Time for some German? Pre-Training a Transformer-based Temporal Tagger for German

Satya Almasian[1,0], Dennis Aumiller[1,0] and Michael Gertz[1]

[1] *Institute of Computer Science, Heidelberg University, Heidelberg, Germany*

[0] *These authors contributed equally to this work*

Abstract

Non-English languages are notorious for their lack of available resources, and temporal tagging is no exception. In this work, we explore transfer strategies to improve the quality of a German temporal tagger. From a model perspective, we employ a weakly-supervised pre-training strategy to stabilize the convergence of Transformer-based taggers. In addition, we also augment data with automatically translated English resources, which serve as an alternative to commonly used alignments of latent embedding spaces. With this, we provide preliminary empirical evidence that indicates the suitability of transfer approaches to other low-resourced languages: A small number of gold data coupled with an existing data set in a resource-rich language and a weak labeling baseline system may be sufficient to boost performance.

Keywords

Temporal tagging, Weakly-supervised learning, German

1. Introduction

Annotated data has become an essential part of modern-day NLP approaches, but non-English resources remain scarce. In the absence of data, it then becomes increasingly difficult to even transfer existing approaches to a multilingual context. In this work, we particularly focus on the task of Temporal Tagging, which serves a multitude of downstream applications in the area of narrative extraction [1]. For example, more accurate temporal tags can be utilized in timeline summarization [2, 3] or event reasoning [4]. For temporal tagging, too, the largest resources exist without a doubt for English [5, 6, 7, 8]. While some non-English resources do exist [9, 10], they are still scarce, and generally smaller than their English counterparts. Despite attempts to approach the lack of language-specific resources through the lens of multilingual transfer learning [11, 12], Heideltime [13, 14], a rule-based approach extending to multiple languages, remains state-of-the-art. Yet, rule-based approaches generally suffer from a precision-heavy tagging, since slight variations on patterns cannot be successfully detected. By applying state-of-the-art neural models instead, such variations could be covered as well, increasing the overall tagging performance. However, the lack of available data makes the training of data-hungry neural models non-trivial. We illustrate a generic transfer pipeline with German

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'22 Workshop, Stavanger (Norway), 10-April-2022

almasian@informatik.uni-heidelberg.de (S. Almasian); aumiller@informatik.uni-heidelberg.de (D. Aumiller); gertz@informatik.uni-heidelberg.de (M. Gertz)

as an example of a lower-resource language. By using a combination of automatically labeled data for pre-training and additional translated English data, we boost the amount of available training data. With this augmented corpus, we are able to fine-tune Transformer models that improve temporal tagging performance for German.

2. Related Work

The main reference point for temporal tagging of non-English resources is Heideltime [13, 14], which provides automatically transduced rules for other languages; the coverage varies depending on the language's syntactic structure. At the same time, they also provide language-specific rules for a smaller set of languages, including German.

As for datasets, this work relies on the KRAUTS corpus [9], which consists of roughly 1,100 annotations of Tyrolian and German newspaper articles. WikiwarsDE [15] is another German-specific resource, yet, the temporal annotations are not available in the current TIMEX3 format, limiting their applicability for recent models.

Approaches dealing with German include Lange et al. [11], who experimented with adversarially aligned embeddings. While their method beats the automatically translated rule set of Heideltime, it falls short of the language-specific rule set. With a similar strategy, Starý et al. [12] fine-tuned a multilingual version of BERT with OntoNotes data. Both works use KRAUTS data for evaluation, and have the advantage of automatically scaling to several target languages, however, at the cost of language-specific performance.

Another notable multilingual dataset is TimeBank [16, 17, 18, 19], which covers several languages including French, Italian, Portuguese and Romanian. Taggers in low-resource settings are generally limited, but do exist: TipSem [10] and Annotador [20] for Spanish, Bosque-T0 [21] and the work by Costa and Branco [22] for Portuguese, and PET [23] for Persian.

3. A Transfer Pipeline for Temporal Tagging

Temporal tagging is the task of identification of temporal expression, classification of the type and sometimes normalization of temporal values. In the work, we focus on identification and classification of expression in four classes defined by TIMEX3 schema, namely `DATE`, `TIME`, `SET` and `DURATION`. As previously mentioned, language-specific resources tend to perform better than multilingual approaches. Therefore, we set out to construct a language-specific German tagging approach with the help of Transformer-based language models [24]. We utilize monolingual language models in this work, opposed to previously utilized multilingual networks. Specifically, Chan et al. [25] present several iterations of German-specific Transformer networks; we choose the best-performing model, which is based on the ELECTRA [26] architecture, namely GELECTRA-large.

However, successfully employing the Transformer networks requires more data than what is available in KRAUTS dataset [9]. For this purpose, we create a corpus of automatically tagged news articles, using Heideltime's German tagger. This provides around 500,000 temporal expressions for an additional "pre-training step", exceeding the available German tagging data by roughly 2,000 times, albeit at a lower guarantee of annotation quality.

Table 1
Statistics of the training resources with TIMEX3 tag distribution. Note that the values for TempEval refer to tags after automated translation. DATE, SET, DURATION, TIME are the temporal types.

	#Docs	#Expressions	DATE	SET	DURATION	TIME
HeideltimeDE train	64,299	400,824	292,388	2,502	66,867	39,067
HeideltimeDE test	14,768	97,981	66,713	634	13,892	16,742
TempEvalDE train	256	1,782	1,455	30	251	30
KRAUTS *Dolomiten* (train)	142	587	376	19	94	98
KRAUTS *Die Zeit* (test)	50	553	358	39	144	12

We further experiment with automatically translated English data, based on the TempEval-3 corpus [7]. Articles were automatically translated with the help of Google Translate[1], and we were able to retain about 90% of the original annotations in the German version. See Table 1 for a detailed comparison, including the tag distribution.

4. Experiments

For experimentation, we use the KRAUTS *Dolomiten* subset as the training set, and the *Die Zeit* subset for testing. Further, all models were run on three NVIDIA A100 GPUs using the Adam optimizer and linear weight decay. Pre-training was performed for 4 epochs, with a learning rate of $1e$-7 and batch size 16 on each GPU and gradient accumulation step of 4, which took approximately 30 hours. Variants with automatically translated TempEval data were trained an additional 8 epochs with batch size 16 and learning rate of $5e$-5 on a single GPU before the final fine-tuning on *Dolomiten* for another 8 epochs. All metrics on fine-tuned models are averaged for 3 different random seeds; pre-training was run once without pre-determined random seeds. We use the official TempEval-3 script for computing results, which also works with German texts. TempEval generally differentiates between partial ("*relaxed*") and exact ("*strict*") tagging overlap.

4.1. Results

Table 2 contains all available results. Note that the adversarially trained model by Lange et al. [11] has transferred from English data, and seen no explicit German training data, which explains its lower performance. The mBERT NER model [12] does not perform type classification. We identify Heideltime as the best-performing baseline system, where its rule-based nature tends to favor precision over recall.

To investigate the effect of continued pre-training, we report results for both off-the-shelf variants and additionally pre-trained models (denoted by "p"). Pre-training was performed on the automatically labeled portion (HeideltimeDE train). "*+ temp*" denotes fine-tuning on translated TempEval data, and "*+ dolo*" fine-tuning on *Dolomiten* data, respectively. For fine-tuning on both sets together, we first train for 8 epochs on TempEval data, and then for another 8 epochs on *Dolomiten*.

[1] translate.google.com, accessed: 2022-01-14

Table 2
Tagging performance on the KRAUTS *Die Zeit* subset; bold highlights indicate best performance. For mBERT results [12], it is unclear whether the entire KRAUTS dataset was used instead. Lange et al. [11] only report F1 scores for their results, which is why the exact precision and recall scores are unknown. Our own results are averaged across three fine-tuning runs with varying random seeds.

		Strict			Relaxed		Type
Method	**F-1**	**Prec.**	**Recall**	**F-1**	**Prec.**	**Recall**	**F-1**
Heideltime	69.72	**77.11**	63.62	79.30	**87.71**	72.37	75.38
Adversarial BERT [11]	66.53	?	?	77.82	?	?	69.04
mBERT NER [12]	43.15	53.92	35.96	64.94	64.94	54.13	–
GELECTRA + *dolo*	75.51	73.06	78.13	85.88	83.09	**88.87**	78.96
GELECTRA + *temp* + *dolo*	70.71	70.52	70.91	84.25	84.01	84.49	75.85
GELECTRA$_p$	65.45	71.10	60.64	77.90	84.62	72.17	73.82
GELECTRA$_p$ + *dolo*	**76.13**	73.52	**78.93**	85.33	82.41	88.47	**80.06**
GELECTRA$_p$ + *temp* + *dolo*	75.32	74.03	76.68	**86.13**	84.65	87.67	79.49

Overall, our best model for relaxed matching (86.13 F1) is GELECTRA$_p$ + *temp* + *dolo*. However, it appears that the automatically translated data is somewhat misleading for strict matches; GELECTRA$_p$ + *dolo*, which is only trained on *Dolomiten*, has the highest strict match, as well as best type classification performance. Since the teacher, Heideltime, is precision-focused, all pre-trained variants also carry slightly higher precision, implying that the choice of weak labeler for pre-training directly affects the fine-tuning performance as well. Variants without pre-training are in comparison more recall-oriented. It is worth noting that even without any fine-tuning and only pre-training, GELECTRA$_p$ manages to perform close to Heideltime in terms of F1 scores, which also highlights the cross-domain performance of neural methods. Translations of TempEval data have a deteriorating effect on non-pre-trained models. A possible explanation is that pre-training makes the model more stable and resilient to noisy inputs, which is likely for automatic translation data. Overall, it can be observed that there is no singular top-performing model across all metrics. Depending on user preferences, appropriate models choices can then be made.

We also include results of type classification. Note the highly uneven class distribution, which is present in all datasets and makes prediction performance for rare classes a challenging task. Accessing a larger corpora in pre-training also means more frequently encountering rare class instances, which benefits the type prediction in the final evaluation. Correspondingly, pre-trained models outperform their respective model counterparts without pre-training.

Additional training results with GottBERT [27] and GELECTRA-base were omitted for the sake of brevity, but exhibited a worse performance than the presented models.

4.2. Current Limitations

Preliminary results indicate that our fine-tuned models are clearly outperforming the baseline tagger in almost every metric. However, it should be noted that the performance without pre-training is already quite good and close to the pre-trained variants. Given the cost of pre-training, this should be considered as a potential trade-off.

Further, we want to point out the high similarity between German and English. This is particularly relevant for automatically translated resources, where it is much easier to obtain additional high-quality annotations through automated translation.

Finally, the approach still relies on existing resources for the final fine-tuning, which includes both existing monolingual models and datasets. However, we suspect multilingual models would also be suitable after sufficient task-specific pre-training, which makes monolingual models less of a requirement. As for data, the 500 tags used for fine-tuning seem already sufficient to learn a decent system on top of a base model, which is promising for other languages without existing annotations.

5. Conclusion and Future Work

In this work, we have introduced a generic way to fine-tune language-specific temporal taggers, demonstrated at the example of a German tagger. While there are limitations to the current approach, we successfully demonstrate surpassing the current state-of-the-art tagger for German, which is a promising start.

For future work, we are planning to investigate patterns of incorrect labels to determine areas of improvement, and employ bootstrapping with semi-supervised learning to further increase the tagging accuracy for precision-heavy model variants.

References

[1] R. Campos, G. Dias, A. M. Jorge, A. Jatowt, Survey of temporal information retrieval and related applications, ACM Comput. Surv. 47 (2014) 15:1–15:41. URL: https://doi.org/10.1145/2619088. doi:10.1145/2619088.

[2] P. Hausner, D. Aumiller, M. Gertz, Time-centric exploration of court documents, in: R. Campos, A. M. Jorge, A. Jatowt, S. Bhatia (Eds.), Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020, Lisbon, Portugal, April 14th, 2020 [online only], volume 2593 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 31–37. URL: http://ceur-ws.org/Vol-2593/paper4.pdf.

[3] P. Hausner, D. Aumiller, M. Gertz, Ticco: Time-centric content exploration, in: M. d'Aquin, S. Dietze, C. Hauff, E. Curry, P. Cudré-Mauroux (Eds.), CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020, ACM, 2020, pp. 3413–3416. doi:10.1145/3340531.3417432.

[4] S. Vashishtha, A. Poliak, Y. K. Lal, B. Van Durme, A. S. White, Temporal reasoning in natural language inference, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4070–4078. URL: https://aclanthology.org/2020.findings-emnlp.363. doi:10.18653/v1/2020.findings-emnlp.363.

[5] M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, J. Pustejovsky, SemEval-2007 Task 15: TempEval Temporal Relation Identification, in: Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007), Association

for Computational Linguistics, Prague, Czech Republic, 2007, pp. 75–80. URL: https://www.aclweb.org/anthology/S07-1014.

[6] M. Verhagen, R. Saurí, T. Caselli, J. Pustejovsky, SemEval-2010 task 13: TempEval-2, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 57–62. URL: https://www.aclweb.org/anthology/S10-1010.

[7] N. UzZaman, H. Llorens, L. Derczynski, J. Allen, M. Verhagen, J. Pustejovsky, 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), Association for Computational Linguistics, Atlanta, Georgia, USA, 2013, pp. 1–9. URL: https://www.aclweb.org/anthology/S13-2001.

[8] X. Zhong, A. Sun, E. Cambria, Time Expression Analysis and Recognition Using Syntactic Token Types and General Heuristic Rules, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 420–429. URL: https://www.aclweb.org/anthology/P17-1039. doi:10.18653/v1/P17-1039.

[9] J. Strötgen, A. Minard, L. Lange, M. Speranza, B. Magnini, KRAUTS: A german temporally annotated news corpus, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018, European Language Resources Association (ELRA), 2018.

[10] H. Llorens, E. Saquete, B. Navarro, Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2, in: Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010, The Association for Computer Linguistics, 2010, pp. 284–291. URL: https://aclanthology.org/S10-1063/.

[11] L. Lange, A. Iurshina, H. Adel, J. Strötgen, Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text, in: Proceedings of the 5th Workshop on Representation Learning for NLP, Association for Computational Linguistics, Online, 2020, pp. 103–109. URL: https://www.aclweb.org/anthology/2020.repl4nlp-1.14. doi:10.18653/v1/2020.repl4nlp-1.14.

[12] M. Starý, Z. Neverilová, J. Valcík, Multilingual recognition of temporal expressions, in: The 14th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2020, Brno (on-line), Czech Republic, December 8-10, 2020, Tribun EU, 2020, pp. 67–78. URL: http://nlp.fi.muni.cz/raslan/2020/paper2.pdf.

[13] J. Strötgen, M. Gertz, HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 321–324. URL: https://www.aclweb.org/anthology/S10-1071.

[14] J. Strötgen, M. Gertz, A Baseline Temporal Tagger for all Languages, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 541–547. URL: https://www.aclweb.org/anthology/D15-1063. doi:10.18653/v1/D15-1063.

[15] J. Strötgen, M. Gertz, Wikiwarsde: A german corpus of narratives annotated with temporal

expressions, in: Proceedings of the conference of the German society for computational linguistics and language technology (GSCL 2011), Citeseer, 2011, pp. 129–134.

[16] A. Bittar, P. Amsili, P. Denis, L. Danlos, French timebank: An iso-timeml annotated reference corpus, in: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers, The Association for Computer Linguistics, 2011, pp. 130–134. URL: https://aclanthology.org/P11-2023/.

[17] T. Caselli, V. B. Lenzi, R. Sprugnoli, E. Pianta, I. Prodanof, Annotating events, temporal expressions and relations in italian: the it-timeml experience for the ita-timebank, in: Proceedings of the Fifth Linguistic Annotation Workshop, LAW 2011, June 23-24, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 143–151. URL: https://aclanthology.org/W11-0418/.

[18] F. Costa, A. Branco, Timebankpt: A timeml annotated corpus of portuguese, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012, European Language Resources Association (ELRA), 2012, pp. 3727–3734. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/246.html.

[19] C. Forascu, D. Tufis, Romanian timebank: An annotated parallel corpus for temporal information, in: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012, European Language Resources Association (ELRA), 2012, pp. 3762–3766. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/770.html.

[20] M. Navas-Loro, V. Rodríguez-Doncel, Annotador: a temporal tagger for spanish, J. Intell. Fuzzy Syst. 39 (2020) 1979–1991. URL: https://doi.org/10.3233/JIFS-179865. doi:10.3233/JIFS-179865.

[21] L. Real, A. Rademaker, F. Chalub, V. de Paiva, Towards temporal reasoning in portuguese, in: Proceedings of the LREC2018 Workshop Linked Data in Linguistics, 2018.

[22] F. Costa, A. Branco, Extracting temporal information from portuguese texts, in: Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings, volume 7243 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 99–105. URL: https://doi.org/10.1007/978-3-642-28885-2_11. doi:10.1007/978-3-642-28885-2_11.

[23] Y. Yaghoobzadeh, G. Ghassem-Sani, S. A. Mirroshandel, M. Eshaghzadeh, Iso-timeml event extraction in persian text, in: COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India, Indian Institute of Technology Bombay, 2012, pp. 2931–2944. URL: https://aclanthology.org/C12-1179/.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[25] B. Chan, S. Schweter, T. Möller, German's next language model, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on

Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6788–6796. URL: https://aclanthology.org/2020.coling-main.598. doi:10.18653/v1/2020.coling-main.598.

[26] K. Clark, M. Luong, Q. V. Le, C. D. Manning, ELECTRA: pre-training text encoders as discriminators rather than generators, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=r1xMH1BtvB.

[27] R. Scheible, F. Thomczyk, P. Tippmann, V. Jaravine, M. Boeker, GottBERT: a pure German Language Model, arXiv preprint arXiv:2012.02110 (2020).