Temporally-Informed Analysis of Named Entity Recognition

Shruti Rijhwani[*]
Language Technologies Institute
Carnegie Mellon University
srijhwan@cs.cmu.edu

Daniel Preoţiuc-Pietro
Bloomberg
dpreotiucpie@bloomberg.net

Abstract

Natural language processing models often have to make predictions on text data that evolves over time as a result of changes in language use or the information described in the text. However, evaluation results on existing data sets are seldom reported by taking the timestamp of the document into account. We analyze and propose methods that make better use of temporally-diverse training data, with a focus on the task of named entity recognition. To support these experiments, we introduce a novel data set of English tweets annotated with named entities.[1] We empirically demonstrate the effect of temporal drift on performance, and how the temporal information of documents can be used to obtain better models compared to those that disregard temporal information. Our analysis gives insights into why this information is useful, in the hope of informing potential avenues of improvement for named entity recognition as well as other NLP tasks under similar experimental setups.

1 Introduction

Natural language processing models are now deployed on a large scale in many applications and used to drive automatic analyses or for making predictions. The usual setup is that these models are trained and evaluated on the data available at model building time, but are used to make inferences on data coming in at a future time, making models susceptible to data drift. The data distribution of the test set used to measure the model's performance after training may be different from the distribution of data from future time periods (Huang and Paul, 2018). This temporal drift in data often results in lower performance during inference. Drift is especially prevalent in information extraction tasks, such as named entity recognition (NER), where the context and the target entities differ across time as a result of changes in language use or the events being discussed (Derczynski et al., 2016).

Despite its intuitive value, there has been little research on using the temporal information contained in text documents to inform modeling of a task (Huang and Paul, 2018; He et al., 2018), and no past research on modeling sequence labeling tasks in particular. Since sequence labeling models are currently trained and evaluated by randomly splitting the available data, performance is measured in an artificially temporal drift-free scenario that is not realistic or similar to how models are used in practice (Dredze et al., 2010). When splitting the available training data temporally and testing on the data from the most recent time period, we formulate the following hypotheses:

a) models trained on data from a closer time to the test set obtain better results, assuming the same model and data size are used;

b) models trained on the combined data from all time periods outperform models trained on subsets of the data, as more data usually leads to better models. In these cases, the commonly used setup of pooling all the data for training while disregarding temporal information may lead to sub-optimal performance.

In this paper, we study the temporal aspects of text data, focusing on the information extraction task of named entity recognition in the Twitter domain. We make the following contributions:

a) a new data set for Twitter Named Entity Recognition consisting of 12,000 English tweets evenly distributed across six years;

b) experimental results that demonstrate the performance drift of models trained on data from

[*] Work done during an internship at Bloomberg

[1] Our data set is available at https://github.com/shrutirij/temporal-twitter-corpus.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7605–7617
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

different time periods and tested on data from a future interval;

c) extensive analysis of the data that highlights temporal drift in the context of named entities and illustrates future modeling opportunities;

d) simple extensions to state-of-the-art NER models that leverage temporal information associated with the training data, which results in an improvement in F1 score over standard pooling methods.

2 Related Work

Language change is a popular topic of research in linguistics (Stephen, 1962). In natural language processing, using data from online platforms such as Twitter or discussion fora, language change and adoption have been studied at the community level (Danescu-Niculescu-Mizil et al., 2013; Eisenstein et al., 2014; Goel et al., 2016; Stewart and Eisenstein, 2018) and at the individual level (Zhang et al., 2019). In some cases, the senses of the same word are known to shift over time (Wijaya and Yeniterzi, 2011), and modeling such changes in word semantics has been explored using diachronic word embeddings (Kulkarni et al., 2015; Hamilton et al., 2016; Kutuzov et al., 2018).

Temporal information has been used to create topic models of better quality, usually by adding smoothing properties (Blei and Lafferty, 2006; Wang et al., 2008). For text classification, the temporal periodicity of Twitter hashtags was modeled in Preoţiuc-Pietro and Cohn (2013) and used as a prior for text classification models for predicting hashtags on future data, which resulted in performance improvements.

Most similar to our experimental setup, Huang and Paul (2018) study the impact of temporal data splits in text classification, finding that performance worsens on data from future periods, and use standard domain adaptation techniques to incorporate time information and improve results. He et al. (2018) introduce a method for training neural networks on data from multiple time intervals while enforcing temporal smoothness between representations. Temporal information has also been used to improve named entity disambiguation on a data set of historical documents (Agarwal et al., 2018). Finally, Huang and Paul (2019) present a model that uses diachronic word embeddings combined with a method inspired by domain adaptation to improve document classification.

A related, but distinct, task built on the assumption of language change with time is automatic prediction of the date on which a document is written (Kanhabua and Nørvåg, 2008; Chambers, 2012; Niculae et al., 2014).

Named entity recognition (NER) is the task of identifying entities such as organizations, persons, and locations in natural language text. NER is a well-studied NLP task over the past 20 years (Nadeau and Sekine, 2007; Yadav and Bethard, 2018) and is a key information extraction task as its used in various downstream applications such as named entity linking (Cucerzan, 2007), relation extraction (Culotta and Sorensen, 2004) and question answering (Krishnamurthy and Mitchell, 2015). On social media text, such as tweets, the performance lags far behind that of standard news corpora (Derczynski et al., 2015b), with data drift as one of the suggested causes (Derczynski et al., 2015a). Agarwal et al. (2020) show that NER models decay substantially on entity mentions from a different distribution than those seen in training.

NER systems struggle to generalize over diverse genres with limited training data (Augenstein et al., 2017). Domain adaptation for NER (Chiticariu et al., 2010; Lin and Lu, 2018; Wang et al., 2020) is related to our task of improving performance over temporal drift, as the data from a future time period can be considered as a target domain with an unknown distribution. However, the relationship between domains is implied from temporal similarity, and temporal information is very fine-grained in contrast to the standard single source to single target domain adaptation setup.

3 Temporal Twitter Data Set

In this paper, we focus on the task of named entity recognition on English tweets as a case study for our hypotheses and analysis regarding model drift with time. Twitter data represents an ideal testbed for our analysis as it contains readily accessible timestamp information for each tweet. Further, users on social media post about current events, which are likely to include entities that change over time. Social media also reflects changes in language use more timely than other sources of data (e.g., newswire), resulting in the potentially rapid evolution of the contexts and ways in which named entities are discussed in natural language. This drift in Twitter data has previously been demonstrated qualitatively in the context of named entity

Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Broad Twitter Corpus	5	127	2,414	275	6,022	–	–	–	–	–
Current Data Set	–	–	–	–	2,000	2,000	2,000	2,000	2,000	2,000

Table 1: Number of tweets from each year in the BTC data set and in the data set introduced in this paper.

recognition (Derczynski et al., 2015a).

Previous research has introduced data sets of tweets annotated with named entities, including the data sets from Finin et al. (2010), Ritter et al. (2011), Liu et al. (2011), the WNUT-17 Corpus (Derczynski et al., 2017), the Microposts NEEL Challenge Corpora (Rowe et al., 2013; Cano et al., 2014; Rizzo et al., 2015; Cano et al., 2016) and the Broad Twitter Corpus (Derczynski et al., 2016). However, these data sets usually consist of tweets collected within a limited time period, making them unsuitable for our proposed work. Of note is the Broad Twitter Corpus (Derczynski et al., 2016), which contains tweets collected over several years, from 2009 to 2014. However, the majority of tweets are from either 2012 or 2014, with fewer than 300 tweets from the other years (details in Table 1). Further, combining existing data sets is challenging, because of the different entity tagging schemes, annotation guidelines and sampling strategies used.

Therefore, we create a new collection of tweets annotated with named entities that attempts to alleviate the lack of temporal diversity in existing Twitter data sets as well as provide us with a suitable experimental setup to study our research questions about temporal entity drift and NER model performance. In this section, we present the details of our data set, including the collection and annotation methodology, as well as an analysis of the named entity mentions in the corpus. The data set can be downloaded at https://github.com/shrutirij/temporal-twitter-corpus.

3.1 Data Collection

The primary goal of creating a new data set is ensuring wide-enough temporal diversity for our work as well as future directions that can leverage timestamp information. We use the public Twitter Search API[2] to sample tweets spanning six years: 2014, 2015, 2016, 2017, 2018 and 2019.

We aim to ensure that the data set is representative of multiple English-speaking locales and a variety of topics, as well as making it comparable to existing data sets. Thus, we follow the same sampling strategy for corpus diversity used by the creators of the Broad Twitter Corpus (Derczynski et al., 2016). Specifically, we collect tweets across six English-speaking regions (the United States, the United Kingdom, New Zealand, Ireland, Canada, and Australia), and focus on two contrasting sets of Twitter handles: a) the *twitterati*, i.e., individuals from array of domains including musicians, journalists and celebrities; b) Twitter accounts for mainstream news organizations, covering both larger networks like CNN and ABC, as well as local news outlets. The Twitter handles correspond to users from the segments F and G of the Broad Twitter Corpus (Derczynski et al., 2016).

Overall, to maintain uniformity across time, we annotated 2,000 tweets for each year from 2014 to 2019 by randomly subsampling tweets from each year. This resulted in a temporally varied and balanced corpus of 12,000 tweets. Table 1 illustrates the temporal data distribution of our data set, as compared to the Broad Twitter Corpus.

3.2 Annotation

In annotating our data with entities, we use a tagset consisting of three entity classes – Organizations (ORG), Persons (PER), and Locations (LOC). This scheme is consistent with some existing data sets for the task (Finin et al., 2010; Derczynski et al., 2016), overlapping with the majority of other general NER datasets in the social media domain (Liu et al., 2011; Rowe et al., 2013) and beyond (Tjong Kim Sang and De Meulder, 2003a).

We use the annotation guidelines used in standard NER data sets (Tjong Kim Sang and De Meulder, 2003a) supplemented with examples that are specific to Twitter data.

Further, we observe in other data sets that usernames are some of the most frequent tokens classified as entities (Ritter et al., 2011; Derczynski et al., 2016). For our experiments, we consider all usernames as non-entities, as otherwise, identifying

[2] https://developer.twitter.com/en/docs/tweets/search/overview

these using character features would be trivial, and typing entities would be similar to the task of Twitter handle classification (McCorriston et al., 2015; Wood-Doughty et al., 2018), which is outside the scope of the current paper.

We preprocess the data set by normalizing URLs, usernames, and Twitter-specific tokens (e.g., RT). We leave hashtags intact as these are often used as words in the context of the tweet, and can be or contain named entities. We use Twokenizer (O'Connor et al., 2010), a Twitter-specific tokenizer to split the tweets into tokens. To limit the impact of imperfect tokenization on the performance of the NER models – especially in the case of hashtags containing multiple tokens (Maddela et al., 2019) – we expanded sub-token annotations to their closest matching token. If multiple sub-token entity annotations match the same token, then we select the label of the first sub-entity in order of appearance.

The data was annotated by multiple annotators that have experience with named entity recognition annotation tasks. Specifically, we used 15 annotators in total, with two annotations per tweet. The inter-annotator agreement is 78.34% on full tweets (same entity types and spans). If the annotators disagree on a tweet in their tagging, we adjudicate in favor of the annotator that had the highest confidence on the task, as judged through measuring their agreement with our annotations on a set of test questions (10% of the total).

In our experiments, we use temporal splits of the data from 2014–2018 for training, and the most recent data (i.e., the tweets from 2019) to evaluate our models, to simulate a "future time period" setup. Thus, we wanted to ensure that the model performance is evaluated on data that has as few annotation errors as possible. Hence, each tweet was checked by either of the authors of the paper, both with significant experience in linguistic annotations, and corrected if needed to ensure additional consistency. This process had the effect of reducing the measurement error of the model performance but ultimately did not affect the conclusions of the experimental results. The type-wise distribution of named entities in for each year in our data set, after annotator adjudication and correction, is shown in Table 2.

Year	PER	ORG	LOC	Total
2014	371	454	350	1,175
2015	363	479	393	1,235
2016	435	501	320	1,256
2017	432	516	314	1,262
2018	468	597	395	1,460
2019	725	881	475	2,081

Table 2: Year-wise number of named entities of each type in the data set introduced in this paper.

4 Base Model Architecture

This section describes the base model architecture we use to perform named entity recognition experiments throughout the paper. We use the same underlying architecture to provide a controlled experimental setup and isolate temporal modeling aspects from other model-related factors.

4.1 Neural Architecture

We use the neural architecture based on a stacked BiLSTM-CRF model introduced in Huang et al. (2015), which is the core model architecture for several state-of-the-art NER results over the past years (Lample et al., 2016; Peters et al., 2018; Akbik et al., 2018). For each sentence, the token representations are fed into two different LSTM layers, each processing the sentence in different directions (one forward and one backward). The output of these two layers are concatenated and passed through a feed-forward layer that produces a distribution over the output tag space. Finally, a Conditional Random Field is applied to the class predictions with the role of jointly assigning predictions to the entire sequence. This also has the function of ensuring that the output tag sequence takes into account the constraints of the IOB2 entity tagging scheme (e.g. I-LOC cannot follow B-ORG) (Tjong Kim Sang and De Meulder, 2003b).

4.2 Embeddings

A key component in the base architecture is how the tokens are represented as inputs. Initial research (Lample et al., 2016) on LSTM-CRF models use static pre-trained word embeddings, such as GloVe (Pennington et al., 2014), to initialize the inputs, which are subsequently fine-tuned on the NER training data. More recently, contextual word embeddings, which represent each token differently based on its context, were shown to obtain an improvement of 2–3 F1 points on the English news CoNLL data set (Peters et al., 2018; Akbik et al., 2018; Devlin et al., 2019). In this paper, we conduct experiments with both the static GloVe embeddings (Pennington et al., 2014) and the state-of-the-art contextual Flair embeddings (Akbik et al., 2018) to test the robustness of our findings to different input representations. All embeddings were trained outside of the time range of our data: the GloVe embeddings were trained on Twitter data before 2014, while the Flair embeddings were trained on the 1-billion word corpus (Chelba et al., 2013) which contains data up to 2012. Exploiting embeddings trained on data more recent than the NER corpus is an avenue of future work.

In addition to token embeddings, we use character embeddings to model subword information that may be indicative of named entities and better represent out-of-vocabulary tokens. We use a character-level BiLSTM with randomly initialized character embeddings to produce the character-based word representations (Lample et al., 2016). These are concatenated to the token embeddings described above and then used as input to the token-level BiLSTM.

4.3 Data Split

We split the data temporally for our experiments. We use the data authored in 2019 as the test data, as this is the most recent data available and best replicates the scenario of making predictions on text from future time periods. We use a random sample of 500 tweets (25%) from the 2019 data as the validation set.

For training, we use data authored between 2014 to 2018 in various temporal splits, depending on the specific experimental setting.

4.4 Implementation and Hyperparameters

We use the PyTorch framework (Paszke et al., 2017) for the implementation of the models. For the model using the GLoVe embeddings, we use the same hyperparameter settings as the original creators of the base models (Lample et al., 2016; Akbik et al., 2018) and ensure the correctness of our implementation by replicating their results on the CoNLL-2003 English NER data set (Tjong Kim Sang and De Meulder, 2003a). Specifically, the character embeddings are of size 32, the character-level LSTM hidden size is 64, and the word-level LSTM has a hidden size of 256. We also use a dropout of 0.5 on the input word embeddings and replace singleton words in the training set with an out-of-vocabulary symbol with a probability of 0.5 to improve robustness to unseen words. We use the flairNLP library (Akbik et al., 2019) for the contextual Flair embedding experiments, using the same hyperparameters as the state-of-the-art result in Akbik et al. (2018). For each experimental setting, we use the training checkpoint with the best performance on the validation set (i.e., early stopping).

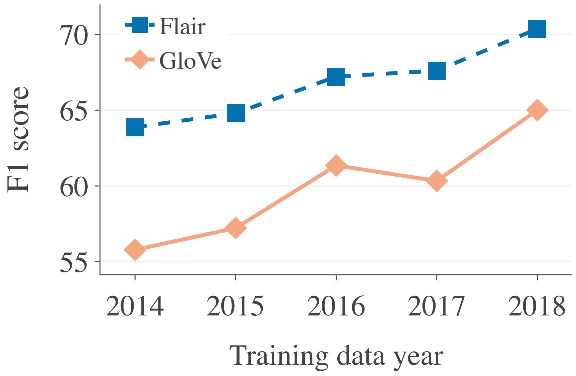


Figure 1: Evaluating the effect of temporal distance: the model is trained on each year individually. F1 score on 2019 data averaged over five random seeds is shown.

Following the recommendation from Reimers and Gurevych (2017), who study the variance of LSTM-CRF models with different random seeds, we report all experimental results as the mean of five runs. The main metric we use for evaluation is span-level named entity F1 score, reported using the official CoNLL evaluation script (Tjong Kim Sang and De Meulder, 2003a).

5 Data Drift

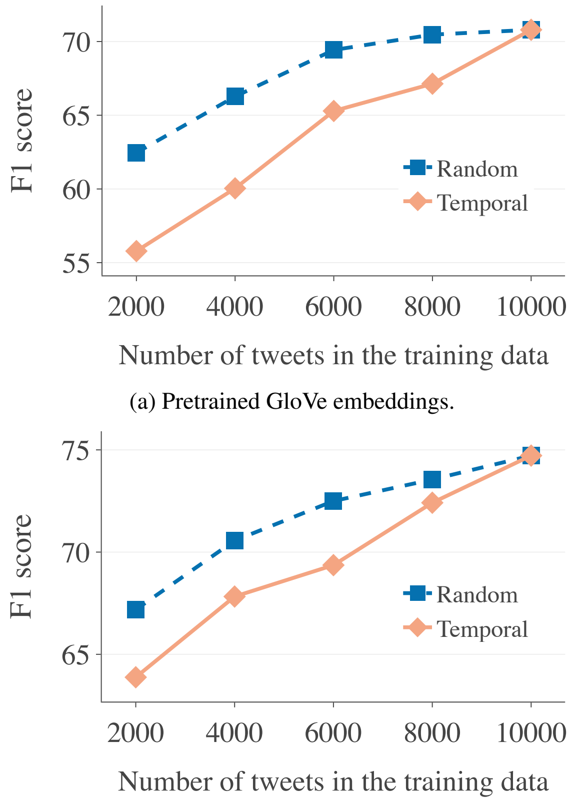
To determine the utility of temporal information, we first attempt to evaluate whether temporal drift in the data affects the performance of NER models. To this end, we conduct experiments to answer the following research questions:

1) What is the effect of the temporal distance between the training and target data sets on NER performance?

2) How do the size and temporal distribution of the training data affect NER performance?

5.1 Effect of Temporal Distance

We empirically study the effect of temporal distance between the training and test data sets by training the base model on each year, from 2014–2018, individually. Based on the design of our data set, each model has access to the same number of



(a) Pretrained GloVe embeddings.

(b) Pretrained Flair embeddings.

Figure 2: Cumulatively training the model on random subsamples of tweets from all the years (2014–2018) compared with temporally adding tweets to the training data, starting from the year 2014 and cumulatively adding data from subsequent years.

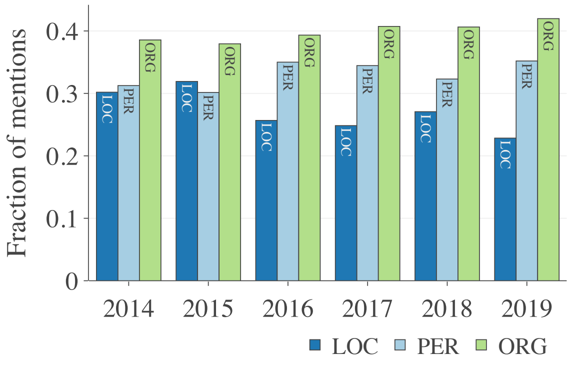


Figure 3: Type distribution across years in our data set.

training instances (2,000 tweets), to remove the impact of this factor in our results.

The results are shown in Figure 1. We observe that the temporal distance between the training and test sets seems to affect NER performance. The F1 score increases as we move temporally closer to the target data, for both the GloVe and Flair embeddings, apart from a slight decrease when moving from 2016 to 2017 when using GloVe embeddings. When using the contextual Flair embeddings, the performance numbers are overall higher, which is consistent with past research (Akbik et al., 2018), as contextual embeddings are more expressive.

5.2 Effect of Data Set Size and Distribution

We now study how the number of instances in the training data and their temporal distribution impact the performance of the model. We first train models on cumulative random samples from the combined training data set (all tweets from 2014–2018), adding 2,000 tweets at each step. Then, we train models starting with the 2,000 tweets from 2014 and incrementally add tweets from subsequent years from 2015 up to 2018.

The NER F1 scores are shown in Figures 2a and 2b, with both “Random” and “Temporal” cumulative compositions of the training data set.

Looking at the “Random” sampling strategy, we see that the performance steadily increases as we add more tweets to the training set – as we would expect for most supervised machine learning models. We see that the “Temporal” model with only the 2014 data (2,000 tweets) has a lower performance than randomly selecting 2,000 tweets across all years. This is indicative of the data drift across time, as training on a random sample of tweets from all the years is more informative and leads to a better NER model than using just the 2014 data.

Moreover, as we add tweets temporally closer to the target into the training data set, the “Temporal” strategy converges with the “Random” strategy. This observation strengthens the hypothesis that temporal information can potentially play an important role while selecting training data and designing model architectures.

5.3 Analysis

To understand why the temporal distribution of the training data impacts the performance of an NER model, we analyze the distribution of entity mentions in our data set to uncover the extent to which data drift occurs at the lexical level.

Type Distribution Figure 3 shows the distribution of entity types across years in our data set. The distribution looks approximately even, with minor differences in the fraction of location (LOC) entities. Since similar types of entities occur in the data set year-wise, this likely does not cause the change in performance across time indicated in the previous sections.

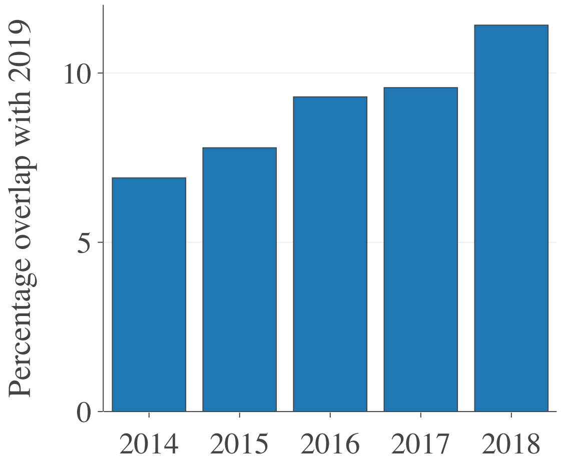


Figure 4: Percentage overlap of unique surface forms of named entity mentions for years 2014 to 2018, with respect to the year 2019. The gradual increase in overlap is an indication of temporal drift.

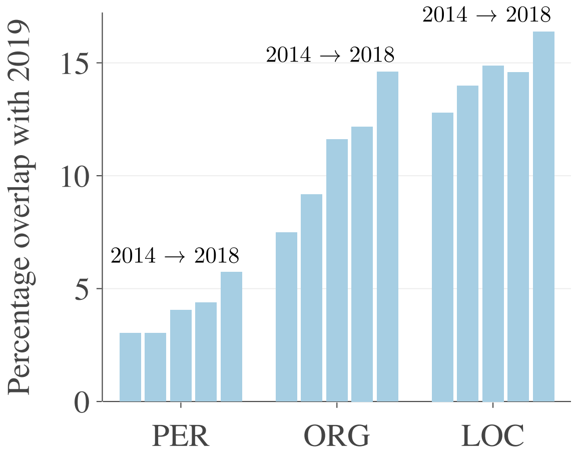


Figure 5: Type-wise percentage overlap of unique mentions for years 2014 to 2018, with respect to the year 2019. For all types, there is a general increase in overlap as we move temporally closer to 2019.

Mention Overlap Figure 4 presents the overlap of unique entity mentions with respect to the test data (2019). There is a clear increase in surface-form overlap as we get temporally closer to the target data, and is potentially an important factor for the F1 score improvement we see in our empirical analysis.

Type-wise Mention Overlap Figure 5 shows the surface-form overlap of entity mentions over types of years 2014 to 2018, with respect to the data from 2019. The figure adds further evidence of temporal data drift at the mention level. For all three entity types (LOC, PER, ORG) in our data set, smaller temporal distance leads to a greater percentage of overlap. Interestingly, the PER overlap is much lower than other types. This is expected, as the people discussed on social media rapidly change with developments in current events (Derczynski et al., 2017). We see that the 2017 data set has a lower overlap for LOC than both 2016 and 2018, which could explain the off-trend performance of the 2017 model in our empirical results (Figure 1).

	GloVe			Flair		
Train	PER	ORG	LOC	PER	ORG	LOC
2014	74.45	41.63	52.78	79.66	53.78	56.90
2015	73.39	45.97	52.14	81.91	52.23	58.77
2016	78.42	49.12	57.60	81.58	58.19	60.85
2017	74.63	51.23	52.97	81.82	60.10	58.41
2018	79.40	56.29	59.25	83.47	61.83	64.90

Table 3: Type-wise F1 when testing on data from 2019. The models are trained on each year individually. The training data sets are of the same size (2,000 tweets).

Type-wise Model Performance Table 3 shows the NER performance by entity type, to gain more insight into which types are affected by data drift.

First, we notice that the improved performance of Flair embeddings seen in previous analyses is caused by better performance across all types. Overall, the PER type obtains the best performance for both models, with an F1 of around 20 points higher than the other two types. This is despite the fact that the PER type has the lowest overall overlap between training and test, which indicates that the model is adequately learning the contexts that PER entities appear in. ORG and LOC show similar absolute performance in both setups.

Next, we study the temporal differences in performance by type. When using GloVe embeddings, the smallest gap between training on different data splits is for PER (4.95 F1), while ORG suffers from substantial drift in performance, resulting in a 14.66 F1 drop on ORG performance. When using Flair embeddings, the most notable difference in performance when training across different years is still for the ORG type (up to 8.05 F1). However, the gap has proportionally tightened the most as compared to when using GloVe embeddings. These observations correspond with the analysis from Figure 5, where we see the largest increase in overlap between mentions from the training data and the test data over the five years ($\tilde{8}$% increase for ORG, compared to 3–4% increase for LOC and PER).

	Unseen Recall	
Train	GloVe	Flair
2014	53.72	56.62
2015	54.22	56.64
2016	58.90	58.98
2017	59.61	57.88
2018	64.36	60.00

Table 4: Recall for named entity mentions in the test data (2019) unseen in the training data, for models trained on each year individually.

We also observe that the slight drop in performance of the model using GloVe embeddings trained on the 2017 data is caused primarily by a decline in performance on the LOC type which holds across both models.

Mentions Unseen in the Training Data In addition to the increase in surface-form overlap across years, we investigate whether mentions unseen in the training data are impacted by the temporal distance between the training and test data. Table 4 shows the recall for these mentions using both the GloVe and Flair embeddings. Notably, for GloVe, the performance steadily improves as the temporal distance decreases, with an almost 5 point improvement in recall when moving from 2017 to 2018. Although less pronounced, there is a similar trend with the Flair embeddings. This indicates that surface-form overlap is not the only factor determining temporal data drift. The model is potentially able to learn more relevant context from the training data of temporally close years, perhaps due to changes in language use over time.

6 Modeling Temporal Information

Supported by the analysis that temporal drift in the training data can impact the performance of NER systems, in this section, we experiment with techniques to account for temporal information while training the NER model. We look at leveraging temporality in two broad ways: a) by altering the architecture of the base model; b) by modifying how the training data set is constructed. These methods are intended to be an initial exploration of using temporal information, with a focus on techniques that do not require significant modification to the base model. We present these in the hope that they will inspire future research on models robust to temporal drift. The specific methods are discussed below, followed by experimental results.

6.1 Methods

Sequential Temporal Training Our analysis from Section 5 showed that using more data is beneficial, irrespective of temporal distance from the target, but individually, the closest data is most useful. Based on this analysis, we attempt to train our model by ordering our training data year-wise such that the model is trained on the temporally closest data last. Specifically, we start with training on the year temporally furthest away from the target data and repeatedly tune the model on the chronological sequence of years (i.e., first train on 2014 data, then 2015 data, and so on up to 2018).

Temporal Fine-tuning The analysis showed that training on the model temporally closest to the target data set obtains the best overall performance. Based on this observation, we decide to train the base model on the entire data set of tweets from the years 2014–2018. Then, we fine-tune the trained model on the data from the year temporally closest to the target (2018). The fine-tuning process is simply retraining the model on the 2018 data with the same hyperparameter settings.

Instance Weighting Previous work in domain adaptation shows that giving higher weights to training instances similar to the target domain can improve performance (Wang et al., 2017). Similarly, we decide to assign a higher weight to tweets temporally closer to the test data (i.e., the 2018 tweets are up-weighted). In our experiments, we up-weight the tweets by a factor of 2.

We note that the above methods do not require any change to the model, making integration of these methods for existing systems very practical.

Year Prediction as an Auxiliary Task Finally, we aim to guide the model to learn temporal features in training. Inspired by related work in domain adaptation (Chen et al., 2018), we enhance the architecture with a multi-task learning component that models an auxiliary task. While training the model for NER, this component uses the LSTM hidden states to predict the year that the tweet was created in. Since the input embeddings and the LSTM are shared between the NER task and the year prediction task, the intuition is that the parameters learned will retain a notion of temporality that

	GloVe	Flair
Base Model	70.80	74.72
Sequential Temporal Training	68.47	74.42
Temporal Fine-tuning	**71.93**	74.95
Instance Weighting	70.59	**75.54**
Year Prediction	71.01	74.70

Table 5: Performance of proposed methods of using temporal information in NER modeling when compared to the base model. Results are F1 scores averaged over five runs with different random seeds. Bold indicates the best F1 score.

can influence the NER prediction. The training objective is the sum of the NER loss and the auxiliary task loss.

6.2 Experimental Results

Table 5 presents the experimental results. The base model combines the training data (2014–2018) without using any temporal information, the current standard setup for most NLP systems.

The results show that we can overall obtain a better performance over the base model by using simple techniques to incorporate temporal information. The margin of improvement is overall lower when using Flair embeddings than with GloVe (+0.82 compared to +1.13). This potentially indicates that semantic drift can be captured partially through contextual embeddings.

Fine-tuning the model on the temporally closest data (i.e., 2018) leads to the best F1 scores when using GloVe embeddings, reaching a 1.13 increase in F1. For the Flair embeddings, we observe that up-weighting the training instances from the year 2018 leads to the best result, a 0.82 improvement in F1 over the base model.

We highlight that these straightforward methods that improve over the base model do not involve any architecture changes, other than a change in how the data is fed to the model. It thus has the potential to both be readily applicable to existing NER implementations as well as generalize to other NLP tasks.

Finally, we find that using an auxiliary task for predicting the year improves the performance slightly when using GloVe embeddings, but has the oposite effect when using Flair embeddings. This is likely because the GloVe embeddings are fine-tuned during the model training and are therefore influenced by the auxiliary loss, while the contextual Flair embeddings are not.

7 Conclusions

This paper studies and models text data drift in the information extraction task of named entity recognition. We introduce a new data set of 12,000 English tweets stratified by time, which allows us to study the effects of drift and evaluate named entity recognition models in a realistic scenario of performing inference on temporally unseen data. By analyzing the data, we quantify the temporal drift in named entity type and mention usage and identify that, as expected, the data distribution is more similar when drawn from closer time intervals. We then use current state-of-the-art approaches for named entity recognition and demonstrated that, through modeling of temporal information, performance can be improved when testing on future data. We expect our data, results, and error analysis to inform the design of similar experimental setups for other NLP tasks beyond NER, such as part-of-speech tagging or relation extraction.

Acknowledgements

We would like to thank Leslie Barrett, Liang-Kang Huang, Prabhanjan Kambadur, Mayank Kulkarni, Amanda Stent, Umut Topkara, Jing Wang, Chuck-Hou Yee and the other members of the Bloomberg AI group. They provided invaluable feedback on the experiments and the paper. We also thank the anonymous reviewers for their valuable suggestions. Shruti Rijhwani is supported by a Bloomberg Data Science Ph.D. Fellowship.

References

Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and Ani Nenkova. 2020. Interpretability analysis for named entity recognition to understand system predictions and how they can improve. *ArXiv*, abs/2004.04564.

Prabal Agarwal, Jannik Strötgen, Luciano del Corro, Johannes Hoffart, and Gerhard Weikum. 2018. diaNED: Time-aware named entity disambiguation for diachronic corpora. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Melbourne, Australia. Association for Computational Linguistics.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019.

FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML, pages 113–120.

Amparo E Cano, Daniel Preotiuc-Pietro, Danica Radovanović, Katrin Weller, and Aba-Sah Dadzie. 2016. # microposts2016: 6th workshop on making sense of microposts: Big things come in small packages. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 1041–1042.

Amparo E Cano, Giuseppe Rizzo, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. 2014. Making sense of microposts:(#microposts2014) named entity extraction & linking challenge. In *Proceedings of the 23rd International Conference Companion on World Wide Web*, volume 1141, pages 54–60.

Nathanael Chambers. 2012. Labeling documents with timestamps: Learning from their time expressions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 98–106, Jeju Island, Korea. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *ArXiv*, abs/1312.3005.

Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. 2018. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 602–607, New Orleans, Louisiana. Association for Computational Linguistics.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012, Cambridge, MA. Association for Computational Linguistics.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 423–429, Barcelona, Spain.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW, pages 307–318.

Leon Derczynski, Isabelle Augenstein, and Kalina Bontcheva. 2015a. USFD: Twitter NER with drift compensation and linked data. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 48–53, Beijing, China. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1169–1179, Osaka, Japan. The COLING 2016 Organizing Committee.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke Van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015b. Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, 51(2):32–49.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Mark Dredze, Tim Oates, and Christine Piatko. 2010. We're not in kansas anymore: Detecting domain changes in streams. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 585–595, Cambridge, MA. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating named entities in twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, Los Angeles. Association for Computational Linguistics.

Rahul Goel, Sandeep Soni, Naman Goyal, John Paparrizos, Hanna Wallach, Fernando Diaz, and Jacob Eisenstein. 2016. The social dynamics of language change in online networks. In *International Conference on Social Informatics*, SocInfo, pages 41–57.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Yu He, Jianxin Li, Yangqiu Song, Mutian He, and Hao Peng. 2018. Time-evolving text classification with deep neural networks. In *IJCAI*, pages 2241–2247.

Xiaolei Huang and Michael J. Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–699, Melbourne, Australia. Association for Computational Linguistics.

Xiaolei Huang and Michael J. Paul. 2019. Neural temporality adaptation for document classification: Diachronic word embeddings and domain adaptation models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4113–4123, Florence, Italy. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *ArXiv*, abs/1508.01991.

Nattiya Kanhabua and Kjetil Nørvåg. 2008. Improving temporal language models for determining time of non-timestamped documents. In *International Conference on Theory and Practice of Digital Libraries*, pages 358–370.

Jayant Krishnamurthy and Tom M. Mitchell. 2015. Learning a compositional semantics for Freebase with an open predicate vocabulary. *Transactions of the Association for Computational Linguistics*, 3:257–270.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW, pages 625–635.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2012–2022, Brussels, Belgium. Association for Computational Linguistics.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 359–367, Portland, Oregon, USA. Association for Computational Linguistics.

Mounica Maddela, Wei Xu, and Daniel Preoţiuc-Pietro. 2019. Multi-task pairwise neural ranking for hashtag segmentation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2538–2549, Florence, Italy. Association for Computational Linguistics.

James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations are users too: Characterizing and detecting the presence of organizations on twitter. In *Ninth International AAAI Conference on Web and Social Media*, ICWSM.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Vlad Niculae, Marcos Zampieri, Liviu Dinu, and Alina Maria Ciobanu. 2014. Temporal text ranking and automatic dating of texts. In *Proceedings of the 14th Conference of the European Chapter of the*

Association for Computational Linguistics, volume 2: Short Papers, pages 17–21, Gothenburg, Sweden. Association for Computational Linguistics.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for Twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*, ICWSM.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*.

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, Doha, Qatar. Association for Computational Linguistics.

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

Daniel Preoţiuc-Pietro and Trevor Cohn. 2013. A temporal model of text periodicities using Gaussian processes. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 977–988, Seattle, Washington, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Giuseppe Rizzo, Amparo Elizabeth Cano Basave, Bianca Pereira, Andrea Varga, Matthew Rowe, Milan Stankovic, and A Dadzie. 2015. Making sense of microposts (microposts2015) named entity recognition and linking (neel) challenge. In *Proceedings of the 24th International Conference Companion on World Wide Web*, pages 44–53.

Matthew Rowe, Milan Stankovic, Aba Sah Dadzie, B.P. Nunes, and Amparo Elizabeth Cano. 2013. Making sense of microposts (#msm2013): Big things come in small packages. In *Proceedings of the 22nd International Conference Companion on World Wide Web*.

Ullmann Stephen. 1962. Semantics: an introduction to the science of meaning.

Ian Stewart and Jacob Eisenstein. 2018. Making "fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370, Brussels, Belgium. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003a. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003b. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Chong Wang, David Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI, pages 579–586.

Jing Wang, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2020. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, Copenhagen, Denmark. Association for Computational Linguistics.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 35–40.

Zach Wood-Doughty, Praateek Mahajan, and Mark Dredze. 2018. Johns Hopkins or johnny-hopkins: Classifying individuals versus organizations on twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 56–61, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding your voice: The linguistic development of mental health counselors. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 936–947, Florence, Italy. Association for Computational Linguistics.