Dynamic Contextualized Word Embeddings

Valentin Hofmann*, Janet B. Pierrehumbert†, Hinrich Schütze‡

*Faculty of Linguistics, University of Oxford
†Department of Engineering Science, University of Oxford
‡Center for Information and Language Processing, LMU Munich
valentin.hofmann@ling-phil.ox.ac.uk

Abstract
Static word embeddings that represent words by a single vector cannot capture the variability of word meaning in different linguistic and extralinguistic contexts. Building on prior work on contextualized and dynamic word embeddings, we introduce dynamic contextualized word embeddings that represent words as a function of both linguistic and extralinguistic context. Based on a pretrained language model (PLM), dynamic contextualized word embeddings model time and social space jointly, which makes them attractive for various tasks in the computational social sciences. We highlight potential applications by means of qualitative and quantitative analyses.

1 Introduction
Over the last decade, word embeddings have revolutionized the field of NLP. Traditional methods such as LSA (Deerwester et al., 1990), word2vec (Mikolov et al., 2013a,b), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017) compute static word embeddings, i.e., they represent words as a single vector. From a theoretical standpoint, this way of modeling lexical semantics is problematic since it ignores the variability of word meaning in different linguistic contexts (e.g., polysemy) as well as different extralinguistic contexts (e.g., temporal and social variation).

The first shortcoming was addressed by the introduction of contextualized word embeddings that represent words as vectors varying across linguistic contexts. This allows them to capture more complex characteristics of word meaning, including polysemy. Contextualized word embeddings have become the de-facto norm in NLP, constituting the semantic backbone of pretrained language models (PLMs) such as ELMo (Peters et al., 2018a), BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), XLNet (Yang et al., 2019), ELECTRA (Clark et al., 2020), and T5 (Raffel et al., 2020).

A concurrent line of work focused on the second shortcoming of static word embeddings, resulting in various types of dynamic word embeddings. Dynamic word embeddings represent words as vectors varying across extralinguistic contexts, in particular time (e.g., Bamler and Mandt, 2017; Rosenfeld and Erk, 2018; Rudolph and Blei, 2018) and social space (e.g., Zeng et al., 2017, 2018).

In this paper, we introduce dynamic contextualized word embeddings that combine the strengths of contextualized word embeddings with the flexibility of dynamic word embeddings. Dynamic contextualized word embeddings mark a departure from existing contextualized word embeddings (which are not dynamic) as well as existing dynamic word embeddings (which are not contextualized). Furthermore, as opposed to all existing dynamic word embedding types, they represent time and social space jointly.
While our general framework for training dynamic contextualized word embeddings is model-agnostic (Figure 1), we present a version using a PLM (BERT) as the contextualizer, which allows for an easy integration within existing architectures. We expect dynamic contextualized word embeddings to be particularly beneficial for the computational social sciences, where they can serve as an analytical tool or be employed for downstream tasks. We illustrate these application areas by examining semantic dynamics on Reddit during the Covid-19 pandemic and performing sentiment classification on Ciao product reviews.

Contributions. We introduce dynamic contextualized word embeddings that represent words as a function of both linguistic and extralinguistic context. Based on a PLM, dynamic contextualized word embeddings model time and social space jointly, which makes them attractive for a range of tasks in the computational social sciences. We showcase potential applications by means of qualitative and quantitative analyses.1

2 Related Work

2.1 Contextualized Word Embeddings

The insight that a distinction must be drawn between the non-contextualized meaning of a word that is stored in the mental lexicon and the sense that is realized in a specific linguistic context lies at the heart of lexical-semantic scholarship (see Geeraerts, 2010 for an overview), going back to at least Paul (1880). In NLP, this two-level structure is closely resembled by contextualized word embeddings that map type-level representations to token-level representations as a function of the linguistic context (McCann et al., 2017; Peters et al., 2017). As part of PLMs (Peters et al., 2018a; Devlin et al., 2019; Radford et al., 2019; Yang et al., 2019; Clark et al., 2020; Raffel et al., 2020), contextualized word embeddings have led to significant performance gains over static word embeddings that only have type-level representations (Deerwester et al., 1990; Mikolov et al., 2013a,b; Pennington et al., 2014; Bojanowski et al., 2017).

Since their introduction, several studies have analyzed the linguistic properties of contextualized word embeddings (Peters et al., 2018b; Goldberg, 2019; Hewitt and Manning, 2019; Jawahar et al., 2019; Lin et al., 2019; Liu et al., 2019; Tenney et al., 2019; Edmiston, 2020; Ettinger, 2020; Hoffmann et al., 2020; Rogers et al., 2020). Regarding lexical semantics, this line of research has shown that contextualized word embeddings are more context-specific in the upper layers of a contextualizer (Ethayarajh, 2019; Mickus et al., 2020; Vulić et al., 2020) and represent different word senses as spatially separated clusters (Peters et al., 2018a; Coenen et al., 2019; Wiedemann et al., 2019).

2.2 Dynamic Word Embeddings

The meaning of a word can also vary across extralinguistic contexts such as time (Bybee, 2015; Koch, 2016) and social space (Robinson, 2010, 2012; Geeraerts, 2018). To capture these phenomena, various types of dynamic word embeddings have been proposed, diachronic word embeddings for temporal semantic change (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016b; Bamler and Mandt, 2017; Rosenfeld and Erk, 2018; Rudolph and Blei, 2018; Dubossarsky et al., 2019; Gong et al., 2020) and personalized word embeddings for social semantic variation (del Tredici and Fernández, 2017; Zeng et al., 2017, 2018; Oba et al., 2019; Welch et al., 2020; Yao et al., 2020). Even though not directly inducing dynamic word embeddings, many other studies have demonstrated that performance on a diverse set of NLP tasks can be increased by using temporal information (Jaidka et al., 2018; Lukes and Søgaard, 2018) and social information (Amir et al., 2016; Hamilton et al., 2016a; Yang et al., 2016; Yang and Eisenstein, 2017; Hazarika et al., 2018; Kolchinski and Potts, 2018; Mishra et al., 2018; del Tredici et al., 2019; Li and Goldwasser, 2019; Mishra et al., 2019).

Sociolinguistics has shown that temporal and social variation in language are tightly interwoven: innovations such as a new word sense in the case of lexical semantics spread through the language community along social ties (Milroy, 1980, 1992; Labov, 2001; Pierrehumbert, 2012). However, none of the proposed dynamic word embedding types can capture more than one dimension of variation. In addition, to capture the full range of lexical-semantic variability, dynamic word embeddings should also be contextualized.2

1We make all our code publicly available upon publication.

2Contextualized word embeddings have been used to investigate semantic change recently (Hu et al., 2019; Giulianelli et al., 2020). However, the word embeddings employed in these studies are not dynamic, i.e., they represent a word in a specific linguistic context by the same contextualized word embedding independent of extralinguistic context.
3 Model

3.1 Model Overview

Given a sequence of words $X = [x^{(1)}, \ldots, x^{(K)}]$ and corresponding non-contextualized embeddings $E = [e^{(1)}, \ldots, e^{(K)}]$, contextualizing language models compute the contextualized embedding of a particular word $x^{(k)}$, $h^{(k)}$, as a function $c$ of its non-contextualized embedding, $e^{(k)}$, and the non-contextualized embeddings of all words in $X$:\footnote{Some contextualizing language models such as GPT-2 (Radford et al., 2019) only operate on a subsequence of $X$.}

$$h^{(k)} = c \left( e^{(k)}, E \right). \quad (1)$$

Crucially, while $h^{(k)}$ is a token-level representation, $e^{(k)}$ is a type-level representation and is modeled as a simple embedding look-up. Here, in order to take the variability of word meaning in different extralinguistic contexts into account, we depart from this practice and model $e^{(k)}$ as a function $d$ that depends not only on the identity of $x^{(k)}$ but also on the social context $s_i$ and the temporal context $t_j$ in which the sequence $X$ occurred,

$$e^{(k)}_{ij} = d \left( x^{(k)}, s_i, t_j \right). \quad (2)$$

Dynamic contextualized word embeddings are hence computed in two stages: words are first mapped to dynamic type-level representations by $d$ and then to contextualized token-level representation by $c$ (Figure 1). This two-stage structure takes into account the fact that extralinguistic information is processed before linguistic information by humans and impacts how the linguistic information is interpreted (Hay et al., 2006).

Since many words in the core vocabulary are semantically stable across social and temporal contexts, we place a Gaussian prior on $e^{(k)}_{ij}$,

$$e^{(k)}_{ij} \sim \mathcal{N} \left( \tilde{e}^{(k)}, \lambda^{-1} I \right), \quad (3)$$

where $\tilde{e}^{(k)}$ denotes a non-dynamic representation of $x^{(k)}$. Similar methods of anchoring dynamic word embeddings have been used before (Hamilton et al., 2016b). Combining Equations 2 and 3, we write the function $d$ as

$$d \left( x^{(k)}, s_i, t_j \right) = \tilde{e}^{(k)} + o^{(k)}_{ij}, \quad (4)$$

where $o^{(k)}_{ij}$ denotes the vector offset from $x^{(k)}$’s non-dynamic embedding $\tilde{e}^{(k)}$, which is stable across social and temporal contexts, to its dynamic embedding $e^{(k)}_{ij}$, which is specific to $s_i$ and $t_j$. The distribution of $o^{(k)}_{ij}$ then follows a Gaussian with

$$o^{(k)}_{ij} \sim \mathcal{N} \left( 0, \lambda^{-1} I \right). \quad (5)$$

We enforce Equation 5 by including a regularization term in the objective function (Section 3.4).

3.2 Contextualizing Component

We leverage a PLM for the function $c$, specifically BERT (Devlin et al., 2019). Denoting with $E_{ij}$ the sequence of dynamic embeddings corresponding to $X$ in $s_i$ and $t_j$, the dynamic version of Equation 1 becomes

$$h^{(k)}_{ij} = \text{BERT} \left( e^{(k)}_{ij}, E_{ij} \right). \quad (6)$$

We also use BERT, specifically its pretrained input embeddings, to initialize the non-dynamic embeddings $\tilde{e}^{(k)}$, which are summed with the vector offsets $o^{(k)}_{ij}$ (Equation 4) and fed into BERT.

Using a PLM for $c$ has the advantage of making it easy to employ dynamic contextualized word embeddings for downstream tasks by adding a task-specific layer on top of the PLM.

3.3 Dynamic Component

We model the vector offset $o^{(k)}_{ij}$ as a function of the word $x^{(k)}$, which we represent by its non-dynamic embedding $\tilde{e}^{(k)}$, as well as the social and temporal context $s_i$ and $t_j$, which we represent by a joint embedding $r_{ij}$. We again use BERT’s pretrained input embeddings for $\tilde{e}^{(k)}$.\footnote{We also tried to learn separate embeddings in the dynamic component, but this led to worse performance.} We combine these representations in a feed-forward network,

$$o^{(k)}_{ij} = \text{FFN}_o \left( \tilde{e}^{(k)} || r_{ij} \right), \quad (7)$$

where $\parallel$ denotes concatenation. The joint representation of time and social context is itself the output of a feed-forward network whose inputs are independent representations of social context $s_i$ and temporal context $t_j$,

$$r_{ij} = \text{FFN}_r \left( s_i \parallel t_j \right). \quad (8)$$

To compute the social embedding $s_i$, we follow common practice in the computational social sciences and represent the social community as a graph $\mathcal{G} = (S, E)$, where $S$ is the set of social
contexts $s_i$ (e.g., authors) in the data, and $E$ is the set of edges between them. We use a graph attention network (GAT) as proposed by Velčković et al. (2018) to encode $G$.

$$s_i = \text{GAT}(\tilde{s}_i, G).$$  

(9)

Following del Tredici et al. (2019), we initialize the input embeddings $\tilde{s}_i$ with node2vec (Grover and Leskovec, 2016) embeddings.

To compute the time embedding $t_j$, we use a feed-forward network,

$$t_j = \text{FFN}_t(\tilde{t}_j).$$  

(10)

For the input embedding $\tilde{t}_j$, we follow previous work (Rosenfeld and Erk, 2018) and rescale the time period covered by a particular dataset by

$$\tilde{t}_j = \frac{t_j - \min(T)}{\max(T) - \min(T)}.$$  

(11)

where $T$ denotes the set of all time points in the data, i.e., the dates of the first and last example correspond to 0 and 1, respectively. $\tilde{t}_j$ is mapped to $\tilde{t}_j$ by

$$\tilde{t}_j = \tanh\left( w_t \tilde{t}_j + b_t \right),$$  

(12)

where $w_t$ and $b_t$ represent a trainable weight and bias vector, respectively.

### 3.4 Model Training

The combination with BERT makes dynamic contextualized word embeddings easily applicable to different tasks by adding a task-specific layer on top of the contextualizing component. For training the model, the overall loss is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{prior}},$$  

(13)

where $\mathcal{L}_{\text{task}}$ is the task-specific loss, and $\mathcal{L}_{\text{prior}}$ is the regularization term that imposes a Gaussian prior on the offset vectors,

$$\mathcal{L}_{\text{prior}} = \frac{\lambda}{2K} \sum_{k=1}^{K} \| o^{(k)}_j \|_2^2.$$  

(14)

Notice that we use a PLM with a fixed set of input embeddings for $c$, and hence the prior is in effect placed only on a part of the vocabulary directly. Given that the input vocabularies of PLMs are constructed by frequency (Bostrom and Durrett, 2020), this is in line with the observation that the vocabulary core constitutes the best anchor for dynamic word embeddings (Hamilton et al., 2016b).

### 4 Experiments

To illustrate potential applications of dynamic contextualized word embeddings, we conduct experiments on two example tasks, masked language modeling and sentiment analysis.

#### 4.1 Data

For masked language modeling, we use data from Reddit, a social media platform hosting discussions about a variety of topics. Reddit is divided into smaller communities, so-called subreddits, which have been shown to be highly conducive to linguistic dynamics (del Tredici and Fernández, 2018). We retrieve all posts between September 2019 and April 2020, which allows us to examine the effects of the rising Covid-19 pandemic on lexical usage patterns. We remove subreddits with less than 100 comments in the examined time period as well as duplicates and posts with less than 10 words, and sample 1,000,000 posts, subsampling frequent subreddits according to the subsampling method proposed by Mikolov et al. (2013b).

For each subreddit, we then compute the set of users that posted at least 10 comments in the examined time period. Based on this, we create a weighted graph $G = (S, E)$, where $S$ is the set of subreddits, and $E$ is the set of pairwise Jaccard similarities between subreddits based on user overlap.

For sentiment analysis, we use the Ciao dataset with 269,809 product reviews (Tang et al., 2012). Users can mark trust relations, which we use to create a directed graph $G = (S, E)$. The dataset covers the time period from May 2000 to September 2011. We remove duplicates, reviews with less than 10 words, and reviews rated as not helpful.

Following Yang and Eisenstein (2017), we convert the five-star rating range into two classes by discarding three-star reviews and treating four/five stars as positive and one/two stars as negative.

#### 4.2 Baselines

As baseline, we compare dynamic contextualized word embeddings against a BERT model using simple contextualized word embeddings. BERT is individually finetuned for the two tasks. We also perform an ablation study in which we train dynamic contextualized word embeddings using only parts of the dynamic component.

---

5We also tried a model with a feed-forward network instead of graph attention, but it consistently performed worse.

6We draw upon the Baumgartner Reddit Corpus, a collection of all public Reddit posts available at https://files.pushshift.io/reddit/comments/.
4.3 Training Details

We randomly split both datasets into 70% training, 10% development, and 20% test. We use the development sets for validation and report performance on the test sets, respectively.

We use BERT BASE (cased) in all experiments. On top of BERT, we add a masking modeling head in the case of masked language modeling and a two-layer feed-forward network in the case of sentiment analysis in which we feed the output for BERT’s classification token. As \(L_{\text{task}}\), we use masked language modeling perplexity (Devlin et al., 2019) for masked language modeling and binary cross-entropy for sentiment analysis. See Appendix A.1 for details about hyperparameters.

5 Results

5.1 Overall Performance

Dynamic contextualized word embeddings outperform the BERT baseline using simple contextualized word embeddings on both tasks, achieving a higher F1 score on sentiment analysis and a lower perplexity score on masked language modeling (Table 1). This indicates that dynamic contextualized word embeddings successfully combine extralinguistic with linguistic information, and that infusing social and temporal information on the lexical level can be useful for NLP tasks.

We perform experiments in which we ablate different parts of the dynamic component. Here, dynamic contextualized word embeddings using only temporal information as extralinguistic context perform slightly better than ones using only social information on masked language modeling, but there is no significant performance difference for sentiment analysis. We also perform an experiment in which we use both social and temporal representations but do not fuse them into a joint representation, i.e., we drop Equation 8 and concatenate \(s_i\), \(t_j\), and \(e^{(k)}\). This setup performs worse than using a joint representation on both tasks, and it is even outperformed by the model using only temporal information on masked language modeling.

5.2 Social and Temporal Dynamics

Besides resulting in performance gains, dynamic contextualized word embeddings have the advantage that they allow for an analysis of social and temporal dynamics in the data. In particular, we are interested whether certain words exhibit more extralinguistically-driven semantic variability than others. To examine this question, we define as \(\sim_{ij}^{(k)}\) the cosine similarity between the non-dynamic embedding of \(x^{(k)}\), \(\tilde{e}^{(k)}\), and the dynamic embeddings of \(x^{(k)}\) given a social context \(s_i\) and temporal context \(t_j\), \(e^{(k)}_{ij}\), i.e.,

\[
\sim_{ij}^{(k)} = \cos \phi_{ij}^{(k)},
\]

where \(\phi_{ij}^{(k)}\) is the angle between \(\tilde{e}^{(k)}\) and \(e^{(k)}_{ij}\).\(^7\) To find words with a high degree of semantic variability, we compute the standard deviation of \(\sim_{ij}^{(k)}\) for all occurrences of \(x^{(k)}\) in the data,

\[
\sigma_{\sim}^{(k)} = \sigma \left( \{ \sim_{ij}^{(k)} \mid (x^{(k)}, s_i, t_j) \in \mathcal{D} \} \right),
\]

where \(\mathcal{D}\) is the development set. Looking at the top-ranked words according to \(\sigma_{\sim}^{(k)}\), we make the following two observations.

Firstly, the top-ranked words tend to have a more general as well as a narrower reading in the data. For Reddit, e.g., many of the top-ranked words denote concepts that have acquired a specialized sense during the Covid-19 pandemic such as “sanitation”, “isolating”, or “testing”. Occurrences with the narrower sense tend to have smaller values of \(\sigma_{\sim}^{(k)}\) than occurrences with the more general sense. We provide examples with linguistic and extralinguistic contexts in Table 2.

Secondly, the distributions of \(\sim_{ij}^{(k)}\) tend to have one or more pronounced peaks that correspond to different senses of \(x^{(k)}\). Given that \(e^{(k)}_{ij}\) is computed based on the identity of \(x^{(k)}\) and the extralinguistic context alone, the dynamic component

\(^7\)In cases where \(x^{(k)}\) is split into several WordPiece tokens by BERT, we follow previous work (Pinter et al., 2020) and take the average of the subword embeddings.
To exemplify further the social and temporal dynamics captured by dynamic contextualized word embeddings, we conduct an analysis of the word “sanitation”, which is the highest-ranked word according to $\sigma_{\text{sim}}^{(k)}$. We are particularly interested to see how the semantic drift of “sanitation” unfolded in the social network. We first choose a sequence of 10 equally distributed time points $T$, generate $e^{(k)}_{ij}$ for all subreddits at these points, and calculate the corresponding values of $\text{sim}^{(k)}_{ij}$. We then calculate for each subreddit the Pearson correlation coefficient $\rho$ between $T$ and $\text{sim}^{(k)}_{ij}$. $\rho$ is close to 0 for no meaning change, positive for meaning change towards a more general meaning (i.e., a word was used with a specialized sense that is now vanishing), and negative for meaning change away from a more general meaning (i.e., a word is becoming associated with a specialized sense). Thus, the lower $\rho$, the further the meaning of “sanitation” has moved away from its general use in a particular subreddit. The analysis shows that even though the meaning of “sanitation” has changed over the entire network, there are cliques that differ in their degree of drift (Figure 2). Interestingly, 7 of the 10 subreddits with the lowest value of $\rho$, i.e., the most pronounced semantic drift, are specifically dedicated to Covid-19.

### 6 Conclusion

We have introduced dynamic contextualized word embeddings that represent words as a function of both linguistic and extralinguistic context. Based on a PLM, specifically BERT, dynamic contextualized word embeddings model time and social space jointly, which makes them advantageous for the computational social sciences. We have applied dynamic contextualized word embeddings in sentiment analysis and masked language modeling, showing that they beat a strong baseline based on BERT. In addition, dynamic contextualized word embeddings allow to track social and temporal variability in word meaning, which we have showcased by an exploratory probe into semantic dynamics during the Covid-19 pandemic.
References

Silvio Amir, Byron C. Wallace, Hao Lyu, Paula Carvalho, and Mário J. Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. In Conference on Computational Natural Language Learning (CoNLL) 20.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In International Conference on Machine Learning (ICML) 34.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In arXiv:2004.03720.

Joan Bybee. 2015. Language change. Cambridge University Press, Cambridge, UK.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In International Conference on Learning Representations (ICLR) 8.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. In Advances in Neural Information Processing Systems (NeurIPS) 33.

Scott Deerwester, Susan T. Dumais, George Furnas, Thomas Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407.

Marco del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In International Conference on Computational Semantics (IWCS) 12.

Marco del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In International Conference on Computational Linguistics (COLING) 27.

Marco del Tredici, Diego Marcheggiani, Sabine Im Schulte Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 17.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In Annual Meeting of the Association for Computational Linguistics (ACL) 57.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. In arXiv 2004.03032.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2019.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. Transactions of the Association for Computational Linguistics, 8:34–48.

Dirk Geeraerts. 2010. Theories of lexical semantics. Oxford University Press, Oxford, UK.

Dirk Geeraerts. 2018. Ten lectures on cognitive sociolinguistics. Brill, Leiden.

Mario Giulianelli, Marco del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In Annual Meeting of the Association for Computational Linguistics (ACL) 58.

Yoav Goldberg. 2019. Assessing BERT’s syntactic abilities. In arXiv 1901.05287.

Hongyu Gong, Suma Bhat, and Pramod Viswanath. 2020. Enriching word embeddings with temporal and spatial information. In Conference on Computational Natural Language Learning (CoNLL) 24.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In International Conference on Knowledge Discovery and Data Mining (KDD) 22.

William Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016a. Inducing domain-specific sentiment lexicons from unlabeled corpora. In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016.

William Hamilton, Jure Leskovec, and Dan Jurafsky. 2016b. Diachronic word embeddings reveal statistical laws of semantic change. In Annual Meeting of the Association for Computational Linguistics (ACL) 54.

Jennifer Hay, Paul Warren, and Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. Journal of Phonetics, 34(4):458–484.

Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, and Rada Mihalcea. 2018. CASCADE: Contextual sarcasm...
John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HTL) 17.

Valentin Hofmann, Janet B. Piechumhert, and Hinrich Schütze. 2020. DagoBERT: Generating derivational morphology with a pretrained language model. In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In Annual Meeting of the Association for Computational Linguistics (ACL) 57.

Kokil Jaidka, Niyati Chhaya, and Lyle H. Ungar. 2018. Diachronically degradation of language models: Insights from social media. In Annual Meeting of the Association for Computational Linguistics (ACL) 56.

Ganesh Jawahar, Benoit Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In Annual Meeting of the Association for Computational Linguistics (ACL) 57.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In Workshop on Language Technologies and Computational Social Science.

Diederik P. Kingma and Jimmy L. Ba. 2015. Adam: A method for stochastic optimization. In International Conference on Learning Representations (ICLR) 3.

Peter Koch. 2016. Meaning change and semantic shifts. In Päivi Juvonen and Maria Koptjevskaja-Tamm, editors, The lexical typology of semantic shifts, pages 21–66. De Gruyter, Berlin.

Y. Alex Kolchinski and Christopher Potts. 2018. Representing social media users for sarcasm detection. In Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In The Web Conference (WWW) 24.

William Labov. 2001. Principles of linguistic change: Social Factors. Blackwell, Malden, MA.

Chang Li and Dan Goldwasser. 2019. Encoding social information with graph convolutional networks for political perspective detection in news media. In Annual Meeting of the Association for Computational Linguistics (ACL) 57.
Hermann Paul. 1880. *Principien der Sprachgeschichte*. Tübingen, Niemeyer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014*.

Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Annual Meeting of the Association for Computational Linguistics (ACL) 55*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 16*.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2018*.

Janet Pierrehumbert. 2012. The dynamic lexicon. In Abigail Cohn, Cécile Fougeron, and Marie Huffman, editors, *The Oxford handbook of laboratory phonology*, pages 173–183. Oxford University Press, Oxford.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020. Will it unblend? In *Findings of Empirical Methods in Natural Language Processing (EMNLP) 2020*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Justyna Robinson. 2010. Awesome insights into semantic variation. In Dirk Geeraerts, Gitte Kristiansen, and Yves Peirsman, editors, *Advances in cognitive sociolinguistics*, pages 85–109. De Gruyter, Berlin.

Justyna Robinson. 2012. A sociolinguistic approach to semantic change. In Kathryn Allan and Justyna Robinson, editors, *Current methods in historical semantics*, pages 199–231. De Gruyter, Berlin.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. In *arXiv 2002.12327*.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 16*, pages 474–484.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *The Web Conference (WWW)*.

Jiliang Tang, Huiji Gao, and Huan Liu. 2012. mTrust: Discerning multi-faceted trust in a connected world. In *International Conference on Web Search and Data Mining (WSDM)*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations (ICLR) 7*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations (ICLR) 6*.

Ivan Vulić, Edoardo M. Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*.

Charles Welch, Jonathan Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *arXiv 2010.02986*.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does BERT make any sense? interpretable word sense disambiguation with contextualized embeddings. In *arXiv 1909.10430*.

Yi Yang, Ming-Wei Chang, and Jacob Eisenstein. 2016. Toward socially-infused information extraction: Embedding authors, mentions, and entities. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016*.

Yi Yang and Jacob Eisenstein. 2017. Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics*, 5:295–307.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)* 33.

Jing Yao, Zhicheng Du, and Ji-Rong Wen. 2020. Employing personal word embeddings for personalized search. In *International Conference on Research and Development in Information Retrieval (SIGIR) 43*. 
A Appendices

A.1 Hyperparameters

The hyperparameters of the contextualizing component are as for BERT\textsubscript{BASE}. In particular, the dimensionality of the input embeddings $\tilde{e}^{(k)}$, $d_e$, is 768. For the dynamic component, the temporal vectors $(t_j, \tilde{t}_j)$ have a dimensionality $d_t$ of 20, and the social vectors $(s_i, \tilde{s}_i)$ have a dimensionality $d_s$ of 50. The node2vec vectors for the initialization of $\tilde{s}_i$ are trained on 10 sampled walks of length 80 per node with a window size of 10. The dimensionality $d_r$ of the joint representation of social and temporal context, $r_{ij}$, equals the sum of $d_s$ and $d_t$, i.e., 70. The GAT has two layers with four attention heads, respectively (activation function: tanh). All feed-forward networks have one layer (activation function: tanh). We do not compute offsets for BERT’s special tokens (e.g., padding tokens).

The search procedure of the remaining hyperparameters differs for the models trained on sentiment analysis and masked language modeling. For sentiment analysis, we use a batch size of 8 and perform grid search for the learning rate $l \in \{1 \times 10^{-6}, 3 \times 10^{-6}\}$ and the number of epochs $n \in \{1, \ldots, 6\}$ (number of hyperparameter search trials: 12). For masked language modeling, we use a batch size of 4 and perform grid search for the learning rate $l \in \{1 \times 10^{-6}, 3 \times 10^{-6}\}$ and the number of epochs $n \in \{1, \ldots, 6\}$ (number of hyperparameter search trials: 12). Models are trained with Adam (Kingma and Ba, 2015).

All experiments are performed on a GeForce GTX 1080 Ti GPU (11GB).