LMU Munich's Neural Machine Translation Systems at WMT 2018

Matthias Huck and **Dario Stojanovski** and **Viktor Hangya** and **Alexander Fraser**

Center for Information and Language Processing
LMU Munich
Munich, Germany

{mhuck,stojanovski,hangyav,fraser}@cis.lmu.de

Abstract

We present the LMU Munich machine translation systems for the English–German language pair. We have built neural machine translation systems for both translation directions (English→German and German→English) and for two different domains (the biomedical domain and the news domain). The systems were used for our participation in the WMT18 biomedical translation task and in the shared task on machine translation of news.[1,2]

The main focus of our recent system development efforts has been on achieving improvements in the biomedical domain over last year's strong biomedical translation engine for English→German (Huck et al., 2017a). Considerable progress has been made in the latter task, which we report on in this paper.

1 Introduction

Domain adaptation is one emphasis of the machine translation research conducted at the Center for Information and Language Processing at LMU Munich. Within the scope of our participation in the EU-funded *HimL* project (Haddow et al., 2017),[3] we were recently working on advancing the quality of machine translation for medical texts. The types of medical texts that we consider range from health information leaflets to professional biomedical research articles.

Some of our latest research towards medical domain adapation of neural translation systems is inspired by the "fine-tuning" approach in combination with high-quality in-domain data. Specifically, we conducted successive optimization runs to domain-adapt a neural translation model. The model was eventually deployed as the core component of the final English→German HimL translation engine in year 3 of the project (Y3).

In this paper, we give a brief technical overview of the HimL Y3 engine's neural translation model for English→German. We will show by how much the translation quality of medical texts improves compared to our previous year's WMT17 biomedical task submission (Huck et al., 2017a). We then proceed to compare with a Transformer model (Vaswani et al., 2017) that we have trained after the end of the HimL project. We find that the Transformer model performs even better than the HimL Y3 engine, which was based on Nematus (Sennrich et al., 2017) with a single hidden layer. The good result encouraged us to try out the Transformer in the other translation direction, German→English. We will also report the German→English results.

In addition to the English–German biomedical task, LMU Munich has participated in the WMT18 English–German news translation task (Bojar et al., 2018) in both translation directions. Our (supervised) news task systems are shortly described towards the end of the paper.[4]

2 Domain Adaptation

Medical texts differ in their style and in their topics from the typical content of many widely used training corpora, such as the parallel Europarl corpus (Koehn, 2005) or most of the large monolingual corpora that are distributed for the WMT shared task on machine translation of news (Bojar et al., 2018, 2017a, 2016, 2015). Medical documents also often contain a large amount of domain-specific technical terms in their vocabulary. Furthermore, sense shifts of words (away

[1] http://www.statmt.org/wmt18/biomedical-translation-task.html

[2] http://www.statmt.org/wmt18/translation-task.html

[3] http://www.himl.eu

[4] LMU's *unsupervised* machine translation system for the news task is described in a separate paper (Stojanovski et al., 2018).

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 648–654

https://doi.org/10.18653/v1/W18-64073

from their respective meaning in out-of-domain corpora) are common (Carpuat et al., 2013; Irvine et al., 2013).

Domain adaptation of conventional phrase-based machine translation systems is a well-explored research area. Several different effective solutions which may be used in order to domain-adapt a phrase-based system have been proposed in the literature. (Inter alia, cf. Huck et al. (2015) for a few interesting empirical results and a list of some major bibliographic references.) Machine translation in academic research labs and also in industry is however going through a paradigm shift away from phrase-based technology and on towards artificial neural network models. Neural machine translation (Sutskever et al., 2014; Bahdanau et al., 2014) is the new state of the art for basically all medium- to high-resource language pairs since around two to three years. The paradigm shift poses new challenges in domain adaptation, since most known techniques are rather specific to the phrase-based translation model and therefore cannot be readily applied to neural systems.

Domain adaptation of neural translation systems is a fresh and active field of scientific inquiry. The most wide-spread practical solution at present is referred to as "fine-tuning". A baseline model is pre-trained by optimizing the neural model parameters on some large general corpus. Subsequently, training is simply continued on an in-domain corpus, usually with a smaller learning rate—i.e., in this second optimization run the parameters are initialized with the trained model parameters from the previous optimization. A crucial aspect is the availability of high-quality in-domain training data, or alternatively, the collection thereof. If a general-domain or out-of-domain neural model from a first optimization run already exists, then fine-tuning allows for quick adjustment of the model to a specific domain by means of a short continued optimization on an in-domain corpus, most often with less data than in the first run.

3 Neural Network Architectures

3.1 GRU Encoder-Decoder

We utilize the Nematus implementation (Sennrich et al., 2017) to build encoder-decoder NMT systems with attention and gated recurrent units (GRUs). Our architecture is flat, it has only one single hidden layer. 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.

3.2 Transformer

We use the Sockeye implementation of the Transformer (Hieber et al., 2017). For the German→English translation direction we train small Transformer models and for English→German big models as outlined in Vaswani et al. (2017). All models have six encoder and decoder layers. The size of the layers and the embeddings is 512 for the small models and 1024 for the big ones. The dimensionality of the feed-forward networks is 2048 (small) and 4096 (big). We use 8 attention heads for the small and 16 for the big models. The models are trained with the Adam optimizer with an initial learning rate of 0.0002. The learning rate is reduced by a factor of 0.7 if not improved for eight checkpoints. We checkpoint the models each 3 000 updates and do early stopping if perplexity has not improved for 32 checkpoints. We apply dropout of 0.1 as used by Vaswani et al. (2017). Additionally, we use label smoothing with a value of 0.1. We also tie the target and output embeddings. All models are trained with a word-level batch size of 4096.

4 Preprocessing

A linguistically informed, cascaded word segmentation technique is applied to the German side of the training data (Huck et al., 2017b). With a linguistically more sound word segmentation, we expect advantages over plain BPE segmentation in three important aspects: vocabulary reduction, reduction of data sparsity, and open vocabulary translation. The NMT system can learn linguistic word formation processes from the segmented data.

We cascade three different word splitting methods on the German side:

1. First we apply a suffix splitter that separates common German morphological suffixes from the word stems. Our suffix splitter is a modification of the German Snow-

ball stemming algorithm that separtates suffixes from the word stem, rather than stripping them.

2. Next we apply the empirical compound splitter as described by Koehn and Knight (2003).
3. We finally apply the Byte Pair Encoding (BPE) technique (Sennrich et al., 2016b) on top of the suffix-split and compound-split data in order to further reduce the vocabulary size.

Special marker symbols allow us to revert the segmentation in postprocessing when German is the target language.

Our linguistically informed word segmentation was already used on the target language side for LMU's participation in the WMT17 shared task on machine translation of news (Huck et al., 2017a). At WMT17, LMU's primary submission was ranked first in the human evaluation (Bojar et al., 2017a). We presume that the high human rating of LMU's WMT17 submission can mostly be attributed to our efforts toward better word segmentation. We anticipate similar benefits in the medical domain. Dedicated methods that tackle rich target-side morphology have also shown good results in phrase-based translation systems previously (Huck et al., 2017c). Future work on neural machine translation could for instance follow a two-step prediction paradigm (Conforti et al., 2018), or improve over our current version of linguistically informed word segmentation by means of a better linguistic analysis (Weissweiler and Fraser, 2017).

In the present work, the linguistically informed word segmentation is not only employed on the target side for English→German machine translation, but in German→English systems also on the source language side.

The English language side is always simply BPE-segmented.

We learn the compound split model and the BPE merge operations from Europarl and use this word segmentation and vocabulary for all corpora.

5 Systems: Medical Translation

5.1 English→German HimL Y3 System

The English→German HimL Y3 engine is based on a shallow GRU encoder-decoder model built with Nematus (Section 3.1). We apply an incremental training regime that is inspired by "fine-tuning" (Section 2). First, we train a model on parallel corpora from the WMT news task. We then successively refine the model and adapt it to the medical domain. Consecutive optimization runs are initialized with the respective previous model parameters. For each refinement step, we replace the training data, first with larger corpora, then with corpora that better match the domain.

The HimL tuning sets are used for validation, and we test separately on the Cochrane and NHS24 parts of the HimL devtest set.[5] The translation quality (in case-sensitive BLEU (Papineni et al., 2002)) of different system setups after several development stages is presented in the top section of Table 1. *WMT_parallel* denotes the Europarl, News Commentary, and Common Crawl parallel training data as provided for WMT17 by the organizers of the news translation shared task. *WMT_backtranslated_news_crawl* denotes Edinburgh's backtranslations of monolingual WMT News Crawl corpora from WMT16.[6] *Y3_base_general_data* is a large collection of English–German bitext used in the HimL project. *Cochrane-selected* and *NHS24-selected* denote synthetic data mixes from HimL whose content is automatically filtered to match the Cochrane or NHS24 use cases. Corpus statistics of the HimL training data and a more detailed description of the data selection procedure are provided by Bojar et al. (2017b) (Section 2.4 of HimL Deliverable D1.1).

We vary the learning rate during system development, as stated in the table. As a last step, we apply n-best list reranking ($n = 50$) with a right-to-left NMT model ("r2l reranking"). Ensembling did not yield any clear gains, so we deployed single models for English→German.

The bottom row of Table 1 contains the BLEU scores of our last year's primary system (Huck et al., 2017a) for the WMT17 biomedical task (Yepes et al., 2017). We improve over it by more than three points.

5.2 English→German Transformer System

We build Transformer models (Section 3.2) in order to evaluate whether they perform better than our Nematus-based HimL Y3 system.

For the English→German Transformer model, we train three separate models and ensemble them.

[5] http://www.himl.eu/test-sets

[6] http://data.statmt.org/rsennrich/wmt16_backtranslations/en-de/

English→German	**Cochrane** BLEU	**NHS24** BLEU
WMT_parallel (*lrate = 0.0001*)	31.5	28.9
+ WMT_parallel, WMT_backtranslated_news_crawl (*lrate = 0.0001*)	29.8	27.6
+ UFAL_medical_shuffled_all (*lrate = 0.0001*)	35.1	28.9
+ Y3_base_general_data (*lrate = 0.00001*)	35.7	29.8
+ Cochrane-selected, NHS24-selected, 10 × UFAL_medical_indomain (*lrate = 0.00001*)	38.6	33.0
+ r2l reranking (= HimL Y3)	**39.6**	**34.0**
Transformer single	37.8	33.3
Transformer ensemble	39.0	34.1
+ r2l reranking	**40.3**	**35.5**
LMU WMT17 biomedical (Huck et al., 2017a)	**35.8**	**30.3**

Table 1: English→German medical translation results on HimL devtest sets (case-sensitive BLEU). Extensions are applied incrementally. Particularly, in the top section of the table, which reports on HimL Y3 system engineering, we conduct successive model refinement by consecutively optimizing on different corpora. The middle section of the table reports on Transformer experiments. The row at the bottom provides the results of our WMT17 biomedical task system.

We also apply right-to-left reranking on these models as well. Because of time constraints we did not train a Transformer right-to-left model. Instead, we generated a 50-best list with the Transformer models and used the already trained Nematus right-to-left models for the reranking.

No incremental training regime or fine-tuning is applied to the Transformer system. We train on the same set of corpora that is also used in the last refinement step of the HimL Y3 system (Cochrane-selected, NHS24-selected, 10 × UFAL_medical_indomain).

The translation results with the English→German Transformer systems are presented in the middle section of Table 1. The Transformer outperforms our other systems.

We submitted three runs to the WMT18 biomedical translation shared task: the r2l-reranked Transformer (run1, primary); a Transformer ensemble without reranking (run2, contrastive); and the HimL Y3 system (run3, contrastive).

5.3 German→English Transformer System

Our German→English Transformer model is an ensemble of three separate models, like in the English→German translation direction. We use the same training corpus, but with source and target side switched. The preprocessing remains the same. Since German is the source language in this setup, our linguistically informed word segmentation technique is applied to the input side here.

The BLEU scores of the German→English Transformer without ensembling (single model) are 53.3 (Cochrane) and 41.7 (NHS24), respectively. The ensemble is reaching BLEU scores of 54.5 (Cochrane) and 42.2 (NHS24), which is a decent gain over the single model.

6 Systems: News Translation

6.1 English→German News Task System

For the shared task on machine translation of news, we did not build any updated system, but participated with our system from WMT17 (Bojar et al., 2017a). The system was trained under "constrained" conditions, employing only permissible resources as defined by the shared task organizers. Huck et al. (2017a) provide a detailed description, along with experimental results. In short, we conducted the following steps in an incremental training regime (with consecutive optimizations, in a similar manner as presented above for the HimL Y3 system):

1. Optimize a Europarl baseline model.
2. Add News Commentary and Common Crawl.
3. Add synthetic training data (Ueffing et al., 2007; Lambert et al., 2011; Huck et al., 2011; Huck and Ney, 2012; Sennrich et al., 2016a).
4. Fine-tune towards the domain of news articles. For that purpose, several `newstest` development sets are employed as a training corpus. The learning rate is decreased.
5. Rerank n-best list with a right-to-left neural model (Liu et al., 2016), which is trained for reverse word order (Freitag et al., 2013).

6.2 German→English News Task System

Finally, for the translation of news articles from German into English, we also trained a basic shallow GRU encoder-decoder system (cf. Section 3.1). The training data is a concatenation of Europarl, News Commentary, Common Crawl, and some synthetic data in the form of backtranslated English news texts. The German source side is preprocessed with our linguistically informed word segmentation (Section 4).

7 Conclusion

In this paper, we have described the steps we took to build a strong neural system for the translation of medical documents. Our English→German translation system was deployed within the HimL project. We used the system to participate in the WMT18 biomedical translation shared task. On HimL devtest sets, our WMT18 biomedical task systems outperforms our WMT17 submission system by more than three BLEU points.

Three aspects make our system effective in our view. (1.) We have high-quality in-domain training data at hand. (2.) A reliable preprocessing pipeline has been developed. (3.) A simple, but well-working domain adaptation method is known for neural machine translation.

The model architecture is also very important, as our additional Transformer experiments show: A less highly engineered Transformer model is on par with our deployed HimL project system.

Additionally to the English→German medical domain system, we have also briefly presented our system for the German→English translation direction and our WMT18 news task submissions.

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, abs/1409.0473. Presented at ICLR 2015.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Ondřej Bojar, Barry Haddow, David Mareček, Roman Sudarikov, Aleš Tamchyna, and Dušan Variš. 2017b. HimL Deliverable D1.1: Report on Building Translation Systems for Public Health Domain. Technical report. http://www.himl.eu/files/D1.1-report-on-building-translation-systems.pdf.

Marine Carpuat, Hal Daume III, Katharine Henry, Ann Irvine, Jagadeesh Jagarlamudi, and Rachel Rudinger. 2013. SenseSpotting: Never let your parallel data tie you to an old domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1435–1445, Sofia, Bulgaria. Association for Computational Linguistics.

Costanza Conforti, Matthias Huck, and Alexander Fraser. 2018. Neural Morphological Tagging of Lemma Sequences for Machine Translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA 2018), vol. 1: MT Research Track*, pages 39–53, Boston, MA, USA.

Markus Freitag, Minwei Feng, Matthias Huck, Stephan Peitz, and Hermann Ney. 2013. Reverse Word Order Models. In *Proceedings of the XIV Machine Translation Summit*, pages 159–166, Nice, France.

Barry Haddow, Alexandra Birch, Ondřej Bojar, Fabienne Braune, Colin Davenport, Alex Fraser, Matthias Huck, Michal Kašpar, Květoslava Kovaříková, Josef Plch, Anita Ramm, Juliane Ried, James Sheary, Aleš Tamchyna, Dušan Variš, Marion Weller, and Phil Williams. 2017. HimL: Health in my Language. In *Proceedings of the EAMT 2017 User Studies and Project/Product Descriptions*, page 33, Prague, Czech Republic.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Matthias Huck, Alexandra Birch, and Barry Haddow. 2015. Mixed-Domain vs. Multi-Domain Statistical Machine Translation. In *Proceedings of MT Summit XV, vol.1: MT Researchers' Track*, pages 240–255, Miami, FL, USA.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017a. LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 315–322, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Huck and Hermann Ney. 2012. Pivot Lightly-Supervised Training for Statistical Machine Translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA 2012)*, San Diego, CA, USA.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017b. Target-side Word Segmentation Strategies for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Huck, Aleš Tamchyna, Ondřej Bojar, and Alexander Fraser. 2017c. Producing Unseen Morphological Variants in Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 369–375, Valencia, Spain. Association for Computational Linguistics.

Matthias Huck, David Vilar, Daniel Stein, and Hermann Ney. 2011. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 91–96, Edinburgh, Scotland. Association for Computational Linguistics.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daumé III, and Dragos Munteanu. 2013. Measuring Machine Translation Errors in New Domains. *Transactions of the Association for Computational Linguistics*, 1:429–440.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit X*, Phuket, Thailand.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 187–194, Budapest, Hungary. Association for Computational Linguistics.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. 2011. Investigations on Translation Model Adaptation Using Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293, Edinburgh, Scotland. Association for Computational Linguistics.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on Target-bidirectional Neural Machine Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 411–416, San Diego, CA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2018. The LMU Munich Unsupervised Machine Translation Systems. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Semi-supervised model adaptation for statistical machine translation. *Machine Translation*, 21(2):77–94.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Leonie Weissweiler and Alexander Fraser. 2017. Developing a Stemmer for German Based on a Comparative Analysis of Publicly Available Stemmers. In *Proceedings of the German Society for Computational Linguistics and Language Technology (GSCL)*, Berlin, Germany.

Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 Biomedical Translation Shared Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.