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).
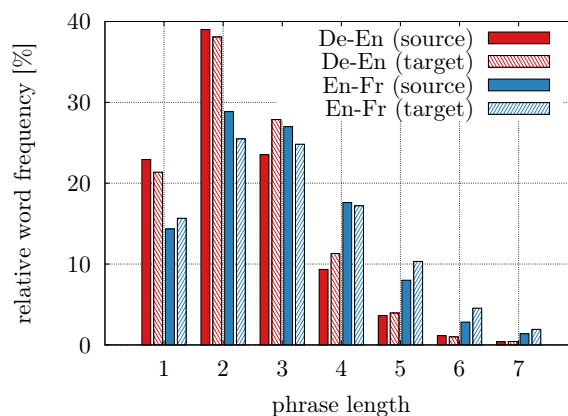

Figure 1: Relative frequency of words dependent on the length of the phrase they were decoded with for the IWSLT `dev2010` German→English and English→French corpora.

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 2014[1] 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.

---

[1] http://www.iwslt2014.org

We address this problem by developing two variants of extended translation models (ETM), the *direct* (EdTM) for the Source→Target and the *inverse* (EiTM) for the Target→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 (Chen and Goodman, 1998) 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. (2006) 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 (Niehues et al., 2011), 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 (Durrani et al., 2011; Durrani et al., 2013a) 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 (Durrani et al., 2013b) 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 (Stewart et al., 2014) and the OSM (Durrani et al., 2014) 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. (2015) 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 (2005) 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. (2013) introduce an reordering model based on sequence labeling techniques by converting the reordering problem into a a tagging task. Zhang et al. (2013) 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 (2013). 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_1^J and its translation e_1^I, EiTM models the *inverse* probability $p(f_1^J|e_1^I)$ and EdTM the *direct* probability $p(e_1^I|f_1^J)$. 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:

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

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 ϕ_i of a target word e_i. It determines the number of source words aligned to the target word e_i:

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

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 $\Delta_{j',j}^{\phi_i}$:

$$\Delta_{j',j}^{\phi_i} = \begin{cases} \downarrow & (\text{'insert'}), & \phi_i = 0 \\ \bullet & (\text{'stay'}), & \phi_i > 0, j = j' \\ \rightarrow & (\text{'forward'}), & \phi_i > 0, j = j'+1 \\ \curvearrowright & (\text{'jump forward'}), & \phi_i > 0, j > j'+1 \\ \leftarrow & (\text{'backward'}), & \phi_i > 0, j = j'-1 \\ \curvearrowleft & (\text{'jump backward'}), & \phi_i > 0, j < j'-1. \end{cases}$$

Figure 2 outlines the jump classes for subsequent target positions i' and i. As shown in Figure 3, for source positions $\bar{j} < j$ which are aligned to the

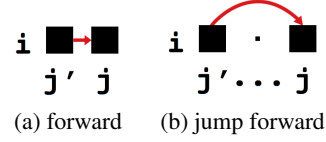(a) forward   (b) jump forward

Figure 3: Overview of the jump classes $\Delta_{\bar{j},j}$.

same target position i, there are two possible jump classes:

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

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_1^I.

3.1 Extended Inverse Translation Model

In order to model the inverse probability $p(f_1^J|e_1^I)$, the unknown inverted alignment b_1^I is introduced as a hidden variable and approximated by the Viterbi alignment.

$$p(f_1^J|e_1^I) = \sum_{b_1^I} p(f_1^J, b_1^I|e_1^I)$$

$$\cong \max_{b_1^I} \left\{ p(f_{b_0}^{b_I}, b_1^I|e_1^I) \right\}$$

$$= \max_{b_1^I} \left\{ p(f_{b_1}^{b_I}, b_1^I|e_1^I) \cdot \underbrace{p(f_{b_0}|f_{b_1}^{b_I}, b_1^I, e_1^I)}_{\text{deletion probability}} \right\}$$

The inverse probability has been decomposed into the deletion probability $p(f_{b_0}|f_{b_1}^{b_I}, b_1^I, e_1^I)$ and the joint probability $p(f_{b_1}^{b_I}, b_1^I|e_1^I)$. The latter is reformulated using the Markov chain rule:

$$p(f_{b_1}^{b_I}, b_1^I|e_1^I) = \prod_{i=1}^{I} p(f_{b_i}, b_i|e_1^I, f_{b_1}^{b_{i-1}}, b_1^{i-1}).$$

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

$$p(f_{b_1}^{b_I}, b_1^I|e_1^I) = \prod_{i=1}^{I} p(f_{b_i}, b_i|e_{i'}, e_i, f_{b_{i'}}, b_{i'}).$$

The conditional joint probability is factorized as

$$p(f_{b_i}, b_i|e_{i'}, e_i, f_{b_{i'}}, b_{i'}) = \\ \underbrace{p(f_{b_i}|e_{i'}, e_i, f_{b_{i'}}, b_{i'}, b_i)}_{\text{lexicon probability}} \cdot \underbrace{p(b_i|e_{i'}, e_i, f_{b_{i'}}, b_{i'})}_{\text{alignment probability}},$$
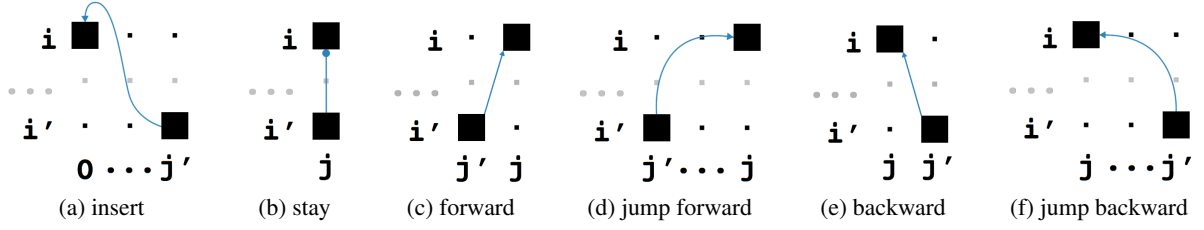
Figure 2: Overview of the jump classes $\Delta_{j',j}^{\phi_i}$.

resulting in the lexicon probability of f_{b_i} and the alignment probability of b_i. In a nutshell, we have decomposed the inverse probability into the following three probabilities:

- deletion: $p(f_{b_0}|f_{b_1}^{b_I}, b_1^I, e_1^I)$

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

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

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 unaligned source words f_{b_0}:

$$p(f_{b_0}|f_{b_1}^{b_I}, b_1^I, e_1^I) = \prod_{j \in b_0} p(f_j|e_0).$$

3.1.2 EiTM: Lexicon Model

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

$$p(f_{b_i}|e_{i'}, e_i, f_{b_{i'}}, b_{i'}, b_i) = \prod_{j \in b_i} p(f_j|e_{i'}, e_i, f_{b_{i'}}, f_{b_i^{<j}}, b_{i'}, b_i)$$

Each source word f_j is dependent on all predecessors $f_{b_i^{<j}}$ aligned to the same target word e_i and all previously aligned source words $f_{b_{i'}}$. If we modelled the probability conditioned on the sets of source words $f_{b_{i'}}$ and $f_{b_i^{<j}}$, 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 $f_{b_{i'}}, f_{b_i^{<j}}$. Without any additional information, we assume all words $f_{b_{i'}}, f_{b_i^{<j}}$ 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 $f_{\bar{j}}$ aligned to the current target word e_i.

Moreover, we reduce the alignments $(b_{i'}, b_i)$ to their corresponding jump classes. As a final result we obtain:

$$p(f_j|e_{i'}, e_i, f_{b_{i'}}, f_{b_i^{<j}}, b_{i'}, b_i) = \frac{1}{\phi_{i'} + \phi_i^{<j}} \left(\sum_{j' \in b_{i'}} p(f_j|e_i, f_{j'}, e_{i'}, \Delta_{j',j}^{\phi_i}) + \sum_{\bar{j} \in b_i^{<j}} p(f_j|e_i, f_{\bar{j}}, \Delta_{\bar{j},j}) \right).$$

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 preceeding source position aligned to the same target word e_i.

$$p(b_i|e_{i'}, e_i, f_{b_{i'}}, b_{i'}) = \prod_{j \in b_i} \frac{1}{\phi_{i'} + \phi_i^{<j}} \left(\sum_{j' \in b_{i'}} p(\Delta_{j',j}^{\phi_i}|e_i, f_{j'}, e_{i'}) + \sum_{\bar{j} \in b_i^{<j}} p(\Delta_{\bar{j},j}|e_i, f_{\bar{j}}) \right).$$

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 $f_{\bar{j}}$ aligned to the current target word. The shape of an arc symbolizes
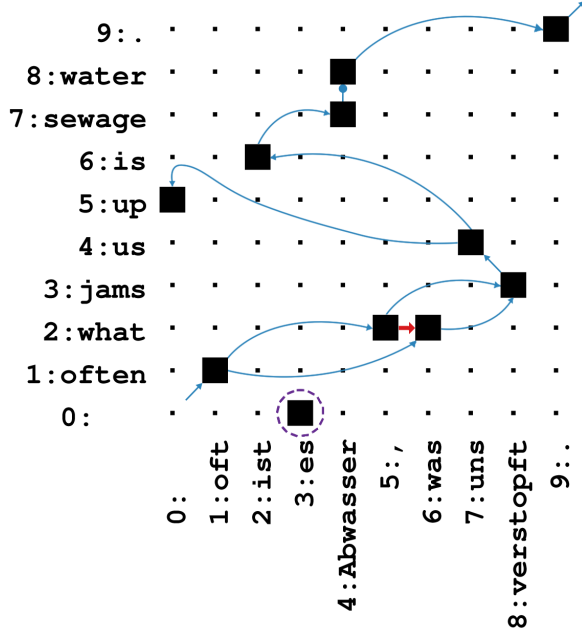
Figure 4: EiTM scoring for a sentence from the IWSLT German→English corpus including the word alignment.

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:

$$p(f_1^9, b_1^9 | e_1^9) =$$

$$p(f_1|e_1,\texttt{<s>},\texttt{<s>},\rightarrow) \cdot p(\rightarrow |e_1,\texttt{<s>},\texttt{<s>})$$

$$\cdot p(f_5|e_2,f_1,e_1,\curvearrowright) \cdot p(\curvearrowright |e_2,f_1,e_1) \qquad (1)$$

$$\cdot \frac{p(f_6|e_2,f_1,e_1,\curvearrowright) + p(f_6|e_2,f_5,\rightarrow)}{2}$$

$$\cdot \frac{p(\curvearrowright |e_2,f_1,e_1) + p(\rightarrow |e_2,f_5)}{2} \qquad (2)$$

$$\cdot \frac{p(f_8|e_3,f_5,e_2,\curvearrowright) + p(f_8|e_3,f_6,e_2,\curvearrowright)}{2}$$

$$\cdot \frac{p(\curvearrowright |e_3,f_5,e_2) + p(\curvearrowright |e_3,f_6,e_2)}{2} \qquad (3)$$

$$\cdot p(f_7|e_4,f_8,e_3,\leftarrow) \cdot p(\leftarrow |e_4,f_8,e_3)$$

$$\cdot p(f_0|e_5,f_7,e_4,\downarrow) \cdot p(\downarrow |e_5,f_7,e_4) \qquad (4)$$

$$\cdot p(f_2|e_6,f_7,e_4,\curvearrowleft) \cdot p(\curvearrowleft |e_6,f_7,e_4) \qquad (5)$$

$$\cdot p(f_4|e_7,f_2,e_6,\curvearrowright) \cdot p(\curvearrowright |e_7,f_2,e_6)$$

$$\cdot p(f_4|e_8,f_4,e_7,\bullet) \cdot p(\bullet |e_8,f_4,e_7) \qquad (6)$$

$$\cdot p(f_9|e_9,f_4,e_8,\curvearrowright) \cdot p(\curvearrowright |e_9,f_4,e_8) \qquad (7)$$

$$\cdot p(\texttt{</s>}|\texttt{</s>},f_9,e_9,\rightarrow) \cdot p(\rightarrow |\texttt{</s>},f_9,e_9)$$

$$\cdot p(f_3|e_0). \qquad (8)$$

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 $e_5 = \texttt{up}$ conditioned on the word pair $(e_4 = \texttt{us}, f_7 = \texttt{uns})$. Note that in (5) there is no dependency of $e_6 = \texttt{is}$ on its predecessor $e_5 = \texttt{up}$ and the empty word f_0, but on its *last aligned* predecessor $e_4 = \texttt{us}$ and the corresponding source word $f_7 = \texttt{uns}$. (6) shows an example of a source word aligned to multiple target words. The deletion probability of the source word $f_3 = \texttt{es}$ is presented in (8).

3.2 Extended Direct Translation Model

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

For a start, the direct probability $p(e_1^I|f_1^J)$ 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 $p(e_{a_0}^{a_J}, a_1^J|f_1^J)$ 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 $e_{i'}$ and the corresponding alignment $b_{i'}$ and source words $f_{b_{i'}}$ 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 $p(e_1^I|f_1^J)$, dependencies beyond phrase boundaries cannot be captured.

Thus, we additionally develop the EdTM which models the direct translation probability $p(e_1^I|f_1^J)$. In comparison to the EiTM trained with swapped corpora and alignments, EdTM incorporates dependencies beyond phrase boundaries by keep-

ing the *inverted* alignment b_1^I instead of using a_1^J. Analogue to the EiTM, the hidden alignment b_1^I is approximated by the Viterbi alignment.

$$p(e_1^I|f_1^J) \cong \max_{b_1^I} \left\{ \underbrace{p(e_0|f_{b_0}^{b_I})}_{\text{deletion probability}} \cdot p(e_1^I, b_1^I|f_{b_0}^{b_I}, e_0) \right\}$$

Applying the Markov chain rule and assuming (e_i, b_i) to be dependent only on the aligned source words f_{b_i}, the previously aligned target word $e_{i'}$ as well as the corresponding alignment $b_{i'}$ and the source words $f_{b_{i'}}$, we obtain:

$$p(e_1^I, b_1^I|f_{b_0}^{b_I}, e_0) = \prod_{i=1}^{I} p(e_i, b_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'}).$$

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

$$p(e_i, b_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'}) = \underbrace{p(e_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'}, b_i)}_{\text{lexicon probability}} \cdot \underbrace{p(b_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'})}_{\text{alignment probability}}$$

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

- deletion: $p(e_0|f_{b_0}^{b_I})$

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

- alignment: $\prod_{i=1}^{I} p(b_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'})$

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:

$$p(e_0|f_{b_0}^{b_I}) = \sum_{j \in b_0} \frac{p(e_0|f_j)}{\phi_0}.$$

3.2.2 EdTM: Lexicon Model

In contrast to the derivation of EiTM, the Markov chain rule cannot be applied at this point, since we do not model the probability of f_{b_i}, but the probability of e_i conditioned on f_{b_i}. Thus, we average

over all aligned source words f_{b_i}, which results in:

$$p(e_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'}, b_i) =$$
$$\frac{1}{\phi_i} \sum_{j \in b_i} \frac{1}{\phi_{i'} + \phi_i^{<j}} \left(\sum_{j' \in b_{i'}} p(e_i|f_j, e_{i'}, f_{j'}, \Delta_{j',j}^{\phi_i}) \right.$$
$$\left. + \sum_{\bar{j} \in b_i^{<j}} p(e_i|f_j, f_{\bar{j}}, \Delta_{\bar{j},j}) \right).$$

3.2.3 EdTM: Alignment Model

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

$$p(b_i|f_{b_{i'}}, f_{b_i}, e_{i'}, b_{i'}) =$$
$$\frac{1}{\phi_i} \sum_{j \in b_i} \frac{1}{\phi_{i'} + \phi_i^{<j}} \left(\sum_{j' \in b_{i'}} p(\Delta_{j',j}^{\phi_i}|f_j, e_{i'}, f_{j'}) \right.$$
$$\left. + \sum_{\bar{j} \in b_i^{<j}} p(\Delta_{\bar{j},j}|f_j, f_{\bar{j}}) \right).$$

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 Δ, 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', Δ) which also includes the jump class. The EdTM lexicon model for dependencies on previously aligned target words is computed as

$$p(e|f, e', f', \Delta) = \frac{p(e, f|e', f', \Delta)}{p(\cdot, f|e', f', \Delta)}, \qquad (9)$$

where $p(e, f|e', f', \Delta)$ is the bigram distribution of (e, f) given its predecessor (e', f', Δ) 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.

| | IWSLT | | IWSLT | | BOLT | | BOLT | |
	German	English	English	French	Chinese	English	Arabic	English
Sentences	4.32M		26.05M		4.08M		0.92M	
Run. Words	108M	109M	698M	810M	78M	86M	14M	16M
Vocabulary	836K	792K	2119K	2139K	384K	817K	285K	203K

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.

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 $h_m(e_1^I, s_1^K, f_1^J)$. The goal of search is to maximize the log-linear feature score (Och and Ney, 2004):

$$\hat{e}_1^{\hat{I}} = \arg\max_{I, e_1^I, s_1^K} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, s_1^K, f_1^J) \right\}, \quad (10)$$

where $s_1^K = s_1 \ldots s_K$ is the hidden phrase alignment. The feature weights λ_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, EiTM and EdTM 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/

288

by the same cross-entropy difference criterion. By concatenating this selection with all available remaining monolingual data, we build an unpruned LM.

The BOLT tasks are evaluated on the "discussion forum" domain. For Chinese→English, the baseline is trained on 4.08M general domain sentence pairs and the 5-gram LM on 2.9 billion running 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) (Galley 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 evaluate 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 translation models in comparison to the phrase-based translation model, we count the number of parameters 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 baseline and the ETM are trained on the same bilingual 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-

model	# parameters
phrase-based translation	57,155,149
EdTM	35,511,396
lexicon	19,899,812
alignment	15,276,718
deletion	334,866
EiTM	34,994,534
lexicon	20,153,114
alignment	14,791,722
deletion	49,698

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

		BLEU	TER
Baseline + HRM		30.7	49.3
+ EiTM	**+ EdTM**		
Ge↔En	*none*	31.4	48.3
none	Ge↔En	31.6	48.1
Ge→En	Ge→En	31.6	48.2
Ge↔En	Ge↔En	31.8	48.2

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

rameters to be trained for the ETM in total. This is slightly more than the 57M parameters for the phrase translation model.

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 described above on the full German→English bilingual 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 phrasal 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

	Ge-En	En-Fr	Zh-En	Ar-En
Baseline	30.6	32.8	16.5	23.8
+ ETM	**31.4**	**33.8**	16.8	**24.1**
+ OSM	**31.6**	**34.1**	**17.3**	**24.1**
+ HRM	30.7	33.1	17.0	24.0
+ ETM	**31.8**	**33.9**	**17.5**	**24.4**
+ OSM	**31.8**	**34.5**	**17.6**	24.1

Table 4: Comparison of ETM to the HRM and OSM measured in Bleu. Statistically significant improvements with ≥99% confidence are printed in boldface.

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.

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

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

References

Michael Auli, Michel Galley, Chris Quirk, and Geoffrey Zweig. 2013. Joint Language and Translation Modeling with Recurrent Neural Networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 1044–1054, Seattle, USA, October.

Peter F. Brown, John Cocke, Stephan A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Rossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85, June.

Peter F. Brown, Stephan A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, June.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *International Workshop on Spoken Language Translation*, pages 2–11, Lake Tahoe, CA, USA, December.

Stanley F. Chen and Joshuo Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, August.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *49th Annual Meeting of the Association for Computational Linguistics:shortpapers*, pages 176–181, Portland, Oregon, June.

Yonggang Deng and William Byrne. 2005. Hmm word and phrase alignment for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 169–176, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and Robust Neural Network Joint Models for Statistical Machine Translation. In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 1370–1380, Baltimore, MD, USA, June.

Nadir Durrani, Helmut Schmid, and Alexander Fraser. 2011. A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1045–1054, Portland, Oregon, USA, June. Association for Computational Linguistics.

Nadir Durrani, Alexander Fraser, and Helmut Schmid. 2013a. Model with minimal translation units, but decode with phrases. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1–11, Atlanta, Georgia, June. Association for Computational Linguistics.

Nadir Durrani, Alexander Fraser, Helmut Schmid, Hieu Hoang, and Philipp Koehn. 2013b. Can markov models over minimal translation units help phrase-based smt? In *Proceedings of the 51st Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 399–405, Sofia, Bulgaria, August. Association for Computational Linguistics.

Nadir Durrani, Philipp Koehn, Helmut Schmid, and Alexander Fraser. 2014. Investigating the usefulness of generalized word representations in smt. In *COLING*, Dublin, Ireland, aug.

Yang Feng and Trevor Cohn. 2013. A markov model of machine translation using non-parametric bayesian inference. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 333–342, Sofia, Bulgaria, August.

Minwei Feng, Jan-Thorsten Peter, and Hermann Ney. 2013. Advancements in reordering models for statistical machine translation. In *Annual Meeting of the Assoc. for Computational Linguistics*, pages 322–332, Sofia, Bulgaria, August.

Yang Feng, Trevor Cohn, and Xinkai Du. 2014. Factored markov translation with robust modeling. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 151–159, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Michel Galley and Christopher D. Manning. 2008. A simple and effective hierarchical phrase reordering model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 848–856, Stroudsburg, PA, USA. Association for Computational Linguistics.

Andreas Guta, Tamer Alkhouli, Jan-Thorsten Peter, Joern Wuebker, and Hermann Ney. 2015. A comparison between count and neural network models based on joint translation and reordering sequences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisboa, Portugal, September. Association for Computational Linguistics. to appear.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria, August.

Yuening Hu, Michael Auli, Qin Gao, and Jianfeng Gao. 2014. Minimum translation modeling with recurrent neural networks. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 20–29, Gothenburg, Sweden, April. Association for Computational Linguistics.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Meeting of the North American chapter of the Association for Computational Linguistics (NAACL-03)*, pages 127–133, Edmonton, Alberta.

Hai Son Le, Alexandre Allauzen, and François Yvon. 2012. Continuous Space Translation Models with Neural Networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–48, Montreal, Canada, June.

José B Mariño, Rafael E Banchs, Josep M Crego, Adrià de Gispert, Patrik Lambert, José A R Fonollosa, and Marta R Costa-jussà. 2006. N-gram-based Machine Translation. *Comput. Linguist.*, 32(4):527–549, December.

R.C. Moore and W. Lewis. 2010. Intelligent Selection of Language Model Training Data. In *ACL (Short Papers)*, pages 220–224, Uppsala, Sweden, July.

Jan Niehues, Teresa Herrmann, Stephan Vogel, and Alex Waibel, 2011. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, chapter Wider Context by Using Bilingual Language Models in Machine Translation, pages 198–206. Association for Computational Linguistics.

Franz J. Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, July.

Franz J. Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och and Hermann Ney. 2004. The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449, December.

Franz J. Och, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD, June.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan, July.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a Method for Automatic Evaluation of Machine Translation. IBM Research Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, September.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*,

292

pages 223–231, Cambridge, Massachusetts, USA, August.

Darelene Stewart, Roland Kuhn, Eric Joanis, and George Foster. 2014. Coarse split and lump bilingual languagemodels for richer source information in smt. In *AMTA*, Vancouver, BC, Canada, oct.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 14–25, Doha, Qatar, October.

Joern Wuebker, Stephan Peitz, Felix Rietig, and Hermann Ney. 2013. Improving statistical machine translation with word class models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1377–1381, Seattle, USA, October.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-Based Statistical Machine Translation. In *25th German Conf. on Artificial Intelligence (KI2002)*, pages 18–32, Aachen, Germany, September. Springer Verlag.

Hui Zhang, Kristina Toutanova, Chris Quirk, and Jianfeng Gao. 2013. Beyond left-to-right: Multiple decomposition structures for smt. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 12–21, Atlanta, Georgia, June. Association for Computational Linguistics.