Embedding Time Differences in Context-sensitive Neural Networks for Learning Time to Event

Nazanin Dehghani[*1], **Hassan Hajipoor**[*2], **Hadi Amiri**[2,3]
[1]IRISA, University of Rennes 1 (ENSSAT), Lannion, France
[2]Department Computer Science, University of Massachusetts, Lowell, USA
[3]Department of Biomedical Informatics, Harvard University, Boston, USA
nazanin.dehghani@irisa.fr, hassan_hajipoor@student.uml.edu,
hadi_amiri@uml.edu

Abstract

We propose an effective context-sensitive neural model for the task of *time to event* (TTE) prediction, which aims to predict the amount of time to/from the occurrence of given events in streaming content. We investigate this problem in the context of a multi-task learning framework, which we enrich with *time difference* embeddings. To conduct this research, we develop a multi-genre dataset of English events about soccer competitions and academy awards ceremonies, as well as their relevant tweets obtained from Twitter. Our model is 1.4 and 3.3 hours more accurate than the current state-of-the-art model in estimating TTE on English and Dutch tweets respectively. We examine different aspects of our model to illustrate its source of improvement.[1]

1 Introduction

The task of time to event (TTE) prediction aims to determine the amount of time to/from the occurrence of a well-defined event. Accurate prediction of this information is important for temporal tasks such as timeline generation (Reimers et al., 2018), news summarization (Born et al., 2020; Huang et al., 2016), and disease onset prediction in medical domain (Zeliger, 2016; Langbehn et al., 2004).

Current approaches mainly focus on news articles and expect at least one temporal expressions in each input data to predict TTE (Chambers et al., 2014; Reimers et al., 2016, 2018; Hürriyetoğlu et al., 2018; Zhou et al., 2020). These approaches cannot be readily applied to streaming content (such as Twitter data) because such data often do not carry any temporal expressions. Figure 1 show

Figure 1: Examples of tweets that don't carry any explicit time expression but indicate a future or past event due to the implicit temporal connotation in "*looking forward to*," "*must be fun to watch*," "*well done*" etc.

two examples of such tweets.[2] In addition, event-related content in data streams are heavily skewed in time distribution as they are often posted in close proximity of their corresponding events.

The above challenges and intuitions inspire our work to develop a context-sensitive model to predict TTE in streaming content. Our approach is a multi-task learning framework that uses a small fraction of temporally-rich neighbors of each input (tweet) and their time differences (learned through *time difference* embeddings) to predict (a): if the tweet has been posted *before*, *at the same time* or *after* the event, and (b): estimate the absolute value of TTE (in hours) with respect to the tweet. We learn time difference embeddings through an effective character-level sequence to sequence model that takes as input two timestamps and predicts the temporal difference between them (in hours).

The contributions of this paper are as follows: (a) an effective multi-task and context-sensitive framework that uses temporally-rich context and time difference embeddings to accurately predict TTE in streaming content, (b) publicizing a time to event dataset that includes different genres of

[*]First and second authors equally contributed to this work.
[1]Our code and data are available at https://github.com/hajipoor/time2event
[2]In fact, 89% of event-related tweets in our dataset do not carry any time expression.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 630–636
August 1–6, 2021.

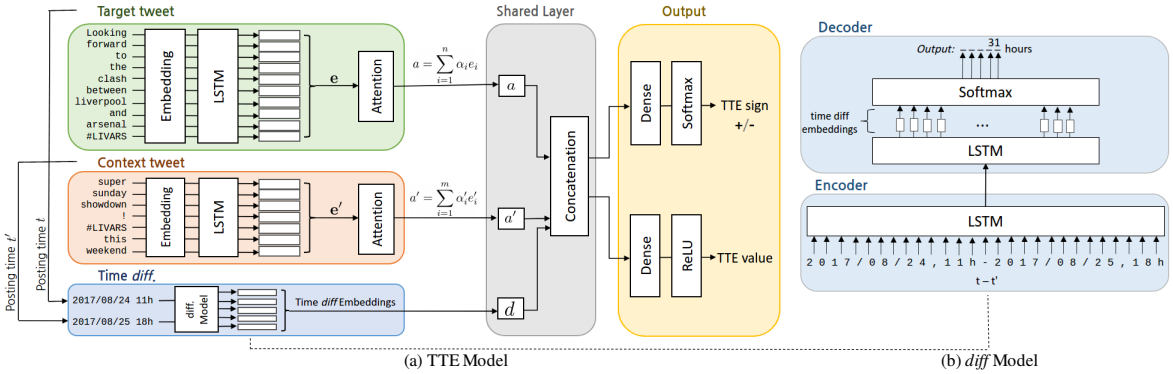


Figure 2: (a) TTE model gives the target tweet and a temporally-rich context tweet, and their time difference embedding learned through *diff* model as input and learns TTE as a combination of regression (TTE value) and classification (TTE sign) tasks. It establishes a common scale between corresponding loss values for effective training and (b) *diff* model gives two times and learns time difference embeddings via sequence to sequence model, which are used in our TTE model.

events (soccer competitions and academy awards ceremonies), their time of occurrence, their relevant tweets as well as TTE information for each tweet.

Our framework is 1.4 and 3.3 hours more accurate than the current state-of-the-art model in estimating TTE on large-scale English and Dutch tweets respectively. In addition, our time difference model achieves an accuracy of 98.3% in terms of creating embeddings that encode temporal differences between given time pairs.

2 Context-sensitive Model

Existing models often assume input data carry explicit temporal information about target events. Although informative, these information may not be available in most textual content, especially in microblogs. We propose to utilize context information (in the form of neighboring tweets) and the relative temporal differences against neighbours to estimate time to event (TTE) for given input texts.

In particular, given a tweet about an event, we propose a multi-task learning framework to predict the absolute value of TTE (in hours) for the tweet, as well as a binary sign which determines if the tweet has been posted before '(+),' or at the same time or after '(-)' the event. Figure 2(a) shows our model for predicting TTE for the target tweet t_i, given its context tweet[3], e.g. a previously posted tweet about the same event, $t_j, j < i$, and their *time difference* embeddings, which encode the time differences between tweet creation times. Our intuition for developing such embeddings is that if context tweets carry useful temporal information about events, then knowing the time differences among tweets could help the model to make more accurate prediction of TTE for the target tweet.

Our model takes as input the concatenation of attention-weighted average embeddings of the target and context tweets (a and a' in Figure 2) and their time difference embedding (d) (see section 2.1). The resulting concatenation are then used to predict *TTE sign* and *TTE value* for the target input. TTE value is a regression task while TTE sign is a binary classification task. To prevent the loss with larger gradient magnitudes dominate the training, we establish a common scale for the different loss magnitudes across the two tasks using the approach proposed in (Kendall et al., 2018), which simultaneously learns classification and regression losses of varying quantities and combines them using homoscedastic uncertainty.

2.1 Time Difference Embeddings

Motivated by recent research on neural numeracy learning (Chen et al., 2019; Wallace et al., 2019), we learn time difference embeddings–`diff embeddings`–as follows: we develop an LSTM-based character-level sequence to sequence model (based on the model presented in (Sutskever et al., 2014)) that takes as input a time pair (t and t') and predicts the difference between them (in hours). The final layer of the model is of size five, where five is determined by the maximum number of digits in the differences of any two timestamps within a 2 years period (i.e., 17520 hours). The final hidden representations of the resulting digits are then concatenated to obtain the diff embeddings, see

[3]Neighboring or context tweets are randomly sampled from the set of previously posted tweets relevant to the target event. Our model can be extended to greater context sizes.

Figure 2(b).

3 Experiments

3.1 Datasets

We develop a dataset from tweets about soccer competitions of the England Premier League (EPL) following the same approach in (Hurriyetoglu et al., 2014; Hürriyetoğlu et al., 2018). We carefully create a list of 42 distinctive hashtags for competitions between seven most famous teams [4]. These matches have the advantage that users tweet about them with distinctive hashtags by convention. We collect tweets that are sent within 14 days of match days between seven popular teams, and obtain the actual time of each event from the EPL schedule.[5]

For the regression task, the tweet label would be the absolute value of the actual time (in hours) to the start of the corresponding event. For the classification task, tweets are labeled as 'before' or 'after' depending on their time of creation against corresponding matches. Our dataset is randomly divided into 80%, 10% and 10% according to events, which are used for training, testing and validation respectively. To study the effect of temporally-rich context in a controlled situation, we divide our dataset of tweets (**All set**) into two disjoint subsets: tweets that carry at least one temporal expression (**T set**), and tweets that have no temporal expression (**N set**), where we use HeidelTime's colloquial temporal tagger (Strötgen and Gertz, 2012, 2013) to extract temporal expressions. We then introduce six new subsets of our data in X-Y format, where $X \in \{$T set, N set, All set$\}$ refers to the type of target tweets and $Y \in \{$T set, N set$\}$ refers to the type of contexts.

To investigate the generalizability of our model on other events, we evaluate our trained model on tweets about the 2018 Academy Awards ceremony. We collected 3K tweets using `#oscars`, `#oscar` and `#academyawards` hashtags in the window of 7 days before and after the date of Oscars 2018. We also use the Dutch dataset to compare our model against the baseline model proposed in (Hurriyetoglu et al., 2014) that developed a hybrid of machine learning and rule-based approach for estimating time to events.

[4]Liverpool, Manchester United, Chelsea, Arsenal, Manchester City, Newcastle United and Tottenham Hotspur

[5]https://www.premierleague.com/

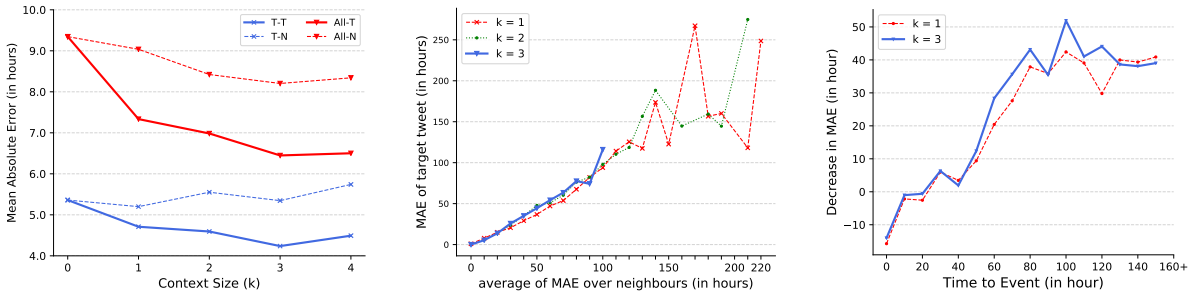
3.2 Settings and Baselines

The hyperparameters of all models are optimized on validation data using random search (Bergstra and Bengio, 2012). We consider **TenseModel** (see below), **Glove**, **BERT**, Event Time Extraction **(ETE)** (Reimers et al., 2018), and **HybridModel** (Hürriyetoğlu et al., 2018) as baselines. TenseModel uses the tense of the outermost verb of a tweet to detect whether it is posted before (+) or after the target event (-). Embedding models are used to represent input tweets and extended to address the time to event task in their last layer. The GLOVE baseline is the model with GLOVE pre-trained embeddings but without context. This baseline has only the attention-weighted average embedding of the target tweet. For BERT baseline, we fine-tuned base version of BERT by adding a linear layer on top for time to event value prediction. ETE uses sentence representation as well as event and position embeddings with a CNN to tackle the target task. They reported a high performance of 84.2% for event status classification on a balanced dataset of news articles. HybridModel is a hybrid of rule-based and data-driven methods focusing on Dutch tweets that carry temporal expressions.

4 Results

4.1 Time to Event Prediction

We compare our context-sensitive model with context size of $k \in \{0, 1, 2, 3\}$ against baseline systems. Mean and Median are heuristic baselines and indicate the mean and median of MAEs of TTE values (i.e., 27.87 and 13.91 respectively). As reported in Table 1, the TenseModel has considerably low performance in distinguishing temporal status of tweets against event times. We attribute this result to the informal language in user-generated content and multi-verb tweets which can challenge the tense model. BERT embeddings slightly improves the performance of other embedding models. However, BERT and ETE's performance are considerably lower than the performance of our model achieved by adding context information ($k \geq 1$). This result indicates that adding neighbouring tweets leads to more accurate prediction of TTE than incorporating better word embeddings. Our model achieves an MAE of 6.43 hours on *All-T* set and 4.24 on *T-T* set (see Table 1). We also compare our model against the Hybrid Model on the



(a) Effect of context size on MAE (b) Accurately-predicted neighbours (c) Farther vs. closer to event instances

Figure 3: (a): Greater context size leads to better estimation of TTE. (b): More accurately-predicted neighbours lead to more accurate estimation of TTE. (c): Context information better help farther to event target instances.

	TTE sign			TTE value
Model	**P**	**R**	**F1**	**MAE** (hours)
Trained and evaluated on the *EPL* dataset				
Mean	-	-	-	27.87
Median	-	-	-	13.91
TenseModel	0.23	0.37	0.28	-
GLOVE (Pennington et al., 2014)	0.73	0.66	0.69	8.43
BERT (Devlin et al., 2019)	0.68	0.79	0.73	7.71
ETE (Reimers et al., 2018)	0.88	0.59	0.70	7.86
Our model ($k = 0$)	0.61	0.52	0.56	9.34
Our model ($k = 1$)	0.73	0.77	0.74	7.31
Our model ($k = 2$)	0.81	**0.87**	0.83	6.98
Our model ($k = 3$)	**0.92**	0.83	**0.87**	**6.43**
Trained on *EPL* and evaluated on the *Oscars* dataset				
BERT	0.46	0.48	0.47	14.2
Our model ($k = 0$)	0.38	0.49	0.43	14.76
Our model ($k = 1$)	0.51	0.57	0.54	13.43
Our model ($k = 2$)	0.55	0.60	0.57	13.37
Our model ($k = 3$)	0.58	0.64	0.61	13.18

Table 1: Model performance in terms of macro precision, recall and F1 for sign classification (TTE sign), and Mean Absolute Error (MAE) for TTE prediction (TTE value) on EPL and Oscars datasets.

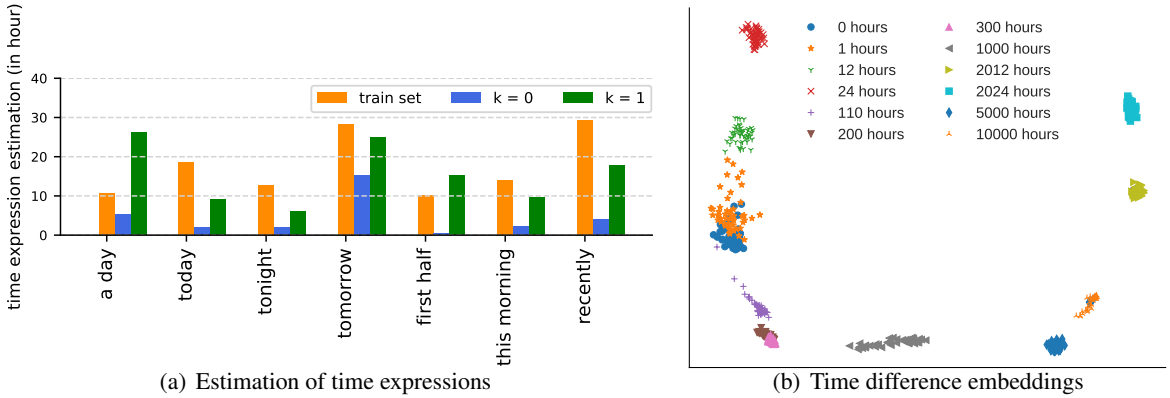
Dutch dataset (see Section 3.1).[6] Using the Dutch embeddings of (Tulkens et al., 2016), our model achieves an MAE of 4.7 hours based on leave-one-out cross validation, while the corresponding value for the Hybrid Model is 8 hours. Evaluation results on Oscars dataset reveals that the model learns how to utilize information of neighbouring tweets and time differences. The lower performance on the Oscar dataset is due to differences in training (EPL) and test (Oscar) data distributions.

Can context information help? To investigate the effect of adding context, we start with a standalone base model that predicts the time to event by just relying on its own content, i.e, $k = 0$. Figure 3(a) illustrates that the performance is higher for tweets that contain at least one time expression (*T set*) compared to *All set*. Accordingly, as we gradually add more context tweets, the performance consistently increases with greater improvement with the *T set* as context. The best preforming model is achieved by adding context of size 3 from *T set*, leading to the lowest time to event estimation error of 4.24 hours.

We also note that context tweets that do not contain any temporal expression (the *N set*) slightly increase the performance; see the dashed lines in Figure 3(a). We conjecture that these tweets add lexical clues that carry implicit temporal information about events. In addition, Figure 3(b) shows a strong correlation between the average error in model prediction performance on context and target tweets. This result shows that neighboring tweets that are more accurately learned by the network are better candidates to use as context for other tweets.

Does context information lead to more accurate estimation of time expressions? To answer this question, we compute the average time to event for each time expression from both training tweets and predictions for test tweets as $H(\text{TIMEX}_i) = \frac{1}{N}\sum_{t_j \in \mathcal{S}, \text{TIMEX}_i \in t_j} \text{TTE}_{t_j}$ where $\mathcal{S} \in \{train, test\}$, TTE_{t_j} indicates time to event for tweet t_j, and N is normalization factor; for training data we use gold values and for test data we use predicted values. Figure 4(a) shows the baseline and estimated values for a range of time expressions. The results show that the value of time expressions are better estimated by adding context. Give that the most frequent time expressions often refer to points in time close to the event (such as *now*) (Hurriyetoglu et al., 2014), our model improves rare time expressions more than the frequent ones, leading to improved prediction of farthest tweets from events.

[6] Note only 71% of 138k tweets are returned by the Twitter API, the rest were deleted or made private by their users.



(a) Estimation of time expressions

(b) Time difference embeddings

Figure 4: (a): Estimation of some selected time expressions. (b): Time differences in the embedding space. Each sample point shows the embedding of time difference between two randomly selected times t_1 and t_2. This figure shows that if $t_1 - t_2 \approx t_1^{'} - t_2^{'}$, their *diff* embeddings are closer in the time difference space.

4.2 Time Difference Embeddings

We generate a synthetic dataset of $2m$ time pairs to evaluate the time difference approach in terms of accurate prediction of time differences between given time pairs, where possible predictions range between 0 to 17520 hours, which corresponds to a maximum difference of two years. Evaluation on $200k$ number of test time pairs shows that the model achieves 98.3% accuracy. In addition, Figure 4(b) shows t-SNE representation of time differences in the embedding space for different randomly selected time pairs. Data points with the same color shows the diff embeddings of the same time differences. The result shows for two random times (t_1, t_2) and $(t_1^{'}, t_2^{'})$, if $t_1 - t_2 \approx t_1^{'} - t_2^{'}$, their *diff* embeddings are very close in the time difference embedding space, indicating the high quality of the resulting space. In addition, Table 2 shows time difference embeddings are useful for TTE estimation since removing them increases the MAE of our full model by a significant amount of 0.8 hours.

4.3 Early prediction

Given that *early* prediction of TTE is more valuable and challenging (due to scarcity of data at earlier times and often imprecise temporal information in earlier tweets), we investigate the performance of our model on target tweets that were posted much earlier than the occurrences of their corresponding events. The results in Figure 3(c) shows that context tweets help farther-to-event instances better than closer ones. This result provides insights for future research on the task of early TTE prediction.

Configuration	MAE (absolute increase)
Full System	6.43
Random diff embeddings	6.64 (+0.21)
No diff embeddings	7.23 (+0.80)
No TTE sign	6.71 (+0.28)

Table 2: Ablation analysis showing changes in Mean Absolute Error (MAE) obtained from removing individual components of the model.

5 Conclusion and Future Work

We developed a context-sensitive neural model that used rich-neighbouring tweets as well as time difference embeddings between target tweets and their neighbors for effective prediction of time to event. We evaluated our and current models on events and tweets of different genres (soccer competitions and academy award ceremonies) and languages (English and Dutch). Future works include expansion to temporal tasks that particularly focus on early prediction of time to events. In addition, it's worth investigating if user or social network information could be helpful for better time to event prediction.

Acknowledgments

We sincerely thank anonymous reviewers for their insightful comments. In addition, this research was completed during the spread of the COVID-19 virus, while the world was in quarantine fearing the epidemic. We would like to dedicate this work to all researchers who contributed to the discovery of the COVID-19 vaccine.

Broader Impact Statement

Our research affects applications that deal with time, and time difference can be an effective feature for them. For example, our work enables automatic creation of calendar of events, which helps keeping individuals informed about potential relevant events. It also help researchers to benchmark their models using our dataset.

In addition, the process of collecting our dataset followed the Twitter policy[7]. We crawled data using the Twitter API and we did not make any attempt to identify any information that have not been volunteered by our user base (e.g., gender, race, wealth, etc.,). We also will just publish the Tweet IDs.

References

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Leo Born, Maximilian Bacher, and Katja Markert. 2020. Dataset reproducibility and ir methods in timeline summarization. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1763–1771.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2019. Numeracy-600k: learning numeracy for detecting exaggerated information in market comments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6307–6313.

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)*, pages 4171–4186.

Ruihong Huang, Ignacio Cases, Dan Jurafsky, Cleo Condoravdi, and Ellen Riloff. 2016. Distinguishing past, on-going, and future events: The eventstatus corpus. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 44–54.

[7]https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

Ali Hurriyetoglu, Nelleke Oostdijk, and Antal van den Bosch. 2014. Estimating time to event from tweets using temporal expressions. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 8–16.

Ali Hürriyetoğlu, Nelleke Oostdijk, and Antal van den Bosch. 2018. Estimating time to event of future events based on linguistic cues on twitter. In *Intelligent Natural Language Processing: Trends and Applications*, pages 67–97. Springer.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Douglas R Langbehn, Ryan R Brinkman, Daniel Falush, Jane S Paulsen, MR Hayden, and an International Huntington's Disease Collaborative Group. 2004. A new model for prediction of the age of onset and penetrance for huntington's disease based on cag length. *Clinical genetics*, 65(4):267–277.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the timebank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2195–2204.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2018. Event time extraction with a decision tree of neural classifiers. *Transactions of the Association of Computational Linguistics*, 6:77–89.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753. ELRA.

Jannik Strötgen and Michael Gertz. 2013. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 3104–3112.

Stephan Tulkens, Chris Emmery, and Walter Daelemans. 2016. Evaluating unsupervised dutch word embeddings as a linguistic resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do nlp models know numbers? probing numeracy in 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)*, pages 5310–5318.

Harold I Zeliger. 2016. Predicting disease onset in clinically healthy people. *Interdisciplinary toxicology*, 9(2):39–54.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589.