Infinite SCAN: An Infinite Model of Diachronic Semantic Change

Seiichi Inoue¹, Mamoru Komachi¹, Toshinobu Ogiso², Hiroya Takamura³ and Daichi Mochihashi⁴

¹Tokyo Metropolitan University  ²The National Institute for Japanese Language and Linguistics  ³The National Institute of Advanced Industrial Science and Technology  ⁴The 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 (Bamber 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

¹Source code is available at https://github.com/seiiichinoue/iscan.
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 Sense ChA\textsc{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 $c_{d} = \{c_{d,1}, \ldots, c_{d,I}\}$ 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 $t \in \{1 \ldots T\}$ is modeled by unigram mixtures at each time point:

- $K$-dimensional multinomial distribution $\phi_{t}$ (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. $\phi$ is obtained by transforming a sampled vector as follows:

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

$\psi$ can be obtained in a similar way. Note that $K$ is assumed to be given (a parameter to be set \textit{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 $\phi$ and sense–word distribution $\psi$ 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).

$$x_{t} | x_{-t}, \kappa \sim \mathcal{N}\left(\frac{1}{2}(x_{t-1} + x_{t+1}), \frac{1}{\kappa}\right), \quad (1)$$

where $x_{-t}$ is a set of $x$ except for $x_{t}$ and $\kappa$ is a precision parameter. The Gaussian distribution, which is a prior distribution of the sense distribution and the sense–word distribution, has precision parameters $\kappa_{\phi}$ and $\kappa_{\psi}$ to control the degree of change, respectively. In particular, the precision parameter of the sense distribution, $\kappa_{\phi}$, 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 $\text{Ga}(a, b)$ denotes the gamma distribution and $\text{Mult}(\theta)$ denotes the multinomial distribution.

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

---

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

Figure 2: A linear chain iGMRF.
A. \( \beta_{t,k} \mid \beta_{-t,k}, \kappa_\psi \sim N \left( \frac{1}{2}(\beta_{t-1,k} + \beta_{t+1,k}), \kappa_\psi^{-1} \right) \)

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

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

i. Draw a sense \( z_d \sim \text{Mult}(%(\phi_t) \)

ii. For context position \( i = 1 \ldots I \)

A. Draw a word \( c_{d,i} \sim \text{Mult}(%t,z_d) \)

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 \( DP(\alpha, G_0) \) is generated as follows:

\[
\pi_k = v_k \prod_{j=1}^{k-1} (1 - v_j), \quad v_k \sim \text{Be}(1, \alpha) \quad (2)
\]

\[
G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k), \quad \theta_k \sim G_0, \quad (3)
\]

where \( \text{Be}(1, \alpha) \) denotes a beta distribution. First, the probability for the \( k \)-th category \( \pi_k \) 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 \( \theta_k \) 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 \( x \in \mathbb{R} \). 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:

\[
\pi(x_k) = \sigma(x_k) \prod_{j=1}^{k-1} (1 - \sigma(x_j)), \quad (4)
\]

\[
G_x = \sum_{k=1}^{\infty} \pi(x_k) \delta(\theta_k). \quad (5)
\]

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).

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 \( c_d = \{c_{d,1}, \ldots, c_{d,I}\} \) 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 \( \phi_t \) and sense–word distributions \( \psi_{t,k} \), each following a first order iGMRF:

\[
\alpha_t \mid \alpha_{-t}, \kappa_\phi \sim N \left( \frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}), \kappa_\phi^{-1} \right). \quad (6)
\]

Here, we modify the generative process of the sense distribution \( \phi_t \) 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:

\[
\phi_{t,k} = \sigma(\alpha_{t,k}) \prod_{j=1}^{k-1} (1 - \sigma(\alpha_{t,j})) \quad (7)
\]

\( (k = 1, \ldots, K) \),
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). 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 $\kappa_\phi$ of the sense distribution is shared across all senses. This is because the sense distribution $\phi_t$ 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 $\phi_t$ 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, $\kappa_\phi$ 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 $\alpha_k$.

4Preliminary 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 $\alpha$ (un-normalized $\phi$), (c) sense–word distributions $\beta$ (un-normalized $\psi$), and (d) the precision parameter $\kappa_\phi$ 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.

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 $\phi$ and $\psi$:

\[
p(z_d | w, t, \phi, \psi) \propto p(z_d | t)p(w | t, z_d) = \phi_z^{(t)} \prod_{w \in w} \psi^{(t,z_d)}.
\]

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 $\alpha$ is computed as follows:

\[
p(\alpha_t | z, \alpha_{-t}, \omega) \propto \mathcal{N}(\omega^{-1} f(c) | \alpha_t) \mathcal{N}(\alpha_t | \alpha_{-t}, \kappa_\phi^{-1}) \propto \mathcal{N}(\alpha_t | \hat{\mu}, \hat{\kappa}_\phi^{-1}).
\]

Here, $\omega$ is an auxiliary variable that is sampled from Pólya-gamma distribution $\omega | z, \alpha_t \sim \text{PG}(N(c_k), \alpha_t)$, 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 $\hat{\mu} = (f(c_k) + \mu_k \kappa_\phi) / \hat{\kappa}_\phi$ and $\hat{\kappa}_\phi = \omega_k + \kappa_\phi$, where $f(c_k) = c_k - N(c_k)/2$. 

\[
\bar{\tau}_k = (f(c_k) + \mu_k \kappa_\phi) / \hat{\kappa}_\phi
\]

\[
\bar{\tau}_k = (f(c_k) + \mu_k \kappa_\phi) / \hat{\kappa}_\phi
\]
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 $\beta$ is as follows:

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

(10)

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 $\kappa_\psi$ is as follows:

$$p(\kappa_\psi^{(k)} \mid \alpha_k, a, b) = \text{Ga} \left( a + \frac{T}{2}, b + \frac{1}{2} \sum_{t=1}^{T} (\alpha_t, k - \bar{\alpha}_k) \right)$$

(11)

where $\bar{\alpha}_k = 1/T \sum_t \alpha_{t,k}$ is the mean of $\alpha$ 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 $\kappa_\psi^{(k)} = 4$ 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.” Finally, we ran the Gibbs sampler for 2,000 iterations and resampled $\kappa_\psi^{(k)}$ for each sense $k$ after every 50 iterations, starting from iteration 150.

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.,

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

(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 at $V = 5,000$, the snippet length at $I = 10$, and changed the number of word senses 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 a q c y c t g w x a h y

$k=3$: d k j d k j d k p d l q d m s d m y d p y e a j e s v

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

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

<table>
<thead>
<tr>
<th>#Senses</th>
<th>SCAN $K = 5$</th>
<th>SCAN $K = 8$</th>
<th>Infinite SCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.523</td>
<td>1.997</td>
<td><strong>0.468</strong></td>
</tr>
<tr>
<td>2</td>
<td>0.335</td>
<td>0.578</td>
<td><strong>0.039</strong></td>
</tr>
<tr>
<td>3</td>
<td>0.216</td>
<td>0.735</td>
<td><strong>0.030</strong></td>
</tr>
<tr>
<td>4</td>
<td>0.212</td>
<td>0.150</td>
<td><strong>0.061</strong></td>
</tr>
<tr>
<td>5</td>
<td>0.004</td>
<td>0.017</td>
<td><strong>0.044</strong></td>
</tr>
</tbody>
</table>

5.2 Evaluation

We evaluated the proposed model on artificial data to validate the model for correctly estimating arbitrary changes and the number of senses. 

In order to evaluate the quality of the estimated sense distribution, we compared the estimated distribution with the actual sense distribution using the Kullback-Leibler divergence.

5.3 Results

The results show that the proposed model consistently estimated the number of senses and the distribution of the generated snippets correctly. The Kullback-Leibler divergence is lower for the proposed model compared to other baselines, indicating better performance.

---

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.

---

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

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

8Most 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 (C COHA) (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 $c_i$ before and after the point where the target word $w$ appeared in the corpus, with a symmetric window width of $\pm 5$ 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. We used base-uncased version of the pre-trained model available at https://github.com/huggingface/transformers.

$^9$ 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.
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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coherence</th>
<th>Diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDP-LDA</td>
<td>0.125</td>
<td>0.821</td>
</tr>
<tr>
<td>SCAN (K = 5)</td>
<td>0.178</td>
<td>0.716</td>
</tr>
<tr>
<td>SCAN (K = 8)</td>
<td>0.171</td>
<td>0.654</td>
</tr>
<tr>
<td>Infinite SCAN</td>
<td><strong>0.181</strong></td>
<td><strong>0.885</strong></td>
</tr>
</tbody>
</table>

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:

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

where $w_i^{(k)}$ is the $i$-th most probable word in sense $k$, and $\eta$ 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.

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:

$$E(S_w) = \exp \left( - \sum_{k=1}^{K} \phi_k \log \phi_k \right),$$  \hspace{1cm} (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 $E(S_w)$ is the perplexity of this distribution. For example, when $\phi = (0.5, 0.5, \ldots, 0)$, $E(S_w)$ is 2 and even when $\phi = (0.49, 0.01, 0.5, \ldots, 0)$, $E(S_w)$ 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.

\(^{11}\)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.
Infinite SCAN, we used the estimated number of clusters as the number of senses, with hyperparameters $\epsilon = 5$ and $\text{min} \_\text{samples} = 2$.

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 differ 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 Frermann 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 (\(\phi_t\) and \(\psi_t\)) at a time point \(t\) can theoretically be estimated even if a small amount of data exists at time \(t\). 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 \(\Delta t\), which controls the granularity of time shift, but for a more detailed analysis, it is necessary to use a large scale time-continuous dataset.

### 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.

<table>
<thead>
<tr>
<th>Samples</th>
<th>#Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>1,000</td>
<td>0 5 10 50 N/A</td>
</tr>
<tr>
<td>2,500</td>
<td>0 5 5 30 40</td>
</tr>
<tr>
<td>5,000</td>
<td>N/A 5 30 30</td>
</tr>
<tr>
<td>10,000</td>
<td>N/A 0 5 10</td>
</tr>
</tbody>
</table>

Table 4 shows the sufficient threshold of low-frequency words at which the model works correctly on artificial data for different numbers 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.

13The 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


A Pseudo-code of MCMC algorithm

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

**Algorithm 1: MCMC algorithm**

1. Initialize $\kappa^{(k)}_\phi = 4.0$ (for all $k$)
2. Initialize $\kappa_\psi = 100.0$
3. Initialize $a = 7.0$, $b = 3.0$
4. for $t = 1 \ldots T$ do
   5. Initialize $\alpha_t \sim \mathcal{N}\left(\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}), \kappa_\phi^{-1}\right)$
   6. Set $\phi_t = \text{LSB}(\alpha_t)$
   7. for $k = 1 \ldots K$ do
      8. Initialize $\beta_{t,k} \sim \mathcal{N}\left(\frac{1}{2}(\beta_{t-1,k} + \beta_{t+1,k}), \kappa_\psi^{-1}\right)$
      9. Set $\psi_{t,k} = \text{Softmax}(\beta_{t,k})$
   10. end
   11. end
   12. for $j = 1 \ldots J$ do
      13. Sample $z$ according to Eq. (8)
      14. for $t = 1 \ldots T$ do
         15. Sample $\phi$ according to posterior in Eq. (9)
         16. Sample $\psi$ according to posterior in Eq. (10)
      17. end
      18. Sample $\kappa_\phi$ according to posterior in Eq. (11)
   19. 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.

<table>
<thead>
<tr>
<th>#Senses</th>
<th>#Words</th>
<th>Samples</th>
<th>#Vocab</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td>15,922</td>
<td>14,403</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>17,085</td>
<td>15,014</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>17,471</td>
<td>15,868</td>
</tr>
<tr>
<td>4</td>
<td>17</td>
<td>18,328</td>
<td>16,388</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>18,178</td>
<td>16,645</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Word</th>
<th>Years</th>
<th>Samples</th>
<th>#Vocab</th>
</tr>
</thead>
<tbody>
<tr>
<td>coach</td>
<td>1811–2009</td>
<td>9,758</td>
<td>11,962</td>
</tr>
<tr>
<td>record</td>
<td>1815–2009</td>
<td>33,992</td>
<td>23,886</td>
</tr>
<tr>
<td>power</td>
<td>1810–2009</td>
<td>142,527</td>
<td>42,932</td>
</tr>
</tbody>
</table>

Table 6: List of target words and snippet statistics.