An Evaluation Method for Diachronic Word Sense Induction

Ashjan Alsulaimani, Erwan Moreau, Carl Vogel

An Evaluation Method for Diachronic Word Sense Induction

Ashjan Alsulaimani
School of Computer Science and Statistics & Trinity Centre for Computing and Language Studies
Trinity College Dublin
alsulaia@tcd.ie

Erwan Moreau
School of Computer Science and Statistics & Adapt Centre
Trinity College Dublin
moreaue@scss.tcd.ie

Carl Vogel
School of Computer Science and Statistics & Trinity Centre for Computing and Language Studies
Trinity College Dublin
vogel@scss.tcd.ie

Abstract

The task of Diachronic Word Sense Induction (DWSI) aims to identify the meaning of words from their context, taking the temporal dimension into account. In this paper we propose an evaluation method based on large-scale time-stamped annotated biomedical data, and a range of evaluation measures suited to the task. The approach is applied to two recent DWSI systems, thus demonstrating its relevance and providing an in-depth analysis of the models.

1 Introduction

Words naturally evolve through time, their meaning may encounter subtle or radical changes resulting in a variety of senses. For example, the word `mouse` only had the meaning of *animal* until it acquired a brand new sense in 1980 as *computer device*. But sense changes are not always so definite, a word usage may drift progressively from its original sense or be affected by historical events. A recent example of this phenomenon is the word `coronavirus`, which has seen a dramatic usage surge in 2020 because of the emergence of its *SARS-CoV-2* variant. Before 2020, the word `coronavirus` was mostly a technical term describing a family of viruses, but it is now used in the mainstream media to mean the specific *SARS-CoV-2*, the related Covid19 disease or even the general health crisis and its consequences.

The dynamic behaviour of words contributes to semantic ambiguity, which is a challenge in many NLP tasks. The ability to detect such changes across time could potentially benefit various applications, such as machine translation and information retrieval. In the biomedical domain, it can improve the quality of the automatic identification of senses in contexts where no complete terminology is available, such as with clinical notes, and to assist indexers who build terminology resources.

Recent research focused on detecting semantic shifts across time (Kutuzov et al., 2018) but also Diachronic Word Sense Induction (Emms and Kumar Jayapal, 2016). The task of Diachronic Word Sense Induction (DWSI) is similar to Word Sense Induction (WSI) in identifying the meaning of words from their context, but also takes the temporal dimension into account.

In §2 we briefly present two Bayesian models that have been proposed for the DWSI task: Emms and Kumar Jayapal (2016) proposed a model which represents the evolution of word senses in order to detect the emergence year of new senses. A different model was proposed by Frermann and Lapata (2016), focusing instead on capturing the subtle meaning changes within a sense over time. However evaluating such models is difficult, as the lack of large scale time-stamped data prevents direct quantitative evaluation.

In this paper we introduce a method which relies on annotated biomedical data to evaluate DWSI.[1] While the general aim of this article is the evaluation of DWSI systems across domains and genres, the biomedical domain is the only one to date which offers suitable data for the task. Our approach leverages the availability of unambiguous manual annotations (and publication years) in the Medline citation database in order to build a large time-stamped dataset, as detailed in §3. In §4 we introduce a range of evaluation measures which can be used to directly and quantitatively measure the performance of a DWSI system on such an annotated dataset. Finally in §5 we compare the two aforementioned models using our evaluation method, which demonstrates the relevance of the approach and allows a deep analysis of the models.

[1] The code is available at https://github.com/AshjanAlsulaimani/DWSI-eval.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3171–3180

2 State of the Art

2.1 Diachronic Word Sense Induction

Most existing work on diachronic meaning change has focused on static methods, in the sense that the learning algorithms are either time-unaware or applied to independent periods of time (Lau et al., 2012; Cook et al., 2014; Mitra et al., 2015). For example, Mitra et al. (2015) split the data into eras and then apply WSI independently on each era subset in order to identify new senses of a word. However, recent approaches have introduced time aware probabilistic models in order to represent the changes in word meaning over time.

2.2 The NEO Model

The model introduced by Emms and Kumar Jayapal (2016), called NEO[2] herein, is a generative Bayesian model that chooses a sense s given a time t (respecting relevant sense-given-time probabilities $P(s|t)$) then chooses context words $\mathbf{w}$ given the sense s (respecting relevant word-given sense probabilities $P(w|s)$). The joint probability distribution over the parameters is defined as in (1).

$$\begin{aligned} &P(t, s, \mathbf{w}; \pi_{1:N}, \theta_{1:K}) \\ &= \prod_t Dirich(\theta_t; \gamma_\pi) \times \prod_k Dirich(\theta_k; \gamma_\theta) \\ &\times P(t; \tau_{1:N}) P(s|t; \pi_{1:N}) \prod_{w_i \in \mathbf{w}} P(w_i|s; \theta_{1:K}) \end{aligned} \tag{1}$$

The authors' aim is to capture sense changes in order to detect the emergence, i.e. origin time, of a novel sense. In this model the probabilities of the context words are represented independently from time, which means that senses can change over time with respect to each other, but the probabilities of the words representing a particular sense are assumed to be constant.

2.3 The SCAN Model

Frermann and Lapata (2016) proposed a generative Bayesian model inspired from dynamic topic modeling (Blei and Lafferty, 2006), hereafter called SCAN, which shares similarities with NEO but is more complex: given a time t, a sense s is chosen following the distribution of the parameter ϕ_t; then given a sense s and a time t, the context words $\mathbf{w}$ are drawn following the distribution of the parameter $\psi_{s,t}$. This design allows the representation of a sense with a different distribution of words at different times, as opposed to NEO. Thus in the SCAN model, time-adjacent representations of a sense are codependent in order to allow capturing the meaning change in a smooth and gradual way. This is made possible by defining their prior as an intrinsic Gaussian Markov Random Field. Following the structural dependencies defined through iGMRF prior, Frermann (2017) expresses the posterior distribution over the latent variables given the input $\mathbf{w}$, parameters $a, b, \kappa^{\mathbf{\Psi}}$ and the choices of the distributions Gamma (Ga), Logistic Normal distribution (N):

$$\begin{aligned} &P(s, \mathbf{\Phi}, \mathbf{\Psi}, \kappa^{\mathbf{\Phi}} | \mathbf{w}, \kappa^{\mathbf{\Psi}}, a, b) \\ &\propto Ga(\kappa^{\mathbf{\Phi}}; a, b) \prod_t \Big[\prod_k \big[N(\mathbf{\Psi}^{t,k} | \kappa^{\mathbf{\Psi}}) \big] \Big] \prod_d \big[\mathbf{\Phi}_s^t \prod_{w^i \in \mathbf{w}} \mathbf{\Psi}_{w^i}^{s,t} \big] \end{aligned} \tag{2}$$

where $\kappa^{\mathbf{\Phi}}$ is drawn from a conjugate Gamma prior and $\kappa^{\mathbf{\Psi}}$ is estimated during inference, which both control the degree of sense-specific word distributions variations over time. Thus the SCAN model is meant to capture changes between senses but also changes of meaning within a sense.

2.4 Existing Evaluation Methods

One way to find the ground truth of sense emergence is by using a dictionary. This approach is taken by many studies (Rohrdantz et al., 2011; Lau et al., 2012; Cook et al., 2014; Mitra et al., 2015).

In (Emms and Kumar Jayapal, 2016), the model is evaluated qualitatively on the Google NGrams corpus (Michel et al., 2011), using a few manually selected target words. The ground truth is obtained by the "tracks-plot" method, which consists in representing a target sense by a few hand-picked co-occurrences (e.g. "screen", "click" for `mouse` as a *computing device*), then tracking these co-occurrences over time and taking the mean of the separate tracks. An emergence detection algorithm "EmergeTime" is proposed in (Jayapal, 2017) to detect the year of emergence either from the "tracks-plot" data (ground truth emergence) or a predicted distribution $P(s|t)$ (predicted emergence). The algorithm checks whether there is a year in the $P(s|t)$ plot which satisfies the following constraints:

- The year is followed by a 10 year window of sufficient increase in probabilities: 85% of the years show a climb in probabilities of 2-3% of the maximum value.
- 80% of the preceding years are lower than 0.1 (i.e. close to zero in probability).

[2]This abbreviation is not provided by the authors of the work. It is used here as a reference for the model.

Emms and Kumar Jayapal (2016) evaluate the quality of the sense clustering qualitatively by inspecting the top 30 ranked words that are associated with a specific sense.

Frermann and Lapata (2016) present four indirect evaluation methods, relying on closely related tasks used as applications of their model:

- "Temporal Dynamic": qualitative evaluation of the appearance of a new sense.
- "Novel Sense Detection": evaluation using Mitra et al. (2015)'s complex approach based on WordNet.[3]
- "Word Meaning Change": evaluation using Gulordava and Baroni (2011)'s method and data for detecting meaning change between two time slices.
- "Task-based Evaluation": extrinsic evaluation on the SemEval Diachronic Text Evaluation task (Popescu and Strapparava, 2015), designed for supervised learning.

Despite the authors's best efforts to compare their results against others, they state that the "scores [that they obtain] are not directly comparable due to the differences in training corpora, focus and reference times, and candidate words" (Frermann and Lapata, 2016, p.39). Additionally, models of both Emms and Kumar Jayapal (2016) and Frermann and Lapata (2016) offer a continuous time representation $P(s|t)$. The sophistication of their systems would deserve a more suitable evaluation framework, since they have to simplify their outcomes in order to compare them against previous works which rely on models which only represent independent time slices.

A recent evaluation framework is proposed by (Schlechtweg et al., 2020) for the task of Unsupervised Lexical Semantic Change Detection (LSC) in SemEval-2020. However, the benchmark datasets contain only two independent periods of time. The subtasks are only designed to capture whether there is a change (subtask 1) or the extent of a change (subtask 2). Precisely, as opposed to the DWSI task, the subtasks do not capture how many distinct senses exist in the data, what kind of change happens over time, to which sense, and the emergence year of a novel sense. Although the annotation process involves clustering senses and computing sense frequency distributions for two independent periods of time, the sense information is neglected. Instead, the target values of the subtasks are based on "change scores" which represent only the existence or degree of LSC. As a result of this simplification, the evaluation methods used in the Unsupervised LSC are incompatible with the WSI and DWSI tasks. The task differs from WSI and DWSI in the sense that it does not either provide a way to predict the sense of an instance or the set of senses of a polysemous target word and their prevalence.

3 A Biomedical Dataset for DWSI

The DWSI task requires not only target words with several senses, but also time-stamped data for every target word. The evaluation of DWSI is challenging because manual annotation of such a large amount of instances (since they span over many years) would be prohibitively costly.[4] In this section, we propose a method to collect diachronic data for ambiguous terms in medical terminologies.

3.1 Data Collection Process

Our method relies on the medical literature and exploits medical terminology resources: Medline[5] is a database referencing most of the biomedical literature (30 millions citations). The citations are annotated with Mesh descriptors. MeSH[6] (Medical Subject Headings) is "the US National Library of Medicine (NLM) controlled vocabulary thesaurus used for indexing articles for PubMed." The Unified Medical Language System (UMLS) Metathesaurus is "a large biomedical thesaurus that is organized by concept, or meaning, and it links similar names for the same concept" (Bodenreider, 2004).[7] Each concept in UMLS is identified by a Concept Unique Id (CUI), and all the terms listed in UMLS are assigned a CUI. Since UMLS includes MeSH terms, there is a partial mapping between MeSH descriptors and UMLS CUIs.

The MSH WSD data (Jimeno-Yepes et al., 2011) consists of 203 ambiguous medical terms, each provided with the list of CUIs which identify the different meanings of the term. This dataset was created for the Word Sense Disambiguation task,

[3] https://wordnet.princeton.edu/

[4] DWSI takes into account the progressive evolution of senses across time, as opposed to other works which consider only two specific points in time, e.g. (Schlechtweg et al., 2020). Thus we chose this biomedical dataset because it has the unique characteristic to contain a large amount of ambiguous instances which are (1) carefully annotated with senses and (2) time-stamped, spanning around 70 years. To our knowledge, there are other datasets which satisfy either condition (1) or (2), but none which satisfies both.

[5] https://www.nlm.nih.gov/bsd/pubmed.html

[6] https://www.ncbi.nlm.nih.gov/mesh

[7] https://www.nlm.nih.gov/research/umls

so the instances it contains are labelled by CUI (sense) but they are not time-stamped. We collect a time-stamped dataset as follows:

1. The MSH WSD data provides us with target terms and CUIS.
2. For every CUI, the corresponding MeSH descriptor is extracted from UMLS.
3. From Medline, all the citations labeled with a particular MeSH descriptor are extracted (title, publication year and abstract if any).
4. When available, the text of the full article is retrieved from PubMed Central.[8]

3.2 Data pre-processing

For every target and every sense (CUI), a collection of documents made of titles, abstracts and full articles is obtained. Every occurrence of the target term in a document is assumed to have the sense given by the CUI.[9] In the interest of maximising the number of instances available for each year, we also collect the full list of terms associated with the CUI from UMLS and substitute every occurence of such a term with the ambiguous target. In both cases of collecting instances, the longest possible term is matched in order to capture the most specific expressions.[10]

SpaCy[11] is used to tokenise the documents into sentences and words. Using a global stopword list based on the tokens frequencies, the most frequent tokens such as non-content words (the, a, however) and punctuation signs (!, %) are removed from the context. Every occurrence of the target in a document is extracted together with its 10-word context (5 words on each side). In order to provide the DWSI systems with sufficient data for every year, we only include the longest consecutive period with at least 4 instances every year across senses.

At the end of the process, the dataset contains 188 target (out of 203 initial targets).[12] 175 targets have two senses, 12 have 3 and one has 5 senses. There are 61,352 instances by sense in average.[13] 102 senses out of 391 have emergence according to the "EmergeTime" method.[14]

4 Evaluation

As explained in §3, the collected dataset contains sense labels which can be used to directly evaluate a DWSI system in a reliable way. Since by definition the ouput of an unsupervised clustering algorithm is unlabeled, we propose in §4.1 a method to match a gold sense with a predicted sense. Thanks to this matching method, a system can be evaluated externally, in a way similar to a supervised WSD system. We propose several evaluation methods, each meant to capture the performance of a DWSI system from a different perspective.

4.1 Global Maximum Matching Method

After estimating the model, the posterior probability is calculated for every instance, according to Eq. (3) for NEO and Eq. (4) for SCAN. The sense corresponding to the maximum probability is assigned to the instance.

$$P(S|t^d, \mathbf{w}^d) = \frac{P(S, t^d, \mathbf{w}^d)}{\sum_{S'} P(S'|t^d, \mathbf{w}^d)} \quad (3)$$

$$P(S|t^d, \mathbf{w}^d) \propto P(S^d|t^d)P(\mathbf{w}^d|t^d, S) \quad (4)$$

The pairs of gold/predicted senses are matched iteratively based on their joint frequency. At every iteration, the pair corresponding to the highest frequency (global maximum) in the table is matched. Once a gold sense is matched with a predicted sense, neither the gold nor the predicted sense can be matched again with another sense. This eliminates the possibility of having two different gold senses matched with the same predicted sense or two different predicted senses matched with the same gold sense, an issue present in the methods used by (Agirre and Soroa, 2007; Manandhar et al., 2010).[15] Moreover, by matching the largest senses first, the number of incorrectly matched instances is minimized. An example is provided in table 1.

4.2 Based on Clusters of Instances

4.2.1 Clustering Classification Measures

Given the true class (i.e. true sense, obtained as explained in §3) and the assigned predicted

[8] https://www.ncbi.nlm.nih.gov/pmc/

[9] This assumption might not be always satisfied, but the noise is likely to be negligible. There might also be a small number of MeSH annotations errors in Medline.

[10] We obtain 3,119,248 instances before substituting the associated terms and 13,791,570 after, that is roughly 4.5 times more instances (these values are only for abstracts, the proportion is probably similar with PMC articles).

[11] https://spacy.io/api/tokenizer

[12] 7 targets are not valid anymore due to UMLS updates that happened since the WSD data was created, 2 are filtered out due to insufficient data across years, and 5 are removed due to a technical incompatibility with one of the two systems tested.

[13] Minimum 8 and maximum 1.6m instances by sense; minimum 778 and maximum 1.7m instances by target.

[14] Details about the dataset and the EmergeTime algorithm are provided in Appendix A.2 and A.1 respectively.

[15] A detailed example is provided in Appendix A.3.

	C0030131	C0030625	C0078944	C0149576	C0429865
0	608	502	4680	352	5171
1	108	191	1963	466	17345
2	131	220	2139	484	16128
3	153	230	2684	637	**26222**
4	1313	1623	885	98	569

	C0030131	C0030625	C0078944	C0149576	C0429865
0	608	502	**4680**	352	-
1	108	191	1963	466	-
2	131	220	2139	484	-
3	-	-	-	-	-
4	1313	1623	885	98	-

Predicted sense	Gold sense
0	C0078944
1	C0030131
2	C0149576
3	C0429865
4	C0030625

Table 1: Global maximum matching example. The top contingency table shows the number of instances for every predicted/gold sense pair (the predicted sense is assigned by calculating the maximum of the posterior probability). At the first iteration, senses C0429865 and 3 are matched based on the global maximum (in bold). The second table shows the remaining frequencies at the second iteration. The bottom table shows the resulting matching at the end of the process.

class (obtained using the matching method presented in §4.1), every instance can be categorised as True/False Positive/Negative for any specific sense s, following the standard classification methodology. This way the standard binary classification measures can be applied at the level of a sense: precision, recall, F1-score. The micro-average and macro-average of these measures are calculated to represent the performance at the level of a target or across targets.

4.2.2 Clustering Mean Absolute Error

The classification measures do not distinguish whether the system is confident in its prediction (e.g. if the posterior probability is 0.99) or not (e.g. if it is 0.51), this is why we also propose to use the mean absolute error (MAE). The intuition behind this measure is that a perfect system should predict probability one for the gold sense and zero for any other sense. Therefore, the further the predicted probability deviates from one, the higher the error. We use the mean absolute error to measure how close to one is the posterior probability of the gold sense in average. The mean absolute error is defined for every sense as in Eq. (5).

$$\frac{1}{|D|}\sum_{d\in D}(1-P(\hat{s_g}|d)) \quad (5)$$

where D represents a set of instances, $\hat{s_g}$ is the sense that matches the gold sense, and the posteriors are defined as mentioned in Eq. (3) and (4). Since the individual error value is unique for a given instance, this measure can be calculated for any set of instances, in particular at the level of a single sense, a target or across the whole data. By contrast to the classification measures which assign a categorical label to an instance, this measure takes into account the potential numerical variations of the probability values. However at the level of a sense it does not capture any information about the false positive cases. As a consequence, classification measures and MAE are susceptible to show complementary aspects of performance.

4.3 Based on the Estimated Parameters

4.3.1 Emergence Classification Measures

Generally the task of emergence detection consists in predicting the year (or period of time) when a new sense emerges. As explained in §2.4, this task is performed by applying the emergence detection algorithm on the inferred $P(s|t)$ parameter. In theory the true answer is the emergence year, but in a classification setting it is reasonable to allow some margin of error. Thus the predictions of an emergence is counted as correct if it falls within the bounds of a 5 year window centered on the true emergence year. Based on this categorisation, the standard precision, recall and F1-score can be calculated across all targets.

4.3.2 Emergence Mean Absolute Error

The binary classification measures restrict the predicted answer to be either inside or outside a window, thus do not take into account the distance between the gold and predicted emergence years. By contrast, a numerical error value can be calculated as follows:

$$e = \begin{cases} 0 & \text{if } \neg g \wedge \neg p \\ M & \text{if } (\neg g \wedge p) \vee (g \wedge \neg p) \\ |y - \bar{y}| & \text{if } g \wedge p \end{cases}$$

where:

- g (resp. p) is true if and only if the gold (resp. predicted) sense has emergence,
- M is the maximum error defined as the number of years of data for a specific target,
- y is the true year of emergence and $\hat{y}$ is the predicted year of emergence.

In order to compare error levels across different targets, a normalised variant is defined as $e_{norm} =$

$\frac{e}{M}$. The MAE is defined over a set of senses S as the mean of their e_{norm} values.

The intuition is that the case where both the gold and the predicted senses have emergence should always be assigned a lower error than when only one of them has emergence, therefore we assign the maximum error in the latter case. Since all the targets do not have the same number of years of data, the maximum individual error is different among targets, this is why a normalised variant is used where the individual value is divided by the total number of years. This allows comparisons of the error level between senses, targets, as well as at the system level.

4.3.3 Time Series Distances

The predicted evolution across time of the sense probability $P(s|t)$ is an essential outcome of the DWSI task. We use distance measures in order to evaluate how far the predicted $P(s|t)$ is from the true probability across time. There are many options available for measuring the distance between two time series. We propose two of them:

- The linear Euclidean distance is a simple measure which assumes that the i^{th} point in one sequence is aligned with the exact i^{th} point in the other one.
- The non-linear Dynamic Time Warping (DTW) distance measure performs an alignment of the two sequences (Berndt and Clifford, 1994; Sardá-Espinosa, 2017). This allows a more flexible comparison of the dissimilarity with respect to the alignment of the two series across time.

The superiority of DTW over Euclidean measure is that DTW is tailored to time shifts, scale and noise and not only defined for series of equal length. In our task, we will compare both Euclidean and DTW results and test whether DTW finds local similarities between sequences which share some patterns but are not fully aligned.

5 Results and Analysis

In this section, we evaluate the NEO and SCAN systems using the dataset presented in §3 and the evaluation methods defined in §4. This allows us to compare the two systems on the same grounds. Additionally this rich annotated dataset allows us to provide an in-depth analysis, thus uncovering the strengths and weaknesses of the two systems.

The DWSI task is unsupervised, so the whole data is used both to estimate the parameters and perform evaluation on the predictions. No parameter has been tuned at any point: the experiments are run using the systems provided by the original authors with their default parameters, except for the number of senses (the true number of senses is used for evey target), one-year time interval, and the size of the context window (10).[16]

5.1 Observations of Posterior Distribution

The graphs in Figure 1 show the frequency of the predicted probabilities that correspond to the matched gold senses and the frequency of the highest predicted probabilities that are assigned for each instance. The predicted probabilities follow a U-shaped distribution, which means the system tends to assign extreme probabilities (close to either zero or one) to the majority of the data. The graphs also show the overlap between the predicted gold sense probabilities and the highest predicted probabilities, which represents the instances where the true sense was predicted correctly. By contrast, the area in red on the left half represents cases where the true sense is predicted with a low probability (false negative), and the blue area which does not overlap represents instances where an incorrect sense is predicted (false positive). In comparison to NEO, SCAN tends to assign even more extreme probabilities. In particular, SCAN tends to make more serious errors: in more than 5 millions cases, the predicted probability is 0 (or close to 0) for the gold sense instead of 1.

Table 2 compares the deciles of the error distribution between NEO and SCAN. For NEO, the error is below 0.1 (near perfect predictions) for more than 30% of the instances while it is above 0.9 (totally incorrect predictions) for slightly less than 20% of the instances. In contrast, SCAN scores correctly more than 40% of the instances while the incorrect predictions are more than 30%.

Overall, NEO performs better than SCAN according to the MAE: 0.425 vs. 0.444. This difference is significant (p-value 0.000024 for Wilcoxon signed rank test at the level of targets).

5.2 Influence of Data Size

It is often expected that performance improves with the amount of data provided. This is not verified in the data, which shows a slight negative correlation level (between -0.1 and -0.3) between data size and performance across targets in both systems.

[16] For details about the parameters, see Appendix A.1.

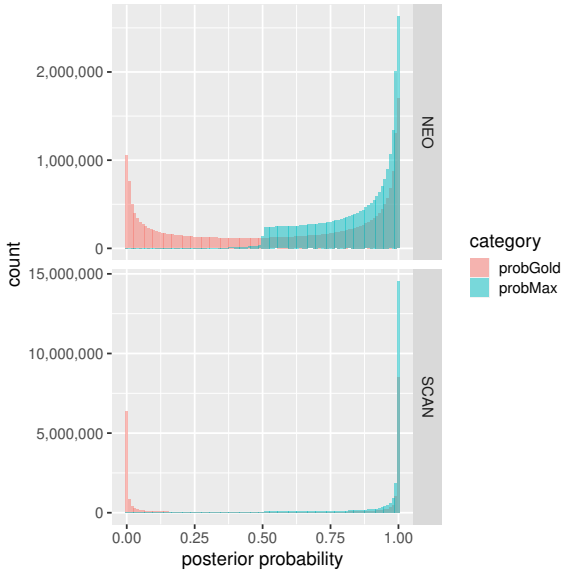

Figure 1: Distribution of the probabilities predicted by NEO and SCAN systems: the red distribution represents the predicted probability of the gold sense for every instance in the data; the blue distribution represents the highest predicted probability for every instance.

Bottom N %	decile (NEO)	decile (SCAN)
10%	0.009	0.0000002
20%	0.039	0.00003
30%	0.095	0.001
40%	0.189	0.016
50%	0.331	0.174
60%	0.518	0.774
70%	0.718	0.985
80%	0.880	0.999
90%	0.973	0.999

Table 2: Deciles for error values for the predicted senses (across all instances) based on the clustering mean absolute error evaluation measure for NEO and SCAN systems.

We investigate how the size of each sense (as opposed to the full target size) contributes to the performance of the model. In other words, we observe the difference between targets where the senses have a similar size and targets where there is a strong imbalance between the senses. For every target, the standard deviation of the sense size proportions is used as a measure of the imbalance across senses. Figure 2 shows the relationship between SD and macro F1-score. There is a clear pattern where higher imbalance between senses is associated with lower performance in general, regardless of the model type.

A detailed analysis shows that SCAN outperforms NEO when the imbalance level is not large between senses within a target, while the two systems perform similarly otherwise. This effect can be observed in the global classification results in table 3. SCAN outperforms NEO at the level of

Figure 2: Relation between gold sense imbalance and performance by target.
Pearson correlation: NEO -0.48, SCAN -0.52

Perf.	NEO			SCAN		
	P	R	F1	P	R	F1
macro	0.548	0.569	0.558	0.562	0.591	0.577
micro	0.595	0.595	0.595	0.558	0.558	0.558

Table 3: Global classification results for NEO and SCAN systems. P/R/F1: Precision/Recall/F1-score

macro results whereas NEO performs better at the level of micro results. However, Wilcoxon rank test shows that the superiority of SCAN at the level of macro F1-score by target is not significant (p-value: 0.354) whereas the superiority of NEO at the level of micro F1-score is (p-value: 1.167e-07). Given that macro scores are based on the average performance across senses independently from their size, this means that SCAN performs better than NEO with the minority class (i.e. sense) and conversely NEO shows better performance with the majority class. Table 4 confirms that the superiority of SCAN for the minority class is not significant yet the superiority of NEO for the majority class is.

Number of Senses	Sense rank	Mean F1-score		Wilcoxon test p-value
		NEO	SCAN	
-	first	0.299	0.321	6.657119e-01
-	last	0.732	0.692	3.503092e-10
2	first	0.315	0.335	6.920240e-01
2	second	0.740	0.6995	1.310836e-09
3	first	0.100	0.143	1.000000e+00
3	second	0.253	0.390	1.220703e-02
3	third	0.629	0.597	2.333984e-01

Table 4: Comparison of the performance by senses, ranked by proportion within a target. The sense rank is organised by the number of senses. It starts from the smallest sense (in proportion; rank first) and increases to the largest (rank last). "-" means the ranking is based on the min and the max senses across all the data. Wilcoxon test is applied on the F1 scores of the senses in order to assess whether the distribution of F1 scores is significantly different between NEO and SCAN by number of senses.

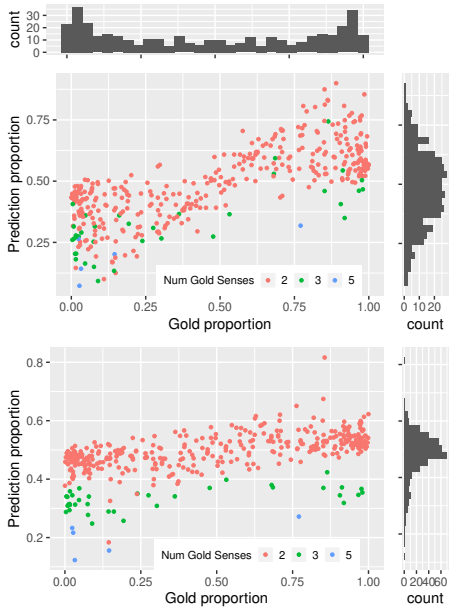

Figure 3: Relation between size of the gold and predicted senses for NEO (top) and SCAN (bottom).

System	Precision	Recall	F1-score
NEO	0.306	0.250	0.275
SCAN	0.126	0.090	0.105

Table 5: Results of NEO and SCAN regarding detecting the emergence of a new sense (5 year window).

Having confirmed that the imbalance between gold senses size has a strong impact on performance, we observe how the two systems behave with respect to the predicted size of the senses. It can be observed on Figure 3 that both systems split the data in favour of the senses with a low proportion, i.e. tend to predict a larger size for small senses and conversely a smaller size for large senses.[17] This tendency is exacerbated for SCAN which splits most senses equally regardless of their true size.

5.3 Evaluation of Emergence

Table 5 shows the global results after applying the emergence algorithm on the predictions of both systems. NEO performs much better than SCAN in predicting the emergence of a new sense, with an F1-score of 0.275 against 0.106 for SCAN.

Figure 4 shows the gold standard and the predicted emergence years for every sense which has emergence in both NEO and SCAN. SCAN tends to have earlier emergence results compared to the gold, while NEO tends to take the opposite direction with an average difference of -17.318 and 0.697 respectively across the senses. This tendency

[17]For the sake of concision, in this analysis we call "small (resp. large) sense" a sense with a low (resp. high) proportion of instances within the target.

System	Global MAE	Normalised Global MAE
NEO	17.076	0.295
SCAN	19.028	0.327

Table 6: Global emergence MAE, based on individual error by sense.

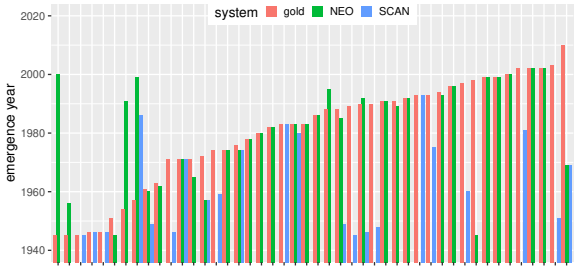

Figure 4: Gold and predicted emergence years for NEO and SCAN, ordered by gold emergence year.

is confimed by the fact that 90% of the amount of the difference error (predicted -gold) is predicted earlier for SCAN while NEO has only 45% of early predictions. The MAE results shown in table 6 are consistent with the classification results, showing a better performance by NEO. The emergence results in both systems are affected by data imbalance: for instance, both systems have a high number of FN cases when senses have a lower proportion of data (< 0.5). Similarly, the FP cases tend to correspond to senses which have a lower proportion.

5.4 Evaluation on $P(s|t)$

Table 7 shows that NEO has less errors by senses across years than SCAN according to the distance measures over $P(s|t)$. This is confirmed by Wilcoxon test, which shows that the errors distributions of the two systems are significantly different.

One would expect that the distance errors have an impact on emergence. By examining the means of two categories, TP cases (when the emergence is predicted within 5 years of the true emergence, see 5.3) as a category and the rest as a second category, one can observe that the means of the errors is lower for the former while its higher for the latter, as shown in table 8.

5.5 Comparing Evaluation Measures

The evaluation measures reflect different types of errors. The correlation values between clustering-based classification and regression measures are -0.71 for NEO and -0.44 for SCAN. This apparent

Distance	NEO	SCAN	Wilcoxon
	Global mean	Global mean	p-value
DTW	0.182	0.222	2.0413e-15
Euclidean	0.124	0.142	5.3543e-06

Table 7: Mean distance errors across senses by DTW and Euclidean algorithms.

	Predicted Emergence	DTW mean	Euclidean mean	Error mean
NEO	TP	0.078	0.0415	0.009
	not TP	0.189	0.130	0.313
SCAN	TP	0.193	0.124	0.016
	not TP	0.222	0.142	0.334

Table 8: Comparison between mean errors by predicted emergence status (error values normalised by the number of years). DTW and Euclidian distance are obtained by comparing the predicted vs. gold $P(s|t)$, whereas the classification status (TP vs. not TP) and normalised error mean are calculated based on the emergence year by sense.

	Distance Measure	Sense level F1-score	Target level macro F1-score
NEO	DTW	-0.270	-0.448
	Euclidean	-0.230	-0.432
SCAN	DTW	-0.313	-0.419
	Euclidean	-0.248	-0.374

Table 9: Correlation between distance measures and classification measures at the level of senses/targets.

discrepancy between the two evaluation measures is explained by several factors, some related to the definition of the measures and some due to the data characteristics. On one hand, the MAE is calculated as the average error across the instances which are labeled only with this particular true sense. On the other hand, in the classification setting, all the instances of a target are taken into account for a specific sense. This implies that the instances of the other senses are also taken into account.

For any given year t, the probability of the parameter $P(s|t)$ is estimated from the proportion of a sense among the instances of this year. This means that the value of the parameter $P(s|t)$ is directly related to the posterior probability used for the evaluation at the level of the instances. Therefore one would expect a quite strong correlation level between the DTW and/or Euclidean distance based on the estimated parameter $P(s|t)$ and the evaluation score based on the instances. However the correlation values observed at the level of senses (e.g. F1-score) is weak, although they are more significant at the level of targets, as shown in table 9.

The low correlation level is primarily due to the fact that the majority of the targets have two senses which are complement of each other, thus the two $P(s|t)$ series are a mirror of each other (i.e. $P(s_1|t) = 1 - P(s_2|t)$), in turn causing the DTW and Euclidean distance values to be the same for both senses. On the contrary, the instance-based evaluation scores tend to be very different for the two senses, especially in the case of strong size imbalance (see 5.2). The difference in correlation between the level of senses and the level of targets is likely due to the fact that the discrepancies in the evaluation between senses are balanced out at the level of targets.

6 Conclusion and Discussion

We have addressed the issue of evaluating DWSI: we evaluated two models, NEO and SCAN, directly on the task itself, independently from any extrinsic related tasks, with a large dataset collected from biomedical resources. We defined and tested various external evaluation measures. Overall, NEO performs significantly better in the tasks of detecting senses and the emergence of new senses, according to most of our evaluation measures.

The design differences between the models and their parameters could potentially have an effect on the amount of data they require, but it turns out that the global data size has no important effect on the accuracy of either system. Both systems are unable to predict the correct size of the clusters: they tend to split the data almost equally between senses irrespective of the true semantic sense represented by the context words, and this impacts the correct detection of the emergence. This issue also explains why the original studies tend to use a high number of senses in order to capture the true senses, even though this causes the clusters to be split and the appearance of "junk senses". We also find that NEO performs better with larger senses while SCAN tends to perform better with smaller senses. This opens the perspective of combining the advantages of the two systems. We acknowledge that the data is domain-specific, however the observed biases of the systems are likely to hold across domains.

Acknowledgements

We would like to thank Dr. Martin Emms and Dr. Lea Frermann for sharing the code of their systems. We are also grateful to the anonymous reviewers for their valuable comments.

The first author is grateful to King Abdullah Scholarship Program from the Saudi Arabian Government for supporting this work.

The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the fourth international workshop on semantic evaluations (semeval-2007)*, pages 7–12.

Donald J Berndt and James Clifford. 1994. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635.

Martin Emms and Arun Kumar Jayapal. 2016. Dynamic generative model for diachronic sense emergence detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1362–1373.

Lea Frermann. 2017. *Bayesian Models of Category Acquistion and Meaning Development*. Phd thesis, University of Edinburgh.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics*, pages 67–71.

Arun Jayapal. 2017. *Finding Sense Changes by Unsupervised Methods*. Phd thesis, Trinity College Dublin.

Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *arXiv preprint arXiv:1806.03537*.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601. Association for Computational Linguistics.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 task 14: Word sense induction &disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden. Association for Computational Linguistics.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

Octavian Popescu and Carlo Strapparava. 2015. Semeval 2015, task 7: Diachronic text evaluation. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 870–878.

Christian Rohrdantz, Annette Hautli, Thomas Mayer, Miriam Butt, Daniel A Keim, and Frans Plank. 2011. Towards tracking semantic change by visual analytics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 305–310. Association for Computational Linguistics.

Alexis Sardá-Espinosa. 2017. Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, 12:41.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.