Visualizing Record China Discourse through Exponential Family Embeddings

Wu Tung-Wen
(Data scientist at Service Development Group, dip Corp.)
March 21, 2024

*This paper reflects opinion of the author as derived from an analysis using quantitative political science and data. It does not represent the views of the organization to which this author belongs.

Abstract

This paper employs exponential family embeddings, a Bayesian machine learning method, to analyze the discourse in articles published by Record China. Specifically, it estimates the meaning of words in Record China articles by using exponential family embeddings. As a result of the estimation, this paper quantitatively reveals China's argument that Chinese democracy is superior and the discourse that the U.S. is a threat. If the amount of data is expanded in the future, it will be possible to visualize changes in Record China's discourse and differences between this discourse and that of the Japanese media in general. (The content of this article is solely the opinion of the author and has nothing to do with dip Corporation, to which the author belongs.)

Introduction

This paper utilizes the full-text data of Record China for the years 2021 and 2022, obtained through data scraping. Professor Maiko Ichihara at Hitotsubashi University provided the data. The paper employs Bayesian machine learning to visualize the underlying thoughts behind the discourse propagated by China, which appears to influence the operation of the Record China website in Japan, as suggested by Ichihara.[1]

Specifically, this paper analyzes the co-occurrence relations between words using the exponential family embeddings proposed by Rudolph et al.[2] For example, by extracting words that appear in contexts similar to ”democracy” or ”America” within Record China articles, this method can visualize how Record China discusses ”democracy” and ”America,” respectively.

Since this paper utilizes the same dataset as a previous analysis of Record China conducted by the author,[3] it omits basic data visualization. Furthermore, similar to the previous study, this paper analyzes particularly articles published by Record China itself or articles republished from the Japanese version of People’s Daily Online.

Methodology

First, I explain the concept of word2vec, an early representative method of embeddings to which exponential family embeddings belong. Then, I describe the exponential family embeddings used in our analysis. However, to omit mathematical descriptions, the following explanations primarily focus on conceptual understanding.

---

[1] Ichihara Maiko, “Is Japan Immune from China’s Media Influence Operations?” *The Diplomat* (December 19, 2020). (https://thediplomat.com/2020/12/is-japan-immune-from-chinas-media-influence-operations/ Accessed 2024-2-2)

[2] Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei, “Exponential Family Embeddings,” *Advances in Neural Information Processing Systems*, 29 (2016).

[3] Wu Tung-Wen, “Analyzing Record China through the Structural Topic Model” *GGR Working Paper No. 8* (March 7, 2024). (https://ggr.hias.hit-u.ac.jp/en/2024/03/07/structural-topic-model-analysis-of-record-china/ Accessed 2024-3-18)

Concept and Usage of word2vec-like Methods

In brief, word2vec is a method that originated in the field of computer science to vectorize words and model their meanings precisely. Two papers by Mikolov et al. are well-known as initial works.[4] Vectorization is a concept similar to setting latent variables in quantitative political science or quantitative economics, assuming that behind a single word are multiple parameters. Among various methods for estimating word meanings using latent variables, this paper primarily explains negative sampling, which is adopted here.

Negative sampling is a concept of teaching language knowledge by having a model learn to distinguish between real data and fake data, by mixing fake data into real data.

As a concrete example, I present a segmented Record China text below.

New Year, China, (???),[5] Chinese people, confidence

Next, insert two words into the brackets. Consider which one is real.

New Year, China, (Rise), Chinese people, confidence
New Year, China, (Philly Cheesesteak), Chinese people, confidence

You would have instantaneously recognized that the former is the genuine content of Record China articles, while the latter is perfunctorily inserted. However, how is it that you are able to judge ”rise” as an appropriate word surrounded by ”New Year,” ”China,” ”Chinese people,” and ”confidence,” whereas ”Philly cheesesteak” is deemed inappropriate? This is because the context of ”New Year,” ”China,” ”Chinese people,” and ”confidence” provides knowledge about this text. More specifically, ”China” and ”confidence” are likely particularly relevant to determining the answer. Given the discourse concerning China’s confidence, it seems plausible for

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space,” *arXiv* preprint, 1301, 3781 (2013). Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, “Distributed Representations of Words and Phrases and Their Compositionality,” *Advances in Neural Information Processing Systems*, 26 (2013).

[5] This paper calls a word in (???) a target word.

"rise" to be included, while the appearance of an American dish name is too abrupt.

In this manner, by predicting the probability of a word based on surrounding words, i.e., context, word2vec enables the model to understand word meanings. As multiple papers like the one by Mikolov et al.[6] have pointed out, the estimated vectors retain word meanings and are used for measuring word similarity. Notably, a paper by Rodriguez et al.[7] is among the primary references regarding the application of the word2vec method and others in political science.

Exponential Family Embeddings

Exponential family embeddings are Bayesian machine learning methods derived from word2vec. Exponential family embeddings define two types of vectors: an embedding vector and a context vector. They then take the inner product of the embedding vector of the target word and the context vector of the word in its context, denoted as η.

$$\eta_{rise} = embedding\ vector_{rise}{}' * (context\ vector_{new\ year} + context\ vector_{China} + context\ vector_{Chinese} + context\ vector_{Confidence})$$

$$\eta_{philly\ cheesesteak} = embedding\ vector_{philly\ cheesesteak}{}' * (context\ vector_{new\ year} + context\ vector_{Chine} + context\ vector_{chinese} + context\ vector_{confidence})$$

η is then sampled as the parameter of a Bernoulli distribution:

$$Flag\ for\ real\ or\ fake \sim Bernoulli\ distribution\ (\eta_{rise})$$

$$Flag\ for\ real\ or\ fake \sim Bernoulli\ Distribution\ (\eta_{Philly\ cheesesteak})$$

If the model performs well, the former flag should be sampled as 1 (real), and the latter flag as 0 (fake).

The greatest advantage of exponential family embeddings lies in their flexibility in combination with other Bayesian statistical and machine learning methods. For instance, by varying embedding vectors by party

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, *op.cit.*

[7] Pedro L. Rodriguez and Arthur Spirling, "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research," *The Journal of Politics*, 84-1 (2022), 101-115.

affiliation, one can precisely capture differences in word meanings by party.[8] Similarly, by varying embedding vectors over time, one can visualize changes in word meanings.[9]

Due to the limited data, hierarchical Bayesian structures are not applied to embedding vectors in this paper. However, if more data become available in the future, slight modifications to the Stan code could enable the visualization of the temporal changes in Record China's discourse, or differences between Record China and general Japanese media discourse.

Data Analysis

Data Preprocessing

Although not significantly different from the previous study,[10] the analysis in this paper also focuses on articles corresponding to Record China or the Japanese version of People's Daily Online, among the dataset provided by Professor Maiko Ichihara. Concerning Japanese text processing, all characters except hiragana, katakana, kanji, and romaji are replaced with spaces.[11] Then, using mecab via RMeCab, only nouns are retained. For word2vec-like methods, it is common to retain more words, but due to limitations in local machine specs and writing time, approximately the same number of words as in the previous article were deleted. Furthermore, all words were id-converted, and a data frame containing three surrounding words as context for the target word was created. Additionally, a data frame containing random data was included, where the context was fixed as mentioned above and only the id of the target word was shuffled. Finally, 5,000 data points from the data frame were kept as validation data and the rest were fed into the exponential family embeddings Bayesian machine learning model created by the author using Stan. The dimensions of the embedding vectors and context vectors were set to 50. Since existing

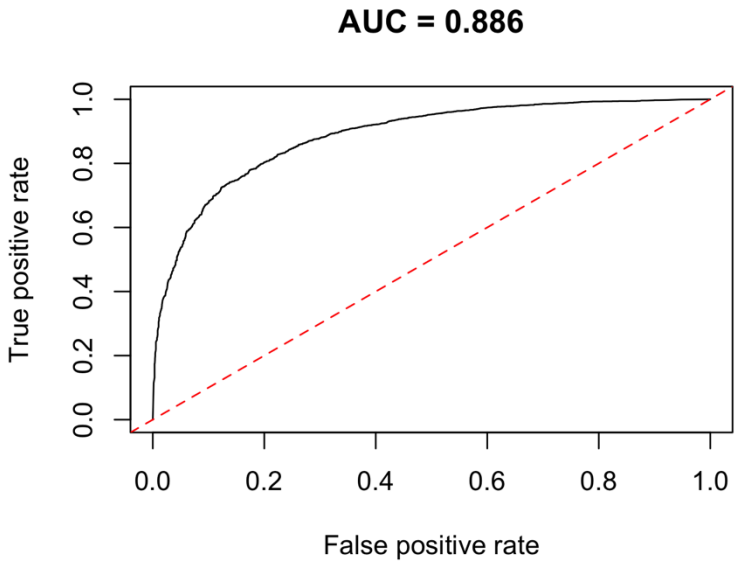
[8] Maja Rudolph, Francisco Ruiz, Susan Athey, and David Blei, "Structured Embedding Models for Grouped Data," *Advances in Neural Information Processing Systems*, 30 (2017).
[9] Maja Rudolph and David Blei, "Dynamic Embeddings for Language Evolution," Proceedings of the 2018 World Wide Web Conference (2018), 1003-011.
[10] Wu, op.cit.
[11] In regular expression, anything matching [^ー-龠ぁ-ンーァ-ケーa-zA-Z] has been replaced with blanks.

packages were not used for this analysis, the Stan code is shown in the Appendix for transparency.

Precision Check

First, I check whether the model can determine the truthfulness of words for validation data that is not passed to the model. The Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is used as a measure of precision. Simply put, the ROC curve indicates that the further it deviates from the 45-degree line, the higher the accuracy, while overlapping with the 45-degree line indicates random classification.

Figure 1: AUC in word truthfulness task



Source: Author's own work

As shown in Figure 1, the AUC for the validation data of this model in word truthfulness task exceeds 0.8, indicating exceptionally high accuracy. Therefore, it is reasonable to conclude that the model learned some patterns in probabilities of words and discourse of Record China.

Discourse Overview from Cosine Similarity

Cosine similarity is an indicator used to measure the similarity of word vectors. With this indicator, one can assess the proximity of word meanings

or the context where words are used.

This analysis utilizes only data from Record China and does not employ so-called ”pre-trained models,” etc., hence the similarity between estimated vectors purely represents the discourse of Record China based on the formulation and preprocessing of the model.

Firstly, starting with the simplest, an overview of the ten words with the highest cosine similarity to ”road” in Record China is below.

Chart 1: The 10 words with the highest cosine similarity to ”road” in Record China

Word (Japanese word)	Cosine similarity
Road (道路)	1.0000000
Railway (鉄道)	0.7952032
Railway section (区間)	0.6528857
Station (駅)	0.6473283
Kilometer (キロメートル)	0.6423741
Large bridge (大橋)	0.6331041
Vehicle (車両)	0.6313493
Train (列車)	0.6180682
Bridge (橋)	0.6118205
Running (走行)	0.6033028

”Railway,” ”station,” ”vehicle,” and other transportation-related words appear.

Next, let’s examine words with high cosine similarity to ”airport.” They are still transportation-related words, but more words related to public transportation appear, and words such as ”big bridge” and ”vehicle” that had high cosine similarity with ”road” are not observed.

Chart 2: The 10 words with the highest cosine similarity to ”airport” in Record China

Word (Japanese word)	Cosine similarity
Airport (空港)	1.0000000
Station (駅)	0.6233985
Arrival (到着)	0.5561386
Aviation (航空)	0.5314159
Railway(鉄道)	0.5304486
Subway (地下鉄)	0.5278712
Mail (郵便)	0.5166218
Passenger (乗客)	0.5165372
Flight (便)	0.5122190
Aircraft (飛行機)	0.5120097

Next, let’s check words with high cosine similarity to ”river.” ”Lake,” ”mountain,” ”fish,” “lake,” and other nature-related words appear.

Chart 3: The 10 words with the highest cosine similarity to ”river” in Record China

Word (Japanese word)	Cosine similarity
River (川)	1.0000000
Lake (池)	0.6394781
Large bridge (大橋)	0.6306532
Yellow river (黄河)	0.5989985
Mountain (山)	0.5945901
Seawater (海水)	0.5860707
Gorge (峡)	0.5855116
Fish (魚)	0.5777902
Watershed (流域)	0.5767979
Lake (湖)	0.5764602

Finally, let’s examine words with high cosine similarity to ”net” (an abbreviation for “internet” in Japanese.

Chart 4: The 10 words with the highest cosine similarity to ”net” in Record China

Word (Japanese word)	Cosine similarity
Net (ネット)	1.0000000
Internet (インターネット)	0.6724772
Post (書き込み)	0.5358616
Video (動画)	0.5171089
Deletion (削除)	0.4966197
Comment (コメント)	0.4766648
Post (投稿)	0.4734309
Content (コンテンツ)	0.4703328
Fact (事実)	0.4548065
Cloud computing (クラウドコンピューティング)	0.4545467

Words related to the Internet such as ”internet,” ”posting,” ”video,” and ”comment” are confirmed to appear.

In this way, both the AUC of the ROC curve and the results for several sample words confirm that this model estimates the usage patterns of words in Record China.

Now, let’s examine significant words in law and political science. Firstly, among words with high cosine similarity to ”democracy,” surprisingly, the terms ”CCP” and ”Chinese Communist Party” appear at the top.

Chart 5: The 10 words with the highest cosine similarity to ”democracy” in Record China

Word (Japanese word)	Cosine similarity
Democracy (民主)	1.0000000
Chinese Communist Party (中共[12])	0.7087134
Politics (政治)	0.6484074
Ethnicity (民族)	0.6299390
Independence (独立)	0.6253643
Opposition party (野党)	0.5998672
Election (選挙)	0.5954480
Chinese Communist Party(中国共産党)	0.5883779
Militarism (軍国)	0.5872631
Party (党)	0.5770189

This quantitatively confirms, as pointed out by Tajimi,[13] that China is promoting the idea of ”Chinese democracy,” which is “much better than American-style democracy,” through Record China. This suggests that the site operators are also trying to promote similar discourse in Japan.

Next, when examining words with high cosine similarity to ”America,” it is found that along with names of American and European countries and regions, terms like ”spy” and ”threat” are also used in similar contexts to ”America.”

[12] Translator’s note: 中共 is an abbreviation for 中国共産党.
[13] Tajimi Makoto, “Endeavors to Realize ‘Chinese Democracy,’” *The Hitotsubashi Journal of Law and International Studies*, 21-2 (2022), pp. 165-199.

Chart 6: The 10 words with the highest cosine similarity to ”America” in Record China

Word (Japanese word)	Cosine similarity
America (アメリカ)	1.0000000
America (米国)	0.5713009
The West (欧米)	0.5136229
Russia (ロシア)	0.4954124
Germany (ドイツ)	0.4741628
America (米)	0.4739354
Spy (スパイ)	0.4701061
Europe (欧州)	0.4692116
ン[14]	0.4647487
Threat (脅威)	0.4597416

It is challenging to determine the rationale behind the model estimation using the exponential family embedding model. However, examining a Record China article[15] where ”America” and ”threat” are simultaneously used, the following is described:

> On the 27th, the US government-affiliated media Voice of America reported that major SNS platforms Facebook and Twitter had removed fake accounts posting pro-US content. The removed accounts were posting criticism of Russia, China, and Iran based on Western perspectives. According to a report by Stanford University and the social media analytics company Graphika, these fake accounts violated the terms of service of Facebook and Twitter and were disseminating pro-Western information to the Middle East and Central Asia using deceptive means.[16]

[14] Translator’s note: ン is one katakana in Japanese. It has no meaning in this context.
[15] Record China, “Feisubukku to tuitta ga ‘shinbei teki nise akaunto wo haijo [Facebook and Twitter removed “pro-US fake account]” (August 29, 2022). (https://www.recordchina.co.jp/b900239-s25-c100-d0198.html Accessed on 2024-2-2)
[16] *Ibid.*

In other words, China claims to be a ”victim” of ”fake news” from America, implying that America is a ”threat.” Furthermore, Chinese discourse asserting America as the enemy also appears in this article.[17]

Lastly, let’s examine the ten words with high cosine similarity to ”Taiwan.”

Chart 6: The 10 words with the highest cosine similarity to ”Taiwan” in Record China

Word (Japanese word)	Cosine similarity
Taiwan (台湾)	1.0000000
Mainland (本土)	0.7017824
Strait (海峡)	0.5893464
Pineapple (パイナップル)	0.5561649
Hong Kong (香港)	0.5517208
Senkaku Islands (尖閣諸島)	0.5513686
Military force (武力)	0.5498675
Continent (大陸)	0.5427460
Yasukuni Shrine (靖国神社)	0.5193581
Repulsion (反発)	0.5146651

The word with the highest cosine similarity is ”mainland,” which could be interpreted as an attempt to propagate the ”One China” discourse claiming that Taiwan is part of China.

Visualization by T-SNE

As a final analysis in this paper, I utilize the T-distributed Stochastic Neighbor Embedding (T-SNE) method [18] to visually comprehend the estimation results of the exponential family embedding model by projecting

[17] Andy Greenberg, “Beikoku no Chuukan senkyo wo nerai ‘Shin chu ha’ niyoru jouhou sousa ga kappatsu ni natteiru [“Pro-China” groups are actively manipulating information targeting the U.S. midterm elections]” (November 1, 2022). (https://wired.jp/article/us-midterm-election-disinformation-dragonbridge/ Accessed on 2024-2-2)

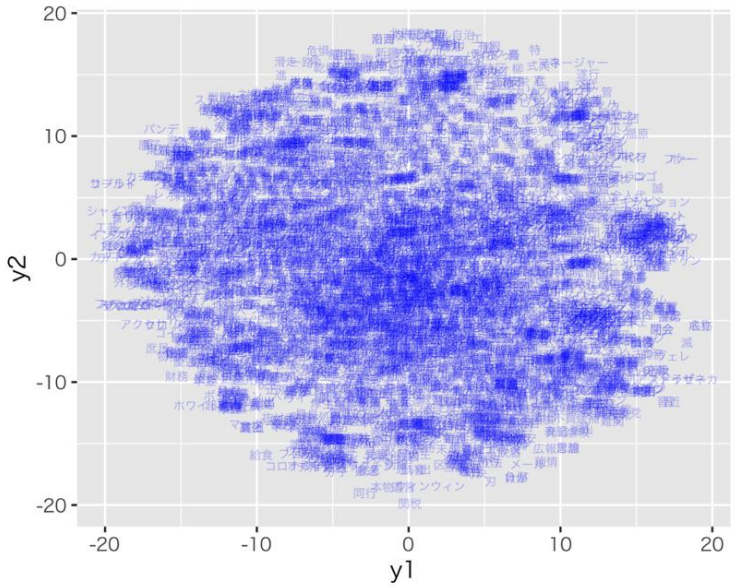
[18] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing Data Using t-SNE,” *Journal of Machine Learning Research,* 9-11 (2008).

vectors onto a two-dimensional space.

Indeed, since T-SNE is a concept distinct from cosine similarity, high cosine similarity does not necessarily mean close distances when visualized using T-SNE. However, humans make various judgments visually, and if the overall estimation results can be visually comprehended, new hypotheses often emerge about points to delve deeper into.

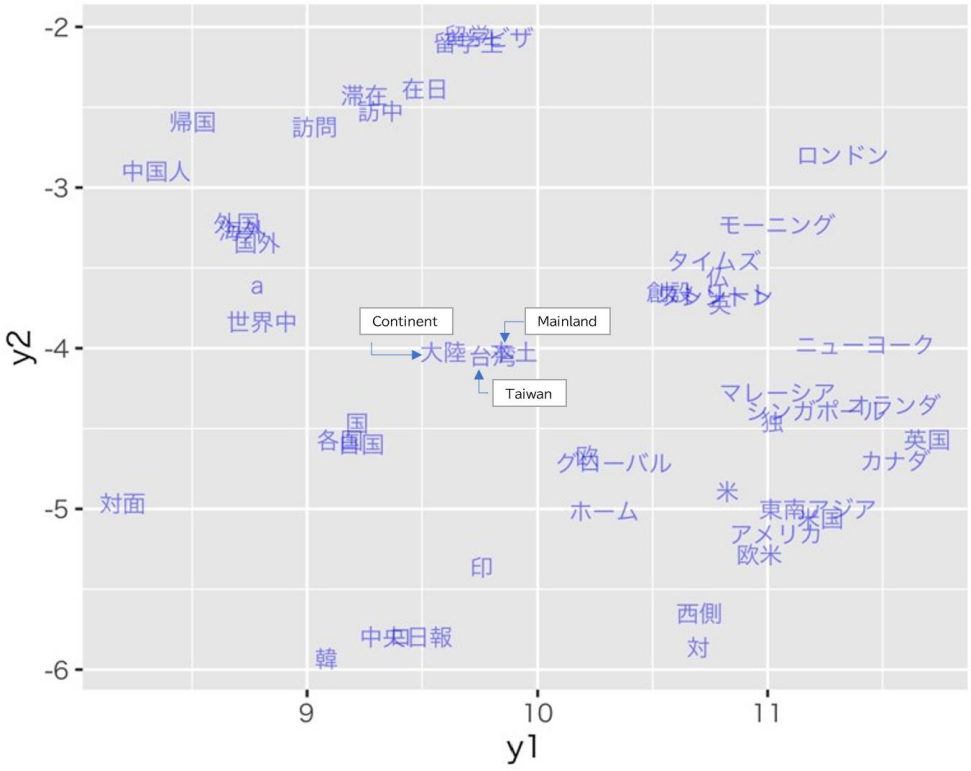
Firstly, Figure 2 depicts the visualization of the entire embedding vectors. Zooming in near ”Taiwan,” as shown in Figure 3, it can be checked that ”continent” and ”mainland” are in close proximity. As pointed out in the analysis using cosine similarity mentioned earlier, this reinforces the analysis results that propagate the discourse of the ”One China” claim that ”Taiwan is part of China.”

Figure 2: T-SNE of embedding vectors in exponential family embeddings



Source: Author’ s own work

Figure 3: T-SNE near ‘Taiwan（台湾）’



Source: Author’s own work revised by the translator

Conclusion

This paper employs exponential family embedding, a Bayesian derivative of word2vec, to analyze the discourse propagated by Record China from the perspective of word meaning. Furthermore, it visualizes China’s discourse praising ”Chinese democracy” compared to Western democracy and asserting America as a threat.

This trend is not captured in the analysis using structural topic models. The analysis also demonstrates that in conducting text-as-data analysis, combining various methods allows for deeper insights to be extracted.

However, similar to the previous study,[19] limitations in data volume prevent using the Bayesian hierarchical structure, an advantage of exponential family embeddings. With an increase in obtainable data in the future, it is expected that a more complex model structure will enable more sophisticated analyses focusing on politically significant aspects such as discourse changes and discrepancies.

[19] Wu, op.cit.

【Translated by】
Takahiro NAKAJIMA
(Bachelor's student, Faculty of Law, Hitotsubashi University)

Appendix

Stan code

```
functions {
real partial_sum_lpmf(
  array[] int result,
  int start, int end,

  array[] int word,

  array[] int word_before_1, array[] int word_after_1,
  array[] int word_before_2, array[] int word_after_2,
  array[] int word_before_3, array[] int word_after_3,

  array[] vector word_embedding,
  array[] vector word_context
){
  vector[end - start + 1] lambda;
  int count = 1;
  for (i in start:end){
    // Rudolph et al.(2016)の式(2)
    // 目標単語の embedding とその文脈(context)にある単語の context vector の和の内積を取る
    lambda[count] = word_embedding[word[i]]' *
    (
      word_context[word_before_1[i]] + word_context[word_after_1[i]] +
      word_context[word_before_2[i]] + word_context[word_after_2[i]] +
      word_context[word_before_3[i]] + word_context[word_after_3[i]]
    );
    count += 1;
  }
```

```
    return (
      // Rudolph et al.(2016)の式(1)のように本物かデタラメかを示すフラグをサンプリング
      bernoulli_logit_lupmf(result | lambda)
    );
  }
}
data {
  int<lower=1> N; //学習データ数
  int<lower=1> K; //embedding 次元数
  int<lower=1> word_type; //単語

  //学習データ
  array[N] int<lower=1,upper=word_type> word; // 目標単語の id
  array[N] int<lower=1,upper=word_type> word_before_1; // 目標単語の 1 個前の単語の id
  array[N] int<lower=1,upper=word_type> word_after_1; // 目標単語の 1 個後の単語の id
  array[N] int<lower=1,upper=word_type> word_before_2; // 目標単語の 2 個前の単語の id
  array[N] int<lower=1,upper=word_type> word_after_2; // 目標単語の 2 個後の単語の id
  array[N] int<lower=1,upper=word_type> word_before_3; // 目標単語の 3 個前の単語の id
  array[N] int<lower=1,upper=word_type> word_after_3; // 目標単語の 3 個後の単語の id
  array[N] int<lower=0,upper=1> result; // 本物かデタラメかを示すフラグ。0:デタラメ、1:本物

  int<lower=0> val_N; //検証データ数
  //検証データ
  array[val_N] int<lower=1,upper=word_type> val_word; // 目標単語の id
  array[val_N] int<lower=1,upper=word_type> val_word_before_1; // 目標単語の 1 個前の単語の id
  array[val_N] int<lower=1,upper=word_type> val_word_after_1; // 目標単語の 1 個後の単語の id
  array[val_N] int<lower=1,upper=word_type> val_word_before_2; // 目標単語の 2 個前の単語の id
  array[val_N] int<lower=1,upper=word_type> val_word_after_2; // 目標単語の 2 個後の単語の id
  array[val_N] int<lower=1,upper=word_type> val_word_before_3; // 目標単語の 3 個前の単語の id
  array[val_N] int<lower=1,upper=word_type> val_word_after_3; // 目標単語の 3 個後の単語の id
}
parameters {
  array[word_type] vector[K] word_embedding; // embedding
  array[word_type] vector[K] word_context; // context vector
}
model {
  // embedding と context vector のサンプリング(標準正規分布)
  for (p in 1:word_type){
```

```
    target += normal_lupdf(word_embedding[p] | 0, 1);
    target += normal_lupdf(word_context[p] | 0, 1);
  }

  int grainsize = 1;

  // Reduce sum
  target += reduce_sum(
    partial_sum_lupmf, result, grainsize,
    word,
    word_before_1, word_after_1,
    word_before_2, word_after_2,
    word_before_3, word_after_3,
    word_embedding,
    word_context
  );
}
generated quantities {
  // 検証データに対する分類結果生成
  array[val_N] int predicted;
  for (n in 1:val_N){
    predicted[n] = bernoulli_logit_rng(
      word_embedding[val_word[n]]' *
      (
        word_context[val_word_before_1[n]] + word_context[val_word_after_1[n]] +
        word_context[val_word_before_2[n]] + word_context[val_word_after_2[n]] +
        word_context[val_word_before_3[n]] + word_context[val_word_after_3[n]]
      )
    );
  }
}
```

Wu Tung-Wen Profile

Wu Tung-Wen is a Data Scientist at Product Development Department, dip Corporation and holds a Master's degree in International and Administrative Policy. After completing the program at School of International and Public Policy, Hitotsubashi University, he joined the workforce as a private sector Data Scientist in Tokyo, where he is currently employed. His expertise includes Text-as-Data analysis, causal inference, and quantitative political science and economics focusing on large-scale Bayesian machine learning. His blog can be found at https://qiita.com/Gotoubun taiwan.