Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text

Lukas Lange[1,2,3] **Anastasiia Iurshina**[1] **Heike Adel**[1] **Jannik Strötgen**[1]
[1] Bosch Center for Artificial Intelligence, Renningen, Germany
[2] Spoken Language Systems (LSV), Saarland University, Saarbrücken, Germany
[3] Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany
{Lukas.Lange,Heike.Adel,Jannik.Stroetgen}@de.bosch.com

Abstract

Although temporal tagging is still dominated by rule-based systems, there have been recent attempts at neural temporal taggers. However, all of them focus on monolingual settings. In this paper, we explore multilingual methods for the extraction of temporal expressions from text and investigate adversarial training for aligning embedding spaces to one common space. With this, we create a single multilingual model that can also be transferred to unseen languages and set the new state of the art in those cross-lingual transfer experiments.

1 Introduction

The extraction of temporal expressions from text is an important processing step for many applications, such as topic detection and questions answering (Strötgen and Gertz, 2016). However, there is a lack of multilingual models for this task. While recent temporal taggers, such as the work by Laparra et al. (2018) focus on English, only little work was dedicated to multilingual temporal tagging so far.

Strötgen and Gertz (2015) proposed to automatically generate language resources for the rule-based temporal tagger HeidelTime, but all of these models are language specific and can only process texts from a fixed language. In this paper, we propose to overcome this limitation by training a single model on multiple languages to extract temporal expressions from text. We experiment with recurrent neural networks using FastText embeddings (Bojanowski et al., 2017) and the multilingual version of BERT (Devlin et al., 2019). In order to process multilingual texts, we investigate an unsupervised alignment technique based on adversarial training, making it applicable to zero- or low-resource scenarios and compare it to standard dictionary-based alternatives (Mikolov et al., 2013).

We demonstrate that it is possible to achieve competitive performance with a single multilingual model trained jointly on English, Spanish and Portuguese. Further, we demonstrate that this multilingual model can be transferred to new languages, for which the model has not seen any labeled sentences during training by applying it to unseen French, Catalan, Basque, and German data. Our model shows superior performance compared to HeidelTime (Strötgen and Gertz, 2015) and sets new state-of-the-art results in the cross-lingual extraction of temporal expressions.

2 Related Work

Temporal Tagging. The current state of the art for temporal tagging are rule-based systems, such as HeidelTime (Strötgen and Gertz, 2013) or SUTime (Chang and Manning, 2012). In particular, HeidelTime uses a different set of rules depending on the language and domain. Strötgen and Gertz (2015) automatically generated HeidelTime rules for more than 200 languages in order to support many languages. However, the quality of these rules does not match the high quality of manually created rules and the models are still language specific. Aside from rule-based systems, Lee et al. (2014) proposed to learn context-dependent semantic parsers for extracting temporal expressions from text. Laparra et al. (2018) made a first step towards neural models by using recurrent neural networks. However, they only performed experiments on English corpora using monolingual models. In contrast, we propose a truly multilingual model.

Multilingual Embeddings. Recently, it became popular to train embedding models on resources from many languages jointly (Lample and Conneau, 2019; Conneau et al., 2019). For example, multilingual BERT (Devlin et al., 2019) was trained on Wikipedia articles from more than 100 languages. Although performance improvements show the possibility to use multilingual BERT in

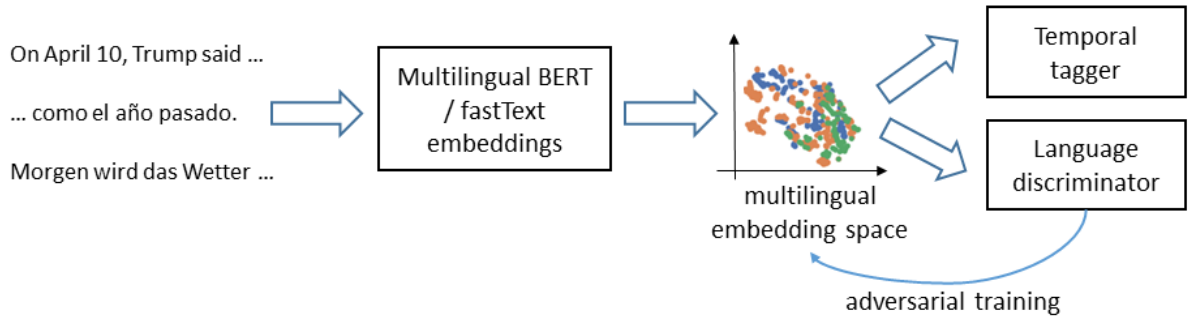


Figure 1: Overview of our multilingual system with adversarial training for improving the embedding space.

monolingual (Hakala and Pyysalo, 2019), multilingual (Tsai et al., 2019) and cross-lingual settings (Wu and Dredze, 2019), it has been questioned whether multilingual BERT is truly multilingual (Pires et al., 2019; Singh et al., 2019; Libovický et al., 2019). Therefore, we will investigate the benefits of aligning its embeddings in our experiments.

Aligning Embedding Spaces. A common method to create multilingual embedding spaces is the alignment of monolingual embeddings (Mikolov et al., 2013; Joulin et al., 2018). Smith et al. (2017) proposed to align embedding spaces by creating orthogonal transformation matrices based on bilingual dictionaries, which we use as baseline alignment method.

It was shown that BERT can also benefit from alignment, i.a. in cross-lingual (Schuster et al., 2019; Liu et al., 2019) or multilingual settings (Cao et al., 2020). In contrast to prior work, we experiment with aligning BERT using adversarial training, which is related to using adversarial training for domain adaptation (Ganin et al., 2016), coping with bias or confounding variables (Li et al., 2018; Raff and Sylvester, 2018; Zhang et al., 2018; Barrett et al., 2019; McHardy et al., 2019) or transferring models from a source to a target language (Zhang et al., 2017; Keung et al., 2019; Wang et al., 2019). Similar to Chen and Cardie (2018), we use a multinomial discriminator in our setting.

3 Methods

We model the task of extracting temporal expressions as a sequence tagging problem and explore the performance of state-of-the-art recurrent neural networks with FastText and BERT embeddings, respectively. In particular, we train multilingual models that process all languages in the same model. To create and improve the multilingual embedding spaces, we propose an unsupervised alignment approach based on adversarial training and compare it to two baseline approaches. Figure 1 provides an overview of the system. The different components are described in detail in the following.

3.1 Temporal Expression Extraction Model

Following previous work, e.g., Lample et al. (2016), we train a bidirectional long-short term memory network (BiLSTM) (Hochreiter and Schmidhuber, 1997) with a conditional random field (CRF) (Lafferty et al., 2001) output layer. As input, we experiment with two embedding methods: (i) pre-trained FastText (Bojanowski et al., 2017) word embeddings from multiple languages,[1] and (ii) multilingual BERT (Devlin et al., 2019) embeddings.[2] For BERT, we use the averaged output of the last four layers as input to the BiLSTM and fine-tune the whole model during the training of temporal information extraction. We also experimented with a BERT setup similar to Devlin et al. (2019) where the embeddings are directly mapped to the label space and the softmax function is used to compute the label probabilities instead of a CRF. However, we found superior performance for the BiLSTM-CRF models.

3.2 Alignment of Embeddings

We propose an unsupervised approach based on adversarial training to align multilingual embeddings in a common space (Section 3.2.2) and compare it with two approaches from related work based on linear transformation matrices (Section 3.2.1).

[1] https://fasttext.cc/docs/en/crawl-vectors.html

[2] https://github.com/google-research/bert/blob/master/multilingual.md

3.2.1 Baseline Alignment

Embedding spaces are typically aligned using a linear transformation based on bilingual dictionaries. We follow the work from Smith et al. (2017), and align embedding spaces based on orthogonal transformation matrices. These matrices can either be constructed in an unsupervised way by using words that appear in the vocabularies from both languages, i.e., equal words that can be identified using string matching, or in a supervised way based on real-world dictionaries (Mikolov et al., 2013; Joulin et al., 2018). For the latter method, we build dictionaries based on translations from wiktionary.[3] For both methods, we reduce the vocabularies to the most frequent 5k words per language and treat English as the pivot language.

3.2.2 Adversarial Alignment

We propose to use gradient reversal training to align embeddings from different (sub)spaces in an unsupervised way. Note that neither dictionaries nor other language resources are needed for this approach, making it applicable to zero- or low-resource scenarios. In particular, we extend the extraction model C with a discriminator D. Both model parts are trained alternately in a multi-task fashion. The *feature extractor* F is shared among them and consists of the embedding layer E, followed by a non-linear mapping: $F(x) = \tanh(W^\top E(x))$ with x being the current word, $W \in \mathbb{R}^{S \times S}$ and S being the embedding dimensionality.

The *discriminator* D is a multinomial non-linear classifier consisting of one hidden layer with ReLU activation (Hahnloser et al., 2000): $D(x) = \text{softmax}(T^\top \text{ReLU}(V^\top F(x)))$ with $V \in \mathbb{R}^{S \times H}$, $T \in \mathbb{R}^{H \times O}$, H being a hyperparameter and O the number of different languages.

In total, we distinguish three sets of parameters: θ_C: the parameters of the downstream classification model (i.e., the temporal tagger), θ_D: the parameters of the discriminator, and θ_F: the parameters of the feature extractor. The loss functions of the temporal tagger L_C and of the discriminator L_D are cross-entropy loss functions. While θ_C and θ_D are updated using standard gradient descent, gradient reversal training updates θ_F as follows:

$$\theta_F = \theta_F - \eta\left(\frac{\partial L_C}{\partial \theta_F} - \lambda \frac{\partial L_D}{\partial \theta_F}\right) \quad (1)$$

[3] https://github.com/open-dsl-dict/wiktionary-dict

Dataset	Train	Dev	Test
English (EN)	3,461/*1,456*	420/*164*	354/*202*
Spanish (ES)	1,705/*972*	189/*122*	332/*199*
Portuguese (PT)	3,501/*948*	389/*100*	481/*172*
French (FR)	-	-	708/*424*
German (DE)	-	-	2,666/*500*
Catalan (CA)	-	-	1,944/*1389*
Basque (EU)	-	-	163/*123*

Table 1: Number of sentences / *temporal expressions* per corpus. The lower part is only used for evaluation.

with η being the learning rate and λ a hyperparameter to control the discriminator influence. Thus, θ_F is updated in the opposite direction of the gradients from the discriminator loss, making the discriminator an adversary. With this, the discriminator is optimized for predicting the correct origin language of a given sentence, but at the same time the feature extractor gets updated with gradient reversal, such that the language detection becomes harder and the discriminator cannot easily distinguish the word representations from different languages.

4 Experiments and Results

4.1 Evaluation Metrics and Datasets

For evaluation, we use the TempEval3 evaluation script and report strict and relaxed extraction F_1 score for complete and partial overlap to gold standard annotations, respectively. We also report the type F_1 score for the classification into the four temporal types: Date, Time, Duration, and Set.

Our multilingual models are trained using the Portuguese TimeBank (Costa and Branco, 2012) and TempEval3 (UzZaman et al., 2013) for Spanish and English (TimeBank subset). To demonstrate that our model is able to generalize to unseen languages, we perform tests using the French (Bittar et al., 2011), Catalan (Saurı and Badia, 2012) and Basque (Altuna et al., 2016) TimeBanks and the Zeit subset of the German KRAUTS corpus (Strötgen et al., 2018). Corpus statistics are shown in Table 1.

4.2 Hyperparameters and Model Training

We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 1e−5 for the BiLSTM-CRF model part and 1e−6 for BERT. The model is trained for a maximum of 50 epochs using early stopping on the development set. The BiLSTM has a hidden size of 128 units per direction. The labels are encoded in the IOB2 format. For

Task	Metric	HeidelTime	FastText unaligned	FastText aligned w/o Dict.	FastText aligned w/ Dict	FastText aligned w/ AT	BERT unaligned	BERT aligned w/ AT
EN	strict	**81.78**	68.36	69.10	70.80	75.63 †	73.09	74.80 †
	relaxed	**90.71**	79.14	79.03	81.21	82.03 †	84.34	86.61 †
	type	**83.27**	72.13	72.18	73.32	72.85 †	75.50	79.53 †
ES	strict	**85.87**	75.67	76.53	77.44	79.64 †	79.11	79.55
	relaxed	**90.13**	82.43	82.45	82.47	84.46 †	84.12	85.71
	type	**87.47**	78.07	78.46	78.24	80.88 †	80.22	80.11
PT	strict	71.59	70.36	70.20	70.48	72.41	74.52	**75.47**
	relaxed	81.18	76.77	75.86	76.29	78.15	80.75	**81.51**
	type	**76.75**	72.29	71.50	72.26	73.84	75.47	76.23

Table 2: Results for multilingual models trained on English, Spanish and Portuguese data jointly. †highlights aligned models with statistical significant differences to the unaligned model (paired permutation test, p=0.05).

regularization, we apply dropout with a rate of 10% after the input embeddings. The discriminator for adversarial training has a hidden size H of 100 units and is trained after every 10^{th} batch of the sequence tagger with λ set to 0.001.

4.3 Results

The results for the multilingual experiments are shown in Table 2. We trained three models with different random seeds and report the performance of the model with median performance on the combined development set of all languages. Current state of the art for English (Lee et al., 2014) achieves 83.1/91.4/85.4 for strict/relaxed/type F_1. However, this is a monolingual model that can only be applied to English.

The effects of aligning FastText embeddings are clearly visible in Table 2. The supervised alignment using a dictionary is always superior compared to the unsupervised alignment without a dictionary or the unaligned embeddings. Our proposed adversarial alignment (w/ AT) leads to the best results across languages. The performance of BERT is close to the best FastText model.[4] Aligning BERT with adversarial training also increases performance. The improvements are smaller compared to FastText but still statistically significant for English.

Table 3 provides transfer results of the models with BERT embeddings to languages without labeled training data.[5] In particular, the model using the Wikipedia data for training the discriminator is effective in generalizing to languages without training resources for temporal expression extraction, as these languages are also aligned during model training. It outperforms the state-of-the-art HeidelTime models by a large margin. The impressive performance of the multilingual BERT in the cross-lingual setting can be explained by the fact that the model has seen many sentences in our target languages during the pre-training phase, which can now be effectively leveraged in this new setting.

Task	Metric	HeidelTime -Auto	BERT unalign.	BERT aligned w/ AT
FR	strict	52.35	60.12	**62.58**
	relaxed	72.02	74.23	**75.46**
	type	**68.70**	61.96	62.07
DE	strict	38.87	63.34	**66.53**
	relaxed	52.11	76.51	**77.82**
	type	50.15	66.95	**69.04**
CA	strict	28.11	63.24	**64.21**
	relaxed	62.81	74.95	**77.00**
	type	60.84	65.66	**67.85**
EU	strict	22.54	43.96	**47.87**
	relaxed	26.76	61.54	**63.83**
	type	23.94	57.14	**58.51**

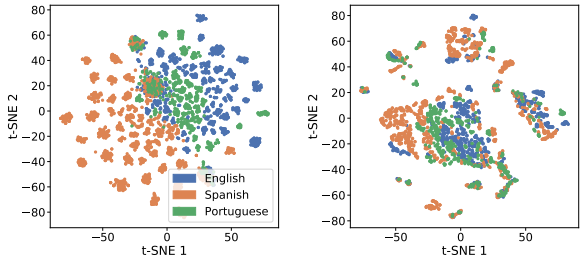
Table 3: Results for the unsupervised cross-lingual setting. We compare to HeidelTime with automatically generated resources, which resembles a similar setting.

4.4 Analysis

The embedding spaces of BERT before and after aligning are shown in Figure 2. The left sub-figure presents the original BERT embeddings without any fine-tuning. In this visualization, clear clusters for each language exist. After fine-tuning on multilingual temporal expression extraction and adversarial alignment (right sub-figure) the clusters for each language mostly disappear.

[4] Additional experiments with the multilingual XLM model (Lample and Conneau, 2019) trained on 100 languages led to similar results as the multilingual BERT model.

[5] The results of the FastText models were considerably lower for cross-lingual transfer.



(a) without training.

(b) after joint temporal extraction and adversarial training.

Figure 2: t-SNE plots of the last BERT layer without any training (left) and after training (right).

5 Conclusion

In this paper, we investigated how a multilingual neural model with FastText or BERT embeddings can be used to extract temporal expressions from text. We investigated adversarial training for creating multilingual embedding spaces. The model can effectively be transferred to unseen languages in a cross-lingual setting and outperforms a state-of-the-art model by a large margin.

Acknowledgments

We would like to thank the members of the BCAI NLP&KRR research group and the anonymous reviewers for their helpful comments.

References

Begoña Altuna, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2016. Adapting timeml to basque: Event annotation. In *International Conference on Intelligent Text Processing and Computational Linguistics*.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French TimeBank: An ISO-TimeML annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Xilun Chen and Claire Cardie. 2018. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Francisco Costa and António Branco. 2012. TimeBankPT: A TimeML annotated corpus of Portuguese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1).

Richard HR Hahnloser, Rahul Sarpeshkar, Misha A Mahowald, Rodney J Douglas, and H Sebastian Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, 405(6789).

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual BERT. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computing*, 9(8).

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Phillip Keung, yichao lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Egoitz Laparra, Dongfang Xu, and Steven Bethard. 2018. From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations. *Transactions of the Association for Computational Linguistics*, 6.

Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. 2014. Context-dependent semantic parsing for time expressions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert? *arXiv preprint arXiv:1911.03310*.

Qianchu Liu, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2019. Investigating cross-lingual alignment methods for contextualized embeddings with token-level evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

E. Raff and J. Sylvester. 2018. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*.

Roser Saurı and Toni Badia. 2012. Catalan timebank 1.0 corpus documentation. Technical report.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*.

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2).

Jannik Strötgen and Michael Gertz. 2015. A baseline temporal tagger for all languages. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Jannik Strötgen and Michael Gertz. 2016. Domain-sensitive temporal tagging. *Synthesis Lectures on Human Language Technologies*, 9(3).

Jannik Strötgen, Anne-Lyse Minard, Lukas Lange, Manuela Speranza, and Bernardo Magnini. 2018. KRAUTS: A German temporally annotated news corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical BERT models for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*.

Haozhou Wang, James Henderson, and Paola Merlo. 2019. Weakly-supervised concept-based adversarial learning for cross-lingual word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.