Infinite SCAN: An Infinite Model of Diachronic Semantic Change

Seiichi Inoue[1], **Mamoru Komachi**[1], **Toshinobu Ogiso**[2], **Hiroya Takamura**[3] and **Daichi Mochihashi**[4]

[1]Tokyo Metropolitan University [2]The National Institute for Japanese Language and Linguistics [3]The National Institute of Advanced Industrial Science and Technology [4]The Institute of Statistical Mathematics

inoue-seiichi@ed.tmu.ac.jp komachi@tmu.ac.jp togiso@ninjal.ac.jp takamura.hiroya@aist.go.jp daichi@ism.ac.jp

Abstract

In this study, we propose a Bayesian model that can jointly estimate the number of senses of words and their changes through time. The model combines a dynamic topic model on Gaussian Markov random fields (Frermann and Lapata, 2016) with a logistic stick-breaking process that realizes the Dirichlet process. In the experiments, we evaluated the proposed model in terms of interpretability, accuracy in estimating the number of senses, and tracking their changes using both artificial data and real data. We quantitatively verified that the model behaves as expected through evaluation using artificial data. Using the CCOHA corpus, we showed that our model outperforms the baseline model and investigated the semantic changes of several well-known target words.

1 Introduction

Words exhibit a range of senses depending on the context in which they are used. These senses can also change over time (Blank and Koch, 1999; Aitchison, 2001). For example, the word *cute* appeared in the early 18th century, which originally meant *clever* or *keen-witted*. By the late 19th century it signified *cunning*, and today, *cute* means *attractive*, *pretty*, or *sweet* (Stevenson, 2010; Frermann and Lapata, 2016). Automatically capturing these semantic changes is an academic contribution to the fields of lexicology and linguistics (Voyles, 1973; Williams, 1976).

In recent years, many methods have been proposed for the detection of semantic changes using distributional methods (Kutuzov et al., 2018). They include word embedding-based methods with alignment of word embedding spaces at different times (Kim et al., 2014; Kulkarni et al., 2015; Hamilton et al., 2016b), without alignment (Yao et al., 2018; Dubossarsky et al., 2019; Aida et al., 2021), and using probabilistic frameworks (Bamler and Mandt, 2017; Rudolph and Blei, 2018). Word embedding-based methods describe semantic change by changes of surrounding words in semantic space. However, the learned embeddings themselves cannot account for the existence of each sense and its relative importance. In contrast, several methods have addressed these issues using a topic model architecture (Frermann and Lapata, 2016; Emms and Jayapal, 2016). These probabilistic models estimate the latent senses explicitly and consider their changes, unlike word embedding models. However, these models have a critical problem in that the number of senses is given and fixed, even though it will vary for each word, which harms the modeling of semantic change. In addition, the number of senses is rarely apparent beforehand, thus it is difficult to set it *a priori*. A recent method of clustering contextualized word embeddings obtained from BERT can similarly track sense changes (Giulianelli et al., 2020), but it does not take time evolution into account, and cannot estimate the number of senses and semantic changes jointly.

Therefore, to address these limitations, we propose a model that can automatically estimate the number of senses and simultaneously capture semantic changes by extending the model proposed by Frermann and Lapata (2016). To this end, we combined a dynamic topic model on Gaussian Markov random fields (GMRF) with a logistic stick-breaking process (Ren et al., 2011) to realize a Dirichlet process in latent Euclidean space.[1] Here, our work can answer the question of how many senses the word has in the context of modeling semantic change, which can be applicable to lexicography. In our experiments, we verified the performance of the proposed model in terms of the estimation accuracy of the number of senses and sense change on artificial data. Then, we evaluated the model performance using real data and

[1]Source code is available at https://github.com/seiichiinoue/iscan.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1605–1616

Year	Example	Snippet
1853	The driver made room for the trunk on the top of the **coach**.	{driver, make, room, trunk}
1900	The chair passed the **coach**, the horses proceeding at a walk.	{chair, pass, horse, proceed, walk}
1949	Tell him if I start **coaching**, it'll be as a head **coach** at a top school.	{tell, start, coach, head, top, school}
2003	Football **coach** and other top school officials have been interviewed.	{football, top, school, official, interview}

Figure 1: Example snippets for input to SCAN for the target word *coach*. The snippets were obtained from the preprocessing step described in Section 6.1.

analyzed the semantic changes of several words. The contributions of this study can be summarized as follows:

- We combine a dynamic topic model on GMRF with a Dirichlet process to propose a model that can jointly estimate the number of senses and semantic change of words.
- We quantitatively and qualitatively show that the proposed model can correctly estimate the number of word senses and semantic changes and outperforms baseline models.

2 Background

2.1 Dynamic Bayesian model of sense change

Frermann and Lapata (2016) proposed a dynamic Bayesian model that captures the diachronic word **S**ense **Ch**A**N**ge (SCAN). In the SCAN framework, one model is constructed for each target word w. The input is a set of snippets, *i.e.* short documents, consisting of context words $\boldsymbol{c}_d = \{c_{d,1}, \dots c_{d,I}\}$ with length I of a sentence containing the target word w, and time label of the year in which each sentence appeared. An example of snippets is shown in Figure 1.

In SCAN, the set of snippets at time $t \in \{1 \dots T\}$ is modeled by unigram mixtures at each time point:

- K-dimensional multinomial distribution ϕ_t (sense distribution) over the senses.
- V-dimensional multinomial distribution $\psi_{t,k}$ (sense–word distribution) over the words for each word sense k.

Also, a Gaussian distribution is assumed for each prior distribution. ϕ is obtained by transforming a sampled vector as follows:

1. Draw a K-dimensional vector $\boldsymbol{\alpha}$ from the multivariate Gaussian distribution.
2. Project this vector to a $K-1$-dimensional simplex, using the softmax transformation $\phi_k = \exp(\alpha_k)/\sum_{k=1}^{K}\exp(\alpha_k)$.

ψ can be obtained in a similar way. Note that K is assumed to be given (a parameter to be set *a priori*), which is a severe problem in practice. Then, they define the first-order intrinsic Gaussian Markov random field (iGMRF) (Rue and Held, 2005) for the prior distribution so that the sense distribution ϕ and sense–word distribution ψ change through time. The iGMRF is a prior distribution such that the value at any location is similar to that of neighbors (graphically shown in Figure 2).[2] For a real vector $\boldsymbol{x} = (x_1, x_2, \dots, x_T)$ and Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the iGMRF is defined as follows:

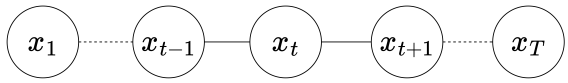


Figure 2: A linear chain iGMRF.

$$x_t \,|\, \boldsymbol{x}_{-t}, \kappa \sim \mathcal{N}\left(\frac{1}{2}(x_{t-1} + x_{t+1}), \frac{1}{\kappa}\right), \qquad (1)$$

where $\boldsymbol{x}_{-t}$ is a set of $\boldsymbol{x}$ except for x_t and κ is a precision parameter. The Gaussian distribution, which is a prior distribution of the sense distribution and the sense–word distribution, has precision parameters κ_ϕ and κ_ψ to control the degree of change, respectively. In particular, the precision parameter of the sense distribution, κ_ϕ, should be estimated from the data because the "speed" of sense change varies depending on the target word w.

Based on the above definition, the generative model of SCAN is described as follows, where $\mathrm{Ga}(a, b)$ denotes the gamma distribution and $\mathrm{Mult}(\theta)$ denotes the multinomial distribution.

1. Draw $\kappa_\phi \sim \mathrm{Ga}(a, b)$
2. For time interval $t = 1 \dots T$
   (a) Draw a sense distribution
      i. $\alpha_t \,|\, \boldsymbol{\alpha}_{-t}, \kappa_\phi$
      $\sim \mathcal{N}\left(\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}), \kappa_\phi^{-1}\right)$
      ii. $\phi_t = \mathrm{Softmax}(\alpha_t)$
   (b) For sense $k = 1 \dots K$
      i. Draw a sense–word distribution

[2] iGMRF can be viewed as a special case of a Gaussian process in which the kernel is restricted to adjacent times.

A. $\beta_{t,k} \mid \boldsymbol{\beta}_{-t}, \kappa_\psi$
$\sim \mathcal{N}\left(\frac{1}{2}(\beta_{t-1,k} + \beta_{t+1,k}), \kappa_\psi^{-1}\right)$

B. $\psi_{t,k} = \mathrm{Softmax}(\beta_{t,k})$

(c) For snippet $d = 1 \ldots D$

i. Draw a sense $z_d \sim \mathrm{Mult}(\phi_t)$

ii. For context position $i = 1 \ldots I$

A. Draw a word $c_{d,i} \sim \mathrm{Mult}(\psi_{t,z_d})$

2.2 Logistic stick-breaking process

The Dirichlet process (Ferguson, 1973; Antoniak, 1974) is an infinite-dimensional generalization of Dirichlet distribution and generates an infinite-dimensional multinomial distribution. The stick-breaking process (SBP) (Sethuraman, 1994) is an example of its realization. In the SBP representation, a probability distribution G that follows the Dirichlet process $\mathrm{DP}(\alpha, G_0)$ is generated as follows:

$$\pi_k = v_k \prod_{j=1}^{k-1}(1 - v_j), \; v_k \sim \mathrm{Be}(1, \alpha) \tag{2}$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k), \; \theta_k \sim G_0, \tag{3}$$

where $\mathrm{Be}(1, \alpha)$ denotes a beta distribution. First, the probability for the k-th category π_k is determined by recursively breaking a stick of length one, which is the sum of the probabilities. Then, the delta function $\delta(\theta_k)$ is set at a location θ_k sampled from the base measure G_0.

Ren et al. (2011) proposed the logistic stick-breaking process (LSBP), which realizes a Dirichlet process in the same way as the original SBP by transforming a random variable with logistic function $\sigma(x) = 1/(1 + e^{-x})$, where each class is associated with a certain covariate[3] $x \in \mathbb{R}$. Let x_k be a random variable that follows a Gaussian distribution for each category; the LSBP generates the probability distribution G_x as follows:

$$\pi(x_k) = \sigma(x_k) \prod_{j=1}^{k-1}(1 - \sigma(x_j)), \tag{4}$$

$$G_x = \sum_{k=1}^{\infty} \pi(x_k)\delta(\theta_k). \tag{5}$$

3 Proposed Method

3.1 Infinite SCAN

We propose an infinite model of diachronic semantic change: Infinite SCAN that extends the architecture of SCAN (introduced in Section 2.1) using LSBP (introduced in Section 2.2) to automatically estimate the number of word senses for each target word. The graphical model of Infinite SCAN is shown in Figure 3. The proposed model extends iGMRF over the time direction in the semantic distribution with LSBP that realizes a Dirichlet process. This makes the sense distribution practically infinite-dimensional, and the number of senses, S_w, appropriate for the target word w can be automatically estimated from the corpus. Here, we note that our idea is similar to the model linking Gaussian process and Dirichlet process for spatial modeling (Duan et al., 2007) and the method using Gaussian process and Pitman-Yor process for image segmentation (Sudderth and Jordan, 2008).

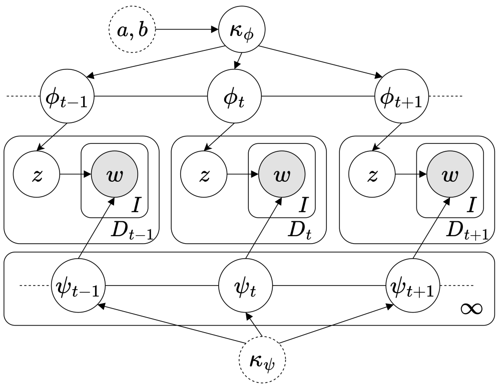


Figure 3: Graphical model of Infinite SCAN for three time steps $\{t-1, t, t+1\}$. Observations are shown as gray nodes, latent variables as clear nodes, and constants as dashed nodes. Adapted from Frermann and Lapata (2016) with the number of senses being infinite.

In Infinite SCAN, as in SCAN, one model is constructed for each target word w. The input is a set of snippets consisting of context words $\boldsymbol{c}_d = \{c_{d,1}, \ldots c_{d,I}\}$ of a sentence containing the target word w, and time label of the year in which each sentence appeared. The set of snippets appearing at time t is represented by the sense distribution ϕ_t and sense–word distributions $\psi_{t,k}$, each following a first order iGMRF:

$$\alpha_t \mid \boldsymbol{\alpha}_{-t}, \kappa_\phi \sim \mathcal{N}\left(\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}), \kappa_\phi^{-1}\right). \tag{6}$$

Here, we modify the generative process of the sense distribution ϕ_t using LSBP as follows, so that the number of senses S_w, which depends on the target word w, can be automatically estimated from the corpus:

$$\phi_{t,k} = \sigma(\alpha_{t,k}) \prod_{j=1}^{k-1}(1 - \sigma(\alpha_{t,j})) \tag{7}$$
$$(k = 1, \ldots, K),$$

[3]In the case of SCAN, the random variables that follow a Gaussian distribution are associated with each sense and each word in the sense.

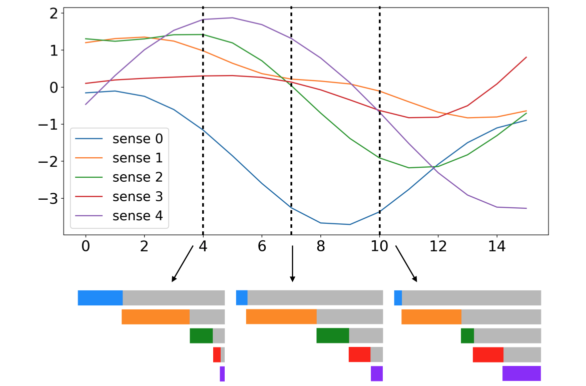


Figure 4: LSBP transformation of random variables following a Gaussian distribution in a sense distribution.

where K is the maximum number of senses considered. Here, LSBP can generate the infinite-dimensional multinomial distribution. However, in practice, if the dimension of the word sense is sufficient to represent the data, using high-dimensional distributions is not necessary. Thus, in this study, we set the maximum number of senses $K = 8$ by referring to the previous study (Frermann and Lapata, 2016).[4] Figure 4 illustrates the LSBP transformation of the sense distribution. The horizontal axis denotes time, and the vertical axis denotes the scale of the random variable following a Gaussian distribution. The probability of each word sense $\{\phi_{t,1}, \ldots, \phi_{t,K}\}$ is obtained by LSBP transformation over the set of random variables $\{\alpha_{t,1}, \ldots, \alpha_{t,K}\}$.

In SCAN, the precision parameter κ_ϕ of the sense distribution is shared across all senses. This is because the sense distribution ϕ_t at time t is constructed by a softmax transformation, which normalizes the distribution by considering all of the senses, so that the variance of all senses is within a certain scale. By contrast, in Infinite SCAN, the sense distribution ϕ_t at time t is constructed by LSBP transformation. In the LSBP transformation, sense k is transformed by a sigmoid function independently of the other senses, such that the scale of the Gaussian random variable corresponding to each sense will differ. Therefore, in the proposed model, κ_ϕ should not be shared across all senses $k \in \{1 \ldots K\}$ unlike SCAN. Instead, we assume and estimate a different precision $\kappa_\phi^{(k)}$ for each Gaussian random variable corresponding to sense α_k.

[4] Preliminary experiments indicated that there are very few words that have more than eight senses.

3.2 Markov Chain Monte Carlo (MCMC) estimation

To estimate the parameters of Infinite SCAN, we used a blocked Gibbs sampler. The parameters to be estimated in Infinite SCAN are (a) the sense assignment z for each snippet, (b) the parameters defined by iGMRF for the semantic distribution α (unnormalized ϕ), (c) sense–word distributions β (unnormalized ψ), and (d) the precision parameter κ_ϕ of the sense distribution following the gamma distribution. In the model estimation, each parameter is sampled from its posterior distribution given the other parameters. The pseudo-code of the MCMC algorithm is shown in Appendix A. Each parameter basically follows Frermann and Lapata (2016), but we changed some parameters; see Section 4 for details.

Sense assignments of snippet The sense assignments of the d-th snippet, z_d, are sampled from the following posterior distribution under the current model parameters ϕ and ψ:

$$\begin{aligned} p(z_d \mid \boldsymbol{w}, t, \phi, \psi) &\propto p(z_d \mid t) p(\boldsymbol{w} \mid t, z_d) \\ &= \phi_{z_d}^{(t)} \prod_{w \in \boldsymbol{w}} \psi_w^{(t, z_d)}. \end{aligned} \tag{8}$$

Sense distribution Because the sense distribution follows a Gaussian distribution, it is not conjugate to the multinomial distribution. Thus, straightforward parameter sampling, such as Dirichlet-multinomial, does not apply. Linderman et al. (2015) proposed a Gibbs sampling for parameters of a multinomial distribution, modeled with a Gaussian prior and the LSBP transformation, by using a Pólya-gamma auxiliary variable (Polson et al., 2013). This approach is used in this study. The posterior distribution of α is computed as follows:

$$\begin{aligned} & p(\alpha_t \mid z, \boldsymbol{\alpha}_{-t}, \omega) \\ & \quad \propto \mathcal{N}\left(\omega^{-1} f(c) \mid \alpha_t\right) \mathcal{N}\left(\alpha_t \mid \boldsymbol{\alpha}_{-t}, \kappa_\phi^{-1}\right) \\ & \quad \propto \mathcal{N}\left(\alpha_t \mid \tilde{\mu}, \tilde{\kappa}_\phi^{-1}\right). \end{aligned} \tag{9}$$

Here, ω is an auxiliary variable that is sampled from Pólya-gamma distribution $\omega \mid z, \alpha_t \sim \mathrm{PG}(N(c_k), \alpha_t)$, where c_k denotes the number of snippet belonging to k-th sense and $N(c_k) = \sum_k c_k - \sum_{j<k} c_j$. Also, the mean and precision of the posterior distribution are computed as $\tilde{\mu} = (f(c_k) + \mu_k \kappa_\phi)/\tilde{\kappa}_\phi$ and $\tilde{\kappa}_\phi = \omega_k + \kappa_\phi$, where $f(c_k) = c_k - N(c_k)/2$.

Sense–word distribution The sense–word distribution, as the sense distribution, follows Gaussian distribution; thus, cannot be applied to such Dirichlet-multinomials. Mimno et al. (2008) proposed a Gibbs sampling for parameters of multinomial distribution modeled with a Gaussian prior and softmax transformation; we estimate the parameters using this approach. Let the number of snippets be D and the snippet length be N_d; the posterior distribution of β is as follows:

$$p(\beta_t \mid z, \boldsymbol{\beta}_{-t}, \kappa_\psi^{-1}) \propto \tag{10}$$
$$\prod_{d=1}^{D} \left(\prod_{n=1}^{N_d} \frac{\exp(\beta_{w_n}^{(t,z_d)})}{\sum_{v=1}^{V} \exp(\beta_v^{(t,z_d)})} \right) \mathcal{N}(\beta_t \mid \boldsymbol{\beta}_{-t}, \kappa_\psi^{-1}).$$

Precision parameter The precision parameter of the Gaussian distribution given its mean follows a gamma posterior distribution. With the shape parameter of the gamma distribution as a and the scale parameter as b, the posterior distribution of κ_ϕ is as follows:

$$\begin{aligned} &p(\kappa_\phi^{(k)} \mid \alpha_k, a, b) \\ &= \mathrm{Ga}\left(a + \frac{T}{2}, b + \frac{1}{2} \sum_{t=1}^{T} (\alpha_{t,k} - \bar{\alpha}_k) \right) \end{aligned} \tag{11}$$

where $\bar{\alpha}_k = 1/T \sum_t \alpha_{t,k}$ is the mean of α corresponding to sense k.

4 Experimental Settings

In the following experiments, we split the data (*i.e.* set of snippets) into period of time slice 1 to T. The time interval for year-labeled data was set to $\Delta t = 20$ years. We used vocabulary with frequency larger than 10.

As model settings, we set the maximum number of senses to $K = 8$, the initial sense precision parameter as $\kappa_\phi^{(k)} = 4$ for each sense k, and the gamma parameters as $a = 7$ and $b = 3$. We set a relatively large value for the word precision parameter, with $\kappa_\psi = 100.0$, following Perrone et al. (2019). This is because we want to capture the sense change of the target word as much as possible in terms of a "shift of the sense distribution" rather than a "shift of the sense–word distribution." [5] Finally, we ran the Gibbs sampler for 2,000 iterations[6] and resampled $\kappa_\phi^{(k)}$ for each sense k after every 50 iterations, starting from iteration 150.

[5]In the former case, sense change is explained only by the shift of the sense–word distribution, resulting in incorrect detection of sense change, making it difficult to estimate the number of senses.

#Senses	SCAN $K=5$	SCAN $K=8$	Infinite SCAN
1	1.523	1.997	**0.468**
2	0.335	0.578	**0.039**
3	0.216	0.735	**0.030**
4	0.212	0.150	**0.061**
5	0.004	0.017	**0.004**

Table 1: Kullback-Leibler divergence (lower is better) between actual sense distribution and sense distribution estimated by each model for the artificial data.

5 Experiments using Artificial Data

Prior to the experiments using real data, we evaluated the proposed model on artificial data to validate the model for correctly estimating arbitrary changes and the number of senses.

5.1 Dataset

When generating artificial data, we first sampled the curve of sense change from a Gaussian process and then transformed it by the LSBP at each time point to obtain the multinomial sense distribution, as shown in Figure 4. Next, we used a Zipfian distribution to generate the sense–word distribution, *i.e.*,

$$f(k \mid s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s} \tag{12}$$

to reproduce Zipf's law of texts (Zipf, 1945) observed in real data. Finally, we randomly generated a set of artificial snippets using these sense distributions and sense–word distributions. In this experiment, we fixed the number of time points in the artificial data at $T = 10$, the original vocabulary size[7] at $V = 5{,}000$, the snippet length at $I = 10$, and changed the number of word senses[8] from $S_w = 1$ to 5. The following example shows the generated snippets with sense $k = 0$ and 3.

$k=0$: `a e y y ag cy et gw xa ahy`

$k=3$: `dkj dkj dkj dkp dlq dms dmy dpy eaj esv`

Here, words are actually expressed as integers from 0 to 5,000, but we use alphabet (base-26 numbers) for interpretability in this example.

[6]The computational time is proportional to sample size, and it took 5 minutes to converge on data with 10,000 samples.

[7]The mode vocabulary size of the corpus used in the experiments was approximately 5,000.

[8]Most polysemous words have five or less senses (Biemann and Nygaard, 2010).

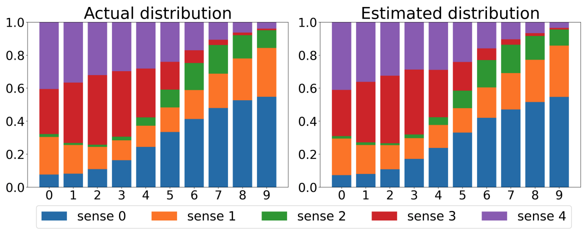


Figure 5: Actual distribution and estimated distribution for the artificial data for the number of senses $S_w = 5$. Senses of the estimated distribution are sorted according to the actual distribution for interpretability.

5.2 Results

Table 1 shows a comparison of the estimation results of SCAN and Infinite SCAN using artificial data with the number of senses ranging from $S_w = 1$ to 5. To quantitatively measure whether the sense distribution is correctly represented, we used the Kullback-Leibler divergence on the space $\mathbb{R}^{T\times K}$ between the estimated and actual sense distributions as an indicator. The results show that Infinite SCAN outperforms SCAN that does not automatically estimate the number of word senses.

Figure 5 shows an example of the actual distribution and the estimation results of Infinite SCAN for the artificial data with the number of word senses $S_w = 5$. Here, the sense of the most dominant word in the estimated sense-word distribution for each sense k is shown in the legend for simplicity.

The figure shows that the proposed model captures the sense change almost precisely and estimates the true number of senses correctly.

6 Experiments using Real Data

In the experiments on real data, we firstly evaluated interpretability of the model output in the same manner as topic models (Section 6.3). We further evaluated the quality of the estimation results of the proposed model in terms of the estimation of the number of senses (Section 6.4) and the sense change (Section 6.5).

6.1 Dataset

We used the Clean Corpus of Historical American English (CCOHA) (Alatrash et al., 2020), a large collection of texts from various genres covering the years 1810–2009. As a preprocessing, we tokenized, lemmatized, and removed stopwords. Moreover, we performed part-of-speech tagging using the Natural Language Toolkit (NLTK) (Bird et al., 2009) and extracted only nouns, verbs, and adjectives. After the above preprocessing, we created the target word-specific input corpora, *i.e.* snippets, for our models. They consisted of a set of context words $\boldsymbol{c}_d$ before and after the point where the target word w appeared in the corpus, with a symmetric window width of ± 5 words.

For quantitative evaluation (Sections 6.3 and 6.4), out of the 4,193 sense-tagged words (noun and verbs), we randomly selected 120 words with five or less senses [9] from OntoNotes (Hovy et al., 2006). The statistics of the randomly selected words are shown in Appendix B. For qualitative evaluation (Section 6.5), we selected the following three words: *coach* (Aida et al., 2021), *record* (Hamilton et al., 2016b), and *power* (Frermann and Lapata, 2016), based on previous studies. The statistics of these words are shown in Appendix C.

6.2 Models

We compared the proposed model with the following previous methods in addition to SCAN.

HDP-LDA We used a Bayesian nonparametric version of LDA (Blei et al., 2003) using hierarchical Dirichlet process (HDP-LDA) (Teh and Jordan, 2010) that can automatically estimate the number of topics as one of the baseline models. Unlike the proposed model, HDP-LDA does not model the temporal evolution of texts. In addition, since a document is represented by a mixture of topics rather than one document with one topic, we used "the set of snippets at time t" instead of "a snippet" as the input unit to estimate the number of senses and semantic changes. The number of topics was initially set to $K = 8$, but estimated adaptively.

BERT + clustering We also compared our model with BERT (Devlin et al., 2019), a method that uses contextualized word embeddings.[10] In line with the previous study (Giulianelli et al., 2020), we first obtained contextualized word embeddings for every sentence containing the target word. Then, we estimated the number of senses and semantic changes by clustering the contextualized word embeddings across time periods. For the clustering methods, we used both k-means (Lloyd, 1982) and DBSCAN (Ester et al., 1996).

[9] Most polysemous words have five or less senses (Biemann and Nygaard, 2010) as noted in Section 5.

[10] We used base-uncased version of the pre-trained model available at `https://github.com/huggingface/transformers`.

Model	Coherence	Diversity
HDP-LDA	0.125	0.821
SCAN ($K\!=\!5$)	0.178	0.716
SCAN ($K\!=\!8$)	0.171	0.654
Infinite SCAN	**0.181**	**0.885**

Table 2: Sense coherence and diversity (higher is better) computed with the baseline models and Infinite SCAN for 120 target words randomly selected from OntoNotes.

6.3 Evaluation of interpretability

Metrics For the evaluation of interpretability of model output, we use sense coherence and diversity to compare the baseline models with the proposed model, following Dieng et al. (2020). Sense coherence C is defined as the average similarity between two words in representative words of each sense:

$$C = \sum_{k=1}^{K} \frac{\eta}{45} \sum_{i=1}^{10} \sum_{j=i+1}^{10} f(w_i^{(k)}, w_j^{(k)}), \qquad (13)$$

where $w_i^{(k)}$ is the i-th most probable word in sense k, and η is the normalization constant of the word sense. Although Dieng et al. (2020) simply set $\eta = 1/K$, in this study, we set $\eta = p(k)$ using the sense probability $p(k)$ of each sense k to legitimately evaluate the sense–word distribution of the sense of very small probability. $f(w, w')$ denotes the similarity of words in the semantic space; we use the cosine similarity calculated through word2vec[11] (Mikolov et al., 2013). We define sense diversity as the proportion of words with no overlap among the top 10 words in all senses. Diversity close to 0 indicates redundancy, and diversity close to 1 indicates less overlap.

Results Table 2 shows the scores calculated using the estimation results of HDP-LDA, SCAN, and Infinite SCAN[12] for the target words. Here, we note that the initial number of senses is fixed for both HDP-LDA and SCAN to match the setting of the proposed model estimating without knowing the number of senses. The results show that Infinite SCAN outperforms the baseline models on both metrics. An improvement in coherence means that the semantic consistency of representative words of the estimated sense–word distribution is high, indicating a high degree of interpretability of estimated sense. Regarding the improvement in diversity, semantic overlap of representative words in the estimated sense–word distribution is smaller, meaning that the number of senses is estimated at an appropriate granularity.

[11] We used the pre-trained model available at https://code.google.com/archive/p/word2vec/.

[12] BERT is not a probabilistic generative model and cannot automatically extract representative words, so it was not comparable in this experiment.

Model	Accuracy	Pearson Corr.
HDP-LDA	0.258	0.019
BERT + k-means	0.217	0.026
BERT + DBSCAN	0.125	−0.070
SCAN ($K\!=\!5$)	0.158	0.141
SCAN ($K\!=\!8$)	0.000	0.087
Infinite SCAN	**0.358**	**0.474**

Table 3: Prediction results for the number of senses computed with the baseline models and Infinite SCAN for 120 target words randomly selected from OntoNotes.

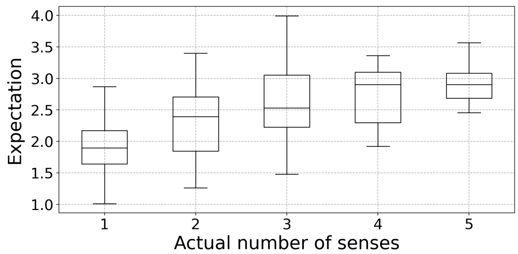


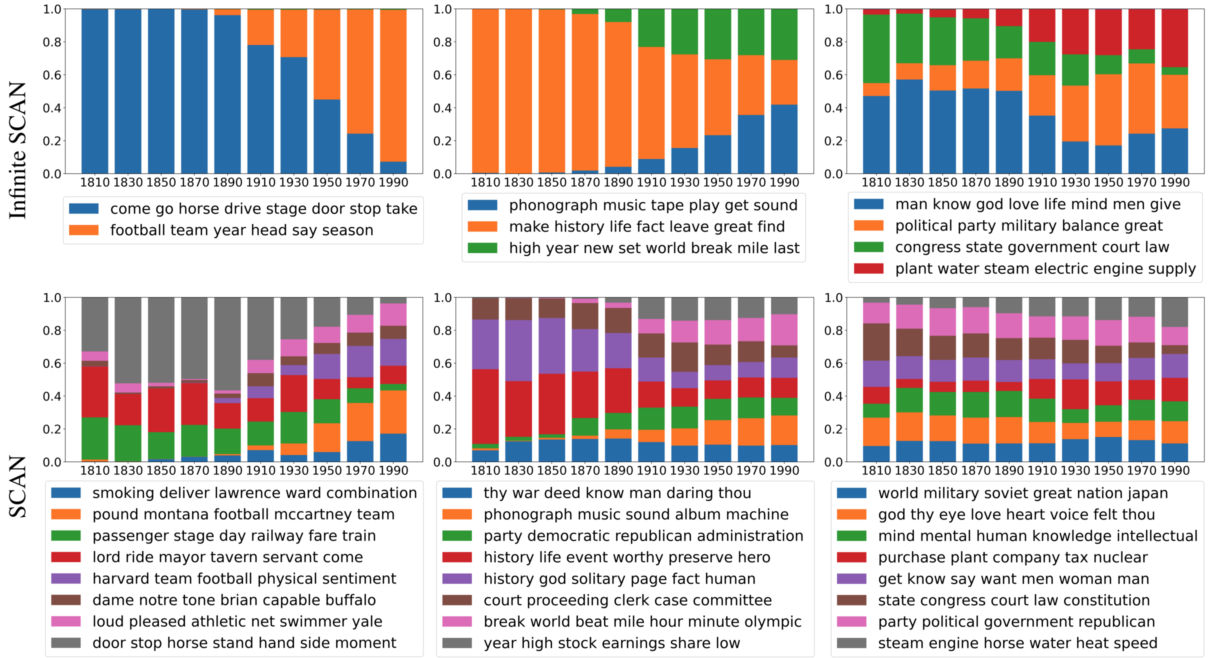
Figure 6: Correlation between actual and expectation of the number of senses computed with Infinite SCAN for 120 target words randomly selected from OntoNotes.

6.4 Evaluation of the number of senses

Metrics For evaluating the number of senses, we calculated the accuracy and Pearson correlation coefficient (PCC) using the number of senses registered with OntoNotes as the gold standard. For SCAN and Infinite SCAN, we calculated the effective number of senses as an expectation of the sense distribution as follows:

$$\mathbb{E}(S_w) = \exp\left(-\textstyle\sum_{k=1}^{K} \phi_k \log \phi_k\right), \qquad (14)$$

where $\phi_k = 1/T \sum_{t=1}^{T} \phi_{t,k}$ is a marginal sense probability. Here, the term within the exponential is an entropy of the sense distribution, meaning that $\mathbb{E}(S_w)$ is the perplexity of this distribution. For example, when $\phi = (0.5, 0.5, \cdots, 0)$, $\mathbb{E}(S_w)$ is 2 and even when $\phi = (0.49, 0.01, 0.5, \cdots, 0)$, $\mathbb{E}(S_w)$ becomes also approximately 2. We used this expectation to calculate PCC, and used its floor value for the accuracy. For HDP-LDA, we directly used the estimated number of topics as the number of senses. For BERT + k-means, we determined the number of senses by selecting the number of clusters that maximizes the silhouette score (Rousseeuw, 1987), following Giulianelli et al. (2020). For BERT +



(a) *coach* (Aida et al., 2021) (b) *record* (Hamilton et al., 2016b) (c) *power* (Fremann and Lapata, 2016)

Figure 7: Estimated senses for the target words *coach*, *record*, and *power*. Each bar shows the proportion of each sense and is labeled with the start year of the respective time interval. Senses are shown as most high NPMI words.

DBSCAN, we used the estimated number of clusters as the number of senses, with hyperparameters `epsilon`=5 and `min_samples`=2.

Results Table 3 shows prediction results of the number of senses for the baseline models and Infinite SCAN. We can see that Infinite SCAN outperforms the other models in terms of accuracy and PCC between the number of gold and estimated senses. Since SCAN has no mechanism for automatically estimating the number of senses, both the accuracy and PCC are quite low. Even though HDP-LDA and BERT+clustering have an architecture for determining the number of senses, PCC is quite low and these methods do not capture trends in the number of senses that differs depending on a word. Here, compared to k-means, DBSCAN is worse because it ignores some examples as noise, which results in sparse clustering and an overestimation of the number of senses. In contrast, Infinite SCAN can estimate the number of senses more appropriately, because it generates a sense distribution in the stick-breaking architecture that automatically estimates the number of senses from data.

Figure 6 shows that the correlation between the actual and expectation of the number of senses, indicating that the proposed model can capture the tendency of the number of senses. Here, we note that it is obviously difficult to estimate the granularity of manually-annotated meanings, since there are some gold labels that have not appeared in the CCOHA corpus we used for estimation.

6.5 Evaluation of sense change

Methodology We qualitatively evaluated the tracking of sense change by visualizing the sense distribution for three target words *coach*, *record*, and *power* estimated using Infinite SCAN and SCAN. We also show several words with high Normalized pointwise mutual information (NPMI) (Bouma, 2009) in the marginalized sense–word distribution $\sum_{t=1}^{T} \psi_{t,k}$ for each sense k.

Results Figure 7 shows the estimation results for three targets; Infinite SCAN acquires senses with more appropriate granularity (coarse-grained) and captures sense transitions more interpretably compared to SCAN. Figure 7(a) shows the results on the target word *coach*. Infinite SCAN captures two senses, and also indicates changes, with the sense *vehicle* (blue) becoming narrower and the sense *teach* (orange) becoming more dominant, which is consistent with the analysis by Aida et al. (2021). SCAN also captures these senses, but there are overlaps in captured sense (e.g., orange and purple), making it difficult to capture the spread of senses. For the target word *record* in Figure 7(b), three senses emerge: *audio record* (blue), *document* or *history* (orange), and *achievement* (green).

According to Hamilton et al. (2016b), the new sense, which is similar to words such as *music* and *tape*, emerged around 1920; they are captured more clearly by Infinite SCAN. For the word *power* in Figure 7(c), our model captures the senses including *mental power* (blue), *authority* (orange), *legal power* (green), and *energy* (red). The latter is an example of "sense birth" (Mitra et al., 2014), described by Fermann and Lapata (2016), and our model captures such a trend. By contrast, in SCAN, the sense *energy* is divided into two senses (red and gray), making it difficult to identify the correct change or birth of the sense.

7 Conclusion

In this study, we proposed a statistical model that can jointly estimate the number of senses and semantic change of words by combining a dynamic topic model on GMRF with a Dirichlet process. In our experiments, we demonstrated that the proposed model correctly estimates the number of word senses and semantic changes in detail, and showed that the proposed model outperforms baseline models.

In the future, we would like to enhance the model by incorporating linguistic knowledge on semantic change (Ghanbarnejad et al., 2014; Feltgen et al., 2017). Furthermore, we would like to work on analyzing semantic change using the proposed method such as classification of change patterns (Hamilton et al., 2016a).

8 Limitations

8.1 Dataset limitation

The proposed model assumes continuous time shift (*i.e.* iGMRF) and existence of time-continuous corpus, although few languages have a large scale diachronic corpus. Because the proposed model is a Bayesian model, the unigram mixtures (ϕ_t and ψ_t) at a time point t can theoretically be estimated even if a small amount of data exists at time t.

Samples	#Senses 1	2	3	4	5
1,000	0	5	10	50	N/A
2,500	0	5	5	30	40
5,000	N/A	5	5	30	30
10,000	N/A	0	0	5	10

Table 4: Sufficient threshold of low-frequency words for the model to correctly estimate the number of senses and sense change for different sample sizes.

However, if there is a time point with no data at all, the estimation is likely to fail because it violates the assumptions of the model. One solution is to adopt a relatively large value for the parameter Δt, which controls the granularity of time shift, but for a more detailed analysis, it is necessary to use a large scale time-continuous dataset.

8.2 Model limitation

We investigated the conditions under which the model can correctly estimate the number of senses and the sense change of words on artificial data generated by different conditions. We showed that Infinite SCAN can estimate the sense change under the hyperparameter conditions for artificial data generation outlined in Section 5. However, estimation does not always work well under all the conditions. To examine these conditions, we searched for a threshold of low-frequency words at which the model works correctly on artificial data for different numbers of senses and samples.[13]

Table 4 shows the sufficient threshold of low-frequency words at which the model works correctly on artificial data for different number of senses and samples (*i.e.* when the vocabulary size was fixed at $V = 5,000$ and the number of samples varied from $D = 1,000$ to $10,000$ for different numbers of senses.) Note that N/A in the table indicates that the estimation fails no matter what the threshold value is. These results indicate that the smaller the sample size is, the larger the required threshold becomes. This is because the data becomes sparser as the number of samples is reduced, and that more low-frequency words must be truncated to capture senses correctly. Additionally, the threshold for low-frequency words increases with the number of senses since data with more senses accelerate data sparsity. Therefore, data must be prepared with a sample size of at least half the vocabulary size, and the threshold must be set appropriately to stabilize the estimation.

These limitations are also present in real data where it is difficult to estimate the number of senses and semantic changes for words with a large number of senses or for data with small sample sizes. This can be solved by appropriately modifying the data distribution (*i.e.* vocabulary) by thresholding. We would like to address the formulation of these heuristics in the future.

[13] The evaluation of the model estimation was performed manually by visualizing the sense distribution.

Acknowledgements

This research was supported by the NINJAL collaborative research project and NINJAL Diachronic Corpus project at the National Institute for Japanese Language and Linguistics, Japan.

References

Taichi Aida, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2021. A comprehensive analysis of PMI-based models for measuring semantic differences. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 21–31, Shanghai, China. Association for Computational Lingustics.

Jean Aitchison. 2001. *Language change: Progress or decay?* Cambridge university press.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Charles E Antoniak. 1974. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, pages 380–389. PMLR.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing wordnet. In *The 5th International Conference of the Global WordNet Association (GWC-2010)*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Andreas Blank and Peter Koch. 1999. *Historical Semantics and Cognition*. 13. Walter de Gruyter.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the German Society for Computational Linguistics and Language Technology*, volume 30, pages 31–40.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Jason A Duan, Michele Guindani, and Alan E Gelfand. 2007. Generalized spatial Dirichlet process models. *Biometrika*, 94(4):809–825.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy. Association for Computational Linguistics.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231.

Quentin Feltgen, Benjamin Fagard, and J-P Nadal. 2017. Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change. *Royal Society open science*, 4(11):170830.

Thomas S Ferguson. 1973. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 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.

Fakhteh Ghanbarnejad, Martin Gerlach, José M Miotto, and Eduardo G Altmann. 2014. Extracting information from s-curves of language change. *Journal of The Royal Society Interface*, 11(101):20141044.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

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

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th international conference on world wide web*, pages 625–635.

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

Scott Linderman, Matthew J Johnson, and Ryan P Adams. 2015. Dependent multinomial models made easy: Stick-breaking with the Pólya-gamma augmentation. *Advances in Neural Information Processing Systems*, 28.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations*.

David Mimno, Hanna Wallach, and Andrew McCallum. 2008. Gibbs sampling for logistic normal topic models with graph-based priors. In *NIPS Workshop on Analyzing Graphs*, volume 61.

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

Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-aware semantic change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 56–66, Florence, Italy. Association for Computational Linguistics.

Nicholas G Polson, James G Scott, and Jesse Windle. 2013. Bayesian inference for logistic models using Pólya-gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Lu Ren, Lan Du, Lawrence Carin, and David B Dunson. 2011. Logistic stick-breaking process. *Journal of Machine Learning Research*, 12(1).

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011.

Havard Rue and Leonhard Held. 2005. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

Jayaram Sethuraman. 1994. A constructive definition of Dirichlet priors. *Statistica sinica*, pages 639–650.

Angus Stevenson. 2010. *The Oxford English Dictionary*. Oxford University Press.

Erik Sudderth and Michael Jordan. 2008. Shared segmentation of natural scenes using dependent Pitman-Yor processes. *Advances in neural information processing systems*, 21.

Yee Whye Teh and Michael I Jordan. 2010. Hierarchical Bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1:158–207.

Joseph B Voyles. 1973. Accounting for semantic change. *Lingua*, 31(2-3):95–124.

Joseph M Williams. 1976. Synaesthetic adjectives: A possible law of semantic change. *Language*, pages 461–478.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the eleventh acm international conference on web search and data mining*, pages 673–681.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of general psychology*, 33(2):251–256.

A Pseudo-code of MCMC algorithm

Algorithm 1 shows the MCMC algorithm for the estimation of Infinite SCAN. In practice, we sample z, ϕ and ψ for 2,000 iterations, and sample $\kappa_\phi^{(k)}$ for each sense k after every 50 iterations, starting from iteration 150.

Algorithm 1: MCMC algorithm

Initialize $\kappa_\phi^{(k)} = 4.0$ (for all k)
Initialize $\kappa_\psi = 100.0$
Initialize $a = 7.0$, $b = 3.0$
for $t = 1 \dots T$ **do**
    Initialize $\alpha_t \sim \mathcal{N}\left(\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}), \kappa_\phi^{-1}\right)$
    Set $\phi_t = \text{LSB}(\alpha_t)$
    for $k = 1 \dots K$ **do**
        Initialize $\beta_{t,k} \sim \mathcal{N}\left(\frac{1}{2}(\beta_{t-1,k} + \beta_{t+1,k}), \kappa_\psi^{-1}\right)$
        Set $\psi_{t,k} = \text{Softmax}(\beta_{t,k})$
    end
end
for $j = 1 \dots J$ **do**
    Sample z according to Eq. (8)
    for $t = 1 \dots T$ **do**
        Sample ϕ according to posterior in Eq. (9)
        Sample ψ according to posterior in Eq. (10)
    end
    Sample κ_ϕ according to posterior in Eq. (11)
end

B Statistics of snippets used for the evaluation of the number of sense

Table 5 shows the statistics of snippets used for the evaluation of the number of senses (Sections 6.3 and 6.4). This table lists the number of words in each sense, the average number of samples, and the vocabulary size.

#Senses	#Words	Samples	#Vocab
1	30	15,922	14,403
2	33	17,085	15,014
3	30	17,471	15,868
4	17	18,328	16,388
5	10	18,178	16,645

Table 5: Snippet statistics of target words randomly selected from OntoNotes. Sample size and vocabulary size are shown as averages.

C Statistics of snippets used for the evaluation of sense change

Table 6 shows the statistics of snippets used for the evaluation of sense change (Section 6.5). The table lists the years, sample size, and vocabulary size for each target word.

Word	Years	Samples	#Vocab
coach	1811–2009	9,758	11,962
record	1815–2009	33,992	23,886
power	1810–2009	142,527	42,932

Table 6: List of target words and snippet statistics.