TIMBERT: Toponym Identifier For The Medical Domain Based on BERT

MohammadReza Davari and **Leila Kosseim** and **Tien D. Bui**
Dept. Computer Science and Software Engineering
Concordia University, Montreal QC H3G 1M8, Canada
mohammadreza.davari@mail.concordia.ca
{leila.kosseim,tien.bui}@concordia.ca

Abstract

In this paper, we propose an approach to automate the process of place name detection in the medical domain to enable epidemiologists to better study and model the spread of viruses. We created a family of **T**oponym **I**dentification **M**odels based on **BERT** (TIMBERT), in order to learn in an end-to-end fashion the mapping from an input sentence to the associated sentence labeled with toponyms. When evaluated with the SemEval 2019 task 12 test set (Weissenbacher et al., 2019), our best TIMBERT model achieves an F1 score of 90.85%, a significant improvement compared to the state-of-the-art of 89.13% (Wang et al., 2019).

1 Introduction

Phylogeographers, who study the geographic distribution of viruses, have long linked the increase in the geographical spread of viruses (Gautret et al., 2012; Green and Roberts, 2000) to the growth in global tourism and international trade of goods. Most notably, in December 2019, a pneumonia-like disease, later dubbed COVID-19, that was detected in the city of Wuhan, China quickly became a world pandemic mainly due to global travels (World Health Organization, 2020b; World Health Organization, 2020a).

Epidemiologists study the global impact of the spread of viruses by considering information on the DNA sequence and structure of viruses, as well as relying on accurate metadata. Although accurate localized geographical data is critical for creating maps of the locations of viruses and their migration paths, most publicly available databases, such as GenBank (Benson et al., 2012), provide insufficient details on the matter, limited to country or state level. In a study by Scotch et al. (2011), it is estimated that 7% of the GenBank records are missing geospatial metadata and 73% lacking detailed records. However, more fine-grained localization information on the viruses may be present in articles that describe the research work. Therefore, a manual inspection of biomedical articles is vital for obtaining more detailed information about the locations of the viruses.

Toponym detection aims to identify the word boundaries of expressions that denote geographic names. Toponym detection has been the focus of much work in recent years (Ardanuy and Sporleder, 2017; DeLozier et al., 2015; Taylor, 2017) and studies have shown that the task is highly dependent on the textual domain (Amitay et al., 2004; Purves et al., 2007; Qin et al., 2010; Kienreich et al., 2006; Garbin and Mani, 2005). The focus of this paper is to propose a competitive deep learning based model for toponym detection that learns in an end-to-end fashion the mapping from an input sentence to the associated sentence with toponym labels.

We evaluated our models using the SemEval 2019 task 12 test set (Weissenbacher et al., 2019) and report their performances based on precision, recall and F1 respectively. These metrics can be measured in two ways: strict or overlapping. The strict measures, consider a prediction to match the gold standard annotation if both point to the exact same span of text at the character level. On the other hand, the overlapping measures are more lenient as they consider a prediction to match the gold standard annotations when they share a common span of text. Since the research community in toponym identification

Proceedings of the 28th International Conference on Computational Linguistics, pages 662–668
Barcelona, Spain (Online), December 8-13, 2020

is more concerned with strict measures (Magge et al., 2018), we only report on the strict measures of precision, recall and F1. As reported in Section 4, our best model achieves a strict F1 score of 90.85% on the SemEval 2019 task 12 test set (Weissenbacher et al., 2019).

2 Previous Work

Categorizing each word of a text as toponym or non-toponym, is the focus of toponym detection. For example, given the sentence: *COVID-19 was first reported in Wuhan, Hubei Province, China.*[1] A toponym detection system should identify *Wuhan*, *Hubei Province*, and *China* as toponyms, and all other words as non-toponyms. Toponym detection tackles ambiguities between toponyms and other classes of Named Entity Recognition as well as metonymic usage of toponyms.

Toponym resolution in the epidemiology domain was the objective of the SemEval 2019 task 12 (Weissenbacher et al., 2019). The majority of current approaches to toponym detection in the medical domain are based on a combination of rule-based techniques, geographical gazetteer approaches, and deep learning models (Magge et al., 2018; Wang et al., 2019; Li et al., 2019; Qi et al., 2019). In our approach, we aim to develop a toponym detection system solely based on deep learning techniques in order to sidestep many design choices and avoid pattern mining that comes along with rule-based techniques and gazetteer approaches.

Contextualized Embeddings Previous attempts to toponym detection in the medical domain have used contextualized embeddings, specifically ELMo (Peters et al., 2018), as the core of their models (Li et al., 2019; Yadav et al., 2019). In this paper, we experiment with a different contextualized embedding model and choose the pretrained BERT model (Devlin et al., 2019) as the backbone of our network architecture.

Linguistic Features Previous works on toponym detection in the medical domain have typically taken advantage of handcrafted features to achieve competitive performance. Most notable features include: (1) orthographic features that capture character level attributes of each token (Magge et al., 2018; Wang et al., 2019; Davari et al., 2019), and (2) part of speech (POS) tags (Wang et al., 2019; Davari et al., 2019; Qi et al., 2019). In Section 4, we will evaluate the influence of these features on our proposed model.

Network Architecture There are two paradigms governing the network architectures used for this task: namely, whether localized contextual information is enough, or all available contextual information should be taken into account when making predictions. The former leads to models that only have access to a sliding window of information such as CNNs or MLPs (Magge et al., 2018; Davari et al., 2019; Magnusson and Dietz, 2019). The latter leads to sequential models operating at the sentence level; among which the BiLSTM-CRF architecture is the most favored and provides state-of-the-art results (Wang et al., 2019; Yadav et al., 2019; Qi et al., 2019; Magnusson and Dietz, 2019). In our experiments, we focused on neural architectures that considered all available contextual information within a sentence. However, we deviated from the trend of Recurrent Neural Networks (RNN) and based our network on the Transformer models (Vaswani et al., 2017) due to its ability to process variable length inputs, while being much more parallelizable in comparison to RNN (see Section 3).

3 Our Proposed Model

The architecture of our toponym recognition model is shown in Figure 1. The WordPiece (Wu et al., 2016) tokenization of a sentence constitutes the input of the model. These tokens are then passed to a pretrained BERT network (Devlin et al., 2019). The output of the network along with certain linguistic features (see Section 4) are then concatenated and passed to a fully connected layer which determines the labels of each token. In our experiments, we used two variations of the BERT model: BERT-Base and BERT-Large (Devlin et al., 2019). The corresponding TIMBERT models are called TIMBERT-Base and TIMBERT-Large. Since the BERT-Large model is much more computationally expensive than BERT-Base, we limit the experiments involving this model.

[1] This example is taken from (Chen et al., 2020)

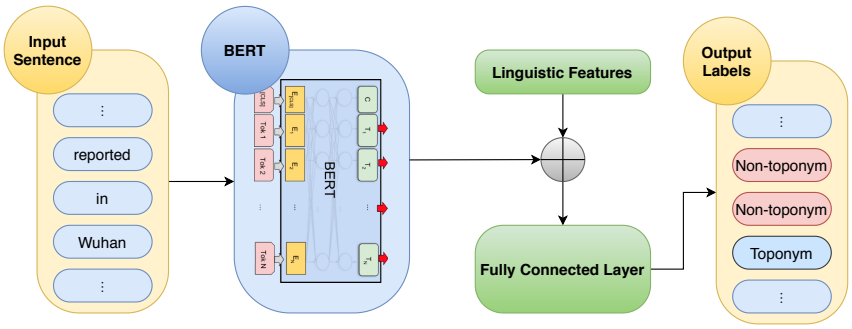

Figure 1: TIMBERT model architecture.

Table 1: Statistics of the dataset.

	Training	Development	Test
Number of articles	90	15	45
Average size of each article (in words)	6422	5191	6146
Average number of toponyms per article	43	44	50

4 Experiments and Results

Our models were evaluated on the SemEval 2019 task 12 (Weissenbacher et al., 2019) dataset. The dataset consists of 150 articles from PubMed annotated with toponym mentions. The dataset was split into 3 sections: training, validation, and test set containing 60%, 10%, and 30% of the dataset respectively. Table 1 shows statistics of the dataset. Table 2 shows the performance of our basic model i.e. TIMBERT-Base without any linguistic features (see row #7) compared to the state-of-the-art system (Wang et al., 2019) (see row #4). A series of experiments was carried out to evaluate the influence of a variety of parameters on the performance of the model, which are described in the following sections.

Stop Words As Table 2 (row #10) shows, the removal of stop words using the NLTK stop words corpus (Bird et al., 2009) of 179 words, significantly worsens the F1 performance of the model (from 81.36% to 72.08%). We hypothesize that some stop words such as *in* do help the system detect toponyms as they provide a learnable structure for detection of toponyms. Moreover, the BERT model is trained to capture contextual information and language structures during pretraining. Therefore, to transfer its knowledge to a downstream task, it should be given fully comprehensible and structured sentences. Stop words removal distorts sentence structure and therefore harms the model performance.

Punctuation We then trained the TIMBERT-Base without any punctuation information. As Table 2 (row #9) shows, the removal of punctuation marks decreased the F1 from 81.36% to 79.39%. A manual error analysis showed that many toponyms appear inside parenthesis, near a dot at the end of a sentence, or after a comma (e.g. *(Kara, Togo)*). Hence, as suggested in (Davari et al., 2019; Gelernter and Balaji, 2013), punctuation is a good indicator of toponyms and should not be ignored.

Table 2: Performance of the TIMBERT based models and the state-of-the-art model (Wang et al., 2019).

#	Model	Precision	Recall	F1
1	TIMBERT-Large-CoNLL-w/-Orthographic-Pruned-10%	90.51%	91.19%	90.85%
2	TIMBERT-Large-CoNLL-w/-Orthographic-Pruned-20%	90.02%	90.72%	90.37%
3	TIMBERT-Large-CoNLL-w/-Orthographic	89.73%	90.23%	89.98%
4	SOTA (Wang et al., 2019)	92.92%	85.64%	89.13%
5	TIMBERT-Large-w/-Orthographic	83.41%	86.88%	85.11%
6	TIMBERT-Base-w/-Orthographic	82.61%	83.19%	82.90%
7	TIMBERT-Base	82.59%	80.17%	81.36%
8	TIMBERT-Base-w/-POS	81.96%	80.08%	81.01%
9	TIMBERT-Base-w/o-Punctuation	80.04%	78.75%	79.39%
10	TIMBERT-Base-w/o-Stop-Words	72.14%	72.01%	72.08%

Part of Speech Tags The majority of the work in toponym detection have indicated that augmenting the model with part of speech (POS) tags results in an improvement in the performance of the model. On the other hand, the BERT model has shown a great ability to capture a number of linguistic features and transferring them to downstream tasks (Clark et al., 2019). Since BERT constitutes the backbone of TIMBERT, we investigated whether or not our model is aware of POS tags. We used the NLTK POS tagger (Bird et al., 2009) which uses the Penn Treebank tagset (Marcus et al., 1993). As indicated in Table 2 (row #8), including the POS tags reduced the performance of the model from the F1 of 81.36% to 81.01%. This suggests that the TIMBERT-Base model is already aware of the POS tags since the augmentation of the model with POS tags does not affect its performance.

Orthographic Features This feature is presented to the model as a one-hot encoded vector capturing whether a word is capitalized (e.g. *Togo*), uncapitalized (e.g. *cell*), or written in uppercase letters (e.g. *UK*). Since in the preprocessing of the data, all tokens are lowercased, the model is unaware of this feature. Our experiments showed that augmenting the model with the orthographic information results in an increase of the F1 performance from 81.36% to 82.90% (see Table 2, row #6).

Backbone Model Experiments on the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018) have shown great performance gains when the backbone model is switched from BERT-Base to BERT-Large (Devlin et al., 2019). Motivated by these findings, we investigated the impact of BERT-Large on our task. We replaced the backbone module with BERT-Large in our best performing model from previous experiments (Table 2 row #6) and observed an improvement in the F1 performance from 82.90% to 85.11% (see Table 2, row #5).

Precursory Fine-tuning Precursory fine-tuning of a language model on an objective comparable to the target task can be seen as a form of distant supervision, where a knowledge resource is exploited to gather possibly noisy training instances (Mintz et al., 2009). We investigated the effect of precursory fine-tuning on the task of toponym detection. We used the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003) and filtered it to only include instances of location names in English and established a training set with 8.5k sentences. We first fine-tuned our best performing model architecture (Table 2 row #5) on this dataset and used early stopping to end the training process. We then further fine-tuned the network with SemEval 2019 task 12 dataset (Weissenbacher et al., 2019) and observed a significant improvement in the F1 performance, from 85.11% to 89.98% (see Table 2, rows #5 and #3).

Pruning and Regularization Parameter pruning of neural networks could be seen as a form of permanent dropout that increases regularization and results in better generalization performance (Srivastava et al., 2014). It can also be seen as a form of L_0 regularization as it encourages sparse model representations that has shown to improve generalization performance (Louizos et al., 2018). We investigated the extent of the parameter pruning paradigm proposed by Frankle and Carbin (2019) on large scale transfer learning for toponym detection. We used a Monte Carlo estimate of the Lottery Ticket[2] to prune and regularize our best performing model from previous experiments (Table 2, row #3). We observed that the compressed models outperformed the original uncompressed model at 10% and 20% pruning level with 90.85% and 90.37% F1 respectively (see Table 2, rows #1 and #2).

5 Discussion

Our search for a toponym identifier for the medical domain with little to no task-specific design choices led us to the development of the TIMBERT models. Our experiments with BERT as the backbone of our models detailed in Section 4 confirmed that certain linguistic insights such as POS tags are seamlessly transferred to downstream tasks while others, such as orthographic features need to be integrated to the model. Our experiments with the typical preprocessing techniques led to poor model performances. This suggests the need for a general agreement between the textual structure of the data during pretraining and fine-tuning.

[2]Publication in preparation.

6 Conclusion and Future Work

In this work, we presented a competitive model for toponym detection in the medical domain that significantly improves the state-of-the-art performance. We developed a family of toponym detection models and used BERT as the backbone of our models. In future studies, we will investigate the effects of using other language models, such as XLNET (Yang et al., 2019), RoBERTa (Liu et al., 2019), and ALBERT (Lan et al., 2019), for the backbone module. We experimented with parameter pruning to regularize our models as well as to reduce their computational complexity. In future studies, we will examine other knowledge distillation techniques in order to find competitive and resource conservative models for toponym identification in the medical domain. Our experiments with precursory fine-tuning resulted in significant performance improvement of our model. Further research can determine if precursory task specific fine-tuning is helpful for other NLP tasks. This can potentially lead to the development of task specific pretrained models for more efficient transfer learning, especially for tasks with small datasets.

Acknowledgements

The authors would like to thank the anonymous reviewers for their feedback on a previous version of this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Einat Amitay, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. Web-a-where: Geotagging web content. In *Proceedings of the 27th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2004)*, pages 273–280, Sheffield, UK, July. ACM.

Mariona Coll Ardanuy and Caroline Sporleder. 2017. Toponym disambiguation in historical documents using semantic and geographic features. In *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, pages 175–180. ACM, June.

Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. 2012. Genbank. *Nucleic Acids Research*, 41(D1):D36–D42, November.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc., June.

Huijun Chen, Juanjuan Guo, Chen Wang, Fan Luo, Xuechen Yu, Wei Zhang, Jiafu Li, Dongchi Zhao, Dan Xu, Qing Gong, et al. 2020. Clinical characteristics and intrauterine vertical transmission potential of COVID-19 infection in nine pregnant women: A retrospective review of medical records. *The Lancet*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP (ACL 2019)*, pages 276–286, Florence, Italy, August.

MohammadReza Davari, Leila Kosseim, and Tien D. Bui. 2019. Toponym identification in epidemiology articles a deep learning approach. In *Proceedings of The 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*, La Rochelle, France, April.

Grant DeLozier, Jason Baldridge, and Loretta London. 2015. Gazetteer-independent toponym resolution using geographic word profiles. In *Proceedings of 29th Annual Conference of Association for the Advancement of Artificial Intelligence (AAAI 2015)*, pages 2382–2388, Austin, USA, February.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*, pages 4171–4186, Minneapolis, USA, June.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of The 7th International Conference on Learning Representations (ICLR 2019)*, New Orleans, USA, May.

Eric Garbin and Inderjeet Mani. 2005. Disambiguating toponyms in news. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pages 363–370, Vancouver, Canada, October.

P. Gautret, E. Botelho-Nevers, P. Brouqui, and P. Parola. 2012. The spread of vaccine-preventable diseases by international travellers: A public-health concern. *Clinical Microbiology and Infection*, 18:77 – 84, October.

Judith Gelernter and Shilpa Balaji. 2013. An algorithm for local geoparsing of microtext. *GeoInformatica*, 17(4):635–667, January.

A. D. Green and K. I. Roberts. 2000. Recent trends in infectious diseases for travellers. *Occupational Medicine*, 50(8):560–565, November.

W. Kienreich, M. Granitzer, and M. Lux. 2006. Geospatial anchoring of encyclopedia articles. In *Proceedings of the 10th International Conference on Information Visualisation (IV 2006)*, pages 211–215, London, UK, July.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, September.

Haonan Li, Minghan Wang, Timothy Baldwin, Martin Tomko, and Maria Vasardani. 2019. UniMelb at SemEval-2019 task 12: Multi-model combination for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 1313–1318, Minneapolis, USA, June.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, July.

Christos Louizos, Max Welling, and Diederik P Kingma. 2018. Learning sparse neural networks through l_0 regularization. In *Proceedings of 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, Canada, April.

Arjun Magge, Davy Weissenbacher, Abeed Sarker, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2018. Deep neural networks and distant supervision for geographic location mention extraction. *Bioinformatics*, 34(13):i565–i573.

Matthew Magnusson and Laura Dietz. 2019. UNH at SemEval-2019 task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 1308–1312, Minneapolis, USA, June.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, June.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL/IJNLP 2009)*, pages 1003–1011, Suntec, Singapore, August.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL/HLT 2018)*, pages 2227–2237, New Orleans, USA, June.

Ross S Purves, Paul Clough, Christopher B Jones, Avi Arampatzis, Benedicte Bucher, David Finch, Gaihua Fu, Hideo Joho, Awase Khirni Syed, Subodh Vaid, et al. 2007. The design and implementation of spirit: A spatially aware search engine for information retrieval on the internet. *International journal of geographical information science*, 21(7):717–745, June.

Tao Qi, Suyu Ge, Chuhan Wu, Yubo Chen, and Yongfeng Huang. 2019. THU_NGN at SemEval-2019 task 12: Toponym detection and disambiguation on scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 1302–1307, Minneapolis, USA, June.

Teng Qin, Rong Xiao, Lei Fang, Xing Xie, and Lei Zhang. 2010. An efficient location extraction algorithm by leveraging web contextual information. In *Proceedings of the 18th International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2010)*, pages 53–60, San Jose, USA, November. ACM.

Matthew Scotch, Indra Neil Sarkar, Changjiang Mei, Robert Leaman, Kei-Hoi Cheung, Pierina Ortiz, Ashutosh Singraur, and Graciela Gonzalez. 2011. Enhancing phylogeography by improving geographical information from genbank. *Journal of biomedical informatics*, 44:S44–S47.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, June.

Mike Taylor. 2017. *Reduced Geographic Scope as a Strategy for Toponym Resolution*. Ph.D. thesis, Department of Engineering and Computer Science, Northern Arizona University, December.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning at (CoNLL 2003)*, pages 142–147, Edmonton, Canada, May.

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 (NIPS 2017)*, pages 5998–6008, Long Beach, USA, January.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November.

Xiaobin Wang, Chunping Ma, Huafei Zheng, Chu Liu, Pengjun Xie, Linlin Li, and Luo Si. 2019. DM_NLP at SemEval-2019 task 12: A pipeline system for toponym resolution. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 917–923, Minneapolis, USA, June.

Davy Weissenbacher, Arjun Magge, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2019. SemEval-2019 task 12: Toponym resolution in scientific papers. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, Minneapolis, USA, June.

World Health Organization. 2020a. Coronavirus disease 2019 (COVID-19): situation report, 46. *Situation Reports*.

World Health Organization. 2020b. Rolling updates on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen, March. [Online; accessed 2020-03-19].

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Vikas Yadav, Egoitz Laparra, Ti-Tai Wang, Mihai Surdeanu, and Steven Bethard. 2019. University of Arizona at SemEval-2019 task 12: Deep-affix named entity recognition of geolocation entities. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*, pages 1319–1323, Minneapolis, USA, June.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NIPS 2019)*, pages 5753–5763. Curran Associates, Inc., Vancouver, Canada, December.