM: Lexicon Model

In contrast to the derivation of ETM, the Markov chain rule cannot be applied at this point, since we do not model the probability of \(f_{b_0}\), but the probability of \(e_i\) conditioned on \(f_{b_0}\). Thus, we average over all aligned source words \(f_{b_0}\), which results in:

\[
p(e_i | f_{b_0}, f_{b_0}, e_i, f_i, b_i) = \frac{1}{\phi_i} \sum_{j \in b_0} \frac{1}{\phi_i} \left( \sum_{j \in b_0} p(e_i | f_j, e_i, f_j, \Delta_i) \right) + \sum_{j \in b_0} p(\Delta_j | f_j, f_j)
\]

### 3.2.3 EdTM: Alignment Model

Applying the same assumptions as for the lexicon model, the EdTM alignment model results in:

\[
p(b_i | f_{b_0}, f_{b_0}, e_i, f_i) = \frac{1}{\phi_i} \sum_{j \in b_0} \frac{1}{\phi_i} \left( \sum_{j \in b_0} p(\Delta_i | f_j, e_i, f_j) \right) + \sum_{j \in b_0} p(\Delta_j | f_j, f_j)
\]

### 3.3 Count Models and Smoothing

So far, we have introduced the ETM and shown how to include unaligned words and multiple word dependencies. However, there are various possibilities to train the lexicon and alignment probabilities derived in Subsections 3.1 and 3.2.

As a starting point, we apply relative frequencies obtained from bilingual training data, where the Viterbi alignment is estimated using GIZA++ (Och and Ney, 2003). In order to address data sparseness, we apply interpolated Kneser-Ney smoothing as described in (Chen and Goodman, 1998). In comparison to monolingual n-grams used in LMs, we lack any clear order of \(e, f, e', f'\) and \(\Delta\), since they include bilingual and reordering information. Similar to the approach taken by Mariño et al. (2006), we model the probability of the bilingual word pair \((e, f)\) given its predecessor \((e', f', \Delta)\) which also includes the jump class. The EdTM lexicon model for dependencies on previously aligned target words is computed as

\[
p(e | f, e', f', \Delta) = \frac{p(e | f, e', f', \Delta)}{p(\cdot | f, e', f', \Delta)},
\]

where \(p(e | f, e', f', \Delta)\) is the bigram distribution of \((e, f)\) given its predecessor \((e', f', \Delta)\) with interpolated Kneser-Ney smoothing. The denominator \(p(\cdot | f, e', f', \Delta)\) is obtained by marginalizing \(p(e | f, e', f', \Delta)\) over all target words \(e\). We follow the same approach for all other models in analogy.
Table 1: Statistics for the bilingual training data of the IWSLT 2014 German→English, English→French and the DARPA BOLT Chinese→English, Arabic→English translation tasks.

<table>
<thead>
<tr>
<th></th>
<th>IWSLT</th>
<th>IWSLT</th>
<th>BOLT</th>
<th>BOLT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>German</td>
<td>English</td>
<td>French</td>
<td>Chinese</td>
</tr>
<tr>
<td>Sentences</td>
<td>4.32M</td>
<td>26.05M</td>
<td>4.08M</td>
<td>0.92M</td>
</tr>
<tr>
<td>Run. Words</td>
<td>108M</td>
<td>109M</td>
<td>698M</td>
<td>810M</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>836K</td>
<td>792K</td>
<td>2119K</td>
<td>2139K</td>
</tr>
</tbody>
</table>

4 Integration into Phrase-based Decoding

In this work, we apply a standard phrase-based translation system (Koehn et al., 2003). The decoding process is implemented as a beam search for the best translation given a set of models \( h_m(e^I_1, s^K_i, f^J_1) \). The goal of search is to maximize the log-linear feature score (Och and Ney, 2004):

\[
\hat{e}^I_1 = \arg\max_{I, e^I_1, s^K_i} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e^I_1, s^K_i, f^J_1) \right\},
\]

where \( s^K_i = s_1 \ldots s_k \) is the hidden phrase alignment. The feature weights \( \lambda_m \) are tuned with minimum error rate training (MERT) (Och, 2003). The models \( h_m \), that are part of all baselines presented in this work, are phrasal and lexical translation scores in both directions, an \( n \)-gram LM, a simple distance-based distortion model and word and phrase penalties. All phrase pairs that are licensed by the word alignment are extracted from the training corpus and their probabilities estimated as relative frequencies. Moreover, the word alignment each phrase pair has been extracted from is memorized in the phrase table.

Our extended translation models are integrated into this framework as additional features \( h_m \). They are trained in both directions on a bilingual corpus and the Viterbi alignment, resulting in four additional features. When training in the Target→Source direction, the alignment direction is also swapped. Thus, EiT and EdT have the advantage of including context beyond phrase boundaries only when trained in the Source→Target direction.

To include the extended translation models into the phrasal decoder, the source position aligned to the last (not inserted) target word of the previously translated phrase has to be memorized in the search state of a partial hypothesis. Although this slightly affects hypothesis recombination and therefore leads to a larger search space, in practice it does not degrade the search accuracy, as experiments with relaxed pruning parameters have shown.

5 Evaluation

We perform experiments on the large-scale IWSLT 2014\(^2\) (Cettolo et al., 2014) German→English, English→French and the large-scale DARPA BOLT Chinese→English, Arabic→English tasks. As mentioned in Section 4, all baseline systems include phrasal and lexical smoothing scores trained in both directions. Word alignments are trained with GIZA++, by sequentially running 5 iterations each for the IBM-1, HMM and IBM-4 alignment models.

The domain of IWSLT consists of lecture-type talks presented at TED conferences which are also available online\(^3\). The baseline systems are trained on all provided bilingual data. All systems are optimized on the dev2010 and evaluated on the test2010 corpus. The ETM is trained on the TED portions of the data: 138K sentences for German→English and 185K sentences for English→French.

For German→English, to estimate the 4-gram LM, we additionally make use of parts of the Shuffled News, LDC English Gigaword and \( 10^9 \)-French-English corpora, selected by a cross-entropy difference criterion (Moore and Lewis, 2010). In total, 1.7 billion running words are taken for LM training. For English→French, we use a large general domain 5-gram LM and an in-domain 5-gram LM. Both are estimated with the KenLM toolkit (Heafield et al., 2013) using interpolated Kneser-Ney smoothing. For the general domain LM, we first select \( \frac{1}{2} \) of the English Shuffled News, \( \frac{1}{4} \) of the French Shuffled News as well as both the English and French Gigaword corpora.

\(^2\)http://www.iwslt2014.org
\(^3\)http://www.ted.com/
by the same cross-entropy difference criterion. By
concatenating this selection with all available re-
remaining monolingual data, we build an unpruned
LM.

The BOLT tasks are evaluated on the "discus-
sion forum" domain. For Chinese→English, the
baseline is trained on 4.08M general domain sen-
tence pairs and the 5-gram LM on 2.9 billion run-
ning words in total. The ETM is trained on an in-
domain subset of 67.8K sentences and the test set
contains 1844 sentences. For the Arabic→English
BOLT task, we use only the in-domain data for
training the baseline and the ETM. The training
and test sets contain text drawn from discussion
forums in Egyptian Arabic. The evaluation set
contains 1510 bilingual sentence pairs.

The baseline systems for all tasks - except the
Arabic→English BOLT task, where preliminary
experiments showed no improvement - contain a
7-gram word cluster language model (Wuebker et
al., 2013) and for comparison, we also experiment
with a hierarchical reordering model (HRM) (Gal-
ley and Manning, 2008). When integrated into a
phrase-based decoder, Durrani et al. (2013b) have
shown the OSM to outperform bilingual LMs on
MTUs. Therefore, we directly compare ourselves
with a 7-gram OSM implemented into our phrase-
based decoder as an additional feature. The OSM
is trained on the same data as the ETM for all
tasks. Bilingual data statistics for all tasks are
shown in Table 1. For each system setting we eval-
uate three MERT runs using multeval (Clark et
al., 2011). Results are reported in BLEU (Papineni
et al., 2001) and TER (Snover et al., 2006). The
optimization criterion for all experiments is BLEU.

5.1 Model parameters

To measure the complexity of the extended trans-
lation models in comparison to the phrase-based
translation model, we count the number of param-
eters to be trained for each.

Table 2 illustrates the phrase-table and ETM
count table entries for the BOLT Arabic→English
translation task, where both the phrase-based base-
line and the ETM are trained on the same bilin-
gual data consisting of 0.92M bilingual sentence
pairs. Here, we only show the numbers for the
Source→Target direction, as the numbers for the
Target→Source direction are similar. The EdTM
and EiTM each have roughly 35M parameters to
be trained, i.e. there are approximately 70M pa-
rameters to be trained for the ETM in total. This
is slightly more than the 57M parameters for the
phrase translation model.

<table>
<thead>
<tr>
<th>model</th>
<th># parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>phrase-based translation</td>
<td>57,155,149</td>
</tr>
<tr>
<td>EdTM</td>
<td>35,511,396</td>
</tr>
<tr>
<td></td>
<td>lexicon</td>
</tr>
<tr>
<td></td>
<td>alignment</td>
</tr>
<tr>
<td></td>
<td>deletion</td>
</tr>
<tr>
<td>EiTM</td>
<td>34,994,534</td>
</tr>
<tr>
<td></td>
<td>lexicon</td>
</tr>
<tr>
<td></td>
<td>alignment</td>
</tr>
<tr>
<td></td>
<td>deletion</td>
</tr>
</tbody>
</table>

Table 2: The number of model parameters for the
BOLT Arabic→English bilingual training data af-
after filtering.

<table>
<thead>
<tr>
<th>BLEU</th>
<th>TER</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ EiTM</td>
<td>none</td>
</tr>
<tr>
<td>Ge↔En</td>
<td>31.4</td>
</tr>
<tr>
<td>none</td>
<td>Ge↔En</td>
</tr>
<tr>
<td>Ge→En</td>
<td>Ge→En</td>
</tr>
<tr>
<td>Ge↔En</td>
<td>Ge↔En</td>
</tr>
</tbody>
</table>

Table 3: Results for the German→English IWSLT
data. The systems are optimized with MERT on the
dev2010 set. All results are statistically signif-
ificant with ≥99% confidence.

5.2 Results

In order to compare the effect of the EiTM and
EdTM used in a phrase-based decoder, we have
trained the baseline including the HRM as de-
scribed above on the full German→English bilin-
gual data of the IWSLT task and the extended
translation models on the TED data. The results
evaluated on test2010 are shown in Table 3.

Including the EiTm trained in both
German→English and English→German
directions into the phrase decoder yields an
absolute improvement of +0.7 BLEU and -1.0
TER, whereas including the EdTM yields +0.9
BLEU and -1.2 TER. This underlines that the
EdTM is more suitable for translation than the
EiTM because it predicts the direct probability
Table 4: Comparison of ETM to the HRM and OSM measured in BLEU. Statistically significant improvements with ≥99% confidence are printed in boldface.

<table>
<thead>
<tr>
<th></th>
<th>Ge-En</th>
<th>En-Fr</th>
<th>Zh-En</th>
<th>Ar-En</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>30.6</td>
<td>32.8</td>
<td>16.5</td>
<td>23.8</td>
</tr>
<tr>
<td>+ ETM</td>
<td>31.4</td>
<td>33.8</td>
<td>16.8</td>
<td>24.1</td>
</tr>
<tr>
<td>+ OSM</td>
<td>31.6</td>
<td>34.1</td>
<td>17.3</td>
<td>24.1</td>
</tr>
<tr>
<td>+ HRM</td>
<td>30.7</td>
<td>33.1</td>
<td>17.0</td>
<td>24.0</td>
</tr>
<tr>
<td>+ ETM</td>
<td>31.8</td>
<td>33.9</td>
<td>17.5</td>
<td>24.4</td>
</tr>
<tr>
<td>+ OSM</td>
<td>31.8</td>
<td>34.5</td>
<td>17.6</td>
<td>24.1</td>
</tr>
</tbody>
</table>

We also compare the ETM to the OSM on top of a PBT system that also includes the HRM, which is shown in the last two lines of Table 4. The performance of the ETM benefits from the information introduced by the HRM, as the gain of using the ETM is further increased by +0.15 BLEU on average. Overall, the ETM gains consistent and statistically significant improvements of +0.7 BLEU on average for all four language pairs over a state-of-the-art phrase-based decoder including the HRM. On the other hand, OSM seems to have a higher overlap with HRM, as the gain of OSM compared to ETM is reduced to +0.1 BLEU on average. Thus, on top of the phrase-based system including the HRM, the ETM including a bilingual word pair and the corresponding reordering jump class proves to be competitive to a 7-gram OSM.

6 Discussion

We have integrated two variants of a novel extended translation model into a state-of-the-art phrase-based decoder. The ETM captures lexical and reordering context beyond phrase boundaries in both the Source→Target and Target→Source directions. Further, the model potentially captures long-range reorderings and utilizes multiple and empty alignments, allowing for target insertions and source deletions. As an initial step, we have implemented the ETM using relative frequencies with interpolated Kneser-Ney smoothing. Its consistent and statistically significant improvement of up to +1.1 BLEU and -1.1 TER respectively +0.7 BLEU on average has been shown for four large-scale translation tasks, outperforming competitive phrase-based systems that include lexical and phrase translation models and hierarchical reordering models.

Compared to a 7-gram OSM, the ETM is much simpler in design: It uses a smaller vocabulary size, estimates the probability of single words instead of bilingual MTUs, avoids the need of reordering gaps and includes less lexical and reordering context, thus being less sparse. For all that, it performs competitively to a 7-gram OSM on top of phrase-based systems including the HRM. This fact underlines the advantages introduced by the ETM: It operates on words rather than MTUs, explicitly models multiple alignments of a target word, which corresponds to the actual translation direction. Note, that both EiTM and EdTM lose the advantage of modelling dependencies beyond phrase boundaries when trained in the inverse direction English→German. Therefore, we have evaluated their joint performance when trained only in German→English direction, which is similar to the performance of EdTM trained in both directions. This can be due to the fact that even though the EiTM trained in German→English direction incorporates dependencies beyond phrase boundaries, the EdTM trained in English→German direction profits from the better suited direct translation probability. The full ETM, i.e. EiTM and EdTM trained in both directions, yields the best overall performance gain of +1.1 BLEU and -1.1 TER over the baseline.

Moreover, we evaluate the performance of the (full) ETM compared to the HRM and a 7-gram OSM, which are all introduced as additional features into the log-linear framework of the baseline phrase-based decoder. The results are presented in Table 4. The ETM performs similarly to the HRM for the Chinese→English and Arabic→English tasks, resulting in +0.3 BLEU over the PBT baseline. For both IWSLT tasks, the ETM outperforms the HRM by +0.7 BLEU, gaining +0.8 BLEU for the German→English and +1.0 BLEU for the German→English task over the PBT baseline. The context captured by the ETM corresponds roughly to the context captured by a 3-gram OSM. Bearing this in mind, we compare the ETM to a 7-gram OSM, which yields +0.25 BLEU more than the ETM averaged over the four language pairs. Comparing the OSM vocabulary of 1.5M words for the Arabic→English task to the 285K words in the Arabic corpus, this results in an ETM vocabulary 5-times smaller than the OSM vocabulary.

We also compare the ETM to the OSM on top of a PBT system that also includes the HRM, which is shown in the last two lines of Table 4. The performance of the ETM benefits from the information introduced by the HRM, as the gain of using the ETM is further increased by +0.15 BLEU on average. Overall, the ETM gains consistent and statistically significant improvements of +0.7 BLEU on average for all four language pairs over a state-of-the-art phrase-based decoder including the HRM. On the other hand, OSM seems to have a higher overlap with HRM, as the gain of OSM compared to ETM is reduced to +0.1 BLEU on average. Thus, on top of the phrase-based system including the HRM, the ETM including a bilingual word pair and the corresponding reordering jump class proves to be competitive to a 7-gram OSM.
instead of incorporating linear dependencies and models reorderings in a less complex way.

So far we have used the ETM as an additional feature in a phrase-based decoder, but we believe that the usage of such a decoder is a limitation. First, the ETM is estimated on alignments, which themselves are optimized for the IBM models. Second, decoding is performed using phrases that are extracted from the alignments using heuristics. Therefore, the potential of a phrase-based decoder is also limited by these heuristics.

Based on these facts, we believe that the ETM will show its full potential when it is also integrated into the training of the alignment, leading not only to a higher alignment quality, but also to a joint optimization of the alignments and the ETM. Further, directly applying the ETM within a word-based decoder utilizing an extended translation and reordering context will redundantize phrases and thus any extraction heuristics. We believe that a consistent framework where the ETM is applied in both training the alignments and decoding will significantly advance machine translation.

For the short term, we will investigate better smoothing strategies and the possibilities of using neural networks instead of count models.

Acknowledgements

This work has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no 645452 (QT21). This material is partially based upon work supported by the DARPA BOLT project under Contract No. HR0011-12-C-0015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References


Nadir Durrani, Philipp Koehn, Helmut Schmid, and Yang Feng and Trevor Cohn. 2013. A markov


