Incorporating Domain Knowledge into Stance Classification

スタンス分類における領域知識の活用に関する研究

Akira Sasaki

Graduate School of Information Sciences

Tohoku University

A thesis submitted for the degree of

Doctor of Information Science

January 2018

Acknowledgements

本研究を進めるにあたり，研究室のみなさまをはじめとして非常に多くの方にご助言，ご協力いただきました．ここに感謝の意を表します．指導教官の乾健太郎教授には，学部 4 年次から約 6 年間という期間，非常に多くのご指導をいただきました．私はやや特殊な経緯で乾研究室に配属されることとなりましたが，その際に快く受け入れていただいたことが今の成果に繋がっています．また，研究室外での私の活動に関しても常に後押ししてくださり，大変励みになりました．心より感謝申し上げます．ご多忙の中，審査委員を引き受けてくださった篠原歩教授，大町真一郎教授に深く感謝申し上げます．東京工業大学の岡崎直観教授には，ほぼ毎週というペースで親身にご指導いただきました．研究の初期段階でのアイデアや，手法の詳細に至るまで相談させていただき，日々成長することができました．研究の最中は多くの壁にぶつかりましたが，その度に私を導いていただきました．深く感謝申し上げます．情報通信研究機構の水野淳太さんには，研究に関するご指導のほか，サーバ管理に関しても初歩から丁寧に教えていただきました．そのおかげで，特に大きなサーバ障害に悩まされることもなく，無事にサーバ管理を全うできました．深く感謝申し上げます．PKSHA Technology の渡邉陽太郎さんには，学部，博士前期課程の未熟な時期にご指導をいただきました．深く感謝申し上げます．松田耕史さんには，主に博士前期課程で多くの共同研究をさせていただきました．深く感謝申し上げます．塙一晃君には，直近の多くの研究で共同で行わせていただきました．深く感謝申し上げます．

最後に，常に私の選ぶ道を応援し，支えていただいた家族に深く感謝します．今後とも宜しくお願いします．

Abstract

Social networking services (SNS) such as Twitter and Facebook have rapidly sunk into our lives in recent years. SNS is used not only as a kind of communication tools but also as a place to be used for advertisement, expressing individual opinions, and so on. In particular, there are various kinds regarding opinion. For example, there are many opinions about a specific product. These opinions are also abundant in electronic commerce services like *Amazon*. As these opinions play an important role in improving products, many researchers and companies analyze them. Unlike the review of the above-mentioned products, there are relatively few sites in which opinion on political topics and events is compiled. There are debate sites such as *debate.org* and *idebate.org* in English, and *zzhh.jp* in Japanese, but it is overwhelmingly small compared to product reviews. Because of such a background, analysis of opinions in SNS has been actively performed.

In this thesis, we especially addressed the task called stance classification. In this task, the goal is to identify whether a given text agrees or disagrees a certain topic. However, it cannot be said that the performance of the current state-of-the-art of this task is satisfactory. Various causes can be considered for this, but the main reason for this is thought to be lack of knowledge about topics. Under such a background, in this thesis, we aim at acquiring and applying knowledge that contributes to the performance improvement of stance classification.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Social networking services (SNS) such as Twitter and Facebook have rapidly sunk into our lives in recent years. SNS is used not only as a kind of communication tools but also as a place to be used for advertisement, expressing individual opinions, and so on. In particular, there are various kinds regarding opinion. Examples are as follows:

(1) Galaxy note8 is the best smartphone ever.

(2) TPP ruins the future of our country.

(3) I can't wait for Tokyo Olympics!

Here, (1) is a user's opinion on a specific product. As these opinions play an important role in improving products, many researchers and companies analyze them. Such reviews are also abundant on electronic commerce services like Amazon. On the other hand, (2) and (3) are opinions on political topics and events, respectively. Unlike the review on the above-mentioned products, there are relatively few sites in which opinion on political topics and events is compiled. There are debate sites such as *debate.org* and *idebate.org* in English, and *zzhh.jp* in Japanese, but it is overwhelmingly small compared to product reviews. Because of such a background, analysis of opinions in SNS has been actively performed.

In this thesis, we especially addressed the task called stance classification. In this task, the goal is to identify whether a given text agrees or disagrees a certain topic.

Recently, Task 6 of SemEval-2016[1] is held to solve this task, and many researchers participated. Task 6 of SemEval-2016 is divided into two subtasks A and B, each having the following features.

Task 6 A

In this task, the training data for six topics (Atheism, Climate Change a Real Concern, Feminist Movement, Hillary Clinton, and Legalization of Abortion) are given. The goal is to predict stances of the test data.

Task 6 B

In this task, the goal is to predict stances of the test data regarding Donald Trump. Here, it differs from task 6 A in that training data on Donald Trump is not given, and only unlabeled data concerning Donald Trump is given. However, labeled data related to other topics used in task 6 A is freely available.

For both tasks, the input is a pair of a topic (e.g. Atheism) and a text. The goal is to predict the stance (agree/disagree) of the text in regard to the topic. A noteworthy characteristic of this task is that each text does not necessarily include the topic. The text quoted from [Mohammad et al., 2016a] is shown below:

Text Jeb Bush is the only sane candidate in this republican lineup.

Target Donald Trump

In this example, although the text does not include Donald Trump, we can guess that this text disagrees with Donald Trump. This is because this text is very favorable to Jef Bush, who is other candidates in the United States presidential election. In this way, texts may indicate stances toward a specific topic without explicitly mentioning it. Therefore, the method used in targeted sentiment analysis is partially useful, but by itself, it cannot be solved completely.

Regarding the evaluation, the average of F1-score in regard to agreement/disagreement is used in this task. In SemEval-2016, 19 teams participated in task 6 A and 9 teams participated in task 6 B. Surprisingly, in task 6 A, even a Support Vector Machines (SVM) with simple word n-gram (1, 2, 3-gram) and character n-gram (2, 3, 4,

[1] `http://alt.qcri.org/semeval2016/task6/`

5-gram) defeated all the participants. From this result, it is said that this task is still immature and there is much room for improvement in performance.

Among them, we explain the method which was the highest score in each task. Zarrella and Marsh [2016] propose a prediction method by Recurrent Neural Network (RNN). They collected tweets including hashtags related to topics (#climatechange, #climatescam, etc) in advance, then they pre-trained the RNN by predicting which hashtags are included in the tweets. As a result, their model achieved 67.82 in F1-score. In task 6 B, Wei et al. [2016] propose a prediction method by Convolutional Neural Network (CNN). Among them, in order to overcome the lack of training data on Donald Trump, they focused on phrases and hashtags that agree with Donald Trump (e.g. *go trump*, *#MakeAmericaGreatAgain*) and phrases and hashtags that disagree with Donald Trump (e.g. *idiot*, *fired*). They collected tweets containing these phrases and hashtags, then they treat these tweets as pseudo labeled tweets to bring it into the framework of supervised learning.

However, it cannot be said that the performance of the current state-of-the-art of this task is satisfactory. Various causes can be considered for this, but the main reason for this is thought to be lack of knowledge about topics. As an example, consider the following texts:

(4) It is better to promote domestic consumption.

(5) It is better to promote monetary easing.

Although both of these texts express opinions on TPP, they have different stance: (4) disagrees TPP, (5) agrees TPP. This is caused by the difference in the part of “domestic consumption” and “monetary easing”. The reason why a person can accurately identify these texts is considered to be because people have knowledge such as “TPP promotes monetary easing” and “TPP suppresses domestic consumption”.

Under such a background, in this thesis, we aim at acquiring and applying knowledge that contributes to the performance improvement of stance classification.

1.2 Contribution

The contribution of this thesis is roughly divided into the following three points.

1. Propose an idea of introducing PRIOR-SITUATION/EFFECT relations to stance classification and manually annotating them. Then, we improve the accuracy of stance classification by utilizing these annotations. In addition, we annotated Wikipedia to automatically acquire knowledge such as PROMOTE/SUPPRESS. The annotated dataset is publicly available.

2. In addition to the relationship PROMOTE/SUPPRESS, knowledge such as “A person who agrees with A also agrees with B” or “A person who disagrees with A also disagrees B” is also important in overlooking the opinions of people, and considered to contribute to improve the accuracy of stance classification. Therefore, we modeled such knowledge by matrix factorization, which is widely used in the field of item recommendation. In addition, for users who have not expressed any stances of themselves, we proposed the method to predict their stance from their entire posts.

3. By applying the method proposed in above, we confirmed that it contributes to the improvement of the accuracy of stance classification.

1.3 Thesis Overview

In this section, we explain the structure of this thesis. In Chapter 2, we will explain the method aimed at improving the accuracy of stance classification by manually giving relationships such as PRIOR-SITUATION/EFFECT. In Chapter 3, we describe data annotated causal relations (PROMOTE/SUPPRESS) on Wikipedia. The annotated corpus is expected to be training data for automatic recognition of causal relationships. In Chapter 4, we focus on people’s trends of stances as knowledge other than PROMOTE/SUPPRESS and perform modeling by matrix factorization. In addition, in Chapter 5, we will also describe research that extends the method of Chapter 4 so that it can also consider the silent majority. Finally, in Chapter 6, we review the summary of the above research and its contribution.

Chapter 2

Annotating Related Events about Targets to Improve Stance Classification

In this chapter, as a part of introducing knowledge into stance classification, we annotated related events (PRIOR-SITUATION/EFFECT) about topics manually[1] [Sasaki et al., 2016].

2.1 Introduction

One recent trial of stance classification is Task 6 of SemEval-2016[2]. This is a task to detect a stance (favor or against) in relation towards a topic of a tweet. Consider the following example[3]. The task is to detect a stance for a topic in the text. In this example, the underlined part suggests that a stance of the text towards the topic is FAVOR.

[1] ©2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

[2] `http://alt.qcri.org/semeval2016/task6/`

[3] This example is quoted from trial data of Task 6

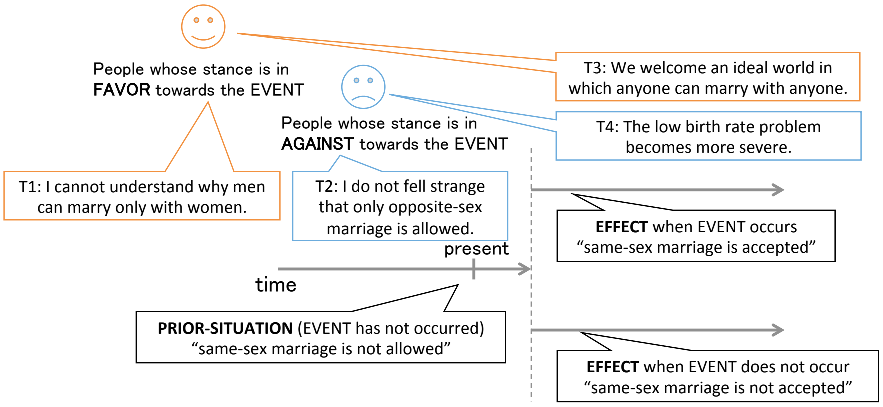

People whose stance is in FAVOR towards the EVENT
People whose stance is in AGAINST towards the EVENT
T3: We welcome an ideal world in which anyone can marry with anyone.
T4: The low birth rate problem becomes more severe.
T1: I cannot understand why men can marry only with women.
T2: I do not fell strange that only opposite-sex marriage is allowed.
present
time
EFFECT when EVENT occurs "same-sex marriage is accepted"
PRIOR-SITUATION (EVENT has not occurred) "same-sex marriage is not allowed"
EFFECT when EVENT does not occur "same-sex marriage is not accepted"

Figure 2.1: Overview of PRIOR-SITUATION and EFFECT of the event "Allowing same-sex marriage."

Text Hillary is our best choice if we truly want to continue being a progressive nation.

Target Hillary Clinton

Stance FAVOR

A popular approach for stance classification uses sentiment polarity towards a topic in a text. The underlined part of the example expresses positive sentiment polarity to "Hillary" corresponding to the topic. This approach is known as targeted sentiment analysis [Mitchell et al., 2013; Zhang et al., 2015].

However, we suffer from a variety of examples for which stance classification is extremely difficult. Figure 2.1 shows four examples for the proposition "allowing same-sex marriage[1]." Note that although our texts are written in Japanese, we provide examples in English for readability. T1 expresses a negative attitude towards the current situation, i.e., "same-sex marriage is not allowed" and T4 expresses negative attitude towards the future situation when "same-sex marriage is accepted." Both texts express negative attitudes, but the stance of T1 is in favor of the proposition and the stance of T4 is against it. To make matters worse, T4 does not even contain keywords representing the topic (e.g., *marriage* nor *marry*) but only a related situation *low birth rate*. Since targeted sentiment analysis methods require that a topic is explicitly mentioned in a

[1] Since same-sex marriage is not allowed under Japanese law, people debate whether it should be permitted or not

Table 2.1: Attitude towards PRIOR-SITUATION/EFFECT and corresponding stance related to the topic.

	PRIOR-SITUATION	EFFECT
positive attitude	against	favorable
negative attitude	favorable	against

text, we cannot directly apply such methods. People often implicitly express their attitudes towards a topic in this way, therefore, it is not a trivial problem. To detect these stances, it is necessary to recognize the situations when the event *occurs* or *does not occur*. We designate the topic as an EVENT (e.g., allowing same-sex marriage) and call the future situation an EFFECT (e.g., same-sex marriage became allowed) hereafter. Note that, the future situation in which the event does not occur can be regarded as the current situation. We designate the current situation as the PRIOR-SITUATION (e.g., same-sex marriage is not allowed).

In this chapter, we focus on situations where the EVENT either occurs or does not occur, and we introduce the first method for annotating such instances. To the best of our knowledge, there is no research focused on such phenomena in either stance classification or sentiment analysis tasks. To predict stances considering these phenomena, we propose a classification method based on machine learning with the PRIOR-SITUATION and EFFECT of EVENT. We first annotate the labels PRIOR-SITUATION and EFFECT to our dataset. Then T1 can be generalized to "I cannot understand why PRIOR-SITUATION" and T4 can be generalized to "The problem of EFFECT becomes more severe." After the generalization, the stance of the text can be detected as favorable when it has a negative attitude to PRIOR-SITUATION (before EVENT occurs) or a positive attitude to EFFECT (after EVENT occurs), or against when the text expresses a positive attitude to PRIOR-SITUATION or a negative attitude to EFFECT (Table 2.1).

Our contributions are two-fold:

1. We propose the concepts of the time variation (i.e., PRIOR-SITUATION and EFFECT) for the first time in the stance classification task and annotated these labels to roughly 3,000 texts.

2. We confirm that the accuracy of stance classification can be improved using these

labels.

2.2 Related Work

Most of the sentiment analysis tasks aim at detecting sentiment polarity (i.e., positive/negative/neutral) of a text or a document without focusing on a specific topic. Kiritchenko et al. [2014] expanded the sentiment lexicon for micro-blogs, and attained better results than the previous sentiment analysis works in the micro-blog domain. Using Long Short-Term Memory (LSTM), Wang et al. [2015] achieved comparable results with data-driven techniques [Hochreiter and Schmidhuber, 1997]. They also showed that tuned word embeddings improve the performance of sentiment analysis. Apart from that, targeted sentiment analysis task set the goal to predict sentiment towards a specific target [Mitchell et al., 2013; Zhang et al., 2015].

As to stance classification task, Murakami and Raymond [2010] and Sridhar et al. [2014] use link-based methods to identify the general positions of users in online debates. Thomas et al. [2006b] classify the speeches of U.S. Congressional floor debates into support of or opposition to proposed legislation. Somasundaran and Wiebe [2009] focus on posts related to debatable topics such as “iPhone vs BlackBerry”, and identify which stance the author of a post is taking. In addition, many works have been undertaken to predict political position (i.e., conservative or liberal) of the text, or of the author of the text [Akoglu, 2014; Bamman and Smith, 2015b; Iyyer et al., 2014; Wong et al., 2013; Zhou et al., 2011]. Chambers et al. [2015] predicted sentiment polarity between each country, which is one kind of targeted sentiment analysis. Furthermore, Tumasjan et al. [2010] conducted analysis of a micro-blog as political sentiment, and predicted the results of the German federal election.

These studies do not consider temporal changes which are caused by an event. This point is a major difference between previous work and ours.

2.3 Stance Classification Task

2.3.1 Data Preparation

Since no dataset for stance classification is available for the Japanese language, we create a dataset from the Japanese debate forum *Zeze-hihi* [1]. Although an existing English dataset is available [Mohammad et al., 2016b], we decided to use Japanese data, because annotating PRIOR-SITUATION and EFFECT requires prior knowledge of each topic[2]. There are widely diverse questions in *Zeze-hihi*, such as politics (*Are you FAVOR or AGAINST accepting same-sex marriage?*), sports (*Which team do you think will win this match, SPAIN or NEDERLAND?*), game (*Do you like watching a gameplay?*), etc. Each question has two choices for voting (e.g. FAVOR/AGAINST, SPAIN/NEDERLAND, LIKE/DON'T LIKE). *Zeze-hihi* users choose questions freely, and answer them. Users can vote and give comment.

We collected questions along with answers from *Zeze-hihi*. The data consists of the following (see also Figure 2.2)

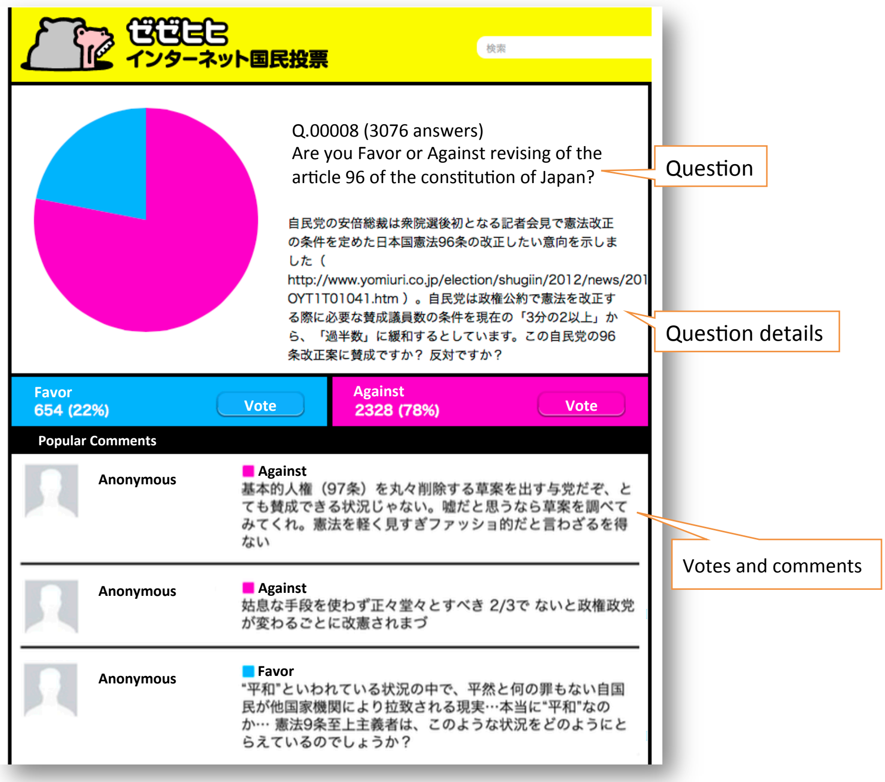
- Questions about debatable topics (e.g. *Are you FAVOR or AGAINST revising article 96 of the Japanese constitution?* [3]).
- Two choices for voting on the question (e.g. FAVOR/AGAINST).
- Votes of users with their comments. (e.g. [FAVOR] *Because article 96 of the Japanese constitution is important, I hope not to revise the article.*) Comments are up to 100 characters.

In this chapter, we set the topic to politics. Therefore, we only address questions that have FAVOR / AGAINST choices. Additionally, we filtered out questions which have less than 150 FAVOR or less than 150 AGAINST. We selected the top 10 most voted questions from them (Table 2.2). Votes with no comment are omitted. To balance them, we randomly select 300 votes (150 FAVOR votes, 150 AGAINST votes) for each questions.

[1] http://zzhh.jp

[2] In this work, an annotator (not the authors) is Japanese

[3] http://zzhh.jp/questions/0008



Q.00008 (3076 answers)
Are you Favor or Against revising of the article 96 of the constitution of Japan?

自民党の安倍総裁は衆院選後初となる記者会見で憲法改正の条件を定めた日本国憲法96条の改正したい意向を示しました（ http://www.yomiuri.co.jp/election/shugiin/2012/news/201 OYT1T01041.htm ）。自民党は政権公約で憲法を改正する際に必要な賛成議員数の条件を現在の「3分の2以上」から、「過半数」に緩和するとしています。この自民党の96条改正案に賛成ですか？ 反対ですか？

Favor
654 (22%) Vote

Against
2328 (78%) Vote

Popular Comments

Anonymous
🟥 Against
基本的人権（97条）を丸々削除する草案を出す与党だぞ、とても賛成できる状況じゃない。嘘だと思うなら草案を調べてみてくれ。憲法を軽く見すぎファッショ的だと言わざるを得ない

Anonymous
🟥 Against
姑息な手段を使わず正々堂々とすべき 2/3で ないと政権政党が変わるごとに改憲されまづ

Anonymous
🟦 Favor
"平和"といわれている状況の中で、平然と何の罪もない自国民が他国家機関により拉致される現実…本当に"平和"なのか… 憲法9条至上主義者は、このような状況をどのようにとらえているのでしょうか？

Figure 2.2: Screen shot of *Zeze-hihi*. For each question, users respond with a vote (e.g. FAVOR/AGAINST) and a comment (up to 100 characters). Note that, we provide an English translation for readability. In addition, we anonymized information of users.

2.3.2 Annotating PRIOR-SITUATION and EFFECT

As described in section 2.1, we adopt concepts of PRIOR-SITUATION and EFFECT to improve the performance of FAVOR/AGAINST classification. In this section, we describe how to annotate PRIOR-SITUATION and EFFECT to comments posted to *Zeze-hihi*.

2.3.2.1 PRIOR-SITUATION

We define the situation before the target event occurs as a PRIOR-SITUATION (i.e., current status). One example is the following.

Table 2.2: Top 10 most voted questions.

Question
Are you FAVOR or AGAINST revising article 9 of the Japanese constitution?
Are you FAVOR or AGAINST revising article 96 of the Japanese constitution?
Are you FAVOR or AGAINST changing constitutional interpretation of the right to collective defense?
Are you FAVOR or AGAINST reducing daily life security expenditures?
Are you FAVOR or AGAINST establishing the system of a husband and wife retaining separate family names?
Are you FAVOR or AGAINST establishing the regulation forbidding gambling using daily life security expenditures?
Are you FAVOR or AGAINST inviting the Olympics to be held in Tokyo?
Are you FAVOR or AGAINST introducing the regional system of division?
Are you FAVOR or AGAINST accepting same-sex marriage?
Are you FAVOR or AGAINST establishing the state secrecy laws?

Question Are you FAVOR or AGAINST revising article 96 of the Japanese constitution?

Event Revising of the article 96 of the Japanese constitution

Vote FAVOR

Comment *The situation in which revising the constitution requires two-thirds agreement of both houses*$_{\text{PRIOR-SITUATION}}$ *is too difficult.*

The underlined part is associated with the current situation related to article 96 of the Japanese constitution.

2.3.2.2 EFFECT

We define the effect of realization of the target event and the effect of NOT realization of the target event as EFFECT. Note that, although EFFECT (Event does not occur) exists in theory, only a few instance correspond to it. Therefore, we do not use this concept in our classification.

The example of EFFECT is the following.

Question Are you FAVOR or AGAINST revising article 96 of the Japanese constitution?

Event Revising of the article 96 of the Japanese constitution

Vote AGAINST

Comment *If the Japanese constitution can be changed easily*$_{\text{EFFECT}}$*, it is useless.*

Although it depends on author's subjectivity, considering the fact that revising article 96 of the Japanese constitution alleviate the condition of revision of the constitution, this substring seems to refer to EFFECT.

Question Are you FAVOR or AGAINST inviting the Olympics to be held in Tokyo?

Event Tokyo's campaign to host the Olympics

Vote AGAINST

Comment *Rather than preparing amusement facilities*$_{\text{EFFECT}}$*, I want the government to restore aging road networks.*

If Tokyo conducts a campaign to host the Olympics, then the Olympics might be held in Tokyo. Then, the government of Japan must prepare amusement facilities for it. In this way, a target event sometime leads to some new events in succession. We also regard an effect of realization of these events as EFFECT.

Using these concepts, we perform annotation of data described in section 2.3.1. This annotation was conducted by one annotator (not the authors). As a result of the annotation, 1,585 answers (52.83%) out of 3,000 answers have at least one of PRIOR-SITUATION/EFFECT. In addition, by defining keywords representing each of the 10 questions [1], we confirmed that 2,108 answers (70.27%) out of 3,000 answers do not have any keywords specific to the question. Among them, 1,090 answers (51.70% of them, namely 36.33% of the whole) have at least one of PRIOR-SITUATION/EFFECT. In other words, compared to the case of only focusing on keywords specific to the question, we can treat 36.33% more answers (66.06% in total) by considering the concepts of PRIOR-SITUATION and EFFECT.

[1] For example, we defined *revise* and *article 96* as keywords for the question *Are you FAVOR or AGAINST revising article 96 of the Japanese constitution?*

2.3.3 FAVOR/AGAINST classification task

We perform the FAVOR/AGAINST classification task using *zeze-hihi* answers that are annotated in section 2.3.2. In this classification task, the input is a comment of an answer with annotated PRIOR-SITUATION/EFFECT labels (e.g. *The situation in which revising constitution requires two-thirds agreement of both houses*$_{\text{PRIOR-SITUATION}}$ *is too difficult.*). The goal of this task is to predict the answer's vote (FAVOR or AGAINST).

2.4 Method

We introduce baseline methods and our proposed method in this section. Because *zeze-hihi* answers are written in Japanese, we tokenize them in advance. We employ `MeCab` (0.996) [Kudo et al., 2004] as a tokenizer, and `mecab-ipadic-neologd` [Sato, 2015] as a dictionary. For example, "*kenpou kaishaku henkou wa muda.*" (Changing constitutional interpretation is useless.)[1] is tokenized as Listing 2.1:

Listing 2.1: Example of tokenization.

```
kenpou/kaishaku/henkou/wa/muda/.
(constitutional/interpretation/changing/is/useless/.)
```

FAVOR/AGAINST classification is a binary classification task. In this chapter, we employ logistic regression to perform a supervised learning, and classify the input text as FAVOR or AGAINST. Note that, although we use the event as a standard for the annotation in section 2.3, we do not use the event as the input. As an implementation of logistic regression, we use `Classias` [2]. When using `Classias`, we set all parameters as default.

2.4.1 Baseline Method

In this section, we explain baseline methods (n-gram baseline, neural network based models, Sentiment lexicon baseline, Nakagawa's model [Nakagawa et al., 2010]) of

[1] In all examples, English follows romanized Japanese

[2] http://www.chokkan.org/software/classias/index.html.en

FAVOR/AGAINST classification. In these baseline methods, we do not use PRIOR-SITUATION/EFFECT labels. We merely use a tokenized answer.

Since our purpose is to investigate the effect of the proposed labels (i.e. PRIOR-SITUATION and EFFECT), we use relatively simple model such as n-gram as baseline methods here.

2.4.1.1 n-gram baseline

We extract n-gram (uni- and bi-grams) from a tokenized answer, and use them as features to perform a supervised learning. For example, Listing 2.2 are extracted from Listing 2.1:

Listing 2.2: Example of n-gram feature.

```
kenpou/kaishaku/henkou/wa/muda/./
kenpou_kaishaku/kaishaku_henkou/
henkou_wa/wa_muda/muda_.
(constitutional/interpretation/changing/is/useless/./
constitutional_interpretation/interpretation_changing/changing_is/is_useless/useless_
    .)
```

Each feature is separated by slash (/), and bi-gram feature is combined by an under score (_).

2.4.1.2 Neural network based models

We employ a variant of neural network models that have been commonly used in sentiment analysis tasks. The implemented models are Long Short-Term Memory [Hochreiter and Schmidhuber, 1997] (LSTM), Bidirectional LSTM (BLSTM), Convolutional Neural Network (CNN) [Kim, 2014], and Neural Attention Model. In all experiments, we set the number of epochs as 20, and the dimension of word embeddings as 128[1]. We set the dimension of hidden layers as 128 (for LSTM, BLSTM, and Neural Attention Model), the number of filters as 64, the width of filter window as 3 (for CNN). All models recieve uni-grams as input. We use `Keras` [2] for implementing these models.

[1] These word embeddings are initialized randomly, and fine-tuned in training.

[2] https://github.com/fchollet/keras

2.4.1.3 Sentiment lexicon baseline

In this baseline, we employ sentiment polarities of words in a tokenized answer, and classify the input text into FAVOR or AGAINST based on these sentiment polarities. The motivation behind the usage of sentiment polarities of words is that sentiment polarities are widely used in sentiment analysis tasks and stance classification tasks [Chambers et al., 2015; Mohammad et al., 2013; Somasundaran and Wiebe, 2009]. We use Japanese Sentiment Polarity Dictionary [Higashiyama et al., 2008; Kobayashi et al., 2007] [1] as a sentiment lexicon. In this lexicon, terms are assigned as positive, negative, or neutral. For example, "*ii*" (good) is assigned positive, "*muda*" (useless) is assigned negative, and "*aisatsu*" (greeting) is assigned neutral. Here, we only use positive terms and negative terms. By counting positive terms and negative terms in the input text, we can define the polarity score as follows:

$$polarity\_score = p_+ - p_- \tag{2.1}$$

Here, p_+ represents the number of positive terms; p_- represents the number of negative terms in the input text. We regard the input text as FAVOR if $polarity\_score$ is greater than zero, otherwise AGAINST. For instance, Listing 2.1 contains the negative term "*muda*". Other terms are not included in the sentiment lexicon. Therefore $polarity\_score = -1$. Then, we classify an answer as FAVOR if its $polarity\_score$ is greater than 0, or classify an answer as AGAINST if its $polarity\_score$ is lower than 0. In terms of answers for which the $polarity\_score$ is 0, we perform classification of two kinds. SEN-P treat these answers as FAVOR, and SEN-N treat these answers as AGAINST.

2.4.1.4 Nakagawa's model

We employ Nakagawa's model [Nakagawa et al., 2010] which is the state-of-the-art method of a Japanese sentiment analysis task. This method is a dependency tree-based. It uses conditional random fields [Lafferty et al., 2001] with hidden variables. As an

[1] http://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resources%2F Japanese%20Sentiment%20Polarity%20Dictionary

implementation of it, we use `extractopinion` [1]. This implementation expects a text as input. The output is a sentiment polarity (positive/negative/neutral). Then, we classify an answer as FAVOR if its sentiment polarity is positive, or classify an answer as AGAINST if its sentiment polarity is negative. Note that, in terms of an answer for which the sentiment polarity is neutral, we perform classification of two kinds. NAK-P treat these answers as FAVOR. NAK-N treat these answers as AGAINST.

2.4.2 Proposed Method

In section 2.4.1, we introduced baseline methods based on previous studies. In this section, we introduce our proposed methods, which use PRIOR-SITUATION/EFFECT labels. Using these labels, we aim to examine whether these labels are effective for FAVOR/AGAINST classification or not.

At first, we replace input texts with PRIOR-SITUATION/EFFECT. We then additionally use a feature of patterns around PRIOR-SITUATION/EFFECT and a feature of sentiment polarities in PRIOR-SITUATION/EFFECT. We introduce them in detail in section 2.4.2.1 to section 2.4.2.3.

2.4.2.1 PRIOR-SITUATION/EFFECT replaced n-gram

For a tokenized text, we replace words labeled PRIOR-SITUATION/EFFECT in section 2.3 with special tokens `%PRIOR-SITUATION%` and `%EFFECT%`. Texts will be simplified by doing this replacement. It is expected that more robust features can be extracted. Consider the following example:

(1) *watashi wa kenpou kaishaku henkou shi te*$_{\mathrm{EFFECT}}$ *hoshii.* (I prefer changing constitutional interpretation$_{\mathrm{EFFECT}}$.)

This text is an answer to the question “Are you FAVOR or AGAINST changing constitutional interpretation of the right to collective defense?”. The event of this question is “changing constitutional interpretation of the right to collective defense”. Then the underlined part of “*kenpou kaishaku henkou shi te*” (changing constitutional interpretation) is annotated EFFECT in section 2.3, because it means the event itself. Next, we tokenize (1) and get Listing 2.3. Then we replace tokens that are in the above underline

[1] https://alaginrc.nict.go.jp/opinion/

with special tokens %EFFECT% (Listing 2.4).

Listing 2.3: Tokenization result of “*watashi wa kenpou kaishaku henkou shi te hoshii.*” (I prefer changing constitutional interpretation.) Note that “*watashi wa*” means “I”.

```
watashi/wa/kenpou/kaishaku/henkou/shi/te/hoshii/.
(I/constitutional/interpretation/changing/prefer/.)
```

Listing 2.4: Replaced tokenization result. Note that, we merged a succession of identical special tokens.

```
watashi/wa/%EFFECT%/hoshii/.
(I/%EFFECT%/prefer/.)
```

When doing this replacement, we merged a succession of identical special tokens (e.g. “%EFFECT%,%EFFECT%,%EFFECT%” becomes “%EFFECT%”). Then we extract n-gram (uni- and bi-grams) from this tokenized text (Listing 2.5).

Listing 2.5: n-gram (uni- and bi-grams) extracted from Listing 2.4

```
watashi/wa/%EFFECT%/hoshii/./
watashi_wa/wa_%EFFECT%/
%EFFECT%_hoshii/hoshii_.
(I/%EFFECT%/prefer/./
I_%EFFECT%/%EFFECT%_prefer/prefer_.)
```

The aim of this replacement is to learn domain-independent features. Consider the following example of the other domain:

(2) *watashi tachi mina koyou zouka shi te*$_{\text{EFFECT}}$ *hoshii.* (We all prefer increasing employment$_{\text{EFFECT}}$.)

This text is an answer to the question “Are you FAVOR or AGAINST inviting the Olympics to be held in Tokyo?”. The event of this question is “inviting the Olympics to be held in Tokyo”, then the underlined part “*koyou zouka shi te*” (increasing employment) is annotated EFFECT in section 2.3, because it is assumed that employment in Tokyo will increase if the Olympics is held in Tokyo. By doing the same procedure as that explained above, we can obtain the following n-gram (Listing 2.6).

Listing 2.6: n-gram extracted from (2) by doing the same procedure as the above. Note that "*watashi tachi mina*" means "We all".

```
watashi/tachi/mina/%EFFECT%/hoshii/./
watashi_tachi/tachi_mina/mina_%EFFECT%/
%EFFECT%_hoshii/hoshii_.
(We/all/%EFFECT%/prefer/
We_all/all_%EFFECT%/%EFFECT%_prefer/prefer_.)
```

Because of the replacement, the feature "%EFFECT%_*hoshii*" (%EFFECT%_prefer) appears in both n-gram of (1) and (2). If the classifier learned (1) is FAVOR, then it can classify (2) as FAVOR using "%EFFECT%_*hoshii*" as a clue. In this manner, it is expected to train the robust classifier by this replacement. Note that, we use only uni-grams for neural network based models.

2.4.2.2 Patterns around PRIOR-SITUATION/EFFECT feature

Because there are various representations among questions, we are concerned about coverage of our training data. Although PRIOR-SITUATION/EFFECT replaced n-gram is aimed at simplifying texts, its classification performance depends heavily on the training data. In contrast, in this method, we semi-automatically gather patterns which tend to indicate FAVOR/AGAINST from the other data. In doing so, it is expected that we can classify texts more correct, even though there are no clues in the training data. Consider the following example:

(3) *san bun no ni*$_{\text{PRIOR-SITUATION}}$ *wa muimi. watashi wa sou giin no kahansuu ga hitsuyou de atte*$_{\text{EFFECT}}$ *hoshii.* (Two-thirds agreements$_{\text{PRIOR-SITUATION}}$ is meaningless. I prefer that revising the constitution requires agreements of the greater part of both houses$_{\text{EFFECT}}$.)

In this example, the author express a negative attitude about PRIOR-SITUATION and positive attitude about EFFECT. However, if we just use sentiment lexicon, then the $polarity\_score$ of this text will be calculated as zero because "*muimi*" (meaningless) is a negative term and "*hoshii*" (prefer) is a positive term. The other terms are neutral. However, because the author's negative attitude related to PRIOR-SITUATION means that he is complaining about the current situation when EVENT is not happening already, this text is apparently FAVOR. Similarly, the author's positive attitudes in relation to EFFECT might indicate that the whole text is FAVOR.

Using this method, we semi-automatically gather patterns which are effective for the classification. When gathering patterns, we use Zeze-hihi's questions except for Table 2.2. Which consists of 93 questions (10,490 answers). These answers also have labels of FAVOR/AGAINST, but PRIOR-SITUATION/EFFECT are not annotated. To gather patterns from these data, we perform the following procedures:

1. Tokenize all 10,490 answers. (The setting is the same as Listing 2.1)

2. Gather sequences of tokens from any noun to the next noun/verb/adjective as pattern, and replace the noun with a special token `%X%`.

3. Sort these patterns by frequency.

4. Select patterns by hand that seem to indicate a positive attitude or negative attitude related to `%X%`. For example, patterns such as "`%X%` *wa muimi*" (`%X%` is meaningless) and "`%X%` *hoshii*" (prefer `%X%`) are selected.

5. Classify these patterns whether indicating a positive attitude or negative attitude related to `%X%`. For example, when "`%X%` *wa muimi*" (`%X%` is meaningless) matches the input text, the text seems to indicate a negative attitude related to `%X%`.

By performing the above procedures, we finally gathered 32 patterns. However, because they are not abstracted, some concern arises that few patterns match. For that reason, we perform the following procedures to gather abstracted patterns.

1. Tokenize all 10,490 answers. (same as above)

2. Gather sequences of tokens from any noun to the next positive/negative term as pattern, replace the noun with a special token `%X%`, and replace the positive/negative term with a special token `%PN%`. The definition of positive/negative terms is the same as the sentiment lexicon baseline.

3. Sort these patterns by frequency.

4. Select patterns by hand that seem to indicate a positive attitude or negative attitude related to `%X%`. For example, patterns such as "`%X%` *wa* `%PN%`" (`%X%` is `%PN%`) are collected. Note that these patterns indicate a positive attitude about `%X%` if `%PN%` is positive term, and indicate a negative attitude with respect to `%X%` if `%PN%` is a negative term.

By performing the above procedures, we finally gathered 23 patterns. When applying these patterns to the input text, `%X%` is assumed to be PRIOR-SITUATION or EFFECT. Then, we activate `%PositiveToX%` when a pattern indicating a positive attitude in relation to `%X%` matches the input text, or we activate `%NegativeToX%` when a pattern indicating a negative attitude about `%X%` matches the input text. Consequently, there are four possible features conditional on `%X%` (`%PositiveToPRIOR-SITUATION%`, `%PositiveToEFFECT%`, `%NegativeToPRIOR-SITUATION%`, and `%NegativeToEFFECT%`). In neural network based models, we concatenated these four binary features and hidden vectors just before the output layer.

Although we extracted patterns semi-automatically here, it is possible to do automatically if a Japanese dataset for targeted sentiment analysis exists. Targeted sentiment analysis is the task that aims at determining the sentiment polarity of a specific topic in text. In this task, the topic is explicitly mentioned in the text. For example, the text "iPhone is awesome." has a positive polarity towards "iPhone", and the text "Don't buy Samsung Galaxy." has a negative polarity towards "Samsung Galaxy". If we had a plenty of training data for this task, we could extract patterns such as "A is awesome" automatically. Although an English dataset exists [Mitchell et al., 2013] [1], there is no existing Japanese dataset. For this reason, it is difficult for us to perform pattern extraction automatically. Since these patterns are independent of the concept of PRIOR-SITUATION/EFFECT, we can improve the method for pattern extraction independently.

Patterns such as these are also used for target sentiment analysis [Chambers et al., 2015].

[1] http://www.m-mitchell.com/code/index.html

2.4.2.3 Sentiment polarity in PRIOR-SITUATION/EFFECT feature

Apart from patterns around the PRIOR-SITUATION/EFFECT, expressions in PRIOR-SITUATION/EFFECT sometimes become an important factor for classifying FAVOR/AGAINST. Consider the following example:

(4) *kokumin touhyou ga naigashiro ni sare sugi*$_{\text{PRIOR-SITUATION}}$.(Referendum is too much neglected$_{\text{PRIOR-SITUATION}}$.)

In this example, although no clue phrases exist for FAVOR/AGAINST classification around the PRIOR-SITUATION/EFFECT, PRIOR-SITUATION itself includes negative term "*naigashiro*" (neglected). Similar to patterns around PRIOR-SITUATION/EFFECT, there is apparently correspondence such that if PRIOR-SITUATION includes a negative attitude, then the whole text is apparently FAVOR, or if EFFECT is containing negative attitude then the whole text is apparently AGAINST. To specify whether PRIOR-SITUATION/EFFECT contains positive or negative attitudes, we do the same way as sentiment lexicon baseline. The only difference between this feature and sentiment lexicon baseline is that this feature takes account of only PRIOR-SITUATION/EFFECT, rather the than whole text. For example, if we calculate $polarity\_score$ of PRIOR-SITUATION as greater than zero (i.e., positive), then we set this feature as `%PositiveInPRIOR-SITUATION%`. Consequently, there are four possible features conditional on $polarity\_score$ (`%PositiveInPRIOR-SITUATION%`, `%PositiveInEFFECT%`, `%NegativeInPRIOR-SITUATION%`, and `%NegativeInEFFECT%`). Note that, if $polarity\_score$ of PRIOR-SITUATION /EFFECT is calculated as zero, then this feature will be not activated. In neural network based models, we concatenated these four binary features and features just before the output layer.

2.5 Evaluation

To evaluate our methods, we measure the accuracy of the FAVOR/AGAINST classification through ten-fold cross validation. For example, we use votes of nine questions except for question "Are you FAVOR or AGAINST accepting same-sex marriage?" as training data. Then we evaluate the classification accuracy on votes of that question.

Then, we calculate the mean of these ten accuracies. Results are presented in Table 2.3. Because of limitations of space, the names of methods are abbreviated as shown below: NGR (n-gram baseline), LSTM (Long Short-Term Memory), BLSTM (Bidirectional LSTM), CNN (Convolutional Neural Network), ATTENTION (Neural Attention Model), SEN-P, SEN-N (Sentiment lexicon baseline, treat neutral as FAVOR, and treat neutral as AGAINST), NAK-P, NAK-N (Nakagawa's model, treat neutral as FAVOR, and treat neutral as AGAINST), REP (PRIOR-SITUATION/EFFECT replaced n-gram), PAT.F (Patterns around PRIOR-SITUATION/EFFECT), SEN.F (Sentiment polarity in PRIOR-SITUATION/EFFECT). Note that, although our proposed methods use PRIOR-SITUATION/EFFECT labels, we can improve the classification accuracy when we use answers that have no PRIOR-SITUATION/EFFECT label in training. Therefore, we use 300 answers of each question in training, and use answers that have at least one PRIOR-SITUATION/EFFECT label in evaluation.

From these results, it can be said that the PRIOR-SITUATION/EFFECT label is effective for FAVOR/AGAINST classification. Specially, REP+PAT.F+SEN.F shows significant improvement over NGR (4.23 point improvement in the classification accuracy). For example, though NGR misclassified following texts, REP+PAT.F+SEN.F correctly classified them.

(1) [gold: FAVOR, system output: FAVOR] Why $\underline{\text{women have to leave the house}}_{\text{PRIOR-SITUATION}}$?

(2) [gold: FAVOR, system output: FAVOR] Because a world-famous event $\underline{\text{enlivens Japan, and it may also make a special demand}}_{\text{EFFECT}}$.

In (1), PRIOR-SITUATION/EFFECT replaced n-gram "Why `%PRIOR-SITUATION%`" makes it possible to correctly classify. In (2), since there are positive terms "enlivens" and "demand" in EFFECT, our proposed method used it as a clue for classifying.

Note that, though Nakagawa's model is the state-of-the-art method of a Japanese sentiment analysis task, its accuracies were lower than baseline methods and proposed methods. This is likely because Nakagawa's model was already trained by corpus of Web data, which is not restricted to the debate domain.

One of the reasons why the classification accuracy remains at about 70% is that the stance classification task is generally very difficult. In Task6 of SemEval-2016, all models were inferior to a baseline method using support vector machine (SVM) with word n-gram and character n-gram [Mohammad et al., 2016b]. Although we tried neural network based models, the accuracies of these models were lower than NGR (logistic regression). This indicates that the text including the concepts of the time variation (i.e., PRIOR-SITUATION and EFFECT) is not easy to solve even for neural network models. On the other hand, neural network models with proposed features show a consistent increase in terms of accuracy, compared to other models without proposed features. Therefore, our proposed features seem to be effective not only for simple models (i.e. n-grams) but also for other sophisticated models (i.e. neural network).

2.5.1 Error Analysis

In this subsection, we investigated texts that were misclassified using the proposed method (REP+PAT.F+SEN.F).

Most errors are caused by not activated PAT.F or SEN.F. The example is the following:

(3) [gold: AGAINST, system output: FAVOR] Because it seems to cause indulge in the Diet$_{\text{EFFECT}}$.

(4) [gold: AGAINST, system output: FAVOR] To avoid changing for the worse, we have to defense that revising the article requires two-thirds of the Diet$_{\text{PRIOR-SITUATION}}$.

In (3), "indulge" indicates a negative attitude, but this term was not in sentiment lexicon. In (4), we would be able to activate `%PositiveToPRIOR-SITUATION%` if "have to defense `%X%`" were in our patterns. These errors seem to decrease if we enrich sentiment lexicon and patterns. This enrichment is left as a subject for our future work.

Next, some errors exists because of multiple opinions included in the text. An example is the following:

(5) [gold: FAVOR, system output: AGAINST] Although changing law easily$_{\text{EFFECT}}$ is bad, simplification of the procedure$_{\text{EFFECT}}$ is needed. Otherwise, old laws will remain.

In (5), the author of the text indicates both a positive attitude and a negative attitude related to `%EFFECT%`. To tackle this problem, one possible solution is to change feature weights according to activated position in the text. For instance, if the author presents a negative attitude in relation to `%EFFECT%` in the first half of the text and presents a positive attitude about `%EFFECT%` in the latter half of the text, then the author is assumed to be in FAVOR of EFFECT all.

Furthermore, some errors are more complicated. One example is the following:

(6) [gold: AGAINST, system output: FAVOR] Because it is important decisions, we must be careful. I wish everyone to consider why people involved in framing article-96 decided that revising the article should require two-thirds of the Diet$_{\text{PRIOR-SITUATION}}$.

2.6 Conclusion

As described herein, we proposed the concepts of the time variation for the first time in the stance classification task and labeled texts collected from different domains. Then, we demonstrated that many texts cannot be classified into FAVOR/AGAINST without PRIOR-SITUATION/EFFECT. Additionally, we performed FAVOR/AGAINST classification with PRIOR-SITUATION/EFFECT, and showed improved classification accuracy. In future work, we plan to gather knowledge related to PRIOR-SITUATION/ EFFECT from Wikipedia, Twitter, and so on, and to use this knowledge to label PRIOR-SITUATION/EFFECT automatically. Consider the following examples:

(8) Same-sex marriage causes the low birth rate problem.

(9) Same-sex marriage changes the situation that man can marry only with a woman.

If we have the prior knowledge that a pattern “A causes B” indicates that B is an EFFECT of A and that a pattern “A changes the situation that B” indicates that B is a PRIOR-SITUATION of A, we can retrieve PRIOR-SITUATION/EFFECT of “same-sex marriage” from (8) and (9). These knowledge are similar to that of Hashimoto et al. [2012b]. They performs acquisition of excitatory templates (e.g. “cause X, increase X”) and inhibitory templates (e.g. “prevent X, discard X”) by bootstrapping. Their method will be an immediate next step for automatic labeling of PRIOR-SITUATION/EFFECT.

Table 2.3: Classification Results of FAVOR/AGAINST Classification (Bold Shows the Best Performance)

Method		Mean Accuracy
Baseline Method	NGR	65.59
	LSTM	56.65
	BLSTM	56.38
	CNN	56.23
	ATTENTION	58.49
	SEN-P	56.71
	SEN-N	57.07
	NAK-P	56.70
	NAK-N	54.92
Proposed Method	REP	67.17
	REP+PAT.F	68.02
	REP+SEN.F	68.85
	REP+PAT.F+SEN.F	**69.82**
Proposed Method (LSTM with)	REP	57.49
	REP+PAT.F	58.28
	REP+SEN.F	58.73
	REP+PAT.F+SEN.F	59.29
Proposed Method (BLSTM with)	REP	58.49
	REP+PAT.F	57.19
	REP+SEN.F	58.52
	REP+PAT.F+SEN.F	58.82
Proposed Method (CNN with)	REP	56.65
	REP+PAT.F	56.94
	REP+SEN.F	56.84
	REP+PAT.F+SEN.F	58.82
Proposed Method (ATTENTION with)	REP	57.64
	REP+PAT.F	58.27
	REP+SEN.F	59.53
	REP+PAT.F+SEN.F	60.87

Chapter 3

Annotating Causal Relation Instances in Wikipedia to Automatically Recognize Causal Relation

In Chapter 2, we focused on the knowledge of PRIOR-SITUATION/EFFECT and showed that the accuracy of stance classification improves by annotating that knowledge manually. However, there is a limit to manually assigning this knowledge to large scale texts, so it is necessary to give them automatically. In this chapter, in order to solve the problem, annotate knowledge of causal relation to the Wikipedia corpus. Hereafter, we treat PROMOTE/SUPPRESS as almost same as EFFECT/PRIOR-SITUATION.

3.1 Introduction

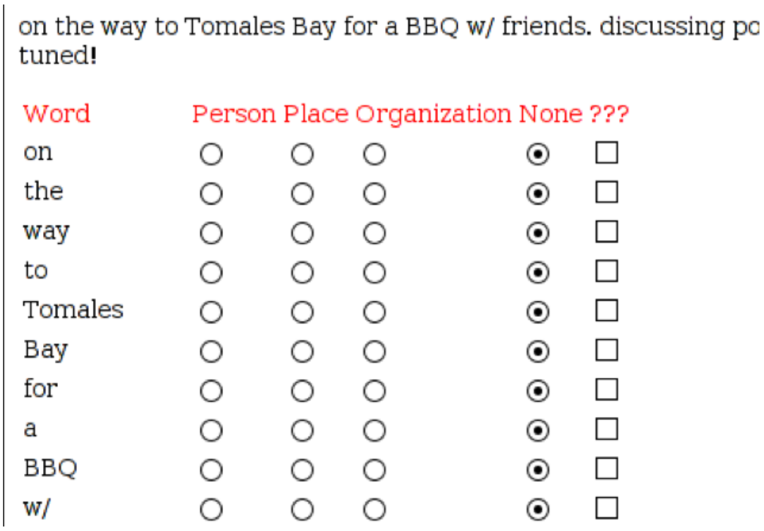
Commonsense knowledge on entity, event and causal relationships plays an important role in recent NLP tasks, such as question answering [Oh et al., 2013, 2016; Sharp et al., 2016], hypothesis generation [Hashimoto et al., 2015a; Radinsky et al., 2012a], and stance classification [Sasaki et al., 2016].

In many previous researches, corpora for acquiring causal relations were built by annotating two text spans (e.g., entities) and their relations in the text [Doddington et al., 2004; Dunietz et al., 2017; Hendrickx et al., 2010; Pyysalo et al., 2015; Rehbein and Ruppenhofer, 2017; Rinaldi et al., 2016]. However, these methods are costly. It involves many tasks, such as choosing a target domain, designing an ontology (semantic classes) of entities, designing an annotation guideline for relations, and annotating the relations between entities. Building such a corpus also requires the annotation efforts of experts. For these reasons, an approach which is scalable to various domains or genres is desired. This chapter presents an approach for annotating causal relation instances to Wikipedia articles via crowdsourcing.

In recent years, crowdsourcing services are used by many researchers in natural language processing [Brew et al., 2010; Finin et al., 2010; Fort et al., 2011; Gormley et al., 2010; Hovy et al., 2014; Jha et al., 2010; Kawahara et al., 2014; Lawson et al., 2010; Takase et al., 2016]. However, it is impossible to make complicated annotations like causal relationships with the existing crowdsourcing frameworks. This is because existing crowdsourcing frameworks were limited to relatively simple input such as multiple choice questions and free descriptions.

In this research, we examine the feasibility of annotating causal relation instances by crowdsourcing. For this reason, we implement a simple micro-task to annotate the parts corresponding to the causal relation in the article on Wikipedia. In addition, we propose a method to link such an annotation system with existing crowdsourcing service. Since we use brat[1] [Stenetorp et al., 2012] widely used in existing research in NLP, our method is applicable to general purpose not limited to causal relation.

We acquired annotations on 95,008 causal relation instances from 8,745 summary sentences[2] contained in 1,494 Wikipedia articles. We publish the annotation system and corpus proposed in this research on the website[3]. Although this corpus was given to a Japanese Wikipedia article, we here use English translations for illustrative purposes.



on the way to Tomales Bay for a BBQ w/ friends. discussing pc
tuned!

Word	Person	Place	Organization	None	???
on	○	○	○	◉	□
the	○	○	○	◉	□
way	○	○	○	◉	□
to	○	○	○	◉	□
Tomales	○	○	○	◉	□
Bay	○	○	○	◉	□
for	○	○	○	◉	□
a	○	○	○	◉	□
BBQ	○	○	○	◉	□
w/	○	○	○	◉	□

Figure 3.1: Named entity annotation by the multiple-choice method [Finin et al., 2010].

3.2 Related work

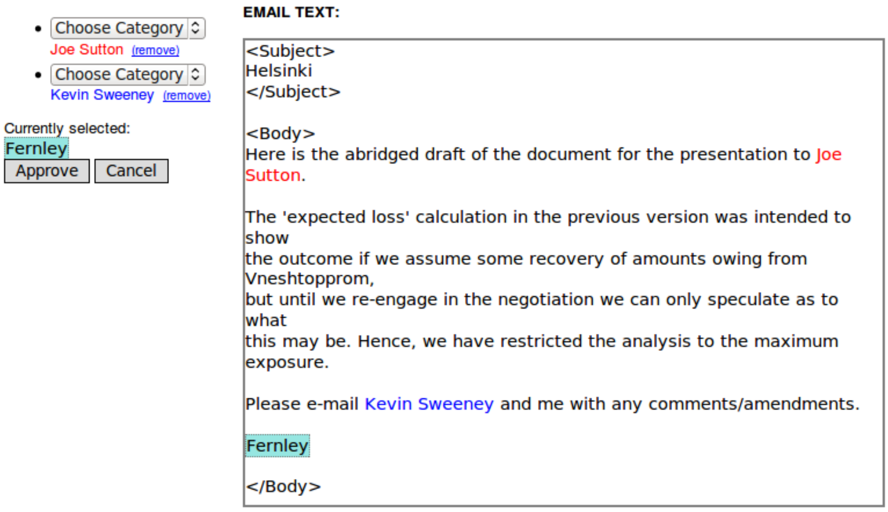
NLP researchers have created corpora in crowdsourcing on a number of tasks. These tasks include part-of-speech tagging [Hovy et al., 2014], PP attachment [Jha et al., 2010], named entity recognition [Finin et al., 2010; Lawson et al., 2010], sentiment classification [Brew et al., 2010], relation extraction [Gormley et al., 2010], semantic modeling of relation patterns [Takase et al., 2016], and discourse parsing [Kawahara et al., 2014]. In most of these tasks, the micro-tasks are designed as multiple-choice problems. For example, Brew et al. [2010] has let the workers annotate positive, negative, or irrelevant on the article. If the target task cannot be shaped like multiple-choice problems, a special approach is required. In particular, labeling text spans cannot be done in multiple-choice problems.

Nevertheless, in some studies, spans have been annotated by crowdsourcing. Finin et al. [2010] annotated the boundary and semantic class of the named entity by transforming the annotation task into the micro-task of multiple-choice problems. They applied the standard interface of Amazon Mechanical Turk (see Figure 3.1). This in-

[1] http://brat.nlplab.org/

[2] The lead paragraph of a Wikipedia article containing a quick summary of the most important points of the article.

[3] http://www.cl.ecei.tohoku.ac.jp/wikipedia_pro_sup/



EMAIL TEXT:

```
<Subject>
Helsinki
</Subject>

<Body>
Here is the abridged draft of the document for the presentation to Joe
Sutton.

The 'expected loss' calculation in the previous version was intended to
show
the outcome if we assume some recovery of amounts owing from
Vneshtopprom,
but until we re-engage in the negotiation we can only speculate as to
what
this may be. Hence, we have restricted the analysis to the maximum
exposure.

Please e-mail Kevin Sweeney and me with any comments/amendments.

Fernley

</Body>
```

Figure 3.2: A custom interface for annotating named entities via crowdsourcing [Lawson et al., 2010].

terface is less readable, and the worker needs to press the radio button for each word. The most relevant to our research is [Lawson et al., 2010] They provided the interface that allows workers to select arbitrary sections in the text and give labels (see Figure 3.2). However, their research focuses on named entity recognition and cannot be generalized to other annotation tasks. In addition, their tools are not published.

In contrast, we propose a framework to facilitate complicated annotation for workers by combining crowdsourcing with brat, an open-source software commonly used in natural language processing. Our proposed method is not limited to causal relations, but corresponds to various annotation tasks that can be performed by brat.

Regarding the causal relations, Dunietz et al. [2017] present the version 2.0 of Bank of Effects and Causes Stated Explicitly (BECauSE). Rehbein and Ruppenhofer [2017] built a German corpus with a similar annotation scheme. Unlike previous research, our research aims to acquire real-world causal knowledge by using Wikipedia.

3.3 Annotating promotion/suppression relations in Wikipedia articles

3.3.1 Labels of causal relations

In this work, we annotate promotion/suppression relations [Fluck et al., 2015; Hashimoto et al., 2012a] in Wikipedia articles. Here, "*X* promotes *Y*" means that *Y* is activated when *X* is activated. Analogously, "*X* suppresses *Y*" means that *Y* is inactivated when *X* is activated.

Here, we focus on the fact that each article in Wikipedia contains knowledge about the article title (*T*, hereafter). Therefore, we consider promotion/suppression relations with *T* as an argument. The annotation task is performed by labeling PRO ("*T* promotes *Y*"), SUP ("*T* suppresses *Y*"), PRO_BY ("*X* promotes *T*"), or SUP_BY ("*X* suppresses *T*") for text spans (denoted by *Y* for PRO and SUP, and denoted by *X* for PRO_BY and SUP_BY) in the article.

We randomly selected 1,494 articles belonging to nine categories and to the subcategories/sub-subcategories: "Social issues", "Disasters", "Diseases and disorders", "Innovation", "Policy", "Finance", "Energy technology", "Biomolecules" and "Nutrients".

3.3.2 Annotation policy

In this chapter, we examined two kinds of units to be annotated: noun phrases and verb phrases. However, none of these units is inadequate to annotate promotion/suppression relations.

In order to explain this, we quote a Wikipedia article about "Nyctalopia"[1].

> Nyctalopia, also called night-blindness, is a condition making it difficult to see in relatively low light. Nyctalopia may exist from birth, or be caused by injury or severe malnutrition.

Here, if we limit the annotation unit to noun phrases, we cannot annotate ⟨SUP, nyctalopia, see in relatively low light⟩. Similarly, if we limit the annotation unit to verb phrases, we cannot annotate ⟨PRO_BY, nyctalopia, injury⟩.

[1] https://en.wikipedia.org/wiki/Nyctalopia

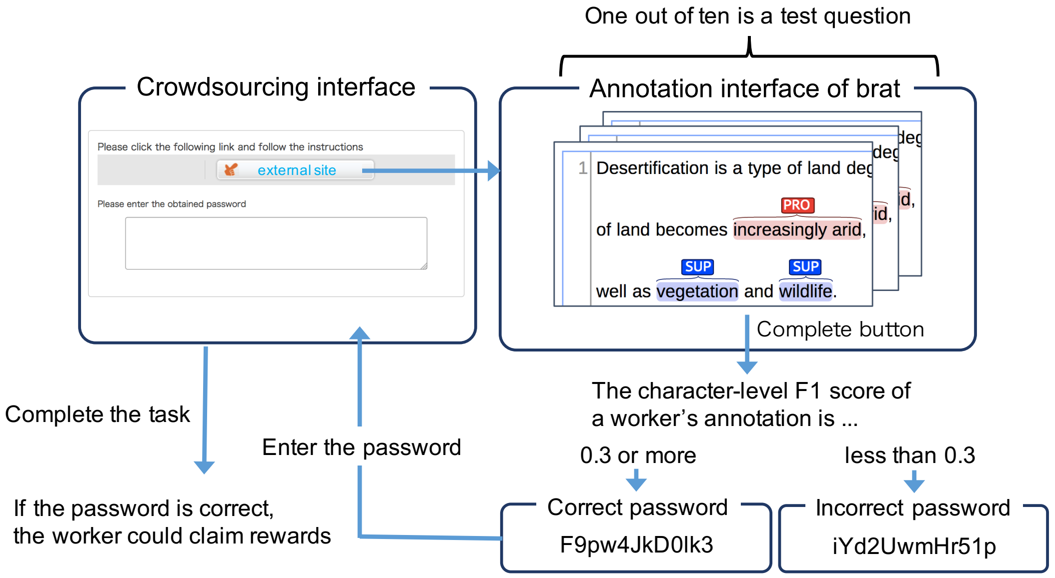

One out of ten is a test question
Crowdsourcing interface
Please click the following link and follow the instructions
external site
Please enter the obtained password
Annotation interface of brat
1 Desertification is a type of land deg
PRO
of land becomes increasingly arid,
SUP
SUP
well as vegetation and wildlife.
Complete button
Complete the task
Enter the password
The character-level F1 score of
a worker's annotation is ...
0.3 or more
less than 0.3
If the password is correct,
the worker could claim rewards
Correct password
F9pw4JkD0lk3
Incorrect password
iYd2UwmHr51p

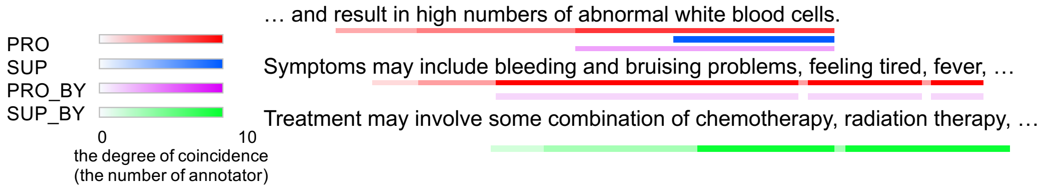
Figure 3.3: Overview of the annotation system integrating Yahoo! crowdsourcing and brat

In addition, there is a problem that segmentation of the annotation cannot be determined uniquely. For example, "severe malnutrition" and "malnutrition" can be regarded as caused by nyctalopia as well. Here, it is difficult to define that one of them is the correct answer. Therefore, we collected annotations by multiple crowd workers without creating detailed instructions of the annotation. As a side effect of this, the more workers annotated the segment, it can be regarded as having high confidence. Thus, this corpus will be useful for improving the annotation scheme for causal relations.

3.3.3 Using brat in crowdsourcing

In crowdsourcing, since annotations are performed by many and unspecified users, quality control is indispensable. In many existing crowdsourcing tasks, quality of a worker was measured by test questions with the correct answers prepared by the task designer.

However, we cannot perform quality control by exact match, because the answer cannot be uniquely determined in our settings. Therefore, we perform quality control



… and result in high numbers of abnormal white blood cells.
PRO
SUP
PRO_BY
SUP_BY
Symptoms may include bleeding and bruising problems, feeling tired, fever, …
Treatment may involve some combination of chemotherapy, radiation therapy, …
0
10
the degree of coincidence
(the number of annotator)

Figure 3.4: An example of annotation results for "Leukemia" on Wikipedia. The color at the bottom of the text shows the relations, and the color intensity shows the number of workers who annotated.

on brat by measuring character-level F1 score between worker's annotation and correct annotation and feeding it back to crowdsourcing service.

Figure 3.3 is an overview of the proposed system. The detailed procedure of annotation is as follows:

1. Workers are led from the crowdsourcing screen to the annotation screen in brat.
2. Workers perform annotations on brat.
3. We measure character-level F1 score between worker's annotation and correct annotation. When F1 score exceeds the threshold (0.3), the correct password is presented, otherwise the incorrect password is presented to the worker.
4. Workers return to the crowdsourcing screen and input the password. At that time, reward is given only when it is correct.

3.4 Annotation results

Using the above system and Yahoo! crowdsourcing service[1], we collected ten annotations for each article. Here, we prepared separate tasks for each promotion/suppression relation PRO, SUP, PRO_BY and SUP_BY. This allows workers to annotate without annoying other relations. Figure 3.4 shows an example of annotation results for "Leukemia" on Wikipedia. Here, for example, it can be seen that many workers judged that leukemia causes "abnormal white blood cells" and "high numbers of abnormal white blood cells".

[1] https://crowdsourcing.yahoo.co.jp/

	PRO	SUP	PRO_BY	SUP_BY
Exact	0.192	0.192	0.132	0.197
Partial	0.448	0.325	0.379	0.380
Character	0.332	0.282	0.309	0.317

Table 3.1: Inter-annotator agreement of each relation (micro F1 score)

3.4.1 Inter-annotator agreement

Table 3.1 shows the average of inter-annotator agreements of each relation. The agreement between two annotations was measured by the F1 score of each of exact match, partial match, and character-level match. We obtained the agreement of annotation for an article by micro-averaging the agreements of all $({}_{10}C_2 = 45)$ pairs of workers. The exact match regards two annotations as matched when the segments are exactly same. The partial match regards two annotation as matched when the segments are partially overlapped. Although these inter-annotator agreements are relatively low, this is reasonable considering the difficulty of the task.

3.5 Conclusion

In this chapter we proposed a method to annotate causal relations to Wikipedia article by crowdsourcing. For this purpose, we developed the system that combines crowdsourcing an brat, an annotation tool widely used in natural language processing. The annotated corpus not only provides training data of models for automatically giving causal relation, and it is also a valuable corpus including confidence based on the number of workers annotated. Moreover, our proposed method can be applied not only to causal relation but also to any annotation task that can be performed using brat.

Chapter 4

Modeling Inter-Topic Preference using Tweets and Matrix Factorization

In Chapter 2 and Chapter 3, we focused on knowledge of causal relationships. In this chapter, we focus on people's trends of stances as knowledge other than causal relationships and perform modeling by matrix factorization.

4.1 Introduction

Social media have changed the way people shape public opinion. The latest survey by the Pew Research Center reported that a majority of US adults (62%) obtain news via social media, and of those, 18% do so often [Gottfried and Shearer, 2016]. Given that news and opinions are shared and amplified by friend networks of individuals [Jamieson and Cappella, 2008], individuals are thereby isolated from information that does not fit well with their opinions [Pariser, 2011]. Ironically, cutting-edge social media technologies promote ideological groups even with its potential to deliver diverse information.

A large number of studies already analyzed discussions, interactions, influences, and communities on social media along the political spectrum from liberal to conservative [Adamic and Glance, 2005; Bakshy et al., 2015; Cohen and Ruths, 2013; Wong et al., 2016; Zhou et al., 2011]. Even though these studies provide intuitive visualizations and interpretations along the liberal-conservative axis, political analysts argue that the axis is flawed and insufficient for representing public opinion and ideologies [Kerlinger, 1984; Maddox and Lilie, 1984].

A potential solution for analyzing multiple axes of the political spectrum on social media is stance classification [Anand et al., 2011; Johnson and Goldwasser, 2016; Mohammad et al., 2016a; Murakami and Raymond, 2010; Somasundaran and Wiebe, 2009; Thomas et al., 2006a; Walker et al., 2012], whose task is to determine whether the author of a text is for, neutral, or against a topic (e.g., *free trade*, *immigration*, *abortion*). However, stance classification across different topics is extremely difficult. Anand et al. [2011] reported that a sophisticated method with topic-dependent features substantially improved the performance of stance classification within a topic, but such an approach could not outperform a baseline method with simple n-gram features when evaluated across topics. More recently, all participants of SemEval-2016 Task 6A (with five topics) could not outperform the baseline supervised method using n-gram features [Mohammad et al., 2016a].

In addition, stance classification encounters difficulties with different user types. Cohen and Ruths [2013] observed that existing methods on stance classification fail on "ordinary" users because such methods primarily obtain training and test data from politically vocal users (e.g., politicians); for example, they found that a stance detector trained on a dataset with politicians achieved 91% accuracy on other politicians but only achieved 54% accuracy on "ordinary" users. Establishing a bridge across different topics and users remains a major challenge not only in stance classification, but also in social media analytics.

An important component in establishing this bridge is commonsense knowledge about topics. For example, consider a topic *a revision of Article 96 of the Japanese Constitution*. We infer that the statement "we should maintain armed forces" tends to favor this topic even without any lexical overlap between the topic and the statement. This inference is reasonable because: the writer of the statement favors *armed forces*; those who favor *armed forces* also favor *a revision of Article 9*[1]; and those who favor *a revision of Article 9* also favor *a revision of Article 96*[2]. In general, this kind of commonsense knowledge can be expressed in the format: *those who agree/disagree*

[1] Article 9 prohibits armed forces in Japan.

[2] Article 96 specifies high requirements for making amendments to Constitution of Japan (including Article 9).

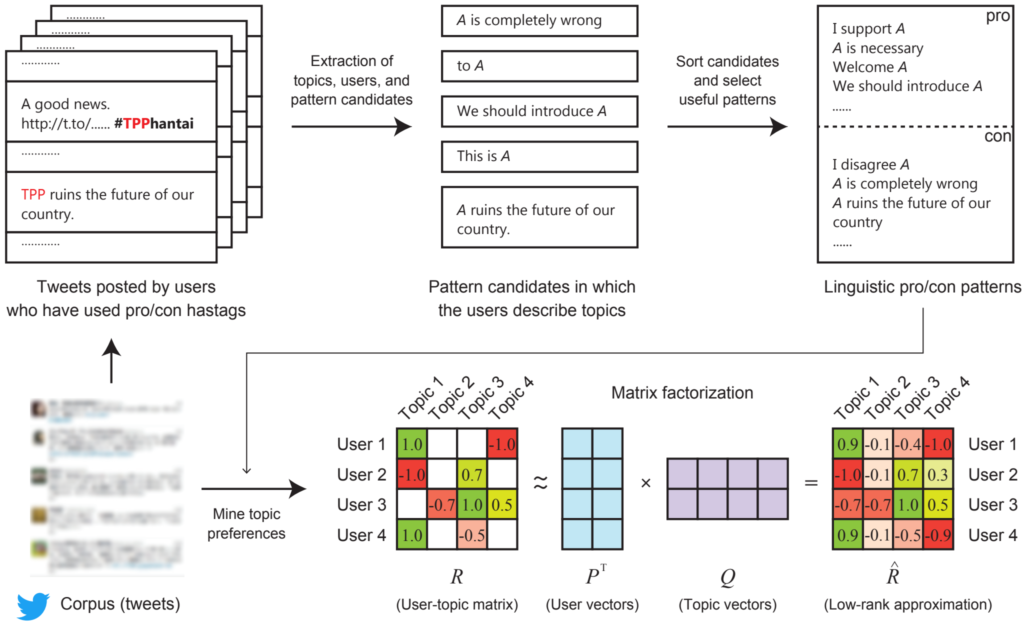


Figure 4.1: An overview of this study.

with topic A also agree/disagree with topic B. We call this kind of knowledge *inter-topic preference* throughout this thesis.

We conjecture that previous work on stance classification indirectly learns inter-topic preferences within the same target through the use of n-gram features on a supervision data. In contrast, in the present chapter, we directly acquire inter-topic preferences from an unlabeled corpus of tweets. This acquired knowledge regarding inter-topic preferences is useful not only for stance classification, but also for various real-world applications including public opinion survey, electoral campaigns, electoral predictions, and online debates.

Figure 4.1 provides an overview of this work. In our system, we extract linguistic patterns in which people agree and disagree about specific topics (e.g., "$\underline{A}$ is completely wrong"); to accomplish this, as described in Section 4.2.1, we make use of hashtags within a large collection of tweets. The patterns are then used to extract instances of users' preferences regarding various topics, as detailed in Section 4.2.2. Inspired by previous work on item recommendation, in Section 4.3, we formalize the task of modeling inter-topic preferences as a matrix factorization: representing a sparse user-topic matrix (i.e., the extracted instances) with the product of low-rank user and topic matrices. These low-rank matrices provide *latent vector representations* of both users and topics. This approach is also useful for completing preferences of "ordinary" (i.e., less vocal) users, which fills the gap between different types of users.

The contributions of this chapter are threefold.

1. To the best of our knowledge, this is the first study that models inter-topic preferences for unlimited targets on real-world data.

2. Our experimental results show that this approach can accurately predict missing topic preferences of users accurately (80–94%).

3. Our experimental results also demonstrate that the latent vector representations of topics successfully encode inter-topic preferences, e.g., *those who agree with nuclear power plants also agree with nuclear fuel cycles*.

This study uses a Japanese Twitter corpus because of its availability from the authors, but the core idea is applicable to any language.

4.2 Mining Topic Preferences of Users

In this section, we describe how we collect statements in which users agree or disagree with various topics on Twitter, which then serves as source data for modeling inter-topic preferences. More formally, we are interested in acquiring a collection of tuples (u, t, v), where: $u \in U$ is a user; U is the set of all users on Twitter; $t \in T$ is a topic; T is the set of all topics; and $v \in \{+1, -1\}$ is $+1$ when the user u agrees with the topic t and -1 otherwise (i.e., disagreement).

Throughout this work, we use a corpus consisting of 35,328,745,115 Japanese tweets (7,340,730 users) crawled from February 6, 2013 to September 30, 2016. We removed retweets from the corpus.

4.2.1 Mining Linguistic Patterns of Agreement and Disagreement

We use linguistic patterns to extract tuples (u, t, v) from the aforementioned corpus. More specifically, when a tweet message matches to one of linguistic patterns of agreement (e.g., "$\underline{t}$ is necessary"), we regard that the author u of the tweet agrees with topic

t. Conversely, a statement of disagreement is identified by linguistic patterns for disagreement (e.g., "$\underline{t}$ is unacceptable").

In order to design linguistic patterns, we focus on hashtags appearing in the corpus that have been popular clues for locating subjective statements such as sentiments [Davidov et al., 2010], emotions [Qadir and Riloff, 2014], and ironies [Van Hee et al., 2016]. Hashtags are also useful for finding strong supporters and critics, as well as their target topics; for example, `#immigrantsWelcome` indicates that the author favors *immigrants*; and `#StopAbortion` is against *abortion*.

Based on this intuition, we design regular expressions for both *pro hashtags* "`#(.+)sansei`"[1] and *con hashtags* "`#(.+)hantai`"[2], where `(.+)` matches a target topic. These regular expressions can find users who have strong preferences to topics. Using this approach, we extracted 31,068 occurrences of pro/con hashtags used by 18,582 users for 4,899 topics. We regard the set of topics found using this procedure as set of target topics T in this study.

Each time we encounter a tweet containing a pro/con hashtag, we searched for corresponding textual statements as follows. Suppose that a tweet includes a hashtag (e.g., `#TPPsansei`) for a topic t (e.g., *TPP*). Assuming that the author of the given tweet does not change their attitude toward a topic over time, we search for other tweets posted by the same author that also have the topic keyword t. This process retrieves tweets like "I support TPP." Then, we replace the topic keyword into a variable A to extract patterns, e.g., "I support $\underline{A}$." Here, the definition of the pattern unit is language specific. For Japanese tweets, we simply recognize a pattern that starts with a variable (i.e., topic) and ends at the end of the sentence[3].

Because this procedure also extracts useless patterns such as "to A" and "this is A", we manually choose useful patterns in a systematic way: sort patterns in descending order of the number of users who use the pattern; and check the sorted list of patterns manually; and remove useless patterns. Using this approach, we obtained 100 pro

[1] Unlike English hashtags, we systematically attach a noun *sansei*, which stands for *pro* (agreement) in Japanese, to a topic, for example, `#TPPsansei`. This thesis uses the alphabetical expression `sansei` only for explanation; the actual pattern uses Chinese characters corresponding to *sansei*.

[2] A Japanese noun *hantai* stands for *con* (disagreement), for example, `#TPPhantai`. This thesis uses the alphabetical expression `hantai` only for explanation; the actual pattern uses Chinese characters corresponding to *hantai*.

[3] In English, this treatment roughly corresponds to extracting a verb phrase with the variable A.

patterns (e.g., “welcome A” and “A is necessary”) and 100 con patterns (“do not let A” and “I don’t want A”).

4.2.2 Extracting Instances of Topic Preferences

By using the pro and con patterns acquired using the approach described in Section 4.2.1, we extract instances of (u, t, v) as follows. When a sentence in a tweet whose author is user u matches one of the pro patterns (e.g., “$\underline{t}$ *is necessary*”) and the topic t is included in the set of target topics T, we recognize this as an instance of $(u, t, +1)$. Similarly, when a sentence matches one of the con patterns (e.g., “I don’t want $\underline{t}$”) and the topic t is included in the set of target topics T, we recognize this as an instance of $(u, t, -1)$. Using this approach, we collected 25,805,909 tuples corresponding to 3,302,613 users and 4,899 topics. Because these collected tuples included comparatively infrequent users and topics, we removed users and topics that appeared less than five times. In addition, there were also meaningless frequent topics such as “of” and “it”. Therefore, we sorted topics in descending order of their co-occurrence frequencies with each of the pro patterns and con patterns, and then removed meaningless topics in the top 100 topics. This resulted in 9,961,509 tuples regarding 273,417 users and 2,323 topics.

4.3 Matrix Factorization

Using the methods described in Section 4.2, from the corpus, we collected a number of instances of users’ preferences regarding various topics. However, Twitter users do not necessarily express preferences for all topics. In addition, it is by nature impossible to predict whether a new (i.e., nonexistent in the data) user agrees or disagrees with given topics. Therefore, in this section, we apply matrix factorization [Koren et al., 2009] in order to predict missing values, inspired by research regarding item recommendation [Bell and Koren, 2007; Dror et al., 2011]. In essence, matrix factorization maps both users and topics onto a latent feature space that abstracts topic preferences of users.

Here, let R be a sparse matrix of $|U| \times |T|$. Only when a user u expresses a preference for topic t do we compute an element of the sparse matrix $r_{u,t}$,

$$r_{u,t} = \frac{\#(u,t,+1) - \#(u,t,-1)}{\#(u,t,+1) + \#(u,t,-1)} \tag{4.1}$$

Here, $\#(u,t,+1)$ and $\#(u,t,-1)$ represent the numbers of occurrences of instances $(u,t,+1)$ and $(u,t,-1)$, respectively. Thus, an element $r_{u,t}$ approaches $+1$ as the user u favors the topic t, and -1 otherwise. If the user u does not make any statement regarding the topic t (i.e., neither $(u,t,+1)$ nor $(u,t,-1)$ exists in the data), we do not fill the corresponding element, leaving it as a missing value.

Matrix factorization decomposes the sparse matrix R into low-dimensional matrices $P \in \mathbb{R}^{k\times|U|}$ and $Q \in \mathbb{R}^{k\times|T|}$, where k is a parameter that specifies the number of dimensions of the latent space. We minimize the following objective function to find the matrices P and Q,

$$\begin{aligned} \min_{P,Q} \sum_{(u,t)\in R} \Bigg(& (r_{u,t} - \boldsymbol{p}_u{}^{\intercal}\boldsymbol{q}_t)^2 \\ & + \lambda_P \left\|\boldsymbol{p}_u\right\|^2 + \lambda_Q \left\|\boldsymbol{q}_t\right\|^2 \Bigg). \end{aligned} \tag{4.2}$$

Here, $(u,t) \in R$ is repeated for elements filled in the sparse matrix R, $\boldsymbol{p}_u \in \mathbb{R}^k$ and $\boldsymbol{q}_v \in \mathbb{R}^k$ are u column vectors of P and v column vectors of Q, respectively, and $\lambda_P \geq 0$ and $\lambda_Q \geq 0$ represent coefficients of regularization terms. We call $\boldsymbol{p}_u$ and $\boldsymbol{q}_t$ the *user vector* and *topic vector*, respectively.

Using these user and topic vectors, we can predict an element $\hat{r}_{u,t}$ that may be missing in the original matrix R,

$$\hat{r}_{u,t} \simeq \boldsymbol{p}_u{}^{\intercal}\boldsymbol{q}_t. \tag{4.3}$$

We use `libmf`[1] [Chin et al., 2015] to solve the optimization problem in Equation 4.2. We set regularization coefficients $\lambda_P = 0.1$ and $\lambda_Q = 0.1$ and use default values for the other parameters of `libmf`.

[1] `https://github.com/cjlin1/libmf`

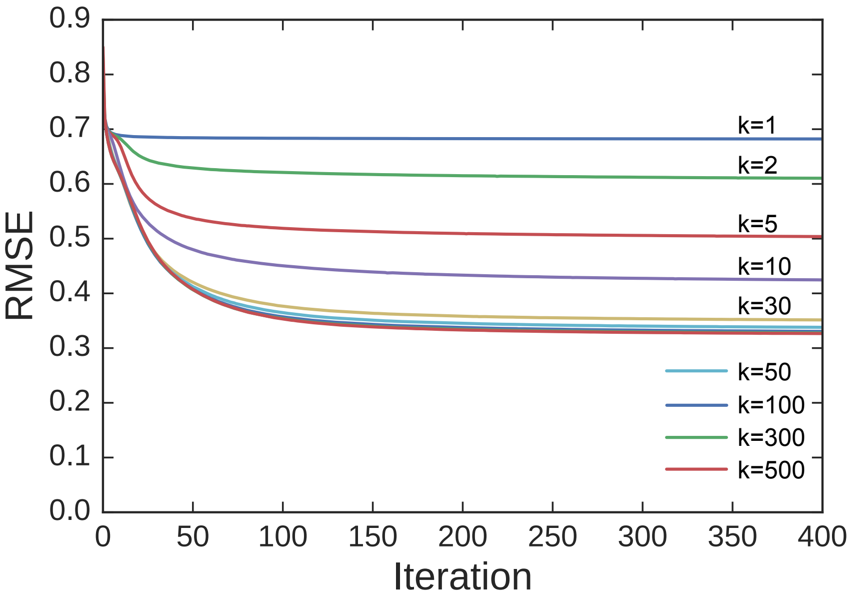


Figure 4.2: Reconstruction error (RMSE) of matrix factorization with different k.

4.4 Evaluation

4.4.1 Determining the Dimension Parameter k

How good is the low-rank approximation found by matrix factorization? And can we find the "sweet spot" for the number of dimensions k of the latent space? We investigate the reconstruction error of matrix factorization using different values of k to answer these questions. We use Root Mean Squared Error (RMSE) to measure error,

$$RMSE = \sqrt{\frac{\sum_{(u,t)\in R} (\boldsymbol{p}_u{}^\intercal \boldsymbol{q}_t - r_{u,t})^2}{N}}. \tag{4.4}$$

Here, N is the number of elements in the sparse matrix R (i.e., the number of known values).

Figure 4.2 shows RMSE values over iterations of `libmf` with the dimension parameter $k \in \{1, 2, 5, 10, 30, 50, 100, 300, 500\}$. We observed that the reconstruction error decreased as the iterative method of `libmf` progressed. The larger the number of dimensions k was, the smaller the reconstruction error became; the lowest reconstruction error was 0.3256 with $k = 500$. We also observed the error with $k = 1$, which corresponds to mapping users and topics onto one dimension similarly to the political spectrum of liberal and conservative. Judging from the relatively high RMSE values with $k = 1$, we conclude that it may be difficult to represent everything in the data using a one-dimensional axis. Based on this result, we concluded that matrix factorization with $k = 100$ is sufficient for reconstructing the original matrix R and therefore used this parameter value for the rest of our experiments.

4.4.2 Predicting Missing Topic Preferences

How accurately can the user and topic vectors predict missing topic preferences? To answer this question, we evaluate the accuracy in predicting hidden preferences in the matrix R as follows. First, we randomly selected 5% of existing elements in R and let

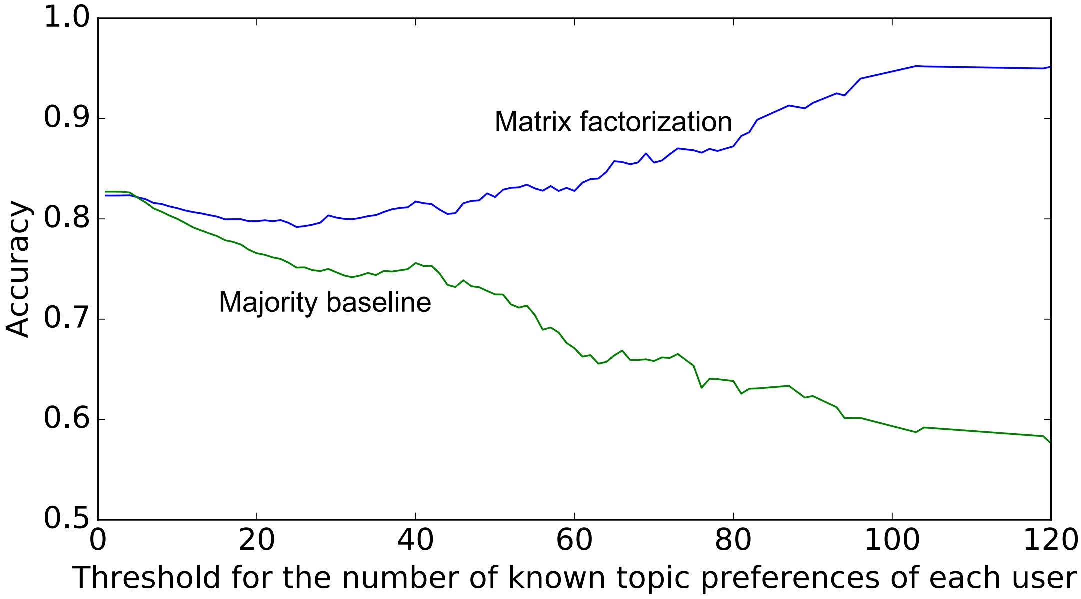


Figure 4.3: Prediction accuracy when changing the threshold for the number of known topic preferences of each user.

Y represent the collection of the selected elements (test set). We then perform matrix factorization on the sparse matrix without the selected elements of Y, that is, only with the remaining 95% elements of R (training set). We define the accuracy of the prediction as

$$\frac{1}{|Y|}\sum_{u,t\in Y}\mathbb{1}\left(\mathrm{sign}(\hat{r}_{u,t})=\mathrm{sign}(r_{u,t})\right) \tag{4.5}$$

Here, $r_{u,t}$ denotes the actual (i.e., self-declared) preference values, $\hat{r}_{u,t}$ represents the preference value predicted by Equation 4.3, $\mathrm{sign}(.)$ represents the sign of the argument, and $\mathbb{1}(.)$ yields 1 only when the condition described in the argument holds and 0 otherwise. In other words, Equation 4.5 computes the proportion of correct predictions to all predictions, assuming zero to be the decision boundary between pro and con.

Figure 4.3 plots prediction accuracy values calculated from different sets of users. Here the x-axis represents a threshold θ, which filters out users whose declarations of topic preferences are no greater than θ topics. In other words, Figure 4.3 shows prediction accuracy when we know user preferences for at least θ topics. For comparison, we also include the majority baseline that predicts pro and con based on the majority of preferences regarding each topic in training set.

Our proposed method was able to predict missing preferences with an 82.1% accuracy for users stating preferences for at least five topics. This accuracy increased as our method received more information regarding the users, reaching a 94.0% accuracy when $\theta = 100$. This result again indicates that our proposed method reasonably utilizes known preferences to complete missing preferences.

In contrast, the performance of the majority baseline decreased as it received more information regarding the users. Because this result was rather counter-intuitive, we examined the cause of this phenomenon. Consequently, this result turned out to be reasonable because preferences of vocal users deviated from those of the average users. Figure 4.4 illustrates this finding, showing the mean of variances of preference values $r_{u,t}$ across self-declared topics. In the figure, the x-axis represents a threshold θ, which filters out users whose statements of topic preferences are no greater than θ topics. We observe that the mean variance increased as we focused on vocal users. Overall, these results demonstrate the usefulness of user and topic vectors in predicting missing preferences.

Table 4.1 shows examples in which missing preferences of two users were predicted from known statements of agreements and disagreements[1]. In the table, pre-

[1] We anonymized user names in these examples. In addition, we removed topics that are too discrim-

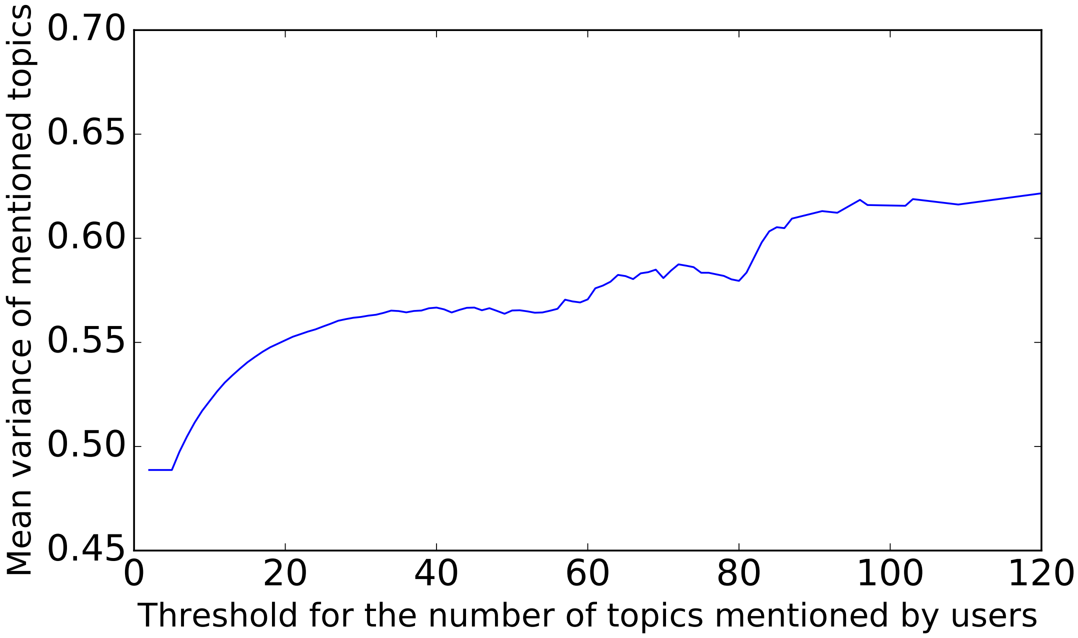


Figure 4.4: Mean variance of preference values of self-declared topics when changing the threshold for the number of self-declared topics.

dicted topics are accompanied by the corresponding $\hat{r}_{u,t}$ value in parentheses. As an example, our proposed method predicted that the user A, who is positive toward *regime change* but negative toward *Okinawa US military base*, may also be positive toward *vote of non-confidence to Cabinet* but negative toward *construction of a new base*.

4.4.3 Inter-topic Preferences

Do the topic vectors obtained by matrix factorization capture inter-topic preferences, such as “People who agree with A also agree with B”?

Because no dataset exists for this evaluation, we created a dataset of pairwise inter-topic preferences by using a crowdsourcing service[1]. Sampling topic pairs randomly, we collected 150 topic pairs whose cosine similarities of topic vectors were below

inatory or aggressive to other countries and races. Even though the experimental results of this chapter do not necessarily reflect our idea, we do not think it is a good idea to distribute politically incorrect ideas through this chapter.

[1] We used Yahoo! Crowdsourcing, a Japanese online service for crowdsourcing. http://crowdsourcing.yahoo.co.jp/

User	Type	Topic
A	Agreement (declared)	regime change, capital relocation
	Disagreement (declared)	Okinawa US military base, nuclear weapons, TPP, Abe Cabinet, Abe government, nuclear cycle, right to collective defense, nuclear power plant, Abenomics
	Agreement (predicted)	same-sex partnership ordinance (0.9697), vote of non-confidence to Cabinet (0.9248), national people's government (0.9157), abolition of tax (0.8978)
	Disagreement (predicted)	steamrollering war bill (-1.0522), worsening dispatch law (-1.0301), Sendai nuclear power plant (-1.0269), war bill (-1.0190), construction of a new base (-1.0186), Abe administration (-1.0173), landfill Henoko (-1.0158), unreasonable arrest (-1.0113)
B	Agreement (declared)	visit shrine, marriage
	Disagreement(declared)	tax increase, conscription, amend Article 9
	Agreement (predicted)	national people's government (0.8467), abolition of tax (0.8300), same-sex partnership ordinance (0.7700), security bills (0.6736)
	Disagreement (predicted)	corporate tax cuts (-1.0439), Liberal Democratic Party's draft constitution (-1.0396), radioactivity (-1.0276), rubble (-1.0159), nuclear cycle (-1.0143)

Table 4.1: Examples of agreement/disagreement topics predicted for two sample users A and B, with predicted score $\hat{r}_{u,v}$ shown in parenthesis.

-0.6, 150 pairs whose cosine similarities were between -0.6 and 0.6, and 150 pairs whose cosine similarities were above 0.6. In this way, we obtained 450 topic pairs for evaluation.

Given a pair of topics A and B, a crowd worker was asked to choose a label from the following three options: (a) *those who agree/disagree with topic A may also agree/disagree with topic B*; (b) *those who agree/disagree with topic A may conversely disagree/agree with topic B*; (c) *otherwise (no association between A and B)*. Creating twenty pairs of topics as gold data, we removed labeling results from workers whose accuracy is less than 90%.

Consequently, we obtained 6–10 human judgements for every topic pair. Regarding (a) as $+1$ point, (b) as -1 point, and (c) as 0 point, we computed the mean of the points (i.e., average human judgements) for each topic pair. Spearman's rank correlation coefficient (ρ) between cosine similarity values of topic vectors and human judgements was 0.2210. We could observe a moderate correlation even though inter-topic preferences collected in this manner were highly subjective.

Topic	Topics with a high degree of cosine similarity
Liberal Democratic Party (LDP)	Abe's LDP (0.3937), resuming nuclear power plant operations (0.3765), bus rapid transit (BRT) (0.3410), hate speech countermeasure law (0.3373), Henoko relocation (0.3353), C-130 (0.3338), Abe administration (0.3248), LDP & Komeito (0.2898), Prime Minister Abe (0.2835)
constitutional amendment	amendment of Article 9 (0.4520), enforcement of specific secret protection law (0.4399), security related law (0.4242), specific confidentiality protection law (0.4022), security bill amendment (0.3977), defense forces (0.3962), my number law (0.3874), collective self-defense rights (0.3687), militarist revival (0.3567)
right of foreigners to vote	human rights law (0.5405), anti-discrimination law (0.5376), hate speech countermeasure law (0.5080), foreigner's life protection (0.4553), immigration refugee (0.4520), co-organized Olympics (0.4379)

Table 4.2: Topics identified as being similar to the three controversial topics shown in the left column.

In addition to the quantitative evaluation, as summarized in Table 4.2, we also checked similar topics for three controversial topics, *Liberal Democratic Party (LDP)*, *constitutional amendment* and *right of foreigners to vote* (Table 4.2). Topics similar to LDP included synonymous ones (e.g., *Abe's LDP* and *Abe administration*) and other topics promoted by the *LDP* (e.g., *resuming nuclear power plant operations*, *bus rapid transit (BRT)* and *hate speech countermeasure law*). Considering that people who support the LDP may also tend to favor its policies, we found these results reasonable. As for the other example, *constitutional amendment* had a feature vector that was similar to that of *amendment of Article 9*, *enforcement of specific secret protection law* and *security related law*. From these results, we concluded that topic vectors were able to capture inter-topic preferences.

4.5 Related Work

In this section, we summarize the related work that spreads across various research fields.

Social Science and Political Science A number of of studies analyze social phenomena regarding political activities, political thoughts, and public opinions on social media. These studies model the political spectrum from liberal to conservative [Adamic and Glance, 2005; Bakshy et al., 2015; Cohen and Ruths, 2013; Wong et al., 2016; Zhou et al., 2011], political parties [Boutet et al., 2013; Makazhanov and Rafiei, 2013; Tumasjan et al., 2010], and elections [Conover et al., 2011; O'Connor et al., 2010].

Employing a single axis (e.g., liberal to conservative) or a few axes (e.g., political parties and candidates of elections), these studies provide intuitive visualizations and interpretations along the respective axes. In contrast, this study is the first attempt to recognize and organize various axes of topics on social media with no prior assumptions regarding the axes. Therefore, we think our study provides a new tool for computational social science and political science that enables researchers to analyze and interpret phenomena on social media.

Next, we describe previous research focused on acquiring lexical knowledge of politics. Sim et al. [2013] measured ideological positions of candidates in US presidential elections from their speeches. The study first constructs "cue lexicons" from political writings labeled with ideologies by domain experts, using sparse additive generative models [Eisenstein et al., 2011]. These constructed cue lexicons were associated with such ideologies as *left*, *center*, and *right*. Representing each speech of a candidate with cue lexicons, they inferred the proportions of ideologies of the candidate. The study requires a predefined set of labels and text data associated with the labels.

Bamman and Smith [2015a] presented an unsupervised method for assessing the political stance of a proposition, such as "global warming is a hoax," along the political spectrum of liberal to conservative. In their work, a proposition was represented by a tuple in the form $\langle \text{subject}, \text{predicate} \rangle$, for example, $\langle \textit{global warming}, \textit{hoax} \rangle$. They presented a generative model for users, subjects, and predicates to find a one-dimensional latent space that corresponded to the political spectrum.

Similar to our present work, their work [Bamman and Smith, 2015a] did not require labeled data to map users and topics (i.e., subjects) onto a latent feature space. In their paper, they reported that the generative model outperformed Principal Component Analysis (PCA), which is a method for matrix factorization. Empirical results here probably reflected the underlying assumptions that PCA treats missing elements as zero and not as missing data. In contrast, in the present work, we properly distinguish missing values from zero, excluding missing elements of the original matrix from the objective function of Equation 4.2. Further, this work demonstrated the usefulness of the latent space, that is, topic and user vectors, in predicting missing topic preferences of users and inter-topic preferences.

Fine-grained Opinion Analysis The method presented in Section 4.2 is an instance of fine-grained opinion analysis [Choi et al., 2006; Deng and Wiebe, 2015; Johansson and Moschitti, 2010; Wiebe et al., 2005; Yang and Cardie, 2013], which extracts a tuple of a subjective opinion, a holder of the opinion, and a target of the opinion from text. Although these previous studies have the potential to improve the quality of the user-topic matrix R, unfortunately, no corpus or resource is available for the Japanese language. We do not currently have a large collection of English tweets, but combining fine-grained opinion analysis with matrix factorization is an immediate future work.

Causality Relation Some of inter-topic preferences in this work can be explained by causality relation, for example, "TPP promotes free trade." A number of previous studies acquire instances of causal relation [Do et al., 2011; Girju, 2003] and promote/-suppress relation [Fluck et al., 2015; Hashimoto et al., 2012b] from text. The causality knowledge is useful for predicting (hypotheses of) future events [Hashimoto et al., 2015b; Radinsky and Davidovich, 2012; Radinsky et al., 2012b].

Inter-topic preferences, however, also include pairs of topics in which causality relation hardly holds. As an example, it is unreasonable to infer that *nuclear plant* and *railroading of bills* have a causal relation, but those who dislike *nuclear plant* also oppose *railroading of bills* because presumably they think the governing political parties rush the bill for resuming a nuclear plant. In this study, we model these inter-topic preferences based on preferences of the public. That said, we have as a promising future direction of our work plans to incorporate approaches to acquire causality knowledge.

4.6 Conclusion

In this chapter, we presented a novel approach for modeling inter-topic preferences of users on Twitter. Designing linguistic patterns for identifying support and opposition statements, we extracted users' preferences regarding various topics from a large collection of tweets. We formalized the task of modeling inter-topic preferences as a matrix factorization that maps both users and topics onto a latent feature space that abstracts users' preferences. Through our experimental results, we demonstrated that our approach was able to accurately predict missing topic preferences of users (80–94%) and that our latent vector representations of topics properly encoded inter-topic preferences.

For our immediate future work, we plan to embed the topic and user vectors to create a cross-topic stance detector. It is possible to generalize our work to model heterogeneous signals, such as interests and behaviors of people, for example, "those who are interested in A also support B," and "those who favor A also vote for B". Therefore, we believe that our work will bring about new applications in the field of NLP and other disciplines.

Chapter 5

Stance Classification with Consideration of the Silent Majority by Factorization Machines

In Chapter 4, we modeled inter-topic preferences such as *those who agree/disagree with topic A also agree/disagree with topic B*. However, this method is not applicable for users who do not declare any stances for targets. In this chapter, to overcome this, we propose a method for predicting stances of users including the silent majority by using factorization machines.

5.1 Introduction

Huge people all over the world use Social networking services (SNS) in recent years. Among such enormous texts, many texts reflecting people's tastes and opinions. Therefore, many researchers tackled tasks such as stance classification, election prediction, and so on.

The silent majority is a big problem related to it. The silent majority refers to who do not declare any stances (agreement/disagreement) on a specific topic. In SNS, it is said that many users belong to the silent majority. In existing research, it is relatively easy to identify stances of the minority people who frequently express their opinions (The noisy minorities). In addition, as to such people, if we know their stances on some topics, we could assume their stances on other topics as well (Chapter 4). On the other hand, the silent majority has not actively been tackled so far. However, if we cannot deal with the silent majority that accounts for the majority of the world, it will not be that we could clarify public opinion.

Therefore, in this chapter, we propose a method to analyze the opinion of the silent majority. Our contribution is as follows:

1. We propose a method for predicting stances by factorization machines which is widely used in the item recommendation.

2. By using features based on users' posted texts in addition to their explicit stances, we confirmed that stance classification performance for each user is improved.

3. In addition, we confirmed that our proposed method can also predict stances of the pseudo silent majority who are expected to have properties close to the actual silent majority.

5.2 Related Work

5.2.1 Social Science & Political Science

Regarding social science and political science, many researchers focus on SNS.

Many works tried to classify people in SNS into liberal/conservative [Adamic and Glance, 2005; Bakshy et al., 2015; Cohen and Ruths, 2013; Wong et al., 2016; Zhou et al., 2011]. Although it is a valuable task regarding reducing the labor of manually looking numerous texts, liberal and conservative are a somewhat rough indicator.

In addition, many users tend to express their own opinions during the election period in SNS. Tumasjan et al. [2010] focused on that point and predicted the result of an election using the tweet data posted during the election period. Besides research that predicts voting results on a candidate or political party basis like this, some researchers predict the user's voting destination on SNS user basis.

In addition to research that analyzes people's opinions on political parties and candidates in this way, there is research that analyzes and predicts public opinion on fine-grained topics such as specific topics and events. In Chapter 4, we modeled stances of

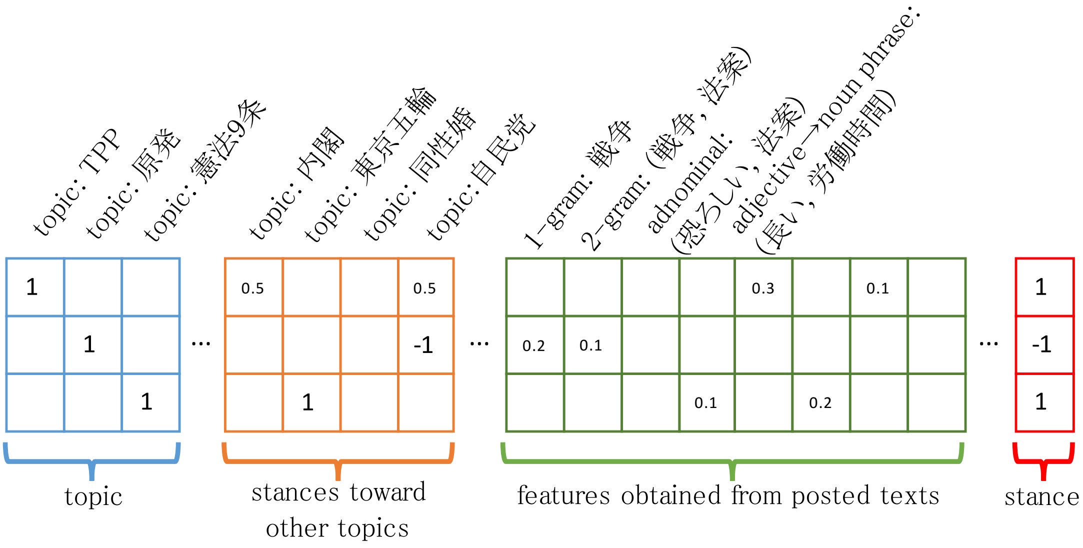


Figure 5.1: An example of an input of factorization machines.

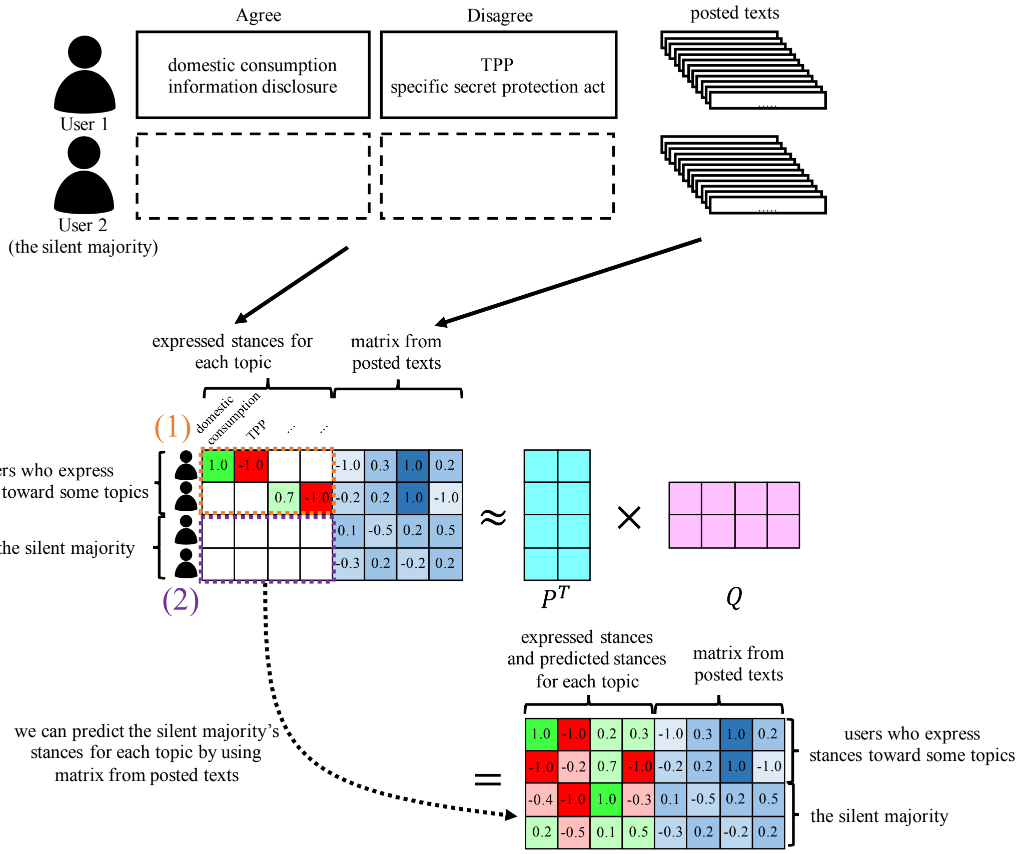
the people such as "Those who agree with A also agree with B" or "Those who disagree with A also disagree with B" by matrix factorization. However, that method can only be applied to users who expressed some stances. Therefore, that method cannot deal with the silent majority who do not explicitly express their stances. In this chapter, on the other hand, we focus on features derived from posted texts of each user.

5.3 Preparation of the Matrix

In this research, we aim to construct a user level stances classification model that can also take the silent majority into account. The problem here is that the silent majority are users who do not explicitly express any stances. Therefore, it is impossible to take a method like Chapter 4 to predict agreement/disagreement to other topics by using some explicit stances. In this case, how can we construct a model that can predict stances including the silent majority?

Therefore, in this research, we focus on posted texts of each user. The overview of the proposed method of this research is shown in Figure 5.2.

Whether it is a user who frequently expresses agreement/disagreement or even the silent majority who do not express stances at all, it is expected to be able to get some


Agree
Disagree
posted texts
domestic consumption
information disclosure
TPP
specific secret protection act
User 1
User 2
(the silent majority)
.....
expressed stances for each topic
matrix from posted texts
(1)
domestic consumption
TPP
...
users who express stances toward some topics
the silent majority
(2)
≈
×
P^T
Q
expressed stances and predicted stances for each topic
matrix from posted texts
we can predict the silent majority's stances for each topic by using matrix from posted texts
=
users who express stances toward some topics
the silent majority

1.0	-1.0			-1.0	0.3	1.0	0.2
		0.7	-1.0	-0.2	0.2	1.0	-1.0
				0.1	-0.5	0.2	0.5
				-0.3	0.2	-0.2	0.2

1.0	-1.0	0.2	0.3	-1.0	0.3	1.0	0.2
-1.0	-0.2	0.7	-1.0	-0.2	0.2	1.0	-1.0
-0.4	-1.0	1.0	-0.3	0.1	-0.5	0.2	0.5
0.2	-0.5	0.1	0.5	-0.3	0.2	-0.2	0.2

Figure 5.2: The overview of the proposed method

texts of them. If it is found that the feature related to agreement/disagreement is included in the text, it is considered to be a clue to predict the stances of the silent majority.

5.3.1 Data

Throughout this work, we use a corpus consisting of 1,763,164,770 Japanese tweets (444,321 users) crawled from February 6, 2013 to September 30, 2016. We removed retweets from the corpus.

Table 5.1: Examples of features based on users' posted texts.

Feature type	Examples
1-gram	戦争
2-gram	(戦争, 法案)
adnominal	(恐ろしい, 法案)
adjective → noun phrase	(長い, 労働時間)
noun phrase → adjective	(電車, 多い)
noun phrase → verb	(給与水準, 戻る)

5.3.2 Retrieve Stances of Users Other Than the Silent Majority

As mentioned earlier, the silent majority are users who do not explicitly express any stances. On the other hand, there are not a few users on Twitter expressing their stances explicitly. How to express stances depends on each user, and many users stances are difficult to extract automatically. Therefore, like Chapter 4, we use a hashtag (*#Xsansei*, *#Yhantai*) expressing stances and a pattern expressing stances (*support X*, *X is terrible*). Then, explicit stances of these users are obtained. We first gathered users using hashtags expressing stances and obtained agreement/disagreement patterns by tracking users' other tweets. In addition, we identify stances of each user by agreement/disagreement patterns and creates a matrix with topics and users as rows and columns. Following Chapter 4, we also create a matrix using patterns. As a result, the (1) part of Figure 5.2 is filled. However, the silent majority's stances ((2) part of 5.2) cannot be filled with this procedure.

5.3.3 Matrix from Words

In this subsection, we explain features from posted texts of users. On Twitter, each user can post many tweets with the limit of 140 characters. The user's usual tweet is considered to reflect the user's preference, opinion, political orientation, and so on.

Here, for example, suppose that the word *war bill* is included in those texts. This word is commonly used mainly by users who disagree *security bill*. In this way, it is considered that word usage is strongly linked with the user's stances. Therefore, in this research, features in each text of users are used as one of the features related to users.

In this work, we used n-gram features (1-gram and 2-gram) and features based on dependency paths. Examples of features are described in Table 5.1.

5.4 Method for Prediction

In the previous section, we explain explicit stances and how to build a matrix based on each user's posted text. How can we predict agreement/disagreement for users including the silent majority by the matrix constructed in this way? In this section, we will explain each method used in this research.

5.4.1 Matrix Factorization

As one of the methods, we use matrix factorization described in 4.3. Note that matrix factorization can treat matrix with only two type features. Thus, we use this method only with user-topic matrix.

5.4.2 Factorization Machines

In the previous section, we explained the procedure of prediction by matrix factorization similar to 4.3. However, in matrix factorization, rows and columns usually have to be one type of feature, like a user and a topic, respectively. In this work, the matrix we build is not suitable for usual matrix factorization because it includes features based on texts of each user in addition to users and topics. Therefore, we need a framework that can flexibly incorporate multiple feature types.

Therefore, we use factorization machines [Rendle, 2010], which is widely used in many tasks such as item recommendation, click-through rate prediction, and so on. In factorization machines, i-th variable has a k dimensional vector v_i (k is a hyperparameter). This vector is trained in a training phase. By using these vectors, we can predict the target value (e.g. rating in item recommendation, a rate in click-through rate prediction) by the following equation:

$$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j \tag{5.1}$$

Here, x_i is the value of the i-th variable, $w_0 \in \mathbb{R}$ is the global bias, $w_i \in \mathbb{R}^n$ models the strength of the i-th variable, and $\langle v_i, v_j \rangle$ (dot product of two vectors) models the interaction between the i-th and j-th variable.

As an implementation, we use `tffm`[1], which implements factorization machines by `tensorflow`.

An example of a matrix of factorization machines is shown in Figure 5.1. Each row of this matrix corresponds to a stance of a user towards a topic. For example, the first row of this matrix means agreement towards TPP. In this work, we treat 1 for agreement and -1 for disagreement. Next, "stances toward other topics" in this matrix indicates stances expressed by the user associated with this row toward other topics. Here, when the user expresses stances toward multiple topics, we divided each value by the number of those topics. Finally, "features obtained from posted texts" means what features the user used in posted texts. Each value of these features is set so that the sum equals to 1 in each row after calculating TF-IDF by regarding a user as a document.

5.5 Evaluation

5.5.1 Predicting Missing Topic Preferences

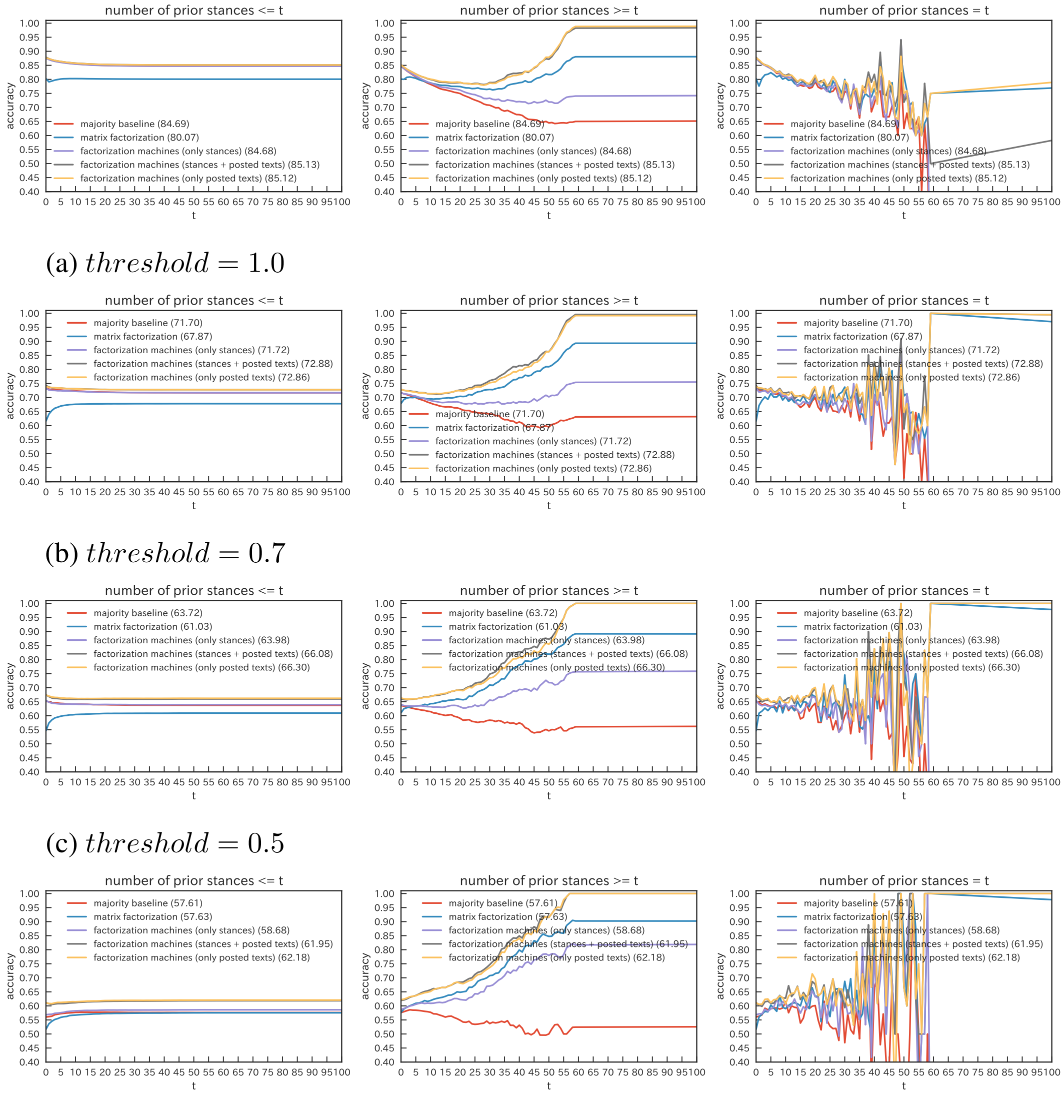
In this subsection, we hide some stances in the matrix and evaluate the accuracy of the prediction for them. Note that, since it is impossible to evaluate users who do not express agreement/disagreement, we only select users who declare stances more than once. As a result, a matrix consisting 326,202 stances was obtained. There are 130,635 users and 1,142 topics in the matrix. In the evaluation, we use 5% of the rows in this matrix as a validation set and use it for a parameter tuning and an early stopping of training. We evaluate the accuracy of 10-fold cross validation with respect to the

[1] `https://github.com/geffy/tffm`

remainder of the matrix (95% of rows).

For comparison, we perform the majority baseline and matrix factorization at the same time. For the majority baseline, we use a large number of agreements or disagreements for the topic in training set for the prediction. For example, if the number of users who expressed disagreement towards "Article 9" exceeds the number of users who expressed agreement towards it, the majority baseline always predict "disagreement" towards "Article 9" in the prediction. For the matrix factorization, we use the same parameters as in Chapter 4 ($k = 100,\ \lambda_P = 0.1,\ \lambda_Q = 0.1$). Note that, matrix factorization used in Chapter 4 cannot incorporate a matrix based on posted texts. Therefore, we use only stances of users in the matrix factorization. In regard to the factorization machines, we performed the evaluation separately in a case when using only stances, using only posted texts, and using both of them. For the parameter tuning of the factorization machines, we measured the accuracy of validation set using one of the training set in 10-fold cross validation dataset.

The evaluation result is shown in Figure 5.3. Here, topics that stances of many users are biased towards an agreement or a disagreement are thought to be not appropriate for the evaluation because stances of those topics can be easily predicted with even a simple model. Therefore, we performed several experiments by selecting targets considering t_{avg} , which is an average stances on the topic in the data. We set some $threshold$ and divided results by $-threshold \leq t_{avg} \leq threshold$. Each figure is divided into three parts: the number of known stances is under t, is above t, and equals to t. From Figure 5.3a, we find that although the matrix factorization exceeds the majority baseline for cells with number of prior stances $t \geq 5$, the result was lower than the majority baseline for the cell with number of prior stances $t \leq 5$. On the other hand, all methods of factorization machines exceeded the majority baseline in each t. In addition, as for the factorization machines, users' posted texts contributed to higher precision than the method only using users' stances. Note that, difference between methods using posted text features and the majority baseline is increased as $threshold$ decreased. This means that the more stances which is difficult to predict, the more useful the features from posted texts.



(a) $threshold = 1.0$

(b) $threshold = 0.7$

(c) $threshold = 0.5$

(d) $threshold = 0.3$

Figure 5.3: Accuracy of prediction. Each row shows results of different threshold. Each column shows a result when the number of known stances is under t, is above t, and equals to t. We show the overall accuracy of methods in parentheses on the legend.

5.5.2 Evaluation for the pseudo silent majority

In 5.5.1, we confirmed the improvement of accuracy by factorization machines using features based on users' posted texts. From this result, it is expected that stances of the user who did not express stances (the silent majority) can also be predicted from

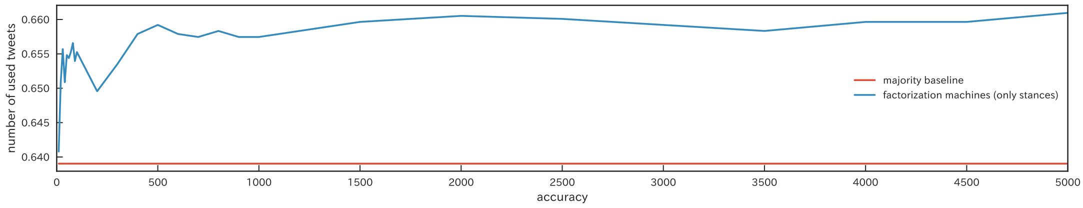


Figure 5.4: Evaluation result of the pseudo silent majority.

their posted texts. Therefore, in this subsection, we treat users with a small number of stances as the pseudo silent majority, and perform evaluations and analysis. Note that, it would be preferable if it was possible to evaluate users who did not express stances at all as silent majority. However, since it is virtually impossible for a third person to annotate stances of such users, we defined the pseudo silent majority as mentioned above.

In this subsection, we changed the number of tweets derived from the pseudo silent majority and measured the change in precision. Here, we fixed $threshold$ to 0.5. The evaluation result is shown in Figure 5.4. From this result, if was found that if we could acquire about 500 tweets of the pseudo silent majority, we can predict stances of those users with accuracy higher than majority baseline. However, there is room for improvement in the future.

5.6 Conclusion

In this work, we focused on posted texts of users on the SNS. In the evaluation, we confirmed that posted texts of users contribute to predicting stances of the users. In addition, our methods are also applicable for predicting stances of the pseudo silent majority. Factorization machines used in this work has the advantage that various kinds of features can be used simultaneously in an input matrix. For future work, we will also consider users' follow-follower relations, retweet relations, or profile of users.

Chapter 6

Conclusion

In this thesis, we try to acquire and to apply knowledge that contributes to the performance improvement of stance classification. In summary, our contributions are as follows:

- We demonstrated that many texts cannot be classified into FAVOR/AGAINST without causal relation knowledge (PRIOR-SITUATION/EFFECT). Then, we performed FAVOR/AGAINST classification with causal relation knowledge and showed improvements in classification accuracy.

- To acquire knowledge, we proposed crowdsourcing-based approach for annotating causal relation instances to Wikipedia articles. The annotated data is publicly available on Web.

- Besides causal relation knowledge, we focused on inter-topic preferences such as "A person who agrees with A also agrees with B" or "A person who disagrees with A also disagrees B". We perform modeling inter-topic preferences by matrix factorization. Through our experimental results, we demonstrated that our approach was able to accurately predict missing topic preferences of users.

- To predict stances of people including the silent majority, we focused on users' texts. By utilizing factorization machines, we demonstrated that stances of the silent majority can be detected by considering their texts. In addition, we showed that features derived from users' texts can improve the classification accuracy in regard to the noisy minority, who frequently express their stances.

References

Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery (LinkKDD 2005)*, pages 36–43, 2005. doi: 10.1145/1134271.1134277. 35, 48, 52

Leman Akoglu. Quantifying political polarity based on bipartite opinion networks. In *Eighth International AAAI Conference on Weblogs and Social Media*, pages 2–11, 2014. 8

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011)*, pages 1–9, 2011. 36

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015. doi: 10.1126/science.aaa1160. 35, 48, 52

David Bamman and Noah A. Smith. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 76–85, 2015a. doi: 10.18653/v1/D15-1008. 48

David Bamman and Noah A. Smith. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85. Association for Computational Linguistics, 2015b. doi: 10.18653/v1/D15-1008. 8

Robert M. Bell and Yehuda Koren. Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75–79, 2007. doi: 10.1145/1345448.1345465. 40

Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. What's in Twitter, I know what parties are popular and who you are supporting now! *Social Network Analysis and Mining (SNAM 2012)*, 3(4):1379–1391, 2013. doi: 10.1109/ASONAM.2012.32. 48

Anthony Brew, Derek Greene, and Pádraig Cunningham. Using crowdsourcing and active learning to track sentiment in online media. In *Proc. of ECAI 2010*, pages 145–150, 2010. 28, 29

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian, Eric Young, Ganesh Harihara, and Eugene Yang. Identifying political sentiment between nation states with social media. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 65–75. Association for Computational Linguistics, 2015. 8, 15, 20

Wei-Sheng Chin, Yong Zhuang, Yu-Chin Juan, and Chih-Jen Lin. A fast parallel stochastic gradient method for matrix factorization in shared memory systems. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(1):2, 2015. doi: 10.1145/2668133. 41

Yejin Choi, Eric Breck, and Claire Cardie. Joint extraction of entities and relations for opinion recognition. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 431–439, 2006. URL `http://aclweb.org/anthology/W06-1651`. 49

Raviv Cohen and Derek Ruths. Classifying political orientation on Twitter: It's not easy! In *Proc. of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*, pages 91–99, 2013. 35, 36, 48, 52

Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *Privacy, 2011 IEEE Third International Conference on Security, Risk and Trust and 2011 IEEE Third Inernational Conference on Social Computing (PASSAT-SocialCom 2011)*, pages 192–199. IEEE, 2011. 48

Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 241–249, 2010. URL `http://aclweb.org/anthology/C10-2028`. 39

Lingjia Deng and Janyce Wiebe. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2015)*, pages 1323–1328, 2015. doi: 10.3115/v1/N15-1146. 49

Quang Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 294–303, 2011. URL `http://aclweb.org/anthology/D11-1027`. 49

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie Strassel, and Ralph M Weischedel. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proc. of LREC 2004*, pages 837–840, 2004. 27

Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The Yahoo! Music dataset and KDD-Cup'11. In *Proceedings of the 2011 International Conference on KDD Cup 2011 (KDDCUP 2011)*, pages 3–18, 2011. 40

Jesse Dunietz, Lori Levin, and Jaime Carbonell. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proc. of the 11th Linguistic Annotation Workshop*, pages 95–104, 2017. 27, 30

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, 2011. 48

Tim Finin, William Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in Twitter data with crowdsourcing. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, 2010. vi, 28, 29

Juliane Fluck, Sumit Madan, Tilia Renate Ellendorff, Theo Mevissen, Simon Clematide, Adrian van der Lek, and Fabio Rinaldi. Track 4 overview: Extraction of causal network information in biological expression language (BEL). In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, pages 333–346, 2015. 31, 49

Karën Fort, Gilles Adda, and K. Bretonnel Cohen. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420, 2011. 28

Roxana Girju. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12*, pages 76–83, 2003. doi: 10.3115/1119312.1119322. 49

Matthew R. Gormley, Adam Gerber, Mary Harper, and Mark Dredze. Non-expert correction of automatically generated relation annotations. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 204–207, 2010. 28, 29

Jeffrey Gottfried and Elisa Shearer. News use across social media platforms 2016. Technical report, Pew Research Center, May 2016. 35

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. Excitatory or inhibitory: a new semantic orientation extracts contradiction and causality from the web. In *Proc. of EMNLP-CoNLL 2012*, pages 619–630, 2012a. 31

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, pages 619–630. Association for Computational Linguistics, 2012b. URL `http://aclweb.org/anthology/D12-1057`. 24, 49

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. Generating event causality hypotheses through semantic relations. In *Proc. of AAAI 2015*, pages 2396–2403, 2015a. 27

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, and Jong-Hoon Oh. Generating event causality hypotheses through semantic relations. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*, pages 2396–2403, 2015b. 49

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 Task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proc. of the 5th International Workshop on Semantic Evaluation*, pages 33–38, 2010. 27

Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. Acquiring noun polarity knowledge using selectional preferences (in Japanese). In *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 584–587, 2008. 15

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 8, 14

Dirk Hovy, Barbara Plank, and Anders Søgaard. Experiments with crowdsourced re-annotation of a POS tagging data set. In *Proc. of ACL 2014*, pages 377–382, 2014. 28, 29

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1113–1122. Association for Computational Linguistics, 2014. 8

Kathleen Hall Jamieson and Joseph N. Cappella. *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment.* Oxford University Press, 2008. 35

Mukund Jha, Jacob Andreas, Kapil Thadani, Sara Rosenthal, and Kathleen McKeown. Corpus creation for new genres: A crowdsourced approach to PP attachment. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 13–20, 2010. 28, 29

Richard Johansson and Alessandro Moschitti. Syntactic and semantic structure for opinion expression detection. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL 2010)*, pages 67–76, 2010. URL `http://aclweb.org/anthology/W10-2910`. 49

Kristen Johnson and Dan Goldwasser. “All I know about politics is what I read in Twitter”: Weakly supervised models for extracting politicians’ stances from twitter. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 2966–2977, 2016. URL `http://aclweb.org/anthology/C16-1279`. 36

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proc. of COLING 2014*, pages 269–278, 2014. 28, 29

Fred N. Kerlinger. *Liberalism and Conservatism: The Nature and Structure of Social Attitudes*. Lawrence Erlbaum Associates, 1984. 36

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D14-1181`. 14

Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014. 8

Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Opinion mining from web documents: Extraction and structurization. *Information and Media Technologies*, 2 (1):326–337, 2007. 15

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009. doi: 10.1109/MC.2009.263. 40

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237. Association for Computational Linguistics, 2004. 13

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., 2001. ISBN 1-55860-778-1. 15

Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz. Annotating large email datasets for named entity recognition with mechanical turk. In *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 71–79, 2010. vi, 28, 29, 30

William S. Maddox and Stuart A. Lilie. *Beyond Liberal and Conservative: Reassessing the Political Spectrum.* Cato Inst, 1984. 36

Aibek Makazhanov and Davood Rafiei. Predicting political preference of twitter users. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 298–305, 2013. doi: 10.1145/2492517.2492527. 48

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654. Association for Computational Linguistics, 2013. 6, 8, 20

Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, pages 321–327. Association for Computational Linguistics, 2013. 15

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the*

10th International Workshop on Semantic Evaluation (SemEval-2016), pages 31–41, 2016a. doi: 10.18653/v1/S16-1003. 2, 36

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016b. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S16-1003`. 9, 23

Akiko Murakami and Rudy Raymond. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 869–875, 2010. URL `http://aclweb.org/anthology/C10-2100`. 8, 36

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 786–794. Association for Computational Linguistics, 2010. 13, 15

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, pages 122–129, 2010. 48

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. Why-question answering using intra- and inter-sentential causal relations. In *Proc. of ACL 2013*, pages 1733–1743, 2013. 27

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. A semi-supervised learning approach to why-question answering. In *Proc. of AAAI-16*, pages 3022–3029, 2016. 27

Eli Pariser. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think.* Penguin Books, 2011. 35

Sampo Pyysalo, Tomoko Ohta, Rafal Rak, Andrew Rowley, Hong-Woo Chun, Sung-Jae Jung, Sung-Pil Choi, Jun'ichi Tsujii, and Sophia Ananiadou. Overview of the cancer genetics and pathway curation tasks of BioNLP shared task 2013. *BMC Bioinformatics*, 16(10):S2, 2015. 27

Ashequl Qadir and Ellen Riloff. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1203–1209, 2014. doi: 10.3115/v1/D14-1127. 39

Kira Radinsky and Sagie Davidovich. Learning to predict from textual data. *Journal of Artificial Intelligence Research (JAIR)*, 45(1):641–684, 2012. 49

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proc. of WWW 2012*, pages 909–918, 2012a. 27

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web (WWW 2012)*, pages 909–918, 2012b. doi: 10.1145/2187836.2187958. 49

Ines Rehbein and Josef Ruppenhofer. Catching the common cause: Extraction and annotation of causal relations and their participants. In *Proc. of the 11th Linguistic Annotation Workshop*, pages 105–114, 2017. 27, 30

Steffen Rendle. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 995–1000. IEEE, 2010. 56

Fabio Rinaldi, Tilia Renate Ellendorff, Sumit Madan, Simon Clematide, Adrian van der Lek, Theo Mevissen, and Juliane Fluck. BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language. *Database: The Journal of Biological Databases and Curation*, page baw067, 2016. 28

Akira Sasaki, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. Stance classification by recognizing related events about targets. In *2016 IEEE/WIC/ACM International*

Conference on Web Intelligence, pages 582–587, Oct 2016. doi: 10.1109/WI.2016.0100. 5, 27

Toshinori Sato. Neologism dictionary based on the language resources on the web for mecab, available from https://github.com/neologd/mecab-ipadic-neologd, 2015. URL `https://github.com/neologd/mecab-ipadic-neologd`. 13

Rebecca Sharp, Mihai Surdeanu, Peter Jansen, Peter Clark, and Michael Hammond. Creating causal embeddings for question answering with minimal supervision. In *Proc. of EMNLP 2016*, pages 138–148, 2016. 27

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. Measuring ideological proportions in political speeches. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 91–101, 2013. URL `http://aclweb.org/anthology/D13-1010`. 48

Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, pages 226–234, 2009. URL `http://aclweb.org/anthology/P09-1026`. 8, 15, 36

Dhanya Sridhar, Lise Getoor, and Marilyn Walker. Collective stance classification of posts in online debate forums. In *Proc. of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 109–117, 2014. 8

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proc. of EACL 2012 (demonstrations)*, pages 102–107, 2012. 28

Sho Takase, Naoaki Okazaki, and Kentaro Inui. Composing distributed representations of relational patterns. In *Proc. of ACL 2016*, pages 2276–2286, 2016. 28, 29

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts.

In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 327–335, 2006a. URL `http://aclweb.org/anthology/W06-1639`. 36

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics, 2006b. 8

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010)*, pages 178–185, 2010. 8, 48, 52

Cynthia Van Hee, Els Lefever, and Veronique Hoste. Monday mornings are my fave :) #not exploring the automatic recognition of irony in english tweets. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, pages 2730–2739, 2016. URL `http://aclweb.org/anthology/C16-1257`. 39

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martell, and Joseph King. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729, 2012. doi: 10.1016/j.dss.2012.05.032. 36

Xin Wang, Yuanchao Liu, Chengjie SUN, Baoxun Wang, and Xiaolong Wang. Predicting polarities of tweets by composing word embeddings with long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1343–1353. Association for Computational Linguistics, 2015. 8

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. pkudblab at semeval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San

Diego, California, June 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S16-1062`. 3

Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2):165–210, 2005. doi: 10.1007/s10579-005-7880-9. 49

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets and retweets. In *Seventh International AAAI Conference on Weblogs and Social Media*, pages 640–649, 2013. doi: 10.1109/TKDE. 2016.2553667. 8

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. Quantifying political leaning from tweets, retweets, and retweeters. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2158–2172, 2016. doi: 10.1109/TKDE. 2016.2553667. 35, 48, 52

Bishan Yang and Claire Cardie. Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1640–1649, 2013. URL `http://aclweb.org/anthology/P13-1161`. 49

Guido Zarrella and Amy Marsh. Mitre at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California, June 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/S16-1074`. 3

Meishan Zhang, Yue Zhang, and Duy Tin Vo. Neural networks for open domain targeted sentiment. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 612–621. Association for Computational Linguistics, 2015. 6, 8

Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. Classifying the political leaning of news articles and users from user votes. In *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, pages 417–424, 2011. 8, 35, 48, 52

List of Publications

Journal Papers (Refereed)

1. Koji Matsuda, Akira Sasaki, Naoaki Okazaki and Kentato Inui. Geographical Entity Annotated Corpus of Japanese Microblogs. Journal of Information Processing Vol. 25, pp.121-130, January 2017.

International Conference/Workshop Papers (Refereed)

1. Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. A Crowdsourcing Approach for Annotating Causal Relation Instances in Wikipedia. In Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC), 10 pages, November 2017.

2. Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki and Kentaro Inui. Other Topics You May Also Agree or Disagree: Modeling Inter-Topic Preferences using Tweets and Matrix Factorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), pp.398-408, August 2017.

3. Akira Sasaki, Junta Mizuno, Naoaki Okazaki and Kentaro Inui. Stance Classification by Recognizing Related Events about Targets. IEEE/WIC/ACM International Conference on Web Intelligence, pp. 582-587, October 2016.

4. Koji Matsuda, Akira Sasaki, Naoaki Okazaki and Kentaro Inui. Annotating Geographical Entities on Microblog Text. In Proceedings of the 9th Linguistic Annotation Workshop (LAW IX 2015), pp.85^^e2^^80^^9394, June 2015.

Other Publications (Not refereed)

1. Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki and Kentaro Inui. Stance Classification Using Causal Relationship Knowledge on Topics. The 12th NLP Symposium for Young Researchers, September 2017.

2. Kazuaki Hanawa, Akira Sasaki, Naoaki Okazaki, Kentaro Inui. A Crowdsourcing Approach for Annotating Causal Relationship Knowledge. In Proceedings of the 23rd Annual Meeting of the Association for Natural Language Processing, pp.406-409, March 2017.

3. Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, Kentaro Inui. Modeling Inter-Topic Preferences Using Pro/Con Patterns and Matrix Factorization. In Proceedings of the 23rd Annual Meeting of the Association for Natural Language Processing, pp.795-798, March 2017.

4. Sonse Shimaoka, Shota Sato, Akira Sasaki, Satoshi Sekine, Kentaro Inui. Future Prediction Information Extraction from Newspaper Using Conditional Random Fields. In Proceedings of the 7th Symposium on Text Mining, pp.57-62, September 2015.

5. Masatoshi Kurihara, Akira Sasaki, Koji Matsuda, Naoaki Okazaki, Kentaro Inui. Extracting Local Resident Demands per Region Using Twitter. In Proceedings of the 29th Annual Conference of the Japanese Society for Artificial Intelligence, 1H3-3, May 2015.

6. Koji Matsuda, Akira Sasaki, Naoaki Okazaki, Kentaro Inui. Annotating Geographical Entities on Microblog Text. In IPSJ SIG Technical Reports, Vol. 2015-NL-220 (12), pp.1-10, January 2015.

7. Youkou Funaki, Akira Sasaki, Naoaki Okazaki, Kentaro Inui, Yosuke Fukada, Ryuichiro Takeshita, Hideaki Tamori, Hiroshi Nozawa. Analyzing the campaigns for the 2013 Japanese House of Councillors election on Twitter. In Proceedings of the 28th Annual Conference of the Japanese Society for Artificial Intelligence, 1K3-2, May 2014.

8. Akira Sasaki, Yuki Igarashi, Yotaro Watanabe, Kentaro Inui. Towards Grounding Geographical Entities. In Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing, pp.177-180, March 2014.

9. Yotaro Watanabe, Akira Sasaki, Yuki Igarashi, Naoaki Okazaki, Kentaro Inui. Information Extraction for Real World Oriented Information Structuring Support. In Proceedings of the 20th Annual Meeting of the Association for Natural Language Processing, pp.1003-1006, March 2014.

10. Naoaki Okazaki, Akira Sasaki, Kentaro Inui, Hirofumi Abe, Nozomi Ishida. Analyzing Twitter for Surveying the Reputation for Peaches Produced in Fukushima. In Proceedings of the SRA Japan The 26th Annual Meeting, November 2013.

11. Akira Sasaki, Junta Mizuno, Naoaki Okazaki, Kentaro Inui. Normalization of Text in Microblogging Based on Machine Learning. In Proceedings of the 27th Annual Conference of the Japanese Society for Artificial Intelligence, 4B1-4, June 2013.