# Improving Diachronic Word Sense Induction with a Nonparametric Bayesian method

**Ashjan Alsulaimani**
School of Computer Science
and Statistics
Trinity College of Dublin
alsulaia@tcd.ie

**Erwan Moreau**
School of Computer Science
and Statistics
Trinity College of Dublin
moreaue@tcd.ie

## Abstract

Diachronic Word Sense Induction (DWSI) is the task of inducing the temporal representations of a word meaning from the context, as a set of senses and their prevalence over time. We introduce two new models for DWSI, based on topic modelling techniques: one is based on Hierarchical Dirichlet Processes (HDP), a nonparametric model; the other is based on the Dynamic Embedded Topic Model (DETM), a recent dynamic neural model. We evaluate these models against two state of the art DWSI models, using a time-stamped labelled dataset from the biomedical domain. We demonstrate that the two proposed models perform better than the state of the art. In particular, the HDP-based model drastically outperforms all the other models, including the dynamic neural models.[1]

## 1 Introduction

Word meanings evolve over time. Recent research works have focused on how to model such dynamic behaviour. The unsupervised task of Diachronic Word Sense Induction (DWSI) aims to capture how the meaning of a word varies continuously over time, in particular when new senses appear or old senses disappear. DWSI takes the time dimension into account and assumes that the data spans over a long continuous period of time in order to model the progressive evolution of senses across time.

The dynamic behaviour of words contributes to semantic ambiguity, which is a challenge in many NLP tasks. DWSI can serve as an analytical tool to help building terminology resources and indexing documents more accurately and therefore can be beneficial for information retrieval tasks.

[1]The code corresponding to this work is available at https://github.com/AshjanAlsulaimani/DWSI-advanced-models

DWSI follows the probabilistic graphical modelling approach to approximate the true meanings from the observed data. Thus, in this paper, we explore the relation of DWSI with topic modelling in general and to the dynamic topic modelling techniques in particular: they both aim to discover a latent variable (sense or topic respectively) from a sequential collection of documents. Despite a close relation between the tasks, topic modelling techniques are not fully explored or compared against in the current state of the art of DWSI.

The state of the art of DWSI consists of only two models: (Emms and Kumar Jayapal, 2016) and (Frermann and Lapata, 2016). They are both designed specifically for DWSI; both are parametric; and both are dynamic, in the sense that they both introduce a time variable into the model in order to capture the evolution of the meaning over time. Emms and Kumar Jayapal (2016) propose a parametric generative model (NEO) where each sense is represented as a $|V|$-dimensional multinomial distribution over the vocabulary $V$, each document is represented as a mixture of senses, and the dependency of the sense proportions on time is represented as a $K$-dimensional multinomial distribution over the $K$ senses. The parameters of the model have finite Dirichlet priors. A more complex model called SCAN (Frermann and Lapata, 2016) allows each sense distribution over the vocabulary to evolve sequentially from adjacent time slices, as well as the senses proportion. The multinomial parameters of words and senses have logistic normal priors.

The two above-mentioned models are parametric, in the sense that the number of senses (which reflects the structure of the hidden meanings in the data) is a hyper-parameter which has to be known a priori. This is not ideal given the nature of the DWSI task, which is meant to infer senses from the



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 8908–8925
July 9-14, 2023 

data. The same issue has been studied for the tasks of topic modelling and WSI; Hierarchical Dirichlet Processes (HDP), a nonparametric hierarchical model introduced by Teh et al. (2006), offer an powerful solution to this problem. HDP extends Latent Dirichlet Allocation (LDA) (Blei et al., 2003) by placing Dirichlet processes priors (DPs) (Ferguson, 1973) on the infinite-dimensional space of multinomial probability distributions. Thus the number of mixture components is infinite a priori and to be inferred from the data. In contrast, LDA posits a predefined number $K$ of topics, each of which is a multinomial distribution over the vocabulary. Each document has specific topic proportions from a Dirichlet prior, and the topics are shared among the documents. Additionally, the HDP model allows sharing topics not only among documents but also across hierarchical levels by the use of multiple DPs.

The intuition behind our approach relies on the fact that the hierarchical DPs allow "new" senses to appear as needed, thanks to the theoretically infinite number of possible senses. Therefore, the hierarchical design of Dirichlet processes can capture the dynamic behaviour of the words, while inferring the optimal number of clusters directly from the data across time.

Word embeddings are another natural direction of potential improvement for DWSI. Introduced by Rumelhart and Abrahamson (1973); Bengio et al. (2003, 2006), they provide a distributed representation where words with similar meanings are close in a lower-dimensional vector space. Recently, various models have been proposed which integrate word embeddings for topic modelling, however these models do not necessarily represent both words and topics using embeddings. Dieng et al. (2019) provide an elegant solution to this problem: Dynamic Embedded Topic Model (DETM) is a parametric generative model inspired by D-LDA (Dynamic LDA) Blei and Lafferty (2006), in which each word is represented with a word embedding, and per-time topics are represented as embeddings as well. Topics and topic proportions evolve sequentially from adjacent time slices. DETM also directly models per-topic conditional probability of a word as the exponentiated inner product between the word embeddings and per-time topic embeddings. This results in a closer semantic correspondence between words and topics, and thus leads to better topics quality.

By contrast to previous contributions in DWSI which were mostly theoretical, this paper is an empirical contribution focusing on adapting different existing topic modelling techniques to DWSI. The aim is to set the state of the art DWSI models up against two serious competitors, in order to check whether they actually fit the task of DWSI optimally. In this perspective, we adapt HDP and DETM to the task of DWSI, describing our approach in §3. We test the ability of these models to detect meaning change over time using the evaluation framework proposed by (Alsulaimani et al., 2020), described in §4: using a large corpus of biomedical time-stamped data, including 188 ambiguous target words, we compare the proposed models with the current state of the art models NEO and SCAN. The results, presented in §5, show that HDP-based models achieve the best results over the dataset, establishing a new state of the art for DWSI.

## 2 Related Work

Topic modelling techniques are hierarchical probabilistic Bayesian models used originally for discovering topics in a collection of documents (Blei et al., 2010). Topic models have also been adopted for the Word Sense Induction (WSI) task, as introduced by (Brody and Lapata, 2009; Yao and Van Durme, 2011): word senses are treated as topics, and a short window around the target word (context) is considered instead of a full document. Topic modelling techniques have been extended further to similar tasks, such as Novel Sense Detection.

Novel Sense Detection (NSD; also called Novel Sense Identification), introduced by Lau et al. (2012), consists of determining whether a target word acquires a new sense over two independent periods of time, separated by a large gap. Several authors have used Hierarchical Dirichlet Processes (HDP) for this task over a small set of target words and/or small set of data (Lau et al., 2012, 2014; Cook et al., 2014). Yao and Van Durme (2011); Lau et al. (2012) show in a preliminary study that HDP is also superior to LDA for WSI, due to its ability to adapt to varying degrees of granularity. Lau et al. (2012) extend this study using an oracle-based method to identify new senses

from HDP predictions for the task of NSD, and for only five target words. Sarsfield and Tayyar Madabushi (2020) used HDP for NSD on a larger dataset (Schlechtweg et al., 2020), which was proposed in a recent shared task about Lexical Semantic Change Detection (LSCD), a refined version of NSD: LSCD intends to answer the question of whether the meaning of a target word has changed between two independent periods of time (also separated by a large time gap). In the LSCD task, methods based on static word embeddings (where the meaning of the word is represented by a single vector) achieved the highest performance.

In contrast to NSD/LSCD, DWSI takes the time dimension into account and thus the task of DWSI is technically broader: it aims to discriminate senses and also models the temporal dynamics of word meaning across a long continuous period of time, e.g. year by year. As a result, DWSI can track the evolution of senses, the emergence of new senses and detect the year where a new sense appears. The DWSI task is introduced independently by Emms and Kumar Jayapal (2016) and Frermann and Lapata (2016); given a target word and a time-stamped corpus, both models estimate two main parameters: the senses as distributions over words, and the senses proportions over time. Frermann and Lapata (2016) extend this by also inferring the subtle meaning changes within a single sense over time, i.e. by allowing different word distributions over time for the same sense.

However, these models are parametric and require the number of senses to be chosen in advance. Previous approaches dealt with this issue by increasing the number of senses. For example, Emms and Kumar Jayapal (2016) vary the number of senses manually for every target word, while Frermann and Lapata (2016) choose an arbitrary fixed large number of senses for all the target words.

Additionally, evaluating and comparing such models on the DWSI task is difficult: the lack of large scale time-stamped and sense-annotated data hinders direct quantitative evaluation. The state of the art models, (Emms and Kumar Jayapal, 2016; Frermann and Lapata, 2016), were originally evaluated only qualitatively on a few hand-picked target words, with a manual investigation of the quality of the associated top words in each cluster; Frermann and Lapata (2016) also evaluated their model on several indirect tasks. Alsulaimani et al. (2020) demonstrate that these evaluation methods are insufficient, and consequently propose a quantitative evaluation of these DWSI models based on a large set of data. In particular, they show that the senses size distribution plays a significant role in capturing the senses representations and emergence of new senses. The number of senses is clearly a crucial hyperparameter for a DWSI model, the choice of which should in theory depend on the characteristics of the data.

## 3 Approach

### 3.1 Parameters Notation

DWSI aims to discover the senses $S$ across time $Y$ for each target word in a sequential collection of documents, where senses are latent variables and the number of senses is unknown a priori. A DWSI model estimates at least two multinomial distributions:

- $P(W|S)$, the word given sense distribution. The changes within senses across time can also be represented as $P(W|S,Y)$, the word given sense and year distribution. These distributions represent the sense.
- $P(S|Y)$, the sense given year distribution. This distribution represents the relative prevalence of a sense over time.

### 3.2 HDP-DWSI

HDP allows senses (i.e. clusters) to appear when a new context occurs, as the number of senses is determined by the data. HDP-DWSI directly relies on this property: in the first step, all the documents, independently from their year, are clustered by HDP. Appendix A provides details about the description of HDP. This means that in this step the documents are assumed to be exchangeable, as opposed to dynamic models in which documents are only exchangeable within a time period. In the second step, the year of the document (observed variable) is reintroduced and the time-related multinomial parameters $P(S = s|Y = y)$ are estimated by marginalising across the documents of each year $j$ independently $\sum_{d \in y} \frac{freq(s_d)}{\sum_{s'} freq(s'_d)}$, where $freq(s_d)$

the number of words predicted as sense $s$ in the document $d$, and $d \in y$ represents the condition that the document $d$ belongs to year $y$.

HDP-DWSI is intended to be used as a nonparametric method, but a parametric mode is also proposed for the purpose of evaluation and comparison against parametric models. In the nonparametric mode, the model parameters are obtained directly as described above. In the parametric mode, an additional step is required to reduce the number of senses because HDP-DWSI tends to induce a higher number of clusters than the gold number of senses, i.e. to split senses into multiple clusters. Depending on the context of the application, it can also be relevant to reduce the number of senses even in the nonparametric mode. This can also be done with the method described below for the parametric mode, called HDP-DWSI$_\text{m}$.

HDP-DWSI$_\text{m}$ consists in merging the predicted senses which are the most semantically similar. Agglomerative hierarchical clustering (Ward Jr, 1963) is used to merge senses, based on a sense cooccurrence matrix obtained from the HDP clustering output.

Pointwise Mutual Information (PMI) is used to represent how strongly two predicted senses are statistically associated, under the assumption of independence:

$$PMI(s_i, s_j) = \log_2 \frac{P(s_i, s_j)}{P(s_i)P(s_j)} \quad (1)$$

where $i \neq j$ and $P(s_i, s_j)$ is the joint probability of observing both $s_i$ and $s_j$ in the same document. $P(s_i)$ (resp. $P(s_j)$) is the probability of a predicted sense with respect to the entire corpus, i.e. an occurrence is counted for every document in which the predicted sense $s_i$ (resp. $s_j$) independently occurs.

Moreover, since a pair of predicted senses with negative PMI is uninformative for the purpose of merging similar senses, Positive Pointwise Mutual Information (PPMI), as defined in Equation 2, is used for constructing the sense cooccurrence matrix.

$$PPMI = \begin{cases} PMI(s_i, s_j) & \text{if } PMI(s_i, s_j) > 0 \\ 0 & \text{else} \end{cases} \quad (2)$$

(P)PMI is sensitive to low frequency events, particularly in the event when one of the predicted senses (or both of them) is/are less frequent with respect to the whole corpus; thus it is possible that two senses mostly cooccur together by chance, yet obtain a high (P)PMI value. In such a case, the two predicted senses are not semantically associated, so this is a potential bias in the merging process. To counter this bias, we use the linkage criterion defined in Equation 3 as the average of the PPMI values weighted by their corresponding joint probabilities. The linkage criterion for two clusters $C_1, C_2$:

$$\sum_{\substack{\forall s_1 \in C_1 \\ \forall s_2 \in C_2}} w(s_1, s_2) \times PPMI(s_1, s_2) \quad (3)$$

$$\text{where } w(s_1, s_2) = \frac{P(s_1, s_2)}{\sum_{\substack{\forall s_1 \in C_1 \\ \forall s_2 \in C_2}} P(s_1, s_2)}$$

The evaluation method proposed by Alsulaimani et al. (2020) (see §4) relies on the gold number of senses, as it is originally intended for parametric methods. In order to compare an HDP-based model against parametric models in an equivalent setting, the HDP-DWSI$_\text{m}$ merging method is used to reduce the predicted number of senses to the gold-standard number of senses.

### 3.3 DETM-DWSI

DETM represents not only the observed words but also latent topics/senses as embeddings, while preserving the traditional representation of a topic/sense as a probability distribution across words. The categorical distributions over the vocabulary is time dependent, i.e. $P(W|S, Y)$ and is derived from the corresponding word embeddings and sense embedding at a given time. DETM also places time-dependent priors over senses proportions: the use of Markov chain over the sense proportions allows smoothness of the variations between the adjacent senses at neighboring times (see Appendix A for the description of DETM). We propose two modes for DETM-DWSI as follows:

- In the regular DETM-DWSI, both the word and sense embeddings are trained simultaneously. This mode does not require any additional resource but the corpus must be large enough for the embeddings to be accurate.
- In DETM-DWSI$_\text{i}$, the model is trained with prefitted word embeddings. This mode leverages the external information contained in the

embeddings, potentially obtaining a more accurate representation of the senses as a consequence. It also allows the application of the model to text containing words not present in the corpus, as long as their embedding is available.

In the experiments described below, the DETM-DWSI$_i$ models are trained using the BioWordVec pretrained word embeddings[2] (Zhang et al., 2019). The fastText subword embedding model (Bojanowski et al., 2017) is a variant of the continuous skip-gram model (Mikolov et al., 2013). The fastText subword embedding can learn a distinct vector for each word while exploiting subword information in a unified n-gram embedding space. BioWordVec embeddings are trained with fastText on the PubMed text and MeSH terms, combined into a unified embedding space. In the biomedical domain, the advantage of a subword embedding model is that it can handle Out of Vocabulary (OOV) words (Zhang et al., 2019).[3] This leads to a more precise word representation, in theory better able to capture the semantics of specialised concepts. We use the *intrinsic* BioWordVec embeddings (as opposed to the extrinsic type), meant to represent the semantic similarity between words (Zhang et al., 2019).

## 4 Experimental Setup

### 4.1 Data

We use the DWSI evaluation framework proposed by Alsulaimani et al. (2020): the biomedical literature is used as a source of labelled and time-stamped data which covers the years 1946 to 2019.[4] The dataset is collected from resources provided by the US National Library of Medicine (NLM): PubMed (a platform which includes the major biomedical literature databases) and MeSH (a controlled vocabulary thesaurus, created manually to index NLM databases).[5] The data is preprocessed as in (Alsulaimani et al., 2020). The data consists of 188 ambiguous target words and 379 gold-standard senses (Jimeno-Yepes et al., 2011): 75 ambiguous target words have 2 senses, 12 have 3 and one has 5 senses. The total data size is $15.36 \times 10^9$ words, and the average number of documents is 61,352 by sense. The input documents for every target word consist of the occurrences of the target word which are provided with a window of 5-word context on each side as well as the year of publication. The gold-standard sense label is also available for evaluation purposes.

### 4.2 Algorithms Settings

- The HDP-DWSI and HDP-DWSI$_m$ models are trained using the official C++ implementation of HDP.[6] No additional preprocessing is needed.
- The DETM-DWSI and DETM-DWSI$_i$ models are trained using the implementation provided by Dieng et al. (2019).[7] The preprocessing is adapted to the DWSI dataset: since the data is strongly imbalanced across time, stratified sampling is used in order to ensure a representative time distribution (with at least one instance by year) across the data partitions. The data is split into 85% of instances for training and 15% for validation. The document frequency thresholds are unused so as to include all the words. For efficiency reasons, during training the number of instances is capped at 2,000 instances per year.

### 4.3 Evaluation Methodology

Since DWSI is an unsupervised task (clustering) and our evaluation is based on the external sense labels, both the estimation of the model and the evaluation are performed on the full set of documents for each target word. The gold-standard number of senses of each ambiguous target word is provided for all the parametric models (excluding HDP-DWSI). The default parameters are used in all the systems,[8] except the number of itera-

[2] https://github.com/ncbi-nlp/BioSentVec.

[3] Note that the PubMed and MeSH terms are biomedical resources, collected from the US National Library of Medicine (NLM) and based on the database of 2019 and 2018 respectively. These are the same version for the DWSI evaluation data.

[4] https://github.com/AshjanAlsulaimani/DWSI-eval

[5] https://www.nlm.nih.gov/

[6] https://github.com/blei-lab/hdp.

[7] https://github.com/adjidieng/DETM.

[8] This means that we do not tune any hyper-parameter for any of the systems. Since DWSI applications would usually not have access to any labelled data, the performance would be unrealistic if the parameters were tuned.

tions/epochs (set to 500 for all the systems),[9] and specifically for DETM-DWSI the batch size is set to 1000 and the dimension of the embeddings is set to 200.

After estimating each model for each ambiguous target word, the posterior probability is calculated for every document. The sense with the highest probability is assigned.

### 4.4 Evaluation Measures

We follow Alsulaimani et al. (2020) for the evaluation measures with some adjustments, detailed below.

The "Global Matching" method, presented by Alsulaimani et al. (2020), consists in determining a one-to-one assignment between predicted senses and gold senses based on their joint frequency: the pair with the highest frequency is matched first, and this process is iterated until all the senses are matched. In the case of HDP-DWSI, the number of predicted senses may be higher than the gold number of senses, and the instances of the predicted senses which remain unmatched are considered as false negative. This allows to compare HDP-DWSI with the parametric models, assuming that in theory the ideal nonparametric model would infer exactly the true number of senses. Of course, HDP-DWSI$_{\text{m}}$ is by definition more appropriate for a comparison in the parametric setting of HDP-based methods.

We also propose to use the V-measure as a different method of evaluation. The V-measure is introduced by Rosenberg and Hirschberg (2007), providing a different way to evaluate a clustering solution. In this case, it evaluates every cluster against every gold sense without relying on a matching method, thus providing an objective assessment even when the number of the clusters is higher than the true number of senses. The V-measure is based on entropy (entropy is a measure of the uncertainty associated with a random variable): it is defined as the harmonic mean of homogeneity and completeness, which are both based on the normalised conditional entropy.

Alsulaimani et al. (2020) also propose to evaluate the emergence of a new sense by considering whether the system predicts the true emergence year of a sense. This requires a method to determine the year from the $P(S|Y)$ distribution, for which the original algorithm "EmergeTime" was proposed in Jayapal (2017). We introduce "LREmergeTime" (see Appendix B Algorithm 1), an improved version of "EmergeTime" using linear regression instead of multiple thresholds within a window. Indeed, the original algorithm is very sensitive to the noise which sometimes occurs in the emergence pattern. Linear regression handles this issue better, since it measures the global trend across the window.[10]

The emergence year is evaluated as in (Alsulaimani et al., 2020): (1) with standard classification measures, considering the sense as correctly predicted if the year is within 5 years of the true emergence year; (2) with (normalized) Mean Absolute Error, representing the average difference in number of years but also penalizing the wrongly predicted presence/absence of emergence.

Finally we also use the distance between the true and predicted evolution of the senses over time ($P(S|Y)$) as an evaluation method for DWSI, again following Alsulaimani et al. (2020).

## 5 Results

### 5.1 Qualitative exploration

We explore the temporal meanings of "SARS-associated coronavirus" over the years (2002-2018) as an example. The ambiguous word has two gold-standard senses described by UMLS concepts $C1175175$ and $C1175743$: *Severe Acute Respiratory Syndrome* (refers to the disease caused by the virus) and *SARS Virus* (refers to the virus related to the Coronavirus family causing the disease) respectively. The top words represented by the inferred parameter word given sense, identified by HDP-

[9] It has been verified that 500 epochs is sufficient for all models to become stable and therefore to achieve their optimal performance.

[10] The superiority of "LREmergeTime" was confirmed using a subset of manually annotated targets (the targets are chosen based on the visual clarity of the emergence pattern). The evaluation results on this subset show that "LREmergeTime" performs closer to the annotated senses. Following the evaluation measures by Alsulaimani et al. (2020), the results of "EmergeTime" and "LREmergeTime" are respectively 0.7 and 0.8 for Fscore, 12.06 and 6.74 for MAE, 0.21 and 0.11 for Normalised MAE. See Appendix C for details of algorithms outputs.

DWSI$_m$ for the first sense are {patients, outbreak, sars, 2003, epidemic, health, case, transmission, hospital} and for the second sense are {cov, sars, coronavirus, patients, infection, protein, respiratory, acute, syndrome, cells}. Figure 1 shows the relative prevalence of the two inferred and gold senses over time, and Table 1 shows the top inferred words/usages associated with sense $C1175175$ at specific times.

In Figure 1, both senses data start in 2002, however the prevalence of sense $C1175175$ was decreasing progressively from 2002 to 2018 since SARS was successfully contained in 2004, while the prevalence of the sense $C1175743$ kept increasing since the research about the *SARS virus* became a priority for the public health around the world.

The temporal changes of the top words within $C1175175$ are highlighted in Table 1. Historically, the first known case of SARS appears in November 2002, causing the 2002-2004 SARS outbreaks in cities and hospitals. Global attention then started and in 2016, for instance, the top words shifted to *facemask, post, era, sars*. Finally, the year 2018 shows the concerns about a second wave of SARS.

| 2002 | 2003 | 2004 | $\Longrightarrow$ | 2016 | 2017 | 2018 |
|---|---|---|---|---|---|---|
| case | patients | patients | | outbreak | outbreak | second |
| outbreak | outbreak | outbreak | | facemask | 2003 | 2003 |
| lessons | case | sars | | post | patients | impact |
| learned | health | transmission | | 2003 | china | epidemic |
| health | 2003 | hospital | | era | data | wave |
| chief | sars | case | | sars | outbreaks | n't |
| falls | hospital | patient | | hong | health | link |

Table 1: Temporal evolution of the top-7 words for the sense *Severe Acute Respiratory Syndrome* learned by HDP-DWSI$_m$, at specific times.

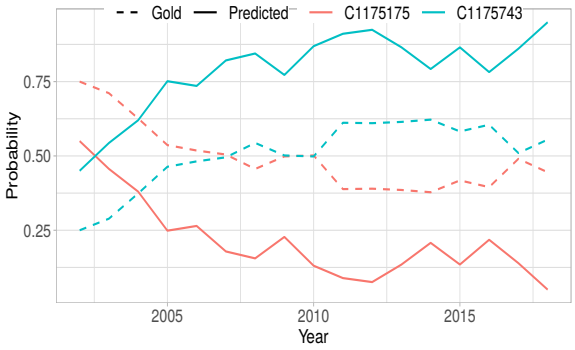



Figure 1: Dynamic representations of "SARS-associated coronavirus". On the Y-axis, $P(S|Y)$ shows the relative prevalence of the gold senses as well as the predicted senses across time estimated by HDP-DWSI$_m$.

## 5.2 Matching-based Evaluation

Table 2 shows the performance of the six models according to standard classification and regression measures using "Global Matching". In general, DWSI models based on HDP perform well compared to NEO or SCAN. In the case of HDP-DWSI, "Global Matching" causes two observable effects: it increases precision, by allowing the system to choose the best predicted clusters matched with the gold senses; but it also decreases recall by introducing a large number of false negative cases due the discarded unmatched predicted clusters. Nevertheless the macro F1 score for HDP-DWSI is much higher than both NEO and SCAN, by 17.7% and 13.8% respectively. This shows that HDP-DWSI can distinguish minority senses significantly better. This can also be seen in Table 3 which shows the mean F1-score by senses size.

| Systems | Macro-average | | | Micro-average | | | |
|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | MAE |
| DETM-DWSI$_i$ | 0.553 | 0.561 | 0.557 | 0.704 | 0.704 | 0.704 | 0.401 |
| DETM-DWSI | 0.559 | 0.590 | 0.574 | 0.650 | 0.650 | 0.650 | 0.379 |
| HDP-DWSI | **0.726** | 0.599 | 0.657 | 0.739 | 0.424 | 0.539 | - |
| HDP-DWSI$_m$ | 0.666 | **0.681** | **0.674** | **0.744** | **0.744** | **0.744** | **0.26** |
| NEO | 0.548 | 0.569 | 0.558 | 0.595 | 0.595 | 0.595 | 0.425 |
| SCAN | 0.562 | 0.591 | 0.577 | 0.558 | 0.558 | 0.558 | 0.444 |

Table 2: Global performance results for all systems using "Global Matching". P/R/F1 stand for Precision/Recall/F1-score (higher is better) MAE stands for Mean Absolute Error (lower is better). Best performance in bold.

The superiority of HDP-DWSI$_m$ is even clearer: the macro F1 score is 20.8% higher than NEO and 16.8% higher than SCAN; the performance difference in micro F1 score is even stronger: 21.0% above DETM-DWSI$_i$, 17.4% higher than DETM-DWSI, 25.0% above NEO and 33.3% above SCAN. Contrary to the differences between NEO and SCAN, HDP-DWSI$_m$ improves performance significantly across the board: both precision and recall are drastically higher, according to both micro and macro scores. This means that HDP-based models are fundamentally much better at discriminating the different senses (with a very significant p-value $< 0.05$), as opposed to strategically favouring large senses for instance. This is confirmed in Table 3.[11]

The two DETM-based models perform very well, in particular achieving micro F1-score much higher than NEO and SCAN. However their macro-average performance is comparable to NEO and

[11] A Wilcoxon rank sum test is applied on the F1-scores of the senses for the results in Table 2 and 3.

SCAN, a clear sign that they do not separate the senses better. Table 3 confirms that the DETM-based models perform closely to NEO and SCAN.

Finally the MAE scores confirm that DETM-DWSI$_i$ and DETM-DWSI peform better than NEO and SCAN, but also that these four models are drastically outperformed by HDP-DWSI$_m$.

| Number of Senses | Sense rank | Mean F1 score | | | | | |
|---|---|---|---|---|---|---|---|
| | | N | S | H | $H_m$ | D | $D_i$ |
| - | first | 0.299 | 0.321 | 0.532 | 0.438 | 0.314 | 0.283 |
| - | last | 0.732 | 0.692 | 0.658 | 0.857 | 0.739 | 0.772 |
| 2 | first | 0.315 | 0.335 | 0.557 | 0.462 | 0.330 | 0.294 |
| 2 | second | 0.740 | 0.6995 | 0.659 | 0.863 | 0.744 | 0.777 |
| 3 | first | 0.100 | 0.143 | 0.224 | 0.132 | 0.111 | 0.134 |
| 3 | second | 0.253 | 0.390 | 0.553 | 0.499 | 0.237 | 0.248 |
| 3 | third | 0.629 | 0.597 | 0.655 | 0.778 | 0.681 | 0.708 |

Table 3: Comparison of the performance by sense according to the "Global Matching" method, ranked by proportion within a target. The sense rank is ordered by the size of senses (in number of instances), from the smallest sense (rank first) to the largest (rank last). "-" means any number of senses (all the data). The systems are referred to by their initials.

### 5.3 Entropy-based Evaluation

| Systems | V-measure | | homogeneity | | completeness | |
|---|---|---|---|---|---|---|
| | Mean | Median | Mean | Median | Mean | Median |
| DETM-DWSI$_i$ | 0.093 | 0.021 | 0.106 | 0.059 | 0.089 | 0.016 |
| DETM-DWSI | 0.092 | 0.026 | 0.111 | 0.059 | 0.085 | 0.020 |
| HDP-DWSI | 0.213 | **0.161** | **0.384** | **0.349** | 0.157 | **0.107** |
| HDP-DWSI$_m$ | **0.272** | 0.110 | 0.289 | 0.154 | **0.268** | 0.094 |
| NEO | 0.046 | 0.018 | 0.053 | 0.026 | 0.043 | 0.014 |
| SCAN | 0.080 | 0.021 | 0.098 | 0.041 | 0.074 | 0.015 |

Table 4: V-measure, homogeneity and completeness for all the systems. Both the mean and median across targets are reported, because the strong differences between targets in terms of size and distribution of the senses may cause a bias with the mean.

Table 4 shows the results of the systems for V-measure, with details about homogeneity and completeness. HDP-DWSI and HDP-DWSI$_m$ perform the best at all three levels, with values far above the other systems. HDP-DWSI has the highest homogeneity mean, because this model produces a higher number of smaller predicted senses; these predicted senses are therefore more homogeneous in general, but also less complete since the gold senses are often split. HDP-DWSI$_m$ merges the senses predicted by HDP-DWSI, thus obtaining lower homogeneity but compensating with higher completeness, leading to higher mean V-measure.

Figure 2 offers a more precise picture of the differences between systems about their V-measure distribution. It confirms that DETM-DWSI, DETM-DWSI$_i$ and SCAN perform very similarly. It

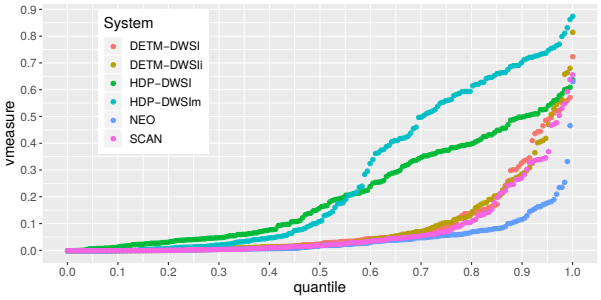



Figure 2: Quantile plot of the V-measure scores by system, with the quantile rank shown on the X axis and the corresponding value on the Y axis. Example: for HDP-DWSI, the median (x=0.5) is y=0.16. The graph is obtained by sorting the values, then normalising their rank between 0 and 1.

shows that the higher performance of DETM-DWSI, DETM-DWSI$_i$ and SCAN compared to NEO is due to a minority of targets, as their 75% lowest scores are almost identical. These targets cause most of the high difference in mean between NEO and SCAN, as the smaller difference in medians shows.

By contrast, HDP-DWSI and HDP-DWSI$_m$ have a much smaller proportion of low scores. Interestingly, HDP-DWSI has higher low scores than HDP-DWSI$_m$, i.e. HDP-DWSI performs better until both systems reach the median. However HDP-DWSI$_m$ skyrockets just after the median and surpasses HDP by having much higher high scores. This explains why the median is slightly lower for HDP-DWSI$_m$ than HDP while the mean is much higher for HDP-DWSI$_m$.[12]

### 5.4 Comparison between Measures

| Measure | V-measure | | | | | |
|---|---|---|---|---|---|---|
| | N | S | H | $H_m$ | D | $D_i$ |
| Macro F-Score | 0.730 | 0.799 | 0.856 | 0.901 | 0.795 | 0.794 |
| Micro F-Score | 0.583 | 0.850 | 0.806 | 0.714 | 0.713 | 0.494 |

Table 5: Pearson correlation coefficients: the relationship between the performance according to different measures. All the results are significantly correlated with p-value <= 5.6e-13. The systems are referred to by their initials.

V-measure can introduce a bias towards systems which predict a number of clusters larger than the number of gold senses. Such systems tend to have very high homogeneity scores and low completeness scores. However, this is not the case for HDP-DWSI. The HDP-DWSI performance is high not only according to the V-measure but also confirmed

[12] This can be verified visually on the quantile plot, because the area under the curve is equal to the mean.

by the F1 scores. The number of senses predicted by HDP-DWSI in average is 8 senses, with the minimum 4 senses and the maximum 13 senses. The Pearson correlation between homogeneity and completeness is 0.853 and with very significant p-value, 2.2e-16. Also, it is found that there is virtually no correlation between the predicted number of senses and either the size of the data or V-measure by target: 0.065, 0.008 (non significant: p-value = 0.3746, 0.261). This indicates that HDP-DWSI is not biased towards generating more senses when the data is larger.

Table 5 shows that all the evaluation measures are significantly correlated. The macro-F1 scores are positively correlated in all four systems. However, the micro F-score favours systems that perform well on the majority sense, whereas the V-measure explicitly evaluates every cluster, taking into account not only the majority sense but also the minority one. Therefore systems which favour the majority sense, like NEO and DETM-DWSI$_i$, have a lower correlation.

### 5.5 Emergence-based Evaluation

| System | precision | recall | F1 score | global mean absolute error | normalised global mean absolute error |
|---|---|---|---|---|---|
| DETM-DWSI$_i$ | **0.500** | 0.009 | 0.019 | 48.713 | 0.812 |
| DETM-DWSI | 0.385 | 0.050 | 0.088 | 45.685 | 0.761 |
| HDP-DWSI$_m$ | 0.371 | 0.254 | 0.301 | **23.148** | 0.403 |
| NEO | 0.383 | **0.397** | **0.390** | 23.967 | **0.399** |
| SCAN | 0.374 | 0.162 | 0.226 | 39.634 | 0.666 |

Table 6: Sense emergence evaluation results for all the systems. The values in bold indicate the best score achieved among the systems.

DWSI systems can also be evaluated based on their ability to predict the year of emergence of a new sense. Table 6 shows the performance of the systems after applying "LREmergeTime" (see §4.4 ) on the predictions of the systems. HDP-DWSI$_m$ and NEO perfom closely to each other and much better than the other systems, according to both classification measures and MAE. NEO was designed and implemented with a focus on detecting sense emergence, this probably explains why it performs particularly well in this task (Jayapal, 2017).

### 5.6 Evaluation based on the predicted evolution over time

Table 7 shows for every system how well their prediction of $P(S|Y)$ matches the true evolution of sense. Among all the systems, HDP-DWSI$_m$ predicts the closest $P(S|Y)$ to the true evolution

| System | Distance Global mean | |
|---|---|---|
| | DTW | Euclidean |
| DETM-DWSI | 0.191 | 0.134 |
| DETM-DWSI$_i$ | 0.165 | 0.106 |
| HDP-DWSI$_m$ | **0.115** | **0.067** |
| NEO | 0.182 | 0.124 |
| SCAN | 0.222 | 0.142 |

Table 7: Mean distance between the true and predicted sense, measured by Dynamic Time Warping (DTW) and Euclidean distance (lower is better). The results in bold indicate the best system.

according to both distance measures. This confirms that not only HDP-DWSI$_m$ produces accurate predictions of the emergence year of novel senses but also predicts accurately the $P(S|Y)$ trends in general, with significantly less errors than the other systems.

## 6 Conclusion and Discussion

In this paper we adapted two topic modelling methods to the task of DWSI and evaluated them against two state of art DWSI systems, NEO and SCAN, using the evaluation framework proposed by Alsulaimani et al. (2020). We also compared using the V-measure, and proposed an improved version of the emergence algorithm.

The results show that HDP-based models are able to fit the data better than the parametric models. The results strongly show that merging HDP-DWSI clusters performs better than the DETM-DWSI models and LDA-like clustering, such as NEO and SCAN. The properties of HDP make it better at accurately fitting the topics/senses, in particular when there is a high imbalance between the senses proportions, i.e. with senses smaller in size (see Table 3). Furthermore, the fact that HDP-DWSI$_m$ outperforms all the other parametric models also demonstrates that these models do not find the optimal separation between the senses. It seems that the additional complexity of the time dimension together with the parametric constraints do not cope well with data imbalance across years.

One could naturally assume that models designed specifically for a task would perform better on it. Implicitly, the research community encourages the creation of new models and tends to reward theoretical contribution over empirical ones. Thus there might be a bias in favor of designing sophisticated ad-hoc models (like NEO and SCAN) rather than adapting existing robust models (like HDP).

## 7 Limitations

### 7.1 Biomedical Domain

The dataset used in these experiments belongs to the biomedical domain and it is in English language. There is no clear reason why the comparison between models would lead to different results on different domains, therefore we would expect the reported results (at least the major tendencies) to be also valid on the general domain.

Nevertheless this assumption would need to be tested experimentally. To our knowledge, there is no equivalent dataset available in the general domain which satisfies the two following conditions:

- Time-stamped documents spanning a relatively long period of time;
- Every document labelled with the sense of the target word.

### 7.2 Duration of the Training Stage

In the table below, we present the computational cost of training the different models presented in this paper. Most of the experiments were carried out on a computing cluster containing 20 to 30 machines with varying characteristics, thus the total duration is approximative.

Computing times are reported in hours of CPU/GPU activity required to train the total of 188 target datasets. It is important to note that the two DETM models are trained on GPUs, whereas all the other models are trained on regular CPUs. Thus in overall computing power, the DETM models are the most costly to train (more than HDP, despite the higher duration).

| System | Duration | Notes |
|---|---|---|
| DETM-DWSI$_i$ | 523.4 | Trained on GPU |
| DETM-DWSI | 474.2 | Trained on GPU |
| HDP-DWSI | 2,471.4 | |
| HDP-DWSI$_m$ | 0.1 | Only the merging process |
| NEO | 25.1 | |
| SCAN | 77.9 | |

## Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. The first author is grateful to the Custodian of the Two Holy Mosques Scholarship Program from the Saudi Arabian Government as well as to Mr. Saud Alsulaimani for supporting this work. This work was conducted using high-performance clusters facilitated by the ADAPT Centre, Trinity College of Dublin.

## 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.

Ashjan Alsulaimani, Erwan Moreau, and Carl Vogel. 2020. An evaluation method for diachronic word sense induction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3171–3180. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. journal of machine learning research, 3 (feb): 1137-1155, 2003. *Google Scholar Google Scholar Digital Library Digital Library*.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Innovations in machine learning. *Neural Probabilistic Language Models*, 194:137–186.

David Blei, Lawrence Carin, and David Dunson. 2010. Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. *IEEE signal processing magazine*, 27(6):55.

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.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 103–111. Association for Computational Linguistics.

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.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

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.

Thomas S. Ferguson. 1973. A bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1:209–230.

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

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.

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 *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

David E Rumelhart and Adele A Abrahamson. 1973. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28.

Eleri Sarsfield and Harish Tayyar Madabushi. 2020. UoB at SemEval-2020 task 1: Automatic identification of novel word senses. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 239–245, Barcelona (online). International Committee for Computational Linguistics.

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*.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Xuchen Yao and Benjamin Van Durme. 2011. Nonparametric bayesian word sense induction. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 10–14. Association for Computational Linguistics.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.

# A Hierarchical Bayesian Models Background

## A.1 Hierarchical Dirichlet Processes

Hierarchical Dirichlet Processes (HDP), introduced by Teh et al. (2006), uses Dirichlet processes priors (DPs), on the infinite-dimensional space of multinomial probability distributions and thus the number of mixture components (senses) is infinite a priori. The Hierarical DPs allow new senses to emerge naturally at any point in time and guarantee the senses are shared within and across the documents. The DP provides a distribution on distributions over an arbitrary space. $H$ is a symmetric Dirichlet on the word simplex and $\gamma$ is a concentration parameter that controls the amount of variability of senses on the base distribution $G_0$, a distribution over senses drawn from a DP. $\alpha$ is also a concentration parameter that controls the amount of variability of per-document senses on $G_d$, a multinomial probability distribution over senses drawn from a DP. Then, for each word $w$ we draw a sense $\beta_{d,n}$ from $G_d$ and finally draw the word $w$ from that sense $\beta_{d,n}$ The graphical model and the generative story of HDP are described in Figure 3.

## A.2 Dynamic Embedded Topic Model

Dynamic Embedded Topic Model (DETM), introduced by Dieng et al. (2019), uses embedding representations of words and topics. For each term $v$, it considers an $L$-dimensional embedding representation $p_v$. It also considers an embedding $\alpha_k^t \in \mathbf{R}^L$ for each topic $k$ at a given time step $t = 1, ..., T$. The topics (i.e. distributions over the vocabulary) are represented by the normalised exponentiated dot product between the embedding represenation of the word and the assigned topic's embedding at every time $t$ for each word in a document $d$: $p(w_{d,n} = v|z_{d,n} = k, \alpha_k^{t_d}) \propto exp\{p_v^T \alpha_k^{(t_d)}\}$. The DETM uses a Markov chain over the topic embeddings $\alpha_k^t$ and thus they evolve under Gaussian noise with variance $\delta^2$. Moreover, DETM posits time-varying prior, the logistic-normal distribution $\mathcal{LN}$ over the topic proportions $\theta_d$, which depends on a latent variable $\eta_{t_d}$.

# B Emergence Algorithm

"LREmergeTime " algorithm in 1 is linear regression based algorithm, an improved version of "EmergeTime" proposed by (Jayapal, 2017).

**ALGORITHM 1**

Emergence Detection algorithm based on linear regression

**Input** $\pi$: $\pi[i]$ is the probability at time $i$, with $1 \leq i \leq N$
**Input** $r$: window size ▷ Value used for window size: 5
**Input** $s$: slope threshold. ▷ Value used for slope threshold: 0.04
**function** LREMERGTIME($\pi, r, s$)
  Surges= $\phi$
  **for** n:=1 to (N-r+1) **do**
    **if** SurgeStart(n,$\pi$,s) **then**
      Surges = Surges $\cup \{n\}$
    **end if**
  **end for**
  **if** Surges $\neq \phi$ **then**
    return min(Surges)
  **else**
    return $\phi$
  **end if**
**end function**

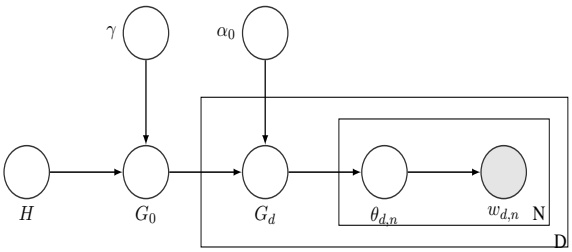
**function** SURGESTART($n, \pi, s$)
  (slope, intercept) = fit linear regression model on $X = [\, n, \ldots, n + r - 1\,]$ and $Y = [\, \pi[n], \ldots, \pi[n + r - 1]\,]$
  **if** slope < s $* max(\pi)$ **then**
    return false
  **end if**
  PrevYears = $\{n' : 1 \leq n' < n\}$
  **if** $|\{n' : n' \in$ PrevYear and $\pi[n'] \leq 0.1 * max(\pi)\}|$ / |PrevYear| $\geq 0.8$ **then**
    return true
  **else**
    return false
  **end if**
**end function**

# C Data: Gold Standard Dataset

The table C below shows the gold standard output (senses and year of emergence), as obtained by the "LREmergeTime" emergence detection algorithm based on the original gold data in (Alsulaimani et al., 2020).

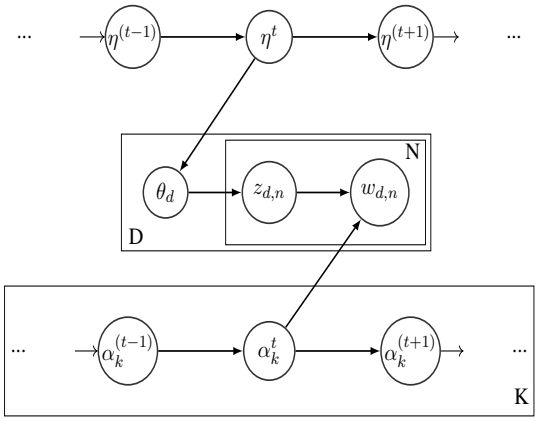
The total number of targets which has emergence is 146 and which has no emergence is 42. This consists of 233 senses with emergence and 158 senses with no emergence. The table includes three type of emergence:

- N: Number of senses
- LRET: "LREmergeTime" emergence year,




- Draw the base distribution over senses $G_0 \sim DP(\gamma, H)$,
- For $d \in 1, ..., D$, draw the per-document distribution over senses $G_d \sim DP(\alpha, G_0)$,
- For each word $w \in 1, ..., N_d$ in each document d,
  - Draw the sense for the word $\beta_{d,n} \sim G_d$
  - Draw the word $w_{d,n} \sim Mult(\beta_{d,n})$

Figure 3: Left: graphical representation of HDP for DWSI. Observed variables represented by shaded nodes and latent variables by clear nodes. Right: the corresponding generative process. Note that in DWSI the sense related variables replace the topic related variables.




- Draw initial sense embedding ${\alpha_k}^{(0)} \sim \mathcal{N}(0, I)$
- Draw initial sense proportion mean $\eta_0 \sim \mathcal{N}(0, I)$
- For time step $t = 1, ...., T$ :
  - Draw sense embeddings ${\alpha_k}^{(t)} \sim \mathcal{N}({\alpha_k}^{(t-1)}, \delta^2 I)\ for\ k = 1, ....., K$
  - Draw sense proportion means $\eta_t \sim \mathcal{N}(\eta_{t-1}, \delta^2 I)$
- For each document $d$:
  - Draw sense proportions $\theta_d \sim \mathcal{LN}(\eta_{t_d}, a^2 I)$
  - For each word n in the document $d$:
    - Draw sense assignment $z_{d,n} \sim Cat(\theta_d)$
    - Draw word $w_{d,n} \sim Cat(softmax(p^T \alpha^{t_d}_{z_{d,n}}))$

Figure 4: Left: graphical representation of DETM for DWSI. Observed variables represented by shaded nodes and latent variables by clear nodes. Right: the corresponding generative process. Note that in DWSI the sense related variables replace the topic related variables.

- ET: "EmergeTime" emergence year,
- FYO: indicates the "First Year Occurrence" of a sense, determined by the start date of each sense in the data,
- MS: indicates the "Manual Surge", i.e. the visual manual annotations by the authors. The value "NA" indicates cases when no emergence found and "?" indicates visually ambiguous cases found during the manual annotation by the authors.

| Target | N | CUI | ET | LRET | FYO | MS |
|---|---|---|---|---|---|---|
| AA | 2 | C0001972 | | 1945 | 1947 | 1947 |
| AA | 2 | C0002520 | | | 1945 | |
| ADA | 2 | C0001457 | | | 1955 | ? |
| ADA | 2 | C0002456 | | | 1955 | ? |
| ADH | 2 | C0001942 | 1978 | 1977 | 1976 | 1979 |
| ADH | 2 | C0003779 | | | 1975 | |
| ADP | 2 | C0001459 | | | 1956 | |
| ADP | 2 | C0004374 | 1958 | 1956 | 1959 | 1959 |
| Adrenal | 2 | C0001625 | | | 1945 | ? |
| Adrenal | 2 | C0014563 | | | 1945 | ? |
| Ala | 3 | C0001898 | | | 1947 | |
| Ala | 3 | C0002563 | 1954 | 1949 | 1953 | 1973 |
| Ala | 3 | C0051405 | | 1982 | 1947 | 1979 |
| ALS | 2 | C0002736 | | | 1948 | |
| ALS | 2 | C0003372 | | 1964 | 1968 | 1968 |
| ANA | 2 | C0002463 | 1962 | 1962 | 1963 | 1963 |

*continued on next column or page*

| Target | N | CUI | ET | LRET | FYO | MS |
|---|---|---|---|---|---|---|
| ANA | 2 | C0003243 | | | 1962 | |
| Astragalus | 2 | C0039277 | | | 1947 | ? |
| Astragalus | 2 | C0330845 | | | 1947 | ? |
| B-Cell Leukemia | 2 | C0023434 | | | 1986 | |
| B-Cell Leukemia | 2 | C2004493 | | 1986 | 1988 | 1988 |
| BAT | 2 | C0006298 | | 1946 | 1949 | 1949 |
| BAT | 2 | C0008139 | | | 1946 | |
| BLM | 2 | C0005740 | | | 1971 | |
| BLM | 2 | C0005859 | | 1978 | 1981 | 1981 |
| Borrelia | 2 | C0006033 | | | 1979 | |
| Borrelia | 2 | C0024198 | | 1980 | 1983 | 1983 |
| BPD | 2 | C0006012 | | | 1980 | ? |
| BPD | 2 | C0006287 | 1980 | 1980 | 1981 | ? |
| BR | 2 | C0006137 | | 1946 | 1946 | ? |
| BR | 2 | C0006222 | | | 1946 | ? |
| Brucella abortus | 2 | C0006304 | | | 1946 | |
| Brucella abortus | 2 | C0302363 | | | 1946 | 1958 |
| BSA | 2 | C0005902 | | | 1952 | ? |
| BSA | 2 | C0036774 | | 1952 | 1952 | ? |
| BSE | 2 | C0085105 | | | 1991 | ? |
| BSE | 2 | C0085209 | 1991 | | 1991 | ? |
| Ca | 3 | C0006675 | | | 1945 | ? |
| Ca | 3 | C0006754 | 1945 | | 1945 | ? |
| Ca | 3 | C0006823 | | | 1945 | ? |
| CAD | 2 | C0011905 | | | 1983 | |
| CAD | 2 | C1956346 | 1983 | 1983 | 1985 | 1985 |
| Callus | 2 | C0006767 | | | 1972 | ? |
| Callus | 2 | C0376154 | | | 1972 | ? |
| CAM | 2 | C0007578 | | | 1981 | |
| CAM | 2 | C0178551 | 2002 | 2001 | 2003 | 2003 |
| CCD | 2 | C0008928 | | | 1965 | |
| CCD | 2 | C0751951 | 1997 | 1995 | 1965 | 1998 |
| CCl4 | 2 | C0007022 | | | 1946 | |
| CCl4 | 2 | C0209338 | 1994 | 1992 | 1991 | 1994 |
| CDA | 2 | C0002876 | | | 1979 | |
| CDA | 2 | C0092801 | 1982 | 1979 | 1983 | 1988 |
| CDR | 2 | C0011485 | | | 1973 | |
| CDR | 2 | C0021024 | | 1994 | 1998 | 1998 |
| Cell | 2 | C0007634 | | | 1969 | |
| Cell | 2 | C1136359 | 2010 | 1998 | 1999 | 2002 |
| Cement | 2 | C0011343 | | | 1957 | ? |
| Cement | 2 | C1706094 | | | 1957 | ? |
| CH | 2 | C0008115 | | | 1946 | ? |
| CH | 2 | C0039021 | 1946 | 1946 | 1946 | ? |
| Cholera | 2 | C0008354 | | | 1945 | |
| Cholera | 2 | C0008359 | | 1945 | 1946 | 1961 |
| CI | 2 | C0008107 | | | 1949 | |
| CI | 2 | C0022326 | | | 1951 | 1955 |
| Cilia | 2 | C0008778 | | 1950 | 1950 | ? |
| Cilia | 2 | C0015422 | | | 1950 | ? |
| CIS | 2 | C0007099 | | | 1972 | |
| CIS | 2 | C0162854 | 1991 | 1989 | 1992 | 1992 |
| CLS | 2 | C0265252 | | 1998 | 2002 | 2002 |
| CLS | 2 | C0343084 | | | 1996 | |
| Coffee | 2 | C0009237 | | | 1960 | |
| Coffee | 2 | C0085952 | 2001 | 1998 | 1962 | 2002 |
| Cold | 3 | C0009264 | | | 1945 | |
| Cold | 3 | C0009443 | 1945 | 1945 | 1945 | |
| Cold | 3 | C0024117 | 1998 | 1997 | 1959 | 1998 |
| Compliance | 2 | C0009563 | | | 1974 | ? |
| Compliance | 2 | C1321605 | | 1974 | 1974 | ? |
| Cortex | 2 | C0001614 | 1948 | 1947 | 1950 | ? |
| Cortex | 2 | C0007776 | | | 1945 | ? |
| Cortical | 3 | C0001613 | 1945 | 1945 | 1945 | |
| Cortical | 3 | C0007776 | | | 1945 | |
| Cortical | 3 | C0022655 | | | 1947 | 1971 |
| CP | 3 | C0007789 | | | 1946 | |
| CP | 3 | C0008925 | | | 1946 | |
| CP | 3 | C0033477 | 1971 | 1969 | 1946 | 1971 |
| CPDD | 2 | C0008838 | 1971 | 1971 | 1972 | 1972 |
| CPDD | 2 | C0553730 | | | 1971 | |
| Crack | 2 | C0040441 | | | 1986 | |
| Crack | 2 | C0085163 | | 1987 | 1990 | 1990 |
| CRF | 2 | C0010132 | | 1954 | 1956 | 1967 |
| CRF | 2 | C0022661 | | | 1954 | |
| cRNA | 2 | C0056208 | 1981 | 1978 | 1982 | 1984 |

*continued on next column or page*

| Target | N | CUI | ET | LRET | FYO | MS |
|---|---|---|---|---|---|---|
| cRNA | 2 | C1321571 | | | 1975 | |
| CTX | 2 | C0010583 | | | 1960 | |
| CTX | 2 | C0238052 | 1997 | 1992 | 1974 | 1996 |
| DAT | 2 | C0002395 | | | 1974 | |
| DAT | 2 | C0114838 | 1989 | 1988 | 1989 | 1991 |
| DBA | 2 | C0025923 | | | 1972 | |
| DBA | 2 | C1260899 | 1999 | 1998 | 2001 | 2001 |
| dC | 2 | C0011485 | 1971 | 1969 | 1973 | 1973 |
| dC | 2 | C0012764 | | | 1966 | |
| DDD | 2 | C0011037 | | | 1962 | |
| DDD | 2 | C0026256 | | | 1963 | 1973 |
| DDS | 3 | C0010980 | | | 1960 | |
| DDS | 3 | C0085104 | 1988 | 1987 | 1990 | 1990 |
| DDS | 3 | C0950121 | 1999 | 1998 | 2001 | 2001 |
| DE | 2 | C0011198 | | | 1945 | ? |
| DE | 2 | C0017480 | | | 1945 | ? |
| DI | 2 | C0011848 | | | 1946 | |
| DI | 2 | C0032246 | | | 1946 | 1976 |
| Digestive | 2 | C0012238 | | | 1945 | ? |
| Digestive | 2 | C0012240 | | | 1945 | ? |
| DON | 2 | C0012020 | | | 1975 | |
| DON | 2 | C0028652 | 1979 | 1978 | 1981 | 1981 |
| drinking | 2 | C0001948 | | | 1946 | ? |
| drinking | 2 | C0684271 | 1946 | | 1946 | ? |
| eCG | 2 | C0018064 | 1989 | 1989 | 1945 | ? |
| eCG | 2 | C1623258 | | | 1945 | ? |
| Eels | 2 | C0013671 | | | 1951 | |
| Eels | 2 | C0677644 | 2003 | 2000 | 2004 | 2004 |
| EGG | 2 | C0013710 | | | 1945 | |
| EGG | 2 | C0029974 | | 1945 | 1945 | 1945 |
| EM | 2 | C0014921 | 1973 | 1972 | 1975 | 1975 |
| EM | 2 | C0026019 | | | 1946 | |
| EMS | 2 | C0013961 | | | 1967 | |
| EMS | 2 | C0015063 | 1974 | 1971 | 1975 | 1975 |
| Epi | 2 | C0014563 | | | 1945 | |
| Epi | 2 | C0014582 | | 1988 | 1980 | 1980 |
| ERP | 2 | C0008310 | 1978 | 1977 | 1978 | 1978 |
| ERP | 2 | C0015214 | | | 1956 | |
| ERUPTION | 2 | C0015230 | | | 1945 | ? |
| ERUPTION | 2 | C1533692 | | | 1945 | ? |
| Erythrocytes | 2 | C0014772 | | | 1945 | |
| Erythrocytes | 2 | C0014792 | | | 1945 | |
| Exercises | 2 | C0015259 | | 1945 | 1945 | ? |
| Exercises | 2 | C0452240 | | | 1945 | ? |
| FA | 2 | C0015625 | | | 1946 | 1975 |
| FA | 2 | C0016410 | | | 1945 | |
| Fe | 2 | C0302583 | | | 1945 | |
| Fe | 2 | C0376520 | 1995 | 1992 | 1946 | 1996 |
| Fish | 2 | C0016163 | | | 1945 | |
| Fish | 2 | C0162789 | 1990 | 1988 | 1953 | 1992 |
| Follicle | 2 | C0018120 | | | 1949 | ? |
| Follicle | 2 | C0221971 | | 1949 | 1949 | ? |
| Follicles | 2 | C0018120 | | | 1949 | ? |
| Follicles | 2 | C0221971 | | 1949 | 1949 | ? |
| FTC | 2 | C0041713 | | | 1982 | |
| FTC | 2 | C0206682 | 1992 | 1989 | 1993 | 1993 |
| GAG | 2 | C0017346 | 1988 | 1986 | 1982 | 1989 |
| GAG | 2 | C0017973 | | | 1949 | |
| Ganglion | 2 | C0017067 | 1946 | | 1946 | |
| Ganglion | 2 | C1258666 | 2006 | 1946 | 1946 | 2002 |
| Gas | 2 | C0016204 | 1945 | | 1945 | ? |
| Gas | 2 | C0017110 | | | 1945 | ? |
| Glycoside | 2 | C0007158 | | | 1946 | ? |
| Glycoside | 2 | C0017977 | 1946 | 1946 | 1946 | ? |
| Haemophilus ducreyi | 2 | C0007947 | | | 1977 | |
| Haemophilus ducreyi | 2 | C0018481 | | 1977 | 1978 | 1978 |
| HCl | 2 | C0020259 | | | 1946 | |
| HCl | 2 | C0023443 | 1975 | 1959 | 1954 | 1976 |
| Hemlock | 2 | C0242872 | 2004 | 2002 | 2002 | ? |
| Hemlock | 2 | C0949851 | | | 2002 | ? |
| Heregulin | 2 | C0626201 | | 1992 | 1994 | 1994 |
| Heregulin | 2 | C0752253 | | | 1992 | |
| HGF | 2 | C0021760 | | 1984 | 1984 | ? |
| HGF | 2 | C0062534 | | | 1984 | ? |
| Hip | 2 | C0019552 | | | 1946 | ? |
| Hip | 2 | C0022122 | | | 1947 | ? |

*continued on next column or page*

| Target | N | CUI | ET | LRET | FYO | MS |
|---|---|---|---|---|---|---|
| HIV | 2 | C0019682 | | | 1985 | |
| HIV | 2 | C0019693 | 1987 | 1985 | 1987 | 1987 |
| HPS | 2 | C0079504 | | 1996 | 2000 | 2000 |
| HPS | 2 | C0242994 | | | 1994 | |
| HR | 2 | C0010343 | | 1947 | 1950 | 1992 |
| HR | 2 | C0018810 | | | 1947 | |
| IA | 2 | C0021487 | 1946 | 1946 | 1946 | 1946 |
| IA | 2 | C0022037 | | | 1946 | |
| Ice | 3 | C0020746 | | | 1946 | |
| Ice | 3 | C0025611 | | 1946 | 1946 | 1946 |
| Ice | 3 | C0534519 | 1990 | 1990 | 1991 | 1991 |
| INDO | 2 | C0021246 | 1961 | 1959 | 1963 | 1963 |
| INDO | 2 | C0021247 | | | 1949 | |
| Ion | 2 | C0022023 | | | 1945 | |
| Ion | 2 | C0022024 | 1945 | 1945 | 1946 | 1946 |
| IP | 2 | C0021069 | 2000 | 1997 | 1989 | 2001 |
| IP | 2 | C0021171 | | | 1986 | |
| Iris | 2 | C0022077 | | | 1945 | |
| Iris | 2 | C1001362 | | 1945 | 1946 | 2001 |
| JP | 2 | C0022341 | | | 1946 | |
| JP | 2 | C0031106 | 1946 | 1946 | 1947 | 1983 |
| LABOR | 2 | C0022864 | 1945 | 1945 | 1945 | 1945 |
| LABOR | 2 | C0043227 | | | 1945 | |
| Lactation | 2 | C0006147 | | | 1945 | ? |
| Lactation | 2 | C0022925 | | | 1945 | ? |
| Language | 2 | C0023008 | | | 1946 | |
| Language | 2 | C0033348 | 1986 | 1954 | 1958 | 1985 |
| Laryngeal | 2 | C0023078 | | | 1945 | ? |
| Laryngeal | 2 | C0023081 | | | 1945 | ? |
| Lawsonia | 2 | C0752045 | | | 2000 | |
| Lawsonia | 2 | C1068388 | | | 2002 | 2002 |
| Leishmaniasis | 2 | C0023281 | | | 1945 | |
| Leishmaniasis | 2 | C1548483 | 2005 | 1996 | 1947 | 2000 |
| lens | 3 | C0023308 | 1951 | 1948 | 1952 | 1978 |
| lens | 3 | C0023317 | | 1945 | 1945 | |
| lens | 3 | C0023318 | | | 1945 | |
| Lupus | 3 | C0024131 | | | 1945 | |
| Lupus | 3 | C0024138 | 1945 | 1945 | 1946 | 1946 |
| Lupus | 3 | C0024141 | | | 1945 | |
| lymphogranulomatosis | 2 | C0019829 | | 1945 | 1945 | 1945 |
| lymphogranulomatosis | 2 | C0036202 | | | 1945 | |
| MAF | 2 | C0079786 | | | 1980 | |
| MAF | 2 | C0919482 | 2001 | 1994 | 1998 | 1998 |
| Malaria | 2 | C0024530 | | | 1945 | |
| Malaria | 2 | C0206255 | 1991 | 1988 | 1945 | 1992 |
| MBP | 2 | C0014063 | | | 1973 | |
| MBP | 2 | C0065661 | 1999 | 1998 | 1984 | 2001 |
| MCC | 2 | C0007129 | | | 1988 | |
| MCC | 2 | C0162804 | 1990 | 1989 | 1991 | 1993 |
| Medullary | 2 | C0001629 | | | 1946 | |
| Medullary | 2 | C0025148 | | | 1947 | 1947 |
| MHC | 2 | C0024518 | | | 1978 | |
| MHC | 2 | C0027100 | | 1991 | 1986 | 1994 |
| Milk | 2 | C0026131 | | 1945 | 1945 | ? |
| Milk | 2 | C0026140 | | | 1945 | ? |
| Moles | 2 | C0027960 | | | 1946 | |
| Moles | 2 | C0324740 | | 1946 | 1946 | 1974 |
| MRS | 2 | C0024487 | | 1959 | 1961 | 1961 |
| MRS | 2 | C0025235 | | | 1950 | |
| NBS | 2 | C0027819 | | | 1947 | |
| NBS | 2 | C0398791 | 2003 | 2002 | 2002 | 2006 |
| NEUROFIBROMA... | 2 | C0085113 | | | 1990 | |
| NEUROFIBROMA... | 2 | C0162678 | 1990 | 1990 | 1991 | 1991 |
| NM | 2 | C0025033 | | | 1946 | |
| NM | 2 | C0027972 | 1963 | 1962 | 1946 | 1946 |
| NPC | 2 | C0028587 | | | 1998 | |
| NPC | 2 | C0220756 | 2005 | 2002 | 2006 | 2006 |
| Nurse | 2 | C0006147 | | | 1945 | ? |
| Nurse | 2 | C0028661 | | | 1945 | ? |
| Nursing | 2 | C0006147 | | | 1945 | ? |
| Nursing | 2 | C0028677 | | | 1945 | ? |
| OCD | 2 | C0028768 | | | 1975 | |
| OCD | 2 | C0029421 | 1983 | 1980 | 1984 | 1984 |
| OH | 2 | C0028905 | | | 1946 | ? |
| OH | 2 | C0063146 | | | 1946 | ? |
| Orf | 2 | C0013570 | | | 1980 | |

*continued on next column or page*

| Target | N | CUI | ET | LRET | FYO | MS |
|---|---|---|---|---|---|---|
| Orf | 2 | C0079941 | 1986 | 1985 | 1982 | 1982 |
| ORI | 2 | C0206601 | | | 1993 | |
| ORI | 2 | C0242961 | 1993 | 1993 | 1993 | 1993 |
| PAC | 2 | C0033036 | | | 1995 | |
| PAC | 2 | C0949780 | | 1997 | 2001 | 2001 |
| PAF | 2 | C0032172 | | | 1979 | |
| PAF | 2 | C0037019 | | | 1980 | 1980 |
| Parotitis | 2 | C0026780 | | 1945 | 1945 | ? |
| Parotitis | 2 | C0030583 | | | 1945 | ? |
| PCA | 5 | C0030131 | 1972 | 1971 | 1974 | 1974 |
| PCA | 5 | C0030625 | | | 1957 | |
| PCA | 5 | C0078944 | 1987 | 1986 | 1989 | 1989 |
| PCA | 5 | C0149576 | 1957 | 1957 | 1957 | 1957 |
| PCA | 5 | C0429865 | 1999 | 1998 | 1960 | 2001 |
| PCB | 2 | C0032447 | | | 1971 | ? |
| PCB | 2 | C0033223 | | | 1971 | ? |
| PCD | 2 | C0022521 | | | 1971 | |
| PCD | 2 | C0162638 | | 1988 | 1991 | 1991 |
| PCP | 2 | C0030855 | | | 1972 | ? |
| PCP | 2 | C0031381 | | 1972 | 1972 | ? |
| PEP | 2 | C0031642 | | | 1971 | |
| PEP | 2 | C0135981 | 1978 | 1976 | 1980 | 1980 |
| PHA | 2 | C0030779 | 2002 | 2007 | 1976 | 1976 |
| PHA | 2 | C0031858 | | 1975 | 1975 | |
| Pharmaceutical | 2 | C0013058 | 1963 | 1962 | 1963 | 1963 |
| Pharmaceutical | 2 | C0031336 | | | 1945 | |
| Phosphorus | 2 | C0031705 | | | 1945 | ? |
| Phosphorus | 2 | C0080014 | | 1945 | 1945 | ? |
| Phosphorylase | 2 | C0017916 | | | 1971 | |
| Phosphorylase | 2 | C0917783 | 2005 | 1998 | 1973 | 2001 |
| pI | 2 | C0022171 | | | 1975 | ? |
| pI | 2 | C0812425 | | | 1975 | ? |
| Plague | 2 | C0032064 | | | 1945 | |
| Plague | 2 | C0032066 | 1959 | 1957 | 1946 | 1960 |
| Plaque | 2 | C0011389 | | | 1950 | ? |
| Plaque | 2 | C0333463 | | | 1950 | ? |
| Platelet | 2 | C0005821 | | 1945 | 1945 | ? |
| Platelet | 2 | C0032181 | | | 1945 | ? |
| Pleuropneumonia | 2 | C0026934 | 1945 | 1945 | 1945 | ? |
| Pleuropneumonia | 2 | C0032241 | | | 1945 | ? |
| POL | 2 | C0017360 | | 1986 | 1989 | 1989 |
| POL | 2 | C0032356 | | | 1946 | |
| posterior pituitary | 2 | C0032009 | | | 1946 | |
| posterior pituitary | 2 | C0032017 | | 1946 | 1947 | 1946 |
| Potassium | 2 | C0032821 | | | 1945 | |
| Potassium | 2 | C0162800 | 1990 | 1989 | 1948 | 1992 |
| PR | 2 | C0034044 | | | 1945 | |
| PR | 2 | C0034833 | 1972 | 1972 | 1973 | 1973 |
| Projection | 2 | C0016538 | | 1970 | 1970 | ? |
| Projection | 2 | C0033363 | | | 1970 | ? |
| PVC | 2 | C0032624 | | | 1974 | |
| PVC | 2 | C0151636 | | 1991 | 1988 | 1992 |
| RA | 3 | C0002893 | | 1945 | 1946 | ? |
| RA | 3 | C0003873 | | | 1945 | ? |
| RA | 3 | C0034625 | | | 1945 | ? |
| Radiation | 2 | C0851346 | | | 1945 | |
| Radiation | 2 | C1522449 | | | 1946 | 1946 |
| RB | 2 | C0035335 | | | 1947 | |
| RB | 2 | C0035930 | | 1947 | 1951 | 1951 |
| RBC | 2 | C0014772 | | | 1945 | ? |
| RBC | 2 | C0014792 | | | 1945 | ? |
| rDNA | 2 | C0012931 | | | 1976 | |
| rDNA | 2 | C0012933 | 1980 | 1978 | 1981 | 1981 |
| Respiration | 2 | C0035203 | | | 1945 | ? |
| Respiration | 2 | C0282636 | | | 1945 | ? |
| Retinal | 2 | C0035298 | | 1945 | 1945 | ? |
| Retinal | 2 | C0035331 | | | 1945 | ? |
| Root | 2 | C0040452 | | 1945 | 1945 | ? |
| Root | 2 | C0242726 | | | 1945 | ? |
| RSV | 2 | C0035236 | | 1957 | 1960 | 1960 |
| RSV | 2 | C0086943 | | | 1955 | |
| SARS | 2 | C1175175 | | | 2002 | |
| SARS | 2 | C1175743 | 2002 | 2002 | 2002 | 2002 |
| SARS-assoc... | 2 | C1175175 | | | 2002 | |
| SARS-assoc... | 2 | C1175743 | 2002 | 2002 | 2002 | 2002 |
| SCD | 2 | C0002895 | | | 1946 | |

*continued on next column or page*

| Target | N | CUI | ET | LRET | FYO | MS |
|---|---|---|---|---|---|---|
| SCD | 2 | C0085298 | 1988 | 1987 | 1950 | 1989 |
| Schistosoma... | 2 | C0036319 | | | 1971 | |
| Schistosoma... | 2 | C0036330 | | 1981 | 1977 | 1985 |
| SLS | 2 | C0037231 | | 1987 | 1991 | 1991 |
| SLS | 2 | C0037506 | | | 1971 | |
| Sodium | 2 | C0037473 | | | 1945 | |
| Sodium | 2 | C0037570 | 1945 | 1945 | 1945 | 1945 |
| SPR | 2 | C0164209 | | | 1981 | |
| SPR | 2 | C0597731 | 1996 | 1994 | 1998 | 1998 |
| SS | 2 | C0039101 | | | 1948 | |
| SS | 2 | C0085077 | 1990 | 1960 | 1964 | 1990 |
| Staph | 2 | C0038160 | | | 1945 | 1945 |
| Staph | 2 | C0038170 | | | 1945 | |
| STEM | 2 | C0162731 | | | 1992 | |
| STEM | 2 | C0242767 | | 1992 | 1994 | 1994 |
| Sterilization | 2 | C0038280 | | 1945 | 1945 | ? |
| Sterilization | 2 | C0038288 | | | 1945 | ? |
| Strep | 2 | C0038395 | | 1945 | 1945 | 1945 |
| Strep | 2 | C0038402 | | | 1945 | |
| Synapsis | 2 | C0039062 | | | 1950 | |
| Synapsis | 2 | C0598501 | 1998 | 1950 | 1951 | 1951 |
| TAT | 3 | C0017375 | 1988 | 1985 | 1989 | 1989 |
| TAT | 3 | C0039341 | 1983 | 1982 | 1985 | 1985 |
| TAT | 3 | C0039756 | | | 1975 | |
| Tax | 2 | C0039371 | | | 1975 | |
| Tax | 2 | C0144576 | 1992 | 1989 | 1983 | 1993 |
| TEM | 2 | C0040975 | | | 2004 | |
| TEM | 2 | C0678118 | | | 2002 | |
| THYMUS | 3 | C0040112 | 1948 | 1946 | 1949 | 1949 |
| THYMUS | 3 | C0040113 | | | 1946 | |
| THYMUS | 3 | C1015036 | | 1946 | 1946 | |
| TLC | 2 | C0008569 | | 1959 | 1959 | ? |
| TLC | 2 | C0040509 | 1974 | 1972 | 1959 | ? |
| TMJ | 2 | C0039493 | | | 1946 | ? |
| TMJ | 2 | C0039496 | | | 1946 | ? |
| TMP | 2 | C0040079 | | 1972 | 1975 | 1975 |
| TMP | 2 | C0041041 | | | 1970 | |
| TNC | 2 | C0076088 | 1983 | 1982 | 1985 | 1985 |
| TNC | 2 | C0077400 | | | 1980 | |
| TNT | 2 | C0041070 | | | 1982 | 1982 |
| TNT | 2 | C0077404 | | | 1981 | |
| Tolerance | 2 | C0013220 | | | 1946 | ? |
| Tolerance | 2 | C0020963 | | 1946 | 1946 | ? |
| tomography | 2 | C0040395 | | | 1947 | ? |
| tomography | 2 | C0040405 | | | 1947 | ? |
| Torula | 2 | C0010414 | | | 1945 | ? |
| Torula | 2 | C0010415 | | | 1945 | ? |
| TPA | 2 | C0032143 | 1983 | 1982 | 1982 | 1985 |
| TPA | 2 | C0039654 | | | 1975 | |
| TPO | 2 | C0021965 | 1974 | 1974 | 1975 | 1975 |
| TPO | 2 | C0040052 | | | 1974 | |
| TRF | 2 | C0021759 | | | 1980 | 1980 |
| TRF | 2 | C0040162 | | | 1968 | |
| TSF | 2 | C0021756 | 1976 | 1974 | 1977 | 1977 |
| TSF | 2 | C0040052 | | | 1974 | |
| TYR | 2 | C0041484 | | | 1945 | ? |
| TYR | 2 | C0041485 | | | 1945 | ? |
| US | 2 | C0041618 | 1971 | 1964 | 1945 | 1966 |
| US | 2 | C0041703 | | | 1945 | |
| Ventricles | 2 | C0007799 | | | 1945 | ? |
| Ventricles | 2 | C0018827 | | | 1945 | ? |
| veterinary | 2 | C0042615 | | | 1945 | |
| veterinary | 2 | C0206212 | | 1959 | 1963 | 1993 |
| Wasp | 2 | C0043041 | | | 1975 | |
| Wasp | 2 | C0258432 | 1993 | 1991 | 1994 | 1994 |
| WBS | 2 | C0004903 | | | 1982 | |
| WBS | 2 | C0175702 | 1994 | 1991 | 1995 | 1995 |
| WT1 | 2 | C0027708 | | | 1946 | |
| WT1 | 2 | C0148873 | 1991 | 1989 | 1991 | 1991 |
| Yellow Fever | 2 | C0043395 | | 1945 | 1945 | ? |
| Yellow Fever | 2 | C0301508 | | | 1945 | ? |

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*Section 7*

☐ A2. Did you discuss any potential risks of your work?
*Not applicable. Left blank.*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*Abstract and Section 1*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*Section 1 and 4 and 5*

☑ B1. Did you cite the creators of artifacts you used?
*Section 4 and 5*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*Section 4*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*Section 4 and 5*

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*Not applicable. The data used in this research is a secondary data which was previously published. The data source files were taken from NML and is made of biomedical scientific publications.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Section 7*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*Section 4*

### C ☑ Did you run computational experiments?

*Section 4 and 5*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*Section 4 and 7*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*Section 5*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*Section 4 and 5*

**D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Not applicable. Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Not applicable. Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*