LMU Munich’s Neural Machine Translation Systems for News Articles and Health Information Texts

Matthias Huck, Fabienne Braune, Alexander Fraser
Center for Information and Language Processing
LMU Munich
Munich, Germany
{mhuck,braune,fraser}@cis.lmu.de

Abstract

This paper describes the LMU Munich English→German machine translation systems. We participated with neural translation engines in the WMT17 shared task on machine translation of news, as well as in the biomedical translation task. LMU Munich’s systems deliver competitive machine translation quality on both news articles and health information texts.

1 Introduction

The Center for Information and Language Processing at LMU Munich has a strong track record at building statistical machine translation (SMT) systems for various language pairs, e.g. for translation between English and German, Czech, Romanian, Russian, or French. LMU has frequently participated in WMT machine translation shared tasks in recent years (Bojar et al., 2016, 2015, 2014, 2013), competing (and also collaborating) internationally in an open evaluation campaign with other leading research labs from both academia and industry.

Research on various different types of machine translation models has previously been conducted at LMU. Core SMT paradigms for LMU’s past shared task participations include phrase-based models (Cap et al., 2015, 2014b; Weller et al., 2013; Sajjad et al., 2013), hierarchical phrase-based models (Huck et al., 2016; Peter et al., 2016), operation sequence models (Durrani et al., 2013), and hybrids of statistical approaches with rule-based and deep syntactic components (Tamchyna et al., 2016b).

At this year’s EMNLP 2017 Second Conference on Machine Translation (WMT17),\(^1\) LMU participated in two shared tasks: the shared task on machine translation of news and the biomedical translation task. We submitted the output of our English→German machine translation systems. The system for the news task was trained under “constrained” conditions, employing only permissible resources as defined by the shared task organizers. The system for the biomedical task builds upon our news task system, but was domain-adapted towards the medical domain via the usage of additional parallel training data from the in-domain sections of the UFAL Medical Corpus v.1.0.

We have trained neural machine translation (NMT) models this year. Neural network models for machine translation (Sutskever et al., 2014; Bahdanau et al., 2014) are now largely successful for many language pairs and domains. This has for instance become apparent with the University of Edinburgh’s excellent results in the WMT16 news translation shared task with neural systems (Sennrich et al., 2016a), which outperformed most other submitted systems, including Edinburgh’s own traditional SMT engines (Williams et al., 2016). LMU’s English→German neural machine translation systems confirm this trend. We have achieved competitive performance—in terms of translation quality as measured with BLEU (Papineni et al., 2002)—in both shared tasks that we participated in.\(^2\)

A unique characteristic of the LMU English→German NMT systems is a linguistically informed, cascaded word segmentation technique that we developed and applied to the German target language side of the training data. Amongst other aspects, SMT research at LMU is

\(^1\)http://www.statmt.org/wmt17/

\(^2\)Our LMU Munich primary system is ranked second in BLEU on the submission website, http://matrix.statmt.org/matrix/systems_list/1869, being outsourced by Edinburgh’s WMT17 NMT setup only. In the human evaluation the LMU Munich primary system is ranked first (Bojar et al., 2017).
focusing on investigating linguistically informed methods that improve machine translation into target languages which exhibit a more complex morphosyntax than English (Huck et al., 2017b; Tamchyna et al., 2016a; Ramm and Fraser, 2016; Weller-Di Marco et al., 2016; Braune et al., 2015; Cap et al., 2014a; Fraser et al., 2012). We are taking advantage of our group’s longstanding experience regarding handling of complex morphosyntax in SMT, now enriching NMT with novel techniques that specifically tackle target-side morphosyntax.

In the following section of this paper (Section 2), we sketch our linguistically motivated target word segmentation technique. Then we describe how we trained and configured our neural machine translation systems (Section 3). Before concluding the paper, we present empirical results on the two translation tasks, which involve machine translation of news articles and of health information texts (Section 4).

2 Target-side Word Segmentation

Compounding and morphological variation are ubiquitous in the German language and have traditionally been challenging for machine translation into German. We believe that specifically targeting complex morphosyntactic phenomena in the output language is not only essential in traditional phrase-based machine translation, but keeps being valuable in NMT. Most previous work in NMT has focused on linguistically agnostic subword splitting, typically with the primary rationale of limiting the vocabulary size, which is required in NMT for efficiency considerations.

LMU is utilizing a more linguistically-informed target word segmentation approach. By doing so, we hope to achieve three major goals: better vocabulary reduction; reduction of data sparsity; and better open vocabulary translation.

We cascade three different word splitting methods on the German target side.

1. First we apply a suffix splitter that separates common German morphological suffixes from the word stems. We modified the German Snowball stemming algorithm from NLTK\(^3\) for that purpose. Rather than stripping suffixes, our modified code splits them off. It otherwise behaves just like the Snowball stemming algorithm.

2. Next we apply the empirical compound splitter as described by Koehn and Knight (2003) and as implemented in the Perl script which is part of the Moses toolkit (Koehn et al., 2007). We choose a fairly aggressive configuration of the compound splitter\(^4\) in order to reduce the vocabulary size more than with its parameters as typically chosen for previous phrase-based translation setups in which German compound splitting was used.

3. Since the vocabulary size is still a bit large after suffix splitting and compound splitting, we adopt segmentation using the Byte Pair Encoding (BPE) technique (Gage, 1994; Sennrich et al., 2016c) on top of the other two word splitters. This last step is performed only for efficiency reasons in NMT. Without BPE, the vocabulary size is still almost 100K. We preferred something around 50K, which is more tractable in practice. Suffix splitting and compound splitting alone are not suitable for arbitrary reduction of the vocabulary size. However, we believe that they are more adequate word segmentation techniques than BPE is. So we prefer to split with those linguistically motivated methods, as far as practicable.

Special marker symbols allow us to revert the segmentation in postprocessing. We also introduce a case marker that is placed before any compound-split word in order to restore upper and lower casing, respectively, since the compound splitting approach modifies the casing of compound parts to the version of each part that (stand-alone) appears most frequently in the corpus.

We were running a comprehensive series of experiments with different target word segmentation strategies on Europarl data beforehand, and we found our cascaded word segmentation to perform clearly better than using BPE only. We furthermore tried prefix splitting, but the results looked less encouraging. Our Europarl results also suggested that the suffix splitting contributes more to improvements in translation quality than the compound splitting does. Huck et al. (2017a) provides further details. For our WMT17 shared task systems, we eventually decided to apply both suffix segmentation.

\(^3\)http://www.nltk.org/_modules/nltk/stem/snowball.html

\(^4\)-min-size 4 -min-count 2 -max-count 999999999
splitting and compound splitting, but to omit prefix splitting.
The English source side is simply BPE-segmented.

3 Neural Translation System Setup

We utilize the Nematus implementation (Sennrich et al., 2017) to build encoder-decoder NMT systems with attention and gated recurrent units. We configure dimensions of 500 for the embeddings and 1024 for the hidden layer. We train with the Adam optimizer (Kingma and Ba, 2015), a learning rate of 0.0001, batch size of 50, and dropout with probability 0.2 applied to the hidden layer, but not to source, target, and embeddings. We validate every 10 000 updates and do early stopping when the validation cost has not decreased over ten consecutive control points.

Our initial baseline NMT system is trained using only data from the Europarl corpus (Koehn, 2005) and no other resources, with the Europarl test2006 set used for validation. We tokenize and frequent-case the data with the standard scripts from the Moses toolkit (Koehn et al., 2007). For our Europarl-trained baseline, sentences of length >50 after tokenization are excluded from the training corpus, all other sentences (1.7 M) are kept in training.

The German compound split model and BPE merge operations are extracted from the Europarl data. In our cascaded word segmentation pipeline, the compound split model is extracted from the training data only after suffix splitting has been applied. Similarly, the BPE operations are extracted after suffix splitting and compound splitting have been applied to the German side of the training corpus. We set the amount of merge operations for BPE to 50K. On the English source side, we apply BPE separately, also with 50K merge operations.

3.1 News Translation Task

For the shared task on machine translation of news (Bojar et al., 2017), we successively improved our initial baseline by incrementally applying the following steps:

1. Adding the News Commentary (NC) and Common Crawl (CC) parallel training data as provided for WMT17 by the organizers of the news translation shared task. We initialize the optimization on the larger corpus with the Europarl-trained baseline model.

2. Adding synthetic training data. The use of automatically translated monolingual data as a supplementary training resource has proved to be effective in SMT for phrase-based, hierarchical, and neural systems (Ueffing et al., 2007; Lambert et al., 2011; Huck et al., 2011; Huck and Ney, 2012; Sennrich et al., 2016b). Sennrich et al. have publicly shared their backtranslations of monolingual WMT News Crawl corpora, which they created for their WMT16 participation (Sennrich et al., 2016a). We exploit the full amount of backtranslations of German data into English.\(^5\) We concatenate the synthetic data and the human-generated parallel training data (Europarl + NC + CC). The optimization is initialized with the pre-trained model from the preceding step.

3. Fine-tuning towards the domain of news articles. We employ the newstest development sets from the years 2008 to 2014 as a training corpus. We reduce the learning rate to 0.000001, initialize with the pre-trained model from the preceding step, and optimize on only the small Devsets2008-14 corpus.

4. Right-to-left reranking. We rerank an \(n\)-best list from the system in the preceding step with a right-to-left (r2l) model, where the order of the target sequence is reversed. Liu et al. (2016) have proposed right-to-left reranking for NMT. Earlier work by Freitag et al. (2013) had already established that reverse word order models can be beneficial in phrase-based and hierarchical phrase-based translation. Freitag et al. (2013) utilized reverse word order models by means of a system combination framework (Freitag et al., 2014), though.

Validation is done on newstest2015 for each of the extended setups. The preprocessing pipeline is not altered when more training data is appended. Particularly, we keep applying the compound split model and BPE operations that have been extracted from only the Europarl corpus, and keep sticking to the vocabulary from Europarl. We force the system to suppress UNK tokens in inference at test time.

\(^5\)http://data.statmt.org/rsennrich/wmt16_backtranslations/en-de/
Table 1: English→German translation results on newstest sets (case-sensitive BLEU). Extensions are applied incrementally.

<table>
<thead>
<tr>
<th>system</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline, Europarl-trained</td>
<td>19.4</td>
<td>22.8</td>
<td>18.3</td>
</tr>
<tr>
<td>+ NC &amp; CC corpora</td>
<td>26.0</td>
<td>30.2</td>
<td>24.5</td>
</tr>
<tr>
<td>+ synthetic data</td>
<td>27.8</td>
<td>32.3</td>
<td>26.1</td>
</tr>
<tr>
<td>+ fine-tuning on Devsets2008-14</td>
<td>28.2</td>
<td>32.3</td>
<td>26.6</td>
</tr>
<tr>
<td>+ r2l reranking</td>
<td>28.6</td>
<td>33.4</td>
<td>27.1</td>
</tr>
</tbody>
</table>

Table 2: English→German translation results on HimL biomedical sets (case-sensitive BLEU).

<table>
<thead>
<tr>
<th>system</th>
<th>Cochrane</th>
<th>NHS24</th>
</tr>
</thead>
<tbody>
<tr>
<td>fine-tuning on Devsets2008-14</td>
<td>29.1</td>
<td>26.3</td>
</tr>
<tr>
<td>fine-tuning on Medical</td>
<td>35.0</td>
<td>29.5</td>
</tr>
<tr>
<td>+ r2l reranking</td>
<td>35.8</td>
<td>30.3</td>
</tr>
</tbody>
</table>

3.2 Biomedical Translation Task

For the biomedical translation task (Yepes et al., 2017), we started off with the pre-trained NMT model after step 2 of our news task system engineering and applied the following steps:

1. Fine-tuning towards the domain of health information texts. We employ the in-domain sections of the UFAL Medical Corpus v.1.0 as a training corpus. Set the learning rate to 0.00001, initialize with the pre-trained model, and optimize on only the in-domain medical data.

2. Right-to-left reranking. An ensemble of domain-adapted r2l models worked best.

The HimL (Haddow et al., 2017) tuning sets are used for validation, and we tested separately on the Cochrane and NHS24 parts of the HimL devtest set.

4 Empirical Results

We evaluate case-sensitive with BLEU (Papineni et al., 2002), computed over postprocessed hypotheses against the raw references with mteval-v13a. The results are reported in Table 1 for the news translation task and in Table 2 for the biomedical translation task.

In the news translation task, the Europarl-trained baseline does not get close to state-of-the-art performance on newstest sets. However, this seems to be mostly due to a domain mismatch (Huck et al., 2015). Once we add in the News Commentary and Common Crawl parallel data, we are able to massively improve the translation quality, by around six to seven BLEU points. Synthetic data gives us a boost of about another two BLEU points. After fine-tuning on Devsets2008-14 towards news articles, we observe a further gain of 0.4 BLEU on newstest2015 but no gain on newstest2016. Reranking with a right-to-left model is effective on all test sets again, with improvements in the range of 0.4 to 1.1 BLEU.

Two LMU submissions have been judged by humans in the manual evaluation for the WMT17 news translation task (Bojar et al., 2017): the output of our final setup with r2l reranking (as a primary submission; “LMU-nmt-reranked”), and the single system output without reranking (as a contrastive submission; “LMU-nmt-single”). Our primary submission is placed first amongst all evaluated systems. We conjecture that our linguistically-informed target word segmentation approach has contributed to a positive assessment by human evaluators. Interestingly, the contrastive submission was rated significantly worse, affirming the utility of r2l reranking.

A few example translations from our primary submission for the news task are shown in Table 3.

For the translation of health information texts, it is again crucial to adapt the NMT system to the domain. When applying the engine fine-tuned on out-of-domain news data (Devsets2008-14) to Cochrane and NHS24 devtest sets, we see quite a gap as compared to fine-tuning on the in-domain sections of the UFAL Medical Corpus. Right-to-left reranking improves the results by 0.8 BLEU for the biomedical task.

5 Conclusion

LMU Munich has participated with English→German neural machine translation systems in the WMT17 shared tasks on machine translation of news and of biomedical texts. A distinctive feature of LMU’s NMT systems is a linguistically informed, cascaded target word segmentation approach. The LMU systems are very competitive in terms of translation quality, achieving top ranks amongst the participants in both tasks. Of all English→German systems manually evaluated in the news task, LMU’s primary submission has received the highest human judgment scores.
<table>
<thead>
<tr>
<th>Source (preproc.)</th>
<th>LMU-nmt (plain)</th>
<th>LMU-nmt (postproc.)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>the Kurdish community in Germany is expecting tens of thousands of people to arrive at short notice in search of protection, fleeing from Turkey to Germany.</td>
<td>die kurdisch $Se Gemeind $e in #U deutsch @@ Land rechnet damit, dass zehntaus $Send $e Mensch $en #L kurz @@ Frist $ig auf der Suche $e nach dem Schutz eintreff $en, der aus der Türkei nach #U deutsch @@ Land gefloh $en ist.</td>
<td>Die kurdische Gemeinde in Deutschland rechnet damit, dass zehntausende Menschen kurzfristig auf der Suche nach dem Schutz eintreffen, der aus der Türkei nach Deutschland geflohen ist.</td>
<td>Die Kurdische Gemeinde Deutschland rechnet kurzfristig mit zehntausenden Schutzsuchenden, die aus der Türkei nach Deutschland flüchten.</td>
</tr>
<tr>
<td>the situation only worsened over the past year when the world’s biggest producer, China, dumped steel into the global market en masse as a result of weakening domestic demand.</td>
<td>die Lag $e verschlechtert $e sich nur im vergang $en $en Jahr, als der #L Welt @@ größt $e Produzent, China, infolg $e der schwä ## chelnd $e #U binn ## r schwä ## chelnd $e Binnennachfrage massenhaft Stahl in den globalen Markt geworfen hat.</td>
<td>Die Lage verschlechterte sich nur im vergangenen Jahr, als der weltgrößte Produzent, China, infolge der schwächelnden Binnennachfrage massenhaft Stahl in den globalen Markt geschafft hat.</td>
<td>Im vergangenen Jahr verschärfte sich die Lage weiter, als das weltgrößte Erzeugerland China angesichts der schwächelnden heimischen Nachfrage massenhaft Stahl auf den Weltmarkt warf.</td>
</tr>
<tr>
<td>analysts fear that separatist groups that had been more or less vanquished in recent years, like the Oromo Liberation Front or the Ogaden National Liberation Front, may try to exploit the turbulence and rearm.</td>
<td>Analyst $en befürcht $en, dass separatist $isch $e Grupp $en, die in den letzt $en Jahr $en mehr oder wenig $er bezwang $en word $en war $en, wie die Oro ## mo Befreiung ## r Front oder der O ## gad $en National $e Befreiung ## r Front, versuch $en kött $en, die Turbulenz $en und die Aufrüst $ung auszunutz $en.</td>
<td>Analysten befürchteten, dass separatistische Gruppen, die in den letzten Jahren mehr oder weniger bezwungen worden waren, wie die Oromo-Befreiungsfront oder der Ogaden Nationale Befreiungsfront, versuchen könnten, die Turbulenzen und die Aufrüstung auszunutzen.</td>
<td>Analytiker befürchten, dass Separatisten wie die Oromo-Befreiungsfront oder die Nationale Befreiungsfront des Ogaden, die in den letzten Jahren mehr oder weniger bezwungen wurden, die Turbulenzen ausnutzen und sich wieder bewaffnen könnten.</td>
</tr>
<tr>
<td>these celebrities are not relatives of famous people, or reality stars, or kids these days who know how to make a good S ## n ## ap ## chat video (although Jenner is all of these things).</td>
<td>dies $e Pro ## minten $en sind kein $e Verwandt $en berühm $er Mensch $en oder Re ## ally ## Star $en oder Kind $er die wiss $en, wie man ein gut $es Sna ## p ## Video mach $en kann (obwohl Jenner all diese Dinge ist).</td>
<td>Diese Prominenten sind nicht mit berühmten Personen verwandt oder Reality Stars oder Kinder dieser Tage, die wissen, wie man ein gutes Snapchat-Video machen kann (obwohl Jenner all diese Dinge ist).</td>
<td>Diese Berühmtheiten sind nicht mit berühmten Personen verwandt oder Reality Stars oder Jugendliche von heute, die wissen, wie man ein gutes Snapchat-Video dreht (auch wenn davon alles auf Jenner zutrifft).</td>
</tr>
<tr>
<td>the specialists from the 3 ## 4th police inspectorate were able to prove that the thieves, who had travelled to Germany to commit the crimes, had committed four crimes.</td>
<td>die Spezialist $en der 34. #U Polizei ## Inspektion konnt $en beweiss $en, dass die Dieb $e, die nach #U deutsch @@ Land gereist war $en, um die Verbrech $en zu begeh $en $en, vier Straftat $en begang $en hat $en.</td>
<td>Die Spezialisten der 34. Polizeiinspektion konnten beweisen, dass die Diebe, die nach Deutschland gereist waren, um die Verbrechen zu begehen, vier Straftaten begangen hatten.</td>
<td>Die Spezialisten des Kriminalkommissariats 34 können den Dieben, die eigens zur Begehung von Straftaten nach Deutschland eingereist waren, vier Taten nachweisen.</td>
</tr>
</tbody>
</table>

Table 3: Example translations, produced with LMU Munich’s primary machine translation system for the news task. The table shows the preprocessed English source, the plain system output, the postprocessed system output, and the German reference translation for every 479th sentence from the newstest2017 evaluation set (excluding the very first of them, sentence 479, since it is too short to be interesting). ## is a BPE split-point, $en is the suffix en, #U and #L are upper and lower case indicators for the first word of compounds, @@ indicates a compound merge-point, @s@ indicates a compound merged with the letter s between the parts, etc.
Acknowledgments

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement № 644402 (HimL). This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement № 640550).

References


