The RWTH Aachen German-English Machine Translation System for WMT 2014

Stephan Peitz, Joern Wuebker, Markus Freitag and Hermann Ney
Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
D-52056 Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

This paper describes the statistical machine translation (SMT) systems developed at RWTH Aachen University for the German→English translation task of the ACL 2014 Eighth Workshop on Statistical Machine Translation (WMT 2014). Both hierarchical and phrase-based SMT systems are applied employing hierarchical phrase reordering and word class language models. For the phrase-based system, we run discriminative phrase training. In addition, we describe our preprocessing pipeline for German→English.

1 Introduction

For the WMT 2014 shared translation task\(^1\) RWTH utilized state-of-the-art phrase-based and hierarchical translation systems. First, we describe our preprocessing pipeline for the language pair German→English in Section 2. Furthermore, we utilize morpho-syntactic analysis to preprocess the data (Section 2.3). In Section 3, we give a survey of the employed systems and the basic methods they implement. More details are given about the discriminative phrase training (Section 3.4) and the hierarchical reordering model for hierarchical machine translation (Section 3.5). Experimental results are discussed in Section 4.

2 Preprocessing

In this section we will describe the modification of our preprocessing pipeline compared to our 2013 WMT German→English setup.

2.1 Categorization

We put some effort in building better categories for digits and written numbers. All written numbers were categorized. In 2013 they were just handled as normal words which leads to a higher number of out-of-vocabulary words. For German→English, in most cases for numbers like '3,000' or '2.34' the decimal mark ',' and the thousands separator '.' has to be inverted. As the training data and also the test sets contain several errors for numbers in the source as well as in the target part, we put more effort into producing correct English numbers.

2.2 Remove Foreign Languages

The WMT German→English corpus contains some bilingual sentence pairs with non-German source or/and non-English target sentences. For this WMT translation task, we filtered all non-matching language pairs (in terms of source language German and target language English) from our bilingual training set.

First, we filtered languages which contain non-ascii characters. For example Chinese, Arabic or Russian can be easily filtered when deleting sentences which contain more than 70 percent non-ascii words. The first examples of Table 1 was filtered due to the fact, that the source sentence contains too many non-ascii characters.

In a second step, we filtered European languages containing ascii characters. We used the WMT monolingual corpora in Czech, French, Spanish, English and German to filter these languages from our bilingual data. We could both delete a sentence pair if it contains a wrong source language or a wrong target language. That is the reason why we even search for English sentences in the source part and for German sentences in the target part. For each language, we built a word count of all words in the monolingual data for each language separately. We removed punctuation which are no indicator of a language. In our experiments, we only considered words with frequency higher than 20 (e.g. to ignore names). Given the word frequency, we removed a bilingual

\(^1\)http://www.statmt.org/wmt14/translation-task.html
Table 1: Examples of sentences removed in preprocessing.

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>remove non-ascii symbols 高效的技以抵消影响。</td>
</tr>
<tr>
<td>zum Bericht Añoveros Triás de Bes</td>
</tr>
<tr>
<td>remove wrong languages from target Honni soit qui mal y pense!</td>
</tr>
<tr>
<td>as you yourself have said: travailler plus pour gagner plus</td>
</tr>
</tbody>
</table>

sentence pair from our training data if more than 70 percent of the words had a higher count in a different language then the one we expected. In Table 1 some example sentences, which were removed, are illustrated.

In Table 2 the amount of sentences and the corresponding vocabulary sizes of partial and totally cleaned data sets are given. Further we provide the number of out-of-vocabulary words (OOVs) for newstest2012. The vocabulary size could be reduced by \( \sim 130k \) words for both source and target side of our bilingual training data while the OOV rate kept the same. Our experiments showed, that the translation quality is the same with or without removing wrong sentences. Nevertheless, we reduced the training data size and also the vocabulary size without any degradation in terms of translation quality.

2.3 Morpho-syntactic Analysis

In order to reduce the source vocabulary size for the German→English translation further, the German text is preprocessed by splitting German compound words with the frequency-based method described in (Koehn and Knight, 2003). To reduce translation complexity, we employ the long-range part-of-speech based reordering rules proposed by Popović and Ney (2006).

3 Translation Systems

In this evaluation, we employ phrase-based translation and hierarchical phrase-based translation. Both approaches are implemented in Jane (Vilar et al., 2012; Wuebker et al., 2012), a statistical machine translation toolkit which has been developed at RWTH Aachen University and is freely available for non-commercial use.\(^2\) In the newest internal version, we use the KenLM Language Model Interface provided by (Heafield, 2011) for both decoders.

3.1 Phrase-based System

In the phrase-based decoder (source cardinality synchronous search, SCSS, Wuebker et al. (2012)), we use the standard set of models with phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based distortion model, an \( n \)-gram target language model and three binary count features. Additional models used in this evaluation are the hierarchical reordering model (HRM) (Galley and Manning, 2008) and a word class language model (wcLM) (Wuebker et al., 2013). The parameter weights are optimized with minimum error rate training (MERT) (Och, 2003). The optimization criterion is BLEU (Papineni et al., 2002).

3.2 Hierarchical Phrase-based System

In hierarchical phrase-based translation (Chiang, 2007), a weighted synchronous context-free grammar is induced from parallel text. In addition to contiguous lexical phrases, hierarchical phrases with up to two gaps are extracted. The search is carried out with a parsing-based procedure. The standard models integrated into our Jane hierarchical systems (Vilar et al., 2010; Huck et al., 2012) are: Phrase translation probabilities and lexical smoothing probabilities in both translation directions, word and phrase penalty, binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, three binary count features, and an \( n \)-gram language model. We utilize the cube pruning algorithm for decoding (Huck et al., 2013a) and optimize the model weights with MERT. The optimization criterion is BLEU.

3.3 Other Tools and Techniques

We employ GIZA++ (Och and Ney, 2003) to train word alignments. The two trained alignments are heuristically merged to obtain a symmetrized word alignment for phrase extraction. All lan-
Table 2: Corpus statistics after each filtering step and compound splitting.

<table>
<thead>
<tr>
<th></th>
<th>Sentences</th>
<th>Vocabulary</th>
<th>OOVs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>German</td>
<td>English</td>
</tr>
<tr>
<td>Preprocessing 2013</td>
<td>4.19M</td>
<td>1.43M</td>
<td>784K</td>
</tr>
<tr>
<td>Preprocessing 2014</td>
<td>4.19M</td>
<td>1.42M</td>
<td>773K</td>
</tr>
<tr>
<td>+ remove non-ascii symbols</td>
<td>4.17M</td>
<td>1.36M</td>
<td>713K</td>
</tr>
<tr>
<td>+ remove wrong languages from target</td>
<td>4.15M</td>
<td>1.34M</td>
<td>675K</td>
</tr>
<tr>
<td>+ remove wrong languages from source</td>
<td>4.08M</td>
<td>1.30M</td>
<td>655K</td>
</tr>
<tr>
<td>+ compound splitting</td>
<td>4.08M</td>
<td>652K</td>
<td>655K</td>
</tr>
</tbody>
</table>

Language models (LMs) are created with the SRILM toolkit (Stolcke, 2002) or with the KenLM language model toolkit (Heafield et al., 2013) and are standard 4-gram LMs with interpolated modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998). We evaluate in true-case with BLEU and TER (Snover et al., 2006).

3.4 Discriminative Phrase Training

In our baseline translation systems the phrase tables are created by a heuristic extraction from word alignments and the probabilities are estimated as relative frequencies, which is still the state-of-the-art for many standard SMT systems. Here, we applied a more sophisticated discriminative phrase training method for the WMT 2014 German—English task. Similar to (He and Deng, 2012), a gradient-based method is used to optimize a maximum expected BLEU objective, for which we define BLEU on the sentence level with smoothed 3-gram and 4-gram precisions. To that end, the training data is decoded to generate 100-best lists. We apply a leave-one-out heuristic (Wuebker et al., 2010) to make better use of the training data. Using these n-best lists, we iteratively perform updates on the phrasal translation scores of the phrase table. After each iteration, we run MERT, evaluate on the development set and select the best performing iteration. In this work, we perform two rounds of discriminative training on two separate data sets. In the first round, training is performed on the concatenation of newstest2008 through newstest2010 and an automatic selection from the News-commentary, Europarl and Common Crawl corpora. The selection is based on cross-entropy difference of language models and IBM-1 models as described by Mansour et al. (2011) and contains 258K sentence pairs. The training took 4.5 hours for 30 iterations. On top of the final phrase-based systems, a second round of discriminative training is run on the full news-commentary corpus concatenated with newstest2008 through newstest2010.

3.5 A Phrase Orientation Model for Hierarchical Machine Translation

In Huck et al. (2013b) a lexicalized reordering model for hierarchical phrase-based machine translation was introduced. The model scores monotone, swap, and discontinuous phrase orientations in the manner of the one presented by (Tillmann, 2004). Since improvements were reported on a Chinese—English translation task, we investigate the impact of this model on a European language pair. As in German the word order is more flexible compared with the target language English, we expect that an additional reordering model could improve the translation quality. In our experiments we use the same settings which worked best in (Huck et al., 2013b).

4 Setup

We trained the phrase-based and the hierarchical translation system on all available bilingual training data. Corpus statistics can be found in the last row of Table 2. The language model are 4-grams trained on the respective target side of the bilingual data, \( \frac{1}{2} \) of the Shuffled News Crawl corpus, \( \frac{1}{4} \) of the 10\(^{10}\) French-English corpus and \( \frac{1}{4} \) of the LDC Gigaword Fifth Edition corpus. The monolingual data selection is based on cross-entropy difference as described in (Moore and Lewis, 2010). For the baseline language model, we trained separate models for each corpus, which were then interpolated. For our final experiments, we also trained a single unpruned language model on the concatenation of all monolingual data with KenLM.

159
4.1 Experimental Results

The results of the phrase-based system (SCSS) as well as the hierarchical phrase-based system (HPBT) are summarized in Table 3.

The phrase-based baseline system, which includes the hierarchical reordering model by (Galley and Manning, 2008) and is tuned on newstest2012, reaches a performance of 25.9% BLEU on newstest2013. Adding the word class language model improves performance by 0.4% BLEU absolute and the first round of discriminative phrase training by 0.5% BLEU absolute. Next, we switched to tuning on a concatenation of newstest2011 and newstest2012, which we expect to be more reliable with respect to unseen data. Although the BLEU score does not improve and TER goes up slightly, we kept this tuning set in the subsequent setups, as it yielded longer translations, which in our experience will usually be preferred by human evaluators. Switching from the interpolated language model to the unpruned language model trained with KenLM on the full concatenated monolingual training data in a single pass gained us another 0.3% BLEU. For the final system, we ran a second round of discriminative training on different training data (cf. Section 3.4), which increased performance by 0.1% BLEU to the final score 27.2.

For the phrase-based system, we also experimented with weighted phrase extraction (Mansour and Ney, 2012), but did not observe improvements.

The hierarchical phrase-based baseline without any additional model is on the same level as the phrase-based system including the word class language model, hierarchical reordering model and discriminative phrase training in terms of BLEU. However, extending the system with a word class language model or the additional reordering models does not seem to help. Even the combination of both models does not improve the translation quality. Note, that the hierarchical system was tuned on the concatenation newstest2011 and newstest2012. The final system employs both word class language model and hierarchical reordering model.

Both phrase-based and hierarchical phrase-based final systems are used in the EU-Bridge system combination (Freitag et al., 2014).

5 Conclusion

For the participation in the WMT 2014 shared translation task, RWTH experimented with both phrase-based and hierarchical translation systems. For both approaches, we applied a hierarchical phrase reordering model and a word class language model. For the phrase-based system we employed discriminative phrase training. Additionally, improvements of our preprocessing pipeline compared to our WMT 2013 setup were described. New introduced categories lead to a lower amount of out-of-vocabulary words. Filtering the corpus for wrong languages gives us lower vocabulary sizes for source and target without loosing any performance.
Acknowledgments

The research leading to these results has partially received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

Furthermore, 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


