Sense of the Day: Short Timeframe Temporal-Aware Word Sense Disambiguation

Yuchen Wei, Milton King
St. Francis Xavier University
Antigonish, Nova Scotia, Canada
x2020fct@stfx.ca, mking@stfx.ca

Abstract

The predominant sense of a lemma can vary based on the timeframe (years, decades, centuries) that the text was written. In our work, we explore the predominant sense of shorter timeframes (days, months, seasons, etc.) and find that different short timeframes can have different predominant senses from each other and from the predominant sense of a corpus. Leveraging the predominant sense and sense distribution of a short timeframe, we design short timeframe temporal-aware word sense disambiguation (WSD) models that outperform a temporal agnostic model. Likewise, author-aware WSD models tend to outperform author agnostic models, therefore we augment our temporal-aware models to leverage knowledge of author-level predominant senses and sense distributions to create temporal and author-aware WSD models. In addition to this, we found that considering recent usages of a lemma by the same author can assist a WSD model. Our approach requires the use of only a small amount of text from authors and timeframes.

Keywords: Word sense disambiguation, Temporal-aware, Personalization

1. Introduction

Gella et al. (2014) found that individual Twitter users tend to favour a sense for a specific lemma, which might not be the same sense for all users. This suggests that a word sense disambiguation (WSD) model should treat authors of text more uniquely by tailoring themselves toward each author. King and Cook (2021) used author-aware personalized WSD models for each author of text by leveraging knowledge of a single author's predominant sense or sense distributions. Their author-aware models outperformed models that were author-agnostic and models that considered the predominant sense or sense distributions of a group of authors. Following the themes of Gella et al. (2014) and King and Cook (2021), we explore if different senses tend to be favoured within specific temporal segments, which we then use to create WSD models that are tailored toward a specific temporal segment (day, month, season).

Many social media platforms include a timestamp with each post that a user publishes, which makes access to temporal information relatively feasible as opposed to metadata that is more difficult to obtain, such as the author's age, gender, and geolocation. We use the same dataset that was used by King and Cook (2021), which contains 1586 sense-annotated instances of nouns contained in timestamped blog posts and was originally used in Schler et al. (2006). Our analysis shows that different temporal segments tend to favour specific senses, which can differ among each temporal segment. Following this finding, we then explore the potential of WSD models that are tailored toward a temporal segment. Specifically, we use the same underline WSD model, known as SensEmBERT (Scarlini et al., 2020a), that King and Cook (2021) used in their personalized WSD model. Alternative WSD models include ARES (Scarlini et al., 2020b) — a WSD model that generates sense embeddings through semi-supervision while performing well on multilingual WSD tasks — and LMMS (Loureiro and Jorge, 2019) — a model that leverages contextual embeddings from BERT to perform WSD. We tailor SensEmBERT to specific temporal segments using similar techniques that were used by King and Cook (2021) along with our proposed techniques, which consider temporal-based predominant senses or sense distributions. Our findings show that our proposed models, which are tailored toward temporal segments, outperform models that do not consider temporal information. However, models that are tailored toward individual authors outperform temporal-based models. Our final set of experiments involves proposing models that consider both author-level and temporal-level information, which achieves our highest performance.

2. Related Work

Many modern knowledge-based WSD models incorporate a Lesk-based approach (Lesk, 1986), which involves comparing the overlap between a target token with the gloss of candidate senses and selecting the sense with the highest amount of overlap. Banerjee et al. (2003) explored extending the gloss of senses with the gloss of similar

LREC-COLING 2024, pages 14676–14686
20-25 May, 2024.

senses, which increased the amount of text representing each sense. Blevins and Zettlemoyer (2020) proposed a bi-encoder model to embed the target word with its context and gloss of senses. This method can reduce the error rate of predicting non-frequent word senses. Basile et al. (2014) incorporated word vectors to embed the context of a target token, whose similarity to the embedding of candidate senses is used to assign a sense. SensEmBERT (Scarlini et al., 2020a) extends this idea of embedding the context of a target token and measuring the similarity to an embedding that represents a sense. They generated a sense-level embedding using BERT (Devlin et al., 2019) to embed the gloss of a sense and concatenating it with the embedding of Wikipedia articles related to the words present in a sense's gloss. Given a target token, it's context is embedded using BERT, which is compared to candidate sense embeddings. The sense that corresponds to the sense-level embedding with the highest similarity is assigned to the target token. Barba et al. (2021) approached WSD as a span extraction problem and trained a transformer-based model to extract the correct definition of a word in context from a string of definitions.

WSD has been explored in a variety of domains, such as editorial, news story, and fiction (Snyder and Palmer, 2004); articles from the Wall street Journal (Pradhan et al., 2007); printed text (Miller et al., 1994); Twitter (Gella et al., 2014); and blog posts (King and Cook, 2021). We perform our experiments with English blog posts, which include timestamps.

The most frequently used sense of a lemma — known as the predominant sense — can change across different temporal periods. For example, the lemma *cell* could more likely be used in the year 1500 with its sense referring to *a room where a prisoner is kept*, but in 2023, it could more likely be used with its sense that refers to *a hand-held mobile radiotelephone for use in an area divided into small sections, each with its own short-range transmitter/receiver*. The change in the predominant sense across temporal periods have been the interest of different groups. Mathew et al. (2017) compared a topic modelling approach (Lau et al., 2014), graph-based model (Mitra et al., 2014), and a statistical-based model (McCarthy et al., 2007) to detect the difference in predominant senses across two corpora of different temporal periods (1987-1995, 2006-2008). Loureiro et al. (2022) presented a task called meaning shift detection along with a dataset called TempoWiC based on Twitter. They focused not only on the semantic representation but also on the change of word senses over time. Similarly, Schlechtweg et al. (2020) proposed a SemEval shared task that involved determining which word types had their number of senses changed between two time-specific corpora and which word types had a larger change in their meaning. Kutuzov et al. (2018) discussed in their a survey a breadth of findings from research related to diachronic word embeddings and how the meaning of words can change over time. They also presented existing challenges and potential direction of research related to the meaning of words changing over time, which includes, but are not limited to: increasing the number of considered languages, methods for smaller datasets, and exploring more detailed analysis beyond the detection of a change in meaning.

Beelen et al. (2021) introduced the task of time-sensitive targeted sense disambiguation, which involves determining if a given target token in timestamped text is related to a given sense. They look at samples from 1760-1920. Piao et al. (2017) considered temporal information in their WSD tool (Historical-Thesaurus-based Semantic Tagger) by limiting the list of candidate senses to only consider senses that occurred in a given temporal window consisting of decades.

Unlike the recently mentioned works that focus on long temporal periods, Gonen et al. (2020) included shorter temporal periods in their experiments. They compared the difference in senses between two corpora by comparing the neighbours of the same lemma in both corpora. Lemmas that have a large difference in neighbours are labelled as having different usages. They evaluated their approach on corpora of tweets that were split on one of the following characteristics: gender, age, occupation, or if the tweet was posted on the weekday/weekend.

3. Dataset

In this section, we discuss the dataset that was used for our experiments.

3.1. Dataset Overview

To study temporal-aware WSD models, it requires sense-annotated text with timestamps. Furthermore, we wanted to evaluate the performance of WSD models with respect to individual authors. Therefore, the dataset that was proposed by King and Cook (2021) is ideal. This dataset contains 1586 sense annotated English blog samples. Each sample contains one sense annotated noun and the sample is associated with an author, and the date that the blog was posted. They selected blog samples that contain at least one lemma from a list of 11 nouns, which resulted in obtaining text from 36 authors. They annotated each sample with

WordNet senses (Miller, 1995) through a crowdsourcing website (Amazon Mechanical Turk).

3.2. Temporal Types

In this work, we explore the benefits of including temporal knowledge in WSD models, therefore, we first analyse the dataset with respect to such information. Each blog sample corresponds to a posting date, which details the day and year that it was posted. We use the posted date to calculate different temporal types (day of month, day of week, month, year, season). For example, the date *December 25, 2003* is a Thursday (day of week) and in the Winter (season) of 2003 (year). The distributions of samples for each temporal type is shown in Figure 1. The number of samples in the different weekday groups ranged from 176 (Sunday) to 316 (Monday), the number of samples in the different season groups ranged from 280 (Fall) to 472 (Summer), and the number of samples in the different month groups ranges from 77 (October) to 237 (July). Regarding the years, the majority of the samples were posted in 2003 and 2004.

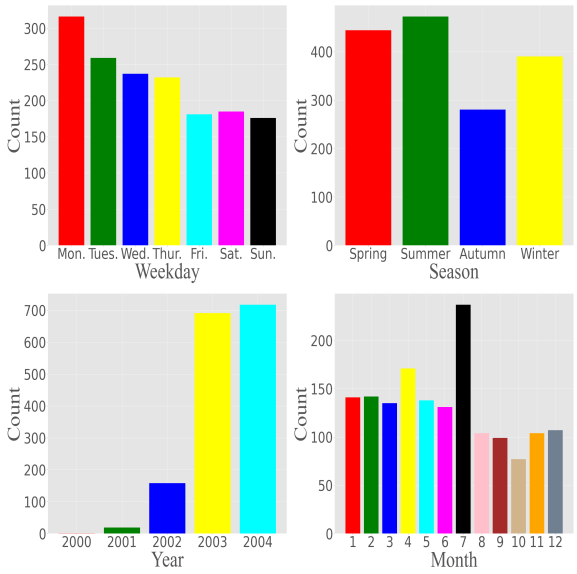

Figure 1: Sample distribution for each temporal type.

3.3. Predominant Senses

Gella et al. (2014) and King and Cook (2021) both showed that different authors may have different predominant senses for the same lemma. In this work, we explore the possibility that different temporal groups may also have different predominant senses for the same lemma. We refer to a group as a value within a temporal type. For example, *Thursday* is a temporal group for the temporal type *day of the week*. Figure 2a shows the proportion of predominant senses of the lemma *case* on different weekdays, which shows that within the seven days, there were four different predominant senses. Furthermore, only the Thursday group and Monday group share the same predominant sense with the whole dataset's predominant sense. Similarly, Figure 2b shows that for the lemma *sign*, only the Winter and Summer groups share the same predominant sense with the whole dataset's predominant sense.

To further analyze the predominant sense of temporal types, we show the percentage of samples in each data group having a different predominant sense than the dataset's predominant sense. Table 1 shows how much each type — including the author as a type — differs from the dataset's predominant sense. Higher values indicate that our temporal-aware models could have a larger potential impact on performance. Table 1 shows that individual authors are more likely to deviate from the dataset's predominant sense than temporal types. We found that some lemmas have values close to 0 for some temporal types (form/season), which indicates that using the predominant senses of these temporal groups may not achieve better performance over using the dataset's predominant sense. We also analyze the percentage of groups that differ from the dataset's predominant sense in Table 2.

Lemma	Author	W_day	Month	Season	Year
paper	0.594	0.641	0.469	0.271	0.557
position	0.568	0.000	0.182	0.000	0.000
sign	0.528	0.239	0.442	0.405	0.620
form	0.083	0.000	0.064	0.000	0.000
case	0.422	0.688	0.474	0.351	0.487
degree	0.521	0.527	0.685	1.000	0.370
track	0.479	0.336	0.363	0.288	0.075
deal	0.550	0.000	0.379	0.000	0.186
field	0.512	0.372	0.372	0.198	0.116
rule	0.414	0.273	0.212	0.263	0.000
charge	0.312	0.118	0.118	0.000	0.000
Mean	0.453	0.290	0.342	0.252	0.219

Table 1: Percentage of samples in the data groups which have predominant senses differing from the dataset predominant sense. *W_day* indicates day of the week.

3.4. Sample Window

The final temporal type that we explore is a more dynamic type, which we call a sample window. In this type, we consider the senses used in a window of time preceding a selected blog sample. We found that 42.1% of the time, the author used the same sense twice in a row for the same lemma.

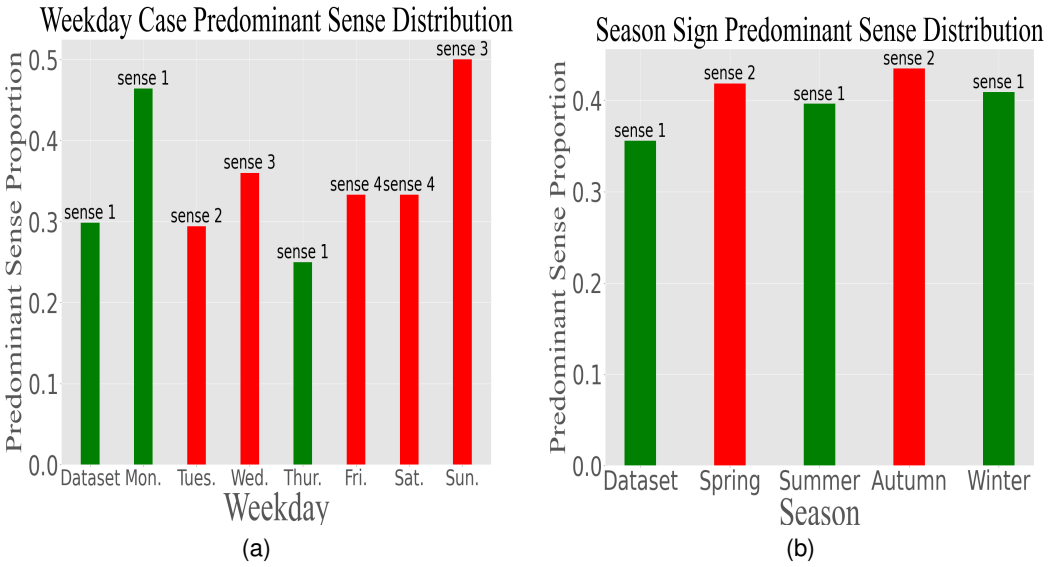

Figure 2: Proportion of predominant senses for the different weekdays (a) and seasons (b) of the lemmas *case* (a) and *sign* (b). Green bar indicates that the predominant sense of the group is the same as the dataset's predominant sense.

Lemma	Author	W_day	Month	Season	Year
paper	0.600	0.714	0.500	0.250	0.500
position	0.600	0.000	0.250	0.000	0.000
sign	0.500	0.286	0.417	0.500	0.750
form	0.100	0.000	0.083	0.000	0.000
case	0.400	0.714	0.583	0.500	0.333
degree	0.600	0.571	0.583	1.000	0.600
track	0.500	0.429	0.333	0.250	0.333
deal	0.500	0.000	0.417	0.000	0.333
field	0.600	0.429	0.417	0.250	0.333
rule	0.400	0.429	0.250	0.250	0.000
charge	0.400	0.286	0.167	0.000	0.000
Average	0.473	0.351	0.364	0.273	0.289

Table 2: Percentage of groups that differ from the dataset predominant sense. *W_day* indicates day of the week.

4. Methods

In this section, we describe our WSD models. Each model will predict the sense of each target token from the dataset.

4.1. Baseline Methods

For our baseline models, we consider two different WSD models: the original SensEmBERT model, and an all-in-one predominant sense model.

4.1.1. SensEmBERT - SEBERT

Our first baseline model uses SenseEmBERT[1] as it was originally proposed (Scarlini et al., 2020a). SensEmBERT provides sense embeddings of synsets in BabelNet by concatenating the embedding of Wikipedia articles about a target token with the embedding of text about the token in BableNet. BERT (Devlin et al., 2019) is used to generate embeddings. The concatenated embeddings are averaged to get the embedding of each synset/sense. Given a token in context, we embed the token using BERT and measure the cosine similarity between the token's embedding and the candidate SensEmBERT sense embeddings. This model will assign the sense with highest cosine similarity. We refer to this model as *SEBERT*.

4.1.2. Predominant Sense - PREDOM

This model always predicts the predominant sense of each data group. We consider the author's predominant sense and the dataset's predominant sense along with our temporal types — collectively we refer to them as data groups. We refer to this model as *PREDOM*.

4.2. SensEmBERT Sense Distribution - SEBERT_SENSE_DISTRI

Following King and Cook (2021), we assume knowledge of sense distributions of a given group and

[1]SensEmBERT is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License.

use their model's design to combine the sense distribution with the methods used in SensEmBERT. This model uses cosine similarity of the BERT embedded target token with SenseEmBERT's sense embeddings to rank the candidate senses. The inverse of each sense's rank is multiplied by the probability of that sense occurring for some data group. The detail is shown in the Equation 1. The difference between our model and King and Cook (2021) is that we consider temporal knowledge. We refer to this model as *SEBERT_SENSE_DISTRI*.

$$score(sense) = p(sense|data\ group) * \frac{1}{sense\ rank} \quad (1)$$

4.3. SensEmBERT Predominant Sense Rank - SEBERT_PREDOM_RANK

For this model, we consider the predominant sense model used by King and Cook (2021). This model assigns the sense with the highest cosine similarity according to SenseEmBERT, unless the predominant sense is in the top *k* ranked senses. This model will assign the predominant sense when the predominant sense is in the top *k* ranked senses. In this model, *k* is a hyperparameter that is tuned. Again, the difference between our model and King and Cook (2021) is that we consider temporal knowledge. We refer to this model as *SEBERT_PREDOM_RANK*.

4.4. SensEmBERT Predominant Sense Cosine Difference - SEBERT_PREDOM_SIM

The previous ranked-based model loses information by only considering the rank of the predicted sense and not the similarity score. In this model, we use the similarity score to assist our sense classification by comparing the difference of SenseEmBERT's highest cosine similarity and the cosine similarity of the predominant sense. Specifically, assume the highest cosine similarity among all candidate senses according to SensEmBERT is *c1* and the cosine similarity of the predominant sense is *c2*. The difference for this sample is $d = |c1 - c2|$. If d is smaller than some threshold t, our model will predict the predominant sense, otherwise, the sense with the highest similarity is assigned. Threshold t is a hyperparameter that is tuned. We refer to this model as *SEBERT_PREDOM_SIM*.

4.5. SensEmBERT Double Predominant Sense Model

In this model, we explore the potential performance gained by considering both the predominant sense of a temporal group and the predominant sense of a specific author. Similar to previous methods, we consider both the rank-based method and the cosine similarity-based method. The model will assign the sense with with highest similarity according to SensEmBERT if both of the two predominant senses (temporal/author) can not satisfy the rank condition or cosine difference condition that we mentioned in earlier methods. Otherwise, the model will assign the predominant sense with the better rank or the smaller similarity difference. We refer to these models as *SEBERT_DPR* and *SEBERT_DPS*, which extend the models *SEBERT_PREDOM_RANK* and *SEBERT_PREDOM_SIM*, respectively.

4.6. SensEmBERT Sample Window Model

In this model, we consider the sample window that was discussed in Section 3.4. This model uses only the predominant sense of the temporally nearest samples of the target sample to support its prediction. The sample window incorporates a windows size, which is adjusted to consider more samples that were posted prior to this sample. We only consider samples that contain annotated instances of the target lemma. The predominant sense of the samples in the window is taken as the input of *SEBERT_PREDOM_SIM* to make the prediction. We refer to this model as *SEBERT_SAMPLE_WIN*.

5. Experimental Results

In this section, we measure the performance of our WSD models on the given dataset by calculating the mean accuracy. We evaluate our models using different types of temporal and author-based predominant senses. The naming of each model includes the method prefixed with the type of knowledge that model is using. For example, *Month SEBERT_PREDOM_RANK* uses the *SEBERT_PREDOM_RANK* model while considering the *month* temporal type, and *Author+Season SEBERT_PREDOM_SIM* uses the *SEBERT_PREDOM_SIM* model while considering the *author* and *season* types.

5.1. Tuning *SEBERT_PREDOM_RANK*

SEBERT_PREDOM_RANK requires tuning of the hyperparameter *k*, which represents the rank threshold. Figure 3 shows the accuracy of each model for each value of *k*. We explored sense rank thresholds from 2 to 8 and see that all models perform as well or better than the SensEmBERT baseline with values for k ranging from 2 to 5. We summarized the highest accuracy of each rank-based model with its corresponding value for *k* in Table 3.

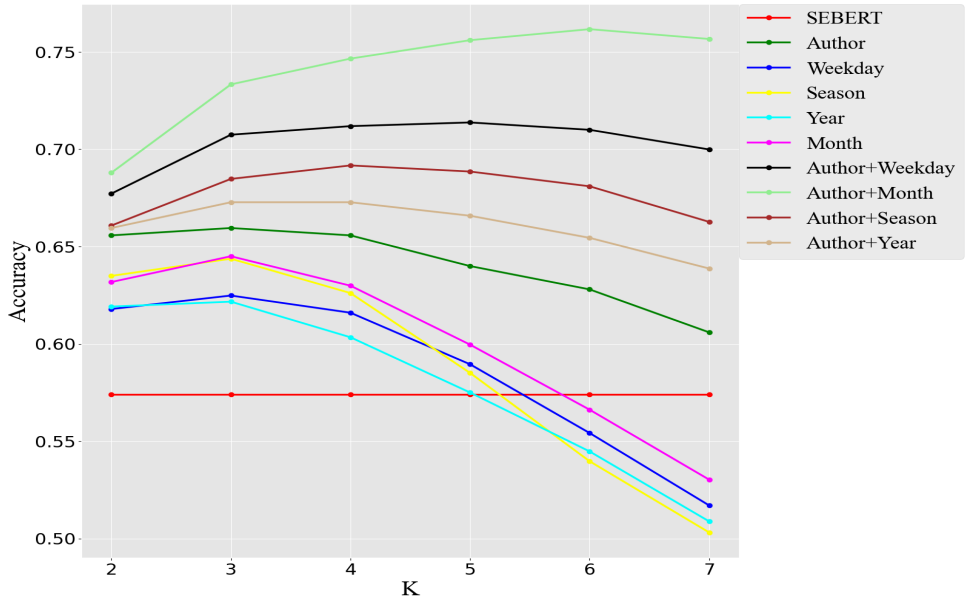

Figure 3: Accuracy for each *SEBERT_PREDOM_RANK* model.

SEBERT_PREDOM_RANK	Accuracy	*k*
Year	0.622	3
Weekday	0.625	3
Season	0.644	3
Month	0.645	3
Author	0.66	3
Author+Year	0.673	3
Author+Season	0.692	4
Author+Weekday	0.714	5
Author+Month	**0.762**	6
SEBERT	0.574	–

Table 3: The highest accuracy of each *SEBERT_PREDOM_RANK* and its corresponding optimal value for *k*.

SEBERT_PREDOM_SIM	Accuracy	*t*
Year	0.625	0.05
Weekday	0.644	0.05
Month	0.645	0.05
Season	0.648	0.05
Author	0.682	0.06
Author+Year	0.704	0.06
Author+Season	0.724	0.06
Author+Weekday	0.754	0.07
Author+Month	**0.799**	0.06
SEBERT	0.574	—

Table 4: The highest accuracy of each model and its corresponding optimal value for *t*.

5.2. Tuning *SEBERT_PREDOM_SIM*

We tune the value for the hyperparameter *t* for *SEBERT_PREDOM_SIM* models on a range from 0.01 to 0.08. Figure 4 shows the accuracy of each model for each parameter *t*. All models outperform the SensEmBERT baseline under all values of *t*. The highest accuracy and its corresponding parameter *t* for each model are shown in Table 4.

5.3. Analysis

Due to the non-uniform distributions of our dataset, accuracy across authors is applied as an important metric. In this section, we evaluate models using accuracy across authors and accuracy across instances.

5.3.1. Performance of Temporal Models

We first compare the performance of models that use our different temporal types (season, month, and weekday). Table 5 shows the accuracy of *PREDOM*, *SEBERT_SENSE_DISTRI*, and *SEBERT_PREDOM* for each temporal type across instances and authors — *SEBERT_PREDOM* represents *SEBERT_PREDOM_RANK* or *SEBERT_PREDOM_SIM*. In terms of accuracy, all *SEBERT_PREDOM* models outperform *PREDOM* models using the same predominant sense and all *SEBERT_SENSE_DISTRI* models outperform the corresponding *SEBERT_PREDOM* models. What's more, all *SEBERT_SENSE_DISTRI* models and *SEBERT_PREDOM* models outperform *SEBERT*. This suggests that the inclusion of author level and temporal level predominant senses has improved the performance of the models. In addition, *temporal SEBERT_PREDOM* models generally outperform *dataset SEBERT_PREDOM* mod-

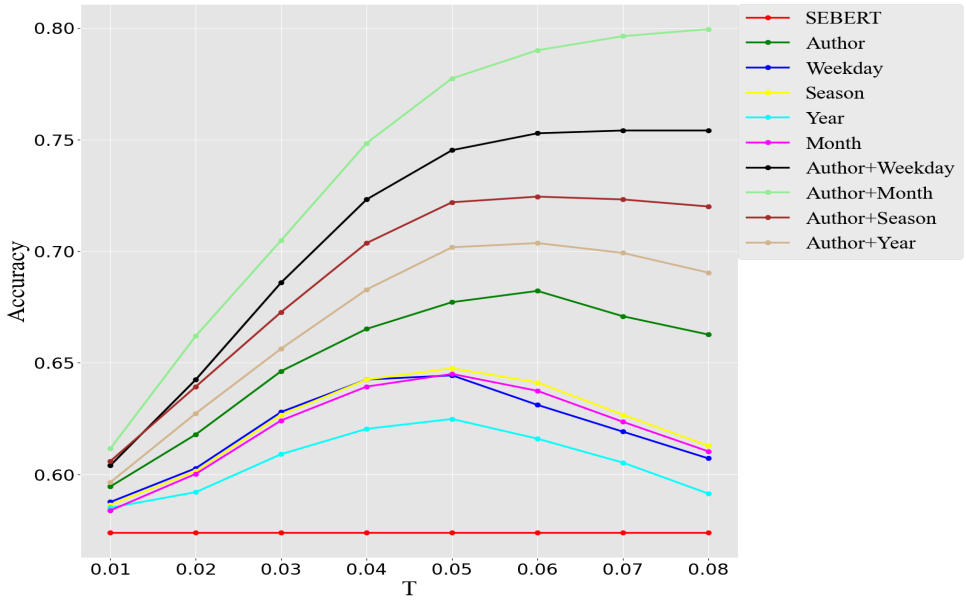

Figure 4: Mean accuracy for each *SEBERT_PREDOM_SIM* model.

els across instances, but only slightly across authors. The personalized author-aware models' performances are better than all temporal models, which indicates the potential gains of acquiring author-level sense knowledge. Most of the *SEBERT_PREDOM_SIM* models are better than the *SEBERT_PREDOM_RANK* models.

To show the fairness of temporal-based models, we measure the performance of the models that consider *month* on each author in Figure 5. The majority of authors with below-average accuracy (less than 55%) received a better performance with a month-aware model than a temporal agnostic model (SEBERT). Likewise, we show the performance for each author using personalized models in Figure 6.

Figure 5: Author-level performance of temporal models using month-level knowledge and *SEBERT*. Authors are in ascending order with respect to the performance of *SEBERT*.

Figure 6: Author-level performance of personalized models using author-level knowledge and *SEBERT*. Authors are in ascending order with respect to the performance of *SEBERT*.

5.3.2. Performance of Personalized Temporal Models

For the personalized temporal models, the two best-performing temporal types are applied in our analysis (Author+Weekday, Author+Month). The accuracy of each personalized temporal model and baselines are shown in Table 6. Similar to temporal models, *SEBERT_PREDOM* outperforms *PREDOM* using the same predominant sense and all *SEBERT_SENSE_DISTRI* outperform the corresponding *SEBERT_PREDOM*. Among these two kinds of author+temporal type models, the Author+Month models achieve better performance.

Figure 7 shows the performance of Author+Month models on individual authors. This

Method	Instance	Author
Author		
PREDOM	0.552	0.569
SEBERT_SENSE_DISTRI	**0.718**	**0.741**
SEBERT_PREDOM_RANK	0.66	0.677
SEBERT_PREDOM_SIM	0.682	0.706
Season		
PREDOM	0.4	0.4
SEBERT_SENSE_DISTRI	**0.67**	**0.668**
SEBERT_PREDOM_RANK	0.644	0.662
SEBERT_PREDOM_SIM	0.648	0.667
Month		
PREDOM	0.441	0.445
SEBERT_SENSE_DISTRI	**0.706**	**0.7**
SEBERT_PREDOM_RANK	0.645	0.667
SEBERT_PREDOM_SIM	0.653	0.669
Weekday		
PREDOM	0.423	0.432
SEBERT_SENSE_DISTRI	**0.693**	**0.709**
SEBERT_PREDOM_RANK	0.625	0.643
SEBERT_PREDOM_SIM	0.644	0.669
Dataset		
PREDOM	0.384	0.396
SEBERT_SENSE_DISTRI	**0.660**	0.664
SEBERT_PREDOM_RANK	0.632	**0.667**
SEBERT_PREDOM_SIM	0.628	0.666
SEBERT	0.574	0.611

Table 5: Accuracy of temporal models averaged across instances and authors. Models are grouped by their prefix (*Author, Season, Month, weekday, Dataset*, which indicates the type of knowledge that the model is leveraging.)

Method	Instance	Author
Author+Month		
PREDOM	0.749	0.771
SEBERT_SENSE_DISTRI	**0.873**	**0.894**
SEBERT_PREDOM_RANK	0.762	0.777
SEBERT_PREDOM_SIM	0.799	0.808
Author+Weekday		
PREDOM	0.682	0.708
SEBERT_SENSE_DISTRI	**0.835**	**0.855**
SEBERT_PREDOM_RANK	0.714	0.745
SEBERT_PREDOM_SIM	0.754	0.786
Author		
PREDOM	0.552	0.569
SEBERT_SENSE_DISTRI	**0.718**	**0.741**
SEBERT_PREDOM_RANK	0.660	0.677
SEBERT_PREDOM_SIM	0.682	0.706
SEBERT	0.574	0.611

Table 6: Accuracy of personalized temporal models across instances and authors. Models are grouped by their prefix (*Author+Month, Author+Weekday, Author*, which indicates the type of knowledge that the model is leveraging.

model performs better than *SEBERT* for all authors up to where *SEBERT* achieves approximately 70%. Although, even for authors whose receive an accuracy higher than 70%, the majority of the authors benefit from the Author+Month models.)

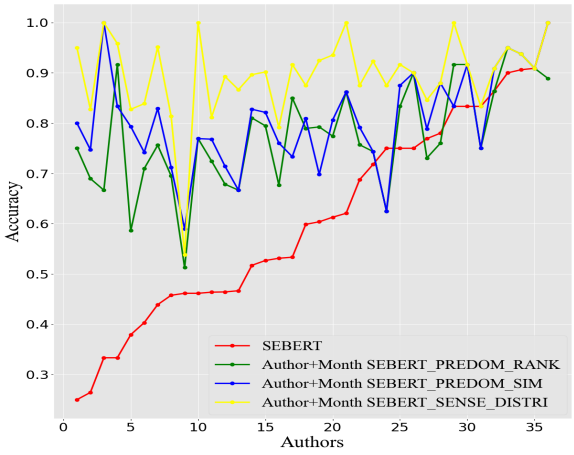

Figure 7: Author-level performance of Author+Month models and *SEBERT*. Authors are in ascending order with respect to the performance of *SEBERT*.

5.4. Performance of Personalized Models with Double Predominant Sense

In this section we compare the performance of double predominant sense personalized models, including *SEBERT_DPR* and *SEBERT_DPS*. Similarly, for *SEBERT_DPR*, we tuned parameter *k* on a range from 1 to 7, and for *SEBERT_DPS*, we tuned parameter *t* on a range from 0.01 to 0.08. Table 7 shows the optimal parameter of each double predominant sense model. We find that the accuracy of *SEBERT_DOUBLE_PREDOM* is somewhere between that of temporal *SEBERT_PREDOM* and author+temporal type *SEBERT_PREDOM*. Compared with both author *SEBERT_PREDOM* and temporal *SEBERT_PREDOM*, *SEBERT_DOUBLE_PREDOM* performs better.

We choose the two best performing data group types (Author&Weekday and Author&Month) and use them for this method. The accuracies of Author&Weekday and Author&Month *SEBERT_DOUBLE_PREDOM* across instances and authors are shown in Table 8. We find that *SEBERT_DPS* outperforms *SEBERT_DPR* by about 0.02 in terms of accuracy. Furthermore, *SEBERT_DPS* achieved far better performance than *PREDOM* and *SEBERT*.

SEBERT_DPR	Accuracy	*k*
Author&Weekday	**0.69**	4
Author&Season	0.672	3
Author&Year	0.668	3
Author&Month	0.677	3
SEBERT_DPS	Accuracy	*t*
Author&Weekday	**0.711**	0.06
Author&Season	0.694	0.06
Author&Year	0.692	0.06
Author&Month	0.694	0.06

Table 7: The highest accuracy of each double predominant sense model using the *Rank* method *SEBERT_DPR* and the *Similarity* method *SEBERT_DPS* with their optimal parameter *k*.

Method	Instance	Author
Month *PREDOM*	0.441	0.445
Weekday *PREDOM*	0.423	0.432
Author *PREDOM*	0.552	0.569
SEBERT	0.574	0.611
Author&Month *SEBERT_DPR*	0.677	0.689
Author&Month *SEBERT_DPS*	0.697	0.711
Author&Weekday *SEBERT_DPR*	0.69	0.713
Author&Weekday *SEBERT_DPS*	**0.711**	**0.735**

Table 8: Mean accuracy across instances and authors for *PREDOM*, *SEBERT*, *SEBERT_DPR*, and *SEBERT_DPS*.

5.5. Performance of *SEBERT_Sample_WIN*

For *SEBERT_Sample_WIN*, we explored different window sizes from 1 to 5 and the threshold t from 0.03 to 0.06. Table 9 shows the accuracies for *SEBERT_Sample_WIN* across instances for each sample window and *t*. Although *SEBERT_Sample_WIN* is superior to *SEBERT*, it is still not comparable to author *SEBERT_PREDOM_SIM*.

w \ *t*	0.03	0.04	0.05	0.06
1	0.61	0.622	0.626	0.614
2	0.61	0.622	0.626	0.614
3	0.62	0.628	0.629	0.62
4	0.625	0.637	**0.645**	0.64
5	0.627	0.636	0.637	0.631

Table 9: Performance of *SEBERT_Sample_WIN* against *w* (window size) and *t*.

6. Conclusion

In this work, we found that short timeframe temporal-aware WSD models outperform temporal-agnostic models, while only requiring a relatively small amount of text. We considered the day of the week, month, and season when the text was posted to assist with labelling a token with its proper sense. We found that personalized models that considered author-level sense information outperformed temporal-aware models, but we were able to achieve our best performances when considering both author and temporal sense-based information. Author-level sense information could be difficult to obtain or estimate due to author-specific text being a relatively-low resource, but obtaining temporal-based sense information could be more obtainable because we can gather text from multiple authors for a specific temporal group. We also considered a sample window approach that looked at how the target lemma was recently used, which outperformed a temporal-agnostic model. We proposed models to incorporate knowledge of predominant senses and sense distributions, which includes models that can incorporate both the author-based sense information and temporal-based sense information. This work assumes knowledge of either the predominant senses or sense distributions of lemmas, therefore, a natural step in future work would be to explore models that estimate the predominant sense of a temporal group or an individual author. Although we focused on SensEmBERT to show that temporal-based information can be leveraged to better performance of an embedding-based WSD model, we believe that the same tailoring techniques can be applied to models such as ARES (Scarlini et al., 2020b) and LMMS (Loureiro and Jorge, 2019), which we will consider in future work. Furthermore, we plan on leveraging different representations of the timestamps, such as the embedding model from Goyal and Durrett (2019), to build better temporal-aware WSD models.

7. Ethical Considerations

Our proposed models have only been tested on English text. Due to the limited information available from the dataset, we aren't able to speak on the dialect of the English. We hope that by considering the idiolect of each author that we mitigate the potential bias toward a dialect, but we are unable to evaluate that hypothesis with this dataset. We don't have any indication of the performance of our models for authors whose age is outside the range of 17 years to 48 years, since that is the range of ages for the authors in the dataset. The dataset contains 10 authors that are female and 26 authors that are male, which we don't consider in our evaluation.

8. References

Satanjeev Banerjee, Ted Pedersen, et al. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Ijcai*, volume 3, pages 805–810.

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. 2014. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1591–1600, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. When time makes sense: A historically-aware approach to targeted sense disambiguation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2751–2761, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss-informed biencoders. *arXiv preprint arXiv:2005.02590*.

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.

Spandana Gella, Paul Cook, and Timothy Baldwin. 2014. One sense per tweeter ... and other lexical semantic tales of Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 215–220, Gothenburg, Sweden.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2019. Embedding time expressions for deep temporal ordering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4400–4406, Florence, Italy. Association for Computational Linguistics.

Milton King and Paul Cook. 2021. Now, it's personal : The need for personalized word sense disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 692–700, Held Online. INCOMA Ltd.

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.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *ACL (1)*, pages 259–270. Citeseer.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*, pages 24–26, Toronto, Canada.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A White, Gabriel Wong, Luis Espinosa Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. Tempowic: An evaluation benchmark for detecting meaning shift in social media. *arXiv preprint arXiv:2209.07216*.

Daniel Loureiro and Alípio Jorge. 2019. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.

Binny Mathew, Suman Kalyan Maity, Pratip Sarkar, Animesh Mukherjee, and Pawan Goyal. 2017. Adapting predominant and novel sense discovery algorithms for identifying corpus-specific sense differences. In *Proceedings*

of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing, pages 11–20. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal1. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, MD.

Scott Piao, Fraser Dallachy, Alistair Baron, Jane Demmen, Steve Wattam, Philip Durkin, James McCracken, Paul Rayson, and Marc Alexander. 2017. A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech & Language*, 46:113–135.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020a. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8758–8765.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020b. With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.