B7IM2048

Master's Thesis

Stance Detection Attending External Knowledge on Topics

Kazuaki Hanawa

November 30, 2018

Graduate School of Information Sciences
Tohoku University

A Master's Thesis
submitted to System Information Sciences,
Graduate School of Information Science,
Tohoku University
in partial fulfillment of the requirements for the degree of
MASTER of Information Science

Kazuaki Hanawa

Thesis Committee:

Professor Kentaro Inui	(Supervisor)
Professor Shinichiro Omachi	
Professor Kazuyuki Tanaka	
Associate Professor Jun Suzuki	(Co-supervisor)

Stance Detection Attending External Knowledge on Topics[*]

Kazuaki Hanawa

Abstract

This paper presents a novel approach to stance detection for unseen topics that takes advantage of external knowledge about the topics. We build a new stance detection dataset consisting of 6,701 tweets on seven topics with associated Wikipedia articles. An analysis of this dataset confirms the necessity of external knowledge for this task. This paper also presents a method of extracting related concepts and events from Wikipedia articles. To incorporate this extracted knowledge into stance detection, we propose a novel neural network model that can attend to such related concepts and events when encoding the given text using bi-directional long short-term memories. Our experimental results demonstrate that the proposed method, using knowledge extracted from Wikipedia, can improve stance detection performance.

Keywords:

natural language processing, stance detection, world knowledge

[*] Master's Thesis, System Information Sciences, Graduate School of Information Sciences, Tohoku University, B7IM2048, November 30, 2018.

トピックに関する外部知識を利用した賛否分類[*]

塙 一晃

内容梗概

本研究では賛否分類においてトピックに関する外部知識を利用するための手法を提案する．Wikipedia記事に紐づいた7トピックに関する6,701件のツイートからなるデータセットを作成し，分析することで賛否分類における外部知識の必要性が明らかとなった．また，本研究ではWikipedia記事から獲得した知識を賛否分類で利用するために，関連する知識を参照しながら文をエンコードすることができるモデルを提案する．Wikipediaから獲得した知識を使用する提案手法は外部知識を使用しないものよりも高い精度で賛否の予測ができることが実験結果より明らかとなった．

キーワード

自然言語処理, 賛否分類, 世界知識

[*] 東北大学 大学院情報科学研究科 システム情報科学専攻 修士論文, B7IM2048, 2018年11月30日.

Contents

List of Figures

List of Tables

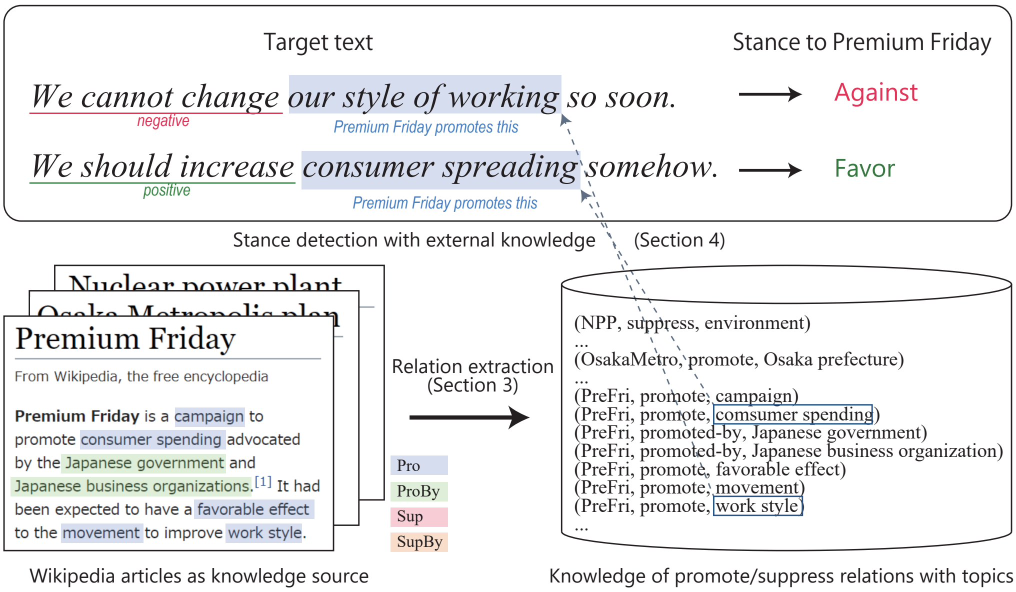

Target text
Stance to Premium Friday
We cannot change our style of working so soon.
negative
Premium Friday promotes this
Against
We should increase consumer spreading somehow.
positive
Premium Friday promotes this
Favor
Stance detection with external knowledge (Section 4)
Nuclear power plant
Osaka Metropolis plan
Premium Friday
From Wikipedia, the free encyclopedia
Premium Friday is a campaign to promote consumer spending advocated by the Japanese government and Japanese business organizations.[1] It had been expected to have a favorable effect to the movement to improve work style.
Relation extraction (Section 3)
Pro
ProBy
Sup
SupBy
(NPP, suppress, environment)
...
(OsakaMetro, promote, Osaka prefecture)
...
(PreFri, promote, campaign)
(PreFri, promote, comsumer spending)
(PreFri, promoted-by, Japanese government)
(PreFri, promoted-by, Japanese business organization)
(PreFri, promote, favorable effect)
(PreFri, promote, movement)
(PreFri, promote, work style)
...
Wikipedia articles as knowledge source
Knowledge of promote/suppress relations with topics

Figure 1: Stance detection using knowledge acquired from Wikipedia articles.

1 Introduction

Stance detection involves inferring whether the attitude of a text's author toward a given topic is positive (for, pro), negative (against, con), or neutral (Mohammad et al., 2016). This task is central to a various applications, such as analyzing on-line debates (Thomas et al., 2006; Murakami and Raymond, 2010; Somasundaran and Wiebe, 2010), identifying opinion groups (Abu-Jbara et al., 2012; Qiu et al., 2013; Hasan and Ng, 2013), predicting election results (Kim and Hovy, 2007; Bermingham and Smeaton, 2011), and detecting fake news (Ferreira and Vlachos, 2016).

Stance detection for unknown topics is important in real-world applications, but its performance is currently significantly lower than that for known topics (Anand et al., 2011; Zarrella and Marsh, 2016; Du et al., 2017). One of the main reasons for this decreased performance is that it often requires external knowledge about the topics involved, as well as about related concepts and events (Sasaki et al., 2016; Boltuzic and Šnajder, 2017; Bar-Haim et al., 2017). For example, consider the sentence "We should adopt free trade," on the topic of the *Trans-Pacific Partnership Agreement* (TPP). It is nontrivial for computers to recognize the author's stance because the text does not contain the topic word TPP. Despite this, humans can easily identify the text's stance as positive if they are aware of the association between the TPP and free trade, namely that the TPP promotes free trade. As seen in Section 2, we frequently encounter cases such as these, where texts do not include topic words directly but instead use related terms.

Finding ways to deal with this issue by incorporating external knowledge into stance detection has proved to be challenging. Sasaki et al. (2016) presented an approach to annotating text spans associated with a given topic, while Boltuzic and Šnajder (2017) proposed representing texts using *microstructures* that express the relation between domain-specific concepts. However, both these studies required the texts to be manually annotated to achieve any improvement in stance detection. Bar-Haim et al. (2017) presented a stance classification method that takes advantage of knowledge about consistent (e.g., similar) and contrastive (e.g., antonym) noun phrase pairs. Even though this approach is promising, they were unable to demonstrate improved performance in stance classification experiments using a standard evaluation metric. We therefore investigate a task where no training data are available for the target topic, but there are data for other topics, as well as external knowledge about the target topic.

In this paper, we break the stance detection task into the following two sub-tasks, as shown in Figure 1.

1. Reading a Wikipedia article about a given topic to learn concepts and events associated with it (knowledge acquisition).

2. Predicting the stance of a given text toward the topic by incorporating the learned concepts and events as external knowledge.

The contributions of this study are fourfold.

1. We build a new stance detection dataset where the topics are associated with Wikipedia articles. This dataset consists of 6,701 tweets on seven topics, and

we show that more than a third of the tweets require topic knowledge for stance detection (Section 2).

2. We construct a corpus of annotated Wikipedia articles to help extract associated concepts and events (Section 3).

3. We propose a novel model that can attend to related concepts and events when encoding a given text using bi-directional long short-term memories (LSTMs) (Section 4).

4. Our experimental results demonstrate that the extracted knowledge improves the F-score for stance classification by about 0.03 (Section 5).

Although the dataset and corpus were built using Japanese texts, the presented method is general and can be applied to other languages. In addition, we use English translations throughout to present readable examples.

2 Building a stance detection dataset

2.1 Goal: surveying public opinion using Twitter data

Our ultimate practical goal is to enhance data journalism (Gray et al., 2012) and automated journalism (Graefe, 2016), which aim to generate stories from data. In particular, we are interested in social listening, namely surveying public opinion using social network service (SNS) data, related to controversial issues in society. We, therefore, focus on topics that are actively discussed on Twitter and have related Wikipedia articles. In this section, we build a new stance detection dataset based around controversial topics with Wikipedia links. This dataset will be useful for assessing the effect of using Wikipedia articles as a knowledge source for stance detection.

2.2 Selecting topics and gathering tweets

We gathered a collection of 26 billion tweets, crawled between April 2015 and June 2017. Then, we computed the TF–IDF scores of each hashtag for weekly intervals, where the term frequency (TF) is the number of times the hashtag occurred during that week and the document frequency (DF) is the number of weeks when the hash tag appeared. Using this procedure, we obtained a list of trending topics for each week during the period.

From this hashtag list, we selected a set of widely discussed topics that had corresponding Wikipedia articles: "Trans-Pacific Partnership" (TPP), "Premium Friday" (PreFri)[1]', "Anti-Conspiracy Bill" (AntiCons), "nuclear power plant" (NPP)', "Osaka Metropolis plan" (OsakaMetro), "2015 Japanese military legislation" (JapanMil), and "right of collective self-defense" (SelfDef). For each topic, we randomly sampled 2,000 tweets from those containing the corresponding hashtag, thus obtaining 14,000 tweets covering seven topics. Finally, we removed the hashtags from the tweets when using them for stance detection.

2.3 Labeling tweet stances

[1] A campaign to finish work at 15:00 on the last Friday of the month and promote consumer spending.

Topic	For	Against	Neutral
TPP	53	802	230
PreFri	153	744	218
AntiCons	86	592	308
NPP	47	783	202
OsakaMetro	239	259	380
JapanMil	168	352	262
SelfDef	160	468	195
Total	906	4,000	1,795

Table 1: Stance label distributions in the stance detection dataset.

To build a stance detection dataset from the tweets, we asked crowd workers to label each tweet as being either *for*, *against*, or *neutral* toward the corresponding topic. After obtaining five annotations for each tweet, we filtered out the tweets that were not assigned the same stance label by no more than three crowd workers. Table 1 shows the number of tweets labeled with each of the three stances for each topic in the dataset.

2.4 Impact of topic-related knowledge on stance detection

Knowledge needed	%	Example statement
None (topic words appear in the tweet)	56.3	*Nuclear power plants* are absolutely necessary.
Promote/suppress (in Wikipedia)	26.3	We should increase the rate of customs duties. (topic: TPP)
Promote/suppress (not in Wikipedia)	13.9	I'm concerned about genetically modified food coming. (topic: TPP)
Other relationships	2.5	Do not revive the Public Security Preservation Law of 1925! (topic: JapanMil)

Table 2: Knowledge needed to detect the stance, focusing on promote and suppress relations.

In this paper, we address the following key questions: what is the impact of topic-related knowledge on stance detection for our dataset? To estimate this, we randomly sampled 491 tweets (10%) from the dataset that had been assigned *for* or *against* labels, and manually associated particular phrases in them with the topics involved. Here, we focus on promote and suppress relations (Hashimoto et al., 2012; Fluck et al., 2015; Hanawa et al., 2017) between the topics and concepts/events in the tweets[2]. Formally, *A promotes B* means that B is activated whenever A is activated, while *A suppresses B* means that B is deactivated whenever A is activated.

Table 2 shows the analysis results. The first row indicates that 56.3% of the tweets included the topic phrase (e.g., "nuclear power plant") in the text. This may be sufficient to perform sentiment analysis with respect to the topic phrase, for example, inferring a *for* stance for the example statement because it consists of a positive sentiment pattern ("X is absolutely necessary") with the variable X filled in by the topic ($X =$ "nuclear power plant").

However, the table also shows that 40.2% (= 26.3% + 13.9%) of the tweets required knowledge about promote and suppress relations between the topics and the terms used. For example, the missing links between the TPP topic and the example tweets are that *TPP suppresses customs duties* and *TPP promotes genetically modified foods*. On further examination, we could often find promote and suppress relations mentioned in the Wikipedia articles, which were helpful in detecting the stances of 26.3% of the tweets. Thus, extracting promote and suppress relations from Wikipedia articles is a promising approach to enhancing stance detection performance.

[2] Promote and suppress relations roughly correspond to consistent and contrastive targets in Bar-Haim et al. (2017). Boltuzic and Šnajder (2017) used eight fine-grained relation types, including promote and suppress relations, in their analyses of claim microstructures. However, when predicting the stance of a claim, these eight relation types can all be reduced to promote and suppress relations; for example, equal(A, B) is can be treated as promote(A, B).

	TPP	PreFri	AntiCons	NPP	OsakaMetro	JapanMil	SelfDef
No. of sentences	333	6	179	257	169	121	39
PRO	257	17	122	190	165	120	42
SUP	67	2	163	74	115	46	25
PROBY	131	7	77	108	64	45	21
SUPBY	145	3	86	96	51	30	6

Table 3: Numbers of sentences in each Wikipedia article and annotated spans of different relations.

3 Acquiring promote and suppress relations from Wikipedia

Based on the analysis in the previous section, we assumed that understanding promote and suppress relations is essential for stance detection. We therefore attempted to obtain them by *reading* Wikipedia articles related to the topics. Following Hanawa et al. (2017), we treated this as a sequential labeling task, namely recognizing text spans in each Wikipedia article that have promote or suppress relations with the article's title. In other words, we identified the relation of a given span to the article's title using the following four directed relation labels.

- PRO: "[*title*] promotes B"
- SUP: "[*title*] suppresses B"
- PROBY: "A promotes [*title*]"
- SUPBY: "A suppresses [*title*]"

Here, A and B are text span placeholders, and [*title*] is the article's title.

3.1 Manually annotating relations in Wikipedia articles

Although Hanawa et al. (2017) has released annotated data giving the promote and suppress relations for the summary sentences of 1,494 Wikipedia articles, we collected additional annotations for the Wikipedia articles corresponding to the seven topics considered here. Specifically, we annotated the articles' promote and suppress relations via crowdsourcing, adding the labels PRO, SUP, PROBY, and SUPBY to text spans in the articles. We used the Yahoo! crowdsourcing service[3] to obtain 10 annotations per article, adopting particular annotations only if at least two out of the 10 workers assigned the same relation to the same span. Table 3 shows the number of sentences in each article and the annotated spans for each relation.

3.2 Automatically extracting relations from Wikipedia articles

We then used the dataset of Wikipedia articles annotated with PRO, SUP, PROBY, and SUPBY labels obtained above as supervised training data in order to extract relation instances automatically from Wikipedia articles. We formalized this task as a sequential labeling problem using IOB2 notation, namely the task of predicting the sequence of labels (e.g., B-PRO, I-PRO, B-SUP, or I-SUP) for a sequence of words in a given article.

We modeled the sequential labeling problem using a bidirectional LSTM with a conditional random field (LSTM-CRF) (Huang et al., 2015). The dimensions of the word embeddings and hidden layers were set to 300, and we initialized the word embeddings to ones that had previously been trained using Japanese Wikipedia articles[4]. We trained the model using a combination of the data released by Hanawa et al. (2017) and the dataset built in Section 3.1.

The predicted IOB2 labels were only adopted when their probability exceeded a threshold α. In this way, we built a knowledge base (KB) $\mathcal{D}$ of tuples consisting of a Wikipedia article title $\mathfrak{t}$, a relation $\mathfrak{r}$, and a mention $\mathfrak{m}$ in the Wikipedia article:

$$(\mathfrak{t}, \mathfrak{r}, \mathfrak{m}) \in \mathcal{D}. \tag{1}$$

The KB $\mathcal{D}$ included the relation extraction results for all Wikipedia articles.

[3] http://crowdsourcing.yahoo.co.jp/

[4] https://github.com/overlast/word-vector-web-api

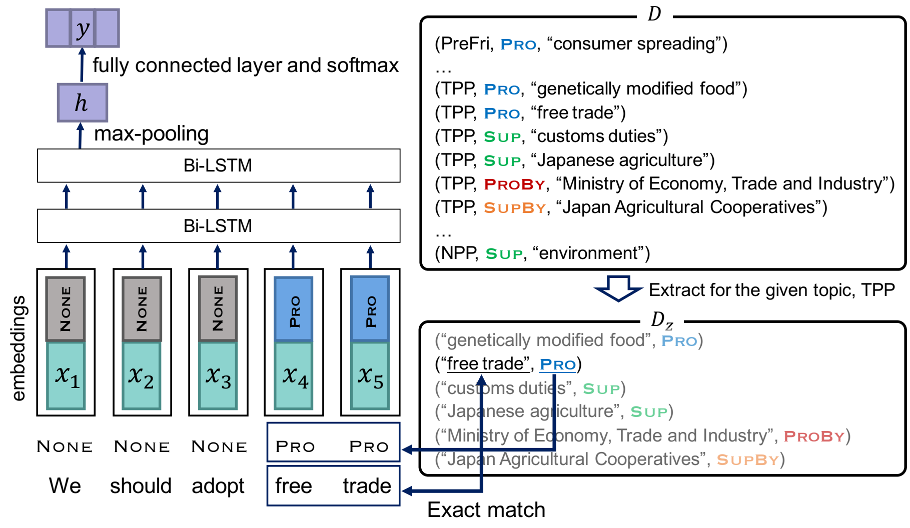

y
fully connected layer and softmax
h
max-pooling
Bi-LSTM
Bi-LSTM
embeddings
NONE NONE NONE PRO PRO
x1 x2 x3 x4 x5
NONE NONE NONE PRO PRO
We should adopt free trade
D
(PreFri, PRO, "consumer spreading")
…
(TPP, PRO, "genetically modified food")
(TPP, PRO, "free trade")
(TPP, SUP, "customs duties")
(TPP, SUP, "Japanese agriculture")
(TPP, PROBY, "Ministry of Economy, Trade and Industry")
(TPP, SUPBY, "Japan Agricultural Cooperatives")
…
(NPP, SUP, "environment")
Extract for the given topic, TPP
D_z
("genetically modified food", PRO)
("free trade", PRO)
("customs duties", SUP)
("Japanese agriculture", SUP)
("Ministry of Economy, Trade and Industry", PROBY)
("Japan Agricultural Cooperatives", SUPBY)
Exact match

Figure 2: Incorporating promote and suppress relations into stance detection with exact matching.

4 Stance detection models

In this section, we propose three stance detection models, two of which take advantage of the promote and suppress relations acquired in Section 3. Given a topic z and an N-word text $s = w_1, w_2, \cdots, w_N$, each model computes the probability distribution over the three stance labels $y \in \mathbb{R}^3$, corresponding to the probabilities that the text s should be classified as *for*, *against*, and *neutral* toward the topic, respectively.

4.1 Baseline model (without external KB)

Our baseline model computes a vector h for the given text s using two-layer bidirectional LSTMs (which can achieve high performance, comparable with those of more complex models, in some tasks) and max pooling. It obtains a word embedding $x_t \in \mathbb{R}^{d_w}$ and ELMo vector $ELMo_t \in \mathbb{R}^{d_e}$ (Peters et al., 2018) for each word w_t, where d_w and d_e denote the dimensionalities of the word embeddings and ELMo, respectively. Then, it concatenates these to arrive at a vector x'_t for each w_t:

$$x'_t = x_t \oplus ELMo_t. \tag{2}$$

Here, $\oplus$ represents vector concatenation. The LSTMs are used to compute the vectors $\overrightarrow{h}_1, \cdots, \overrightarrow{h}_N$ and $\overleftarrow{h}_1, \cdots, \overleftarrow{h}_N$, based on the word vectors $x'_1, \cdots, x'_N$ in the forward and backward directions, as follows:

$$\overrightarrow{h}_t, \overrightarrow{c}_t = \overrightarrow{\mathrm{LSTM}}(x'_t, \overrightarrow{h}_{t-1}, \overrightarrow{c}_{t-1}), \tag{3}$$

$$\overleftarrow{h}_t, \overleftarrow{c}_t = \overleftarrow{\mathrm{LSTM}}(x'_t, \overleftarrow{h}_{t+1}, \overleftarrow{c}_{t+1}). \tag{4}$$

Here, $\overrightarrow{h}_t, \overrightarrow{c}_t \in \mathbb{R}^{d_h}$ $(t = 1, \cdots, N)$ are the hidden states and memory cells of the forward LSTMs ($\overrightarrow{\mathrm{LSTM}}$), respectively, and $\overleftarrow{h}_t, \overleftarrow{c}_t \in \mathbb{R}^{d_h}$ $(t = 1, \cdots, N)$ are those of the backward LSTMs ($\overleftarrow{\mathrm{LSTM}}$). In addition, d_h is the dimensionality of the vectors $\overrightarrow{h}_t$, $\overrightarrow{c}_t$, $\overleftarrow{h}_t$, and $\overleftarrow{c}_t$.

After exploring several methods of constructing text vectors from $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$, e.g., $[\overrightarrow{h}_N; \overleftarrow{h}_1]$, we decided to apply max pooling, as this yielded the highest performance on the validation set. Specifically, the text vector $h \in \mathbb{R}^{2d_h}$ is computed by max pooling over $[\overrightarrow{h}_t; \overleftarrow{h}_t]$ $(t = 1, \cdots, N)$.

Finally, the model computes the probability distribution over the three stance labels from the text vector h as follows:

$$y = \mathrm{softmax}(W \cdot h + b). \tag{5}$$

Here, $W \in \mathbb{R}^{3\times 2d_h}$ and $b \in \mathbb{R}^3$ denote the weight matrix and bias vector, respectively.

4.2 Exact matching (with KB)

A simple way to incorporate additional knowledge into the baseline model is to automatically annotate the spans of a given text s for which there is a relation with the topic z in the KB $\mathcal{D}$ (Figure 2). More concretely, this model finds the (longest) exactly matching text spans that are included in $\mathcal{D}$ for the topic z. First, it obtains the set of tuples from $\mathcal{D}$ where the title $\mathfrak{t}$ matches the topic z:

$$D_z = \{(\mathfrak{m}, \mathfrak{r}) \mid (\mathfrak{t}, \mathfrak{r}, \mathfrak{m}) \in \mathcal{D} \wedge \mathfrak{t} = z\}. \tag{6}$$

Next, it defines the variable p_t to represent the relation by which the word w_t matches a record in D_z. This can take values of either PRO, SUP, PROBY, SUPBY, or NONE, where the latter is a special relation indicating that the word w_t cannot be associated with any record in D_z.

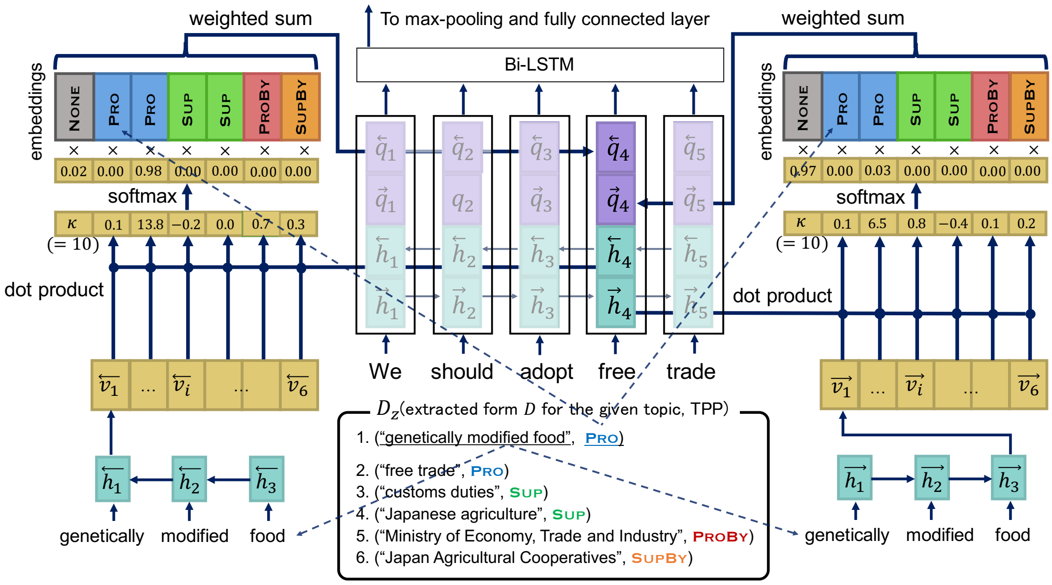

weighted sum
To max-pooling and fully connected layer
weighted sum
Bi-LSTM
embeddings
NONE
PRO
PRO
SUP
SUP
PROBY
SUPBY
×
0.02
0.00
0.98
0.00
0.00
0.00
0.00
softmax
κ
0.1
13.8
−0.2
0.0
0.7
0.3
(= 10)
dot product
$\overleftarrow{q}_1$ $\overleftarrow{q}_2$ $\overleftarrow{q}_3$ $\overleftarrow{q}_4$ $\overleftarrow{q}_5$
$\overrightarrow{q}_1$ q_2 $\overrightarrow{q}_3$ $\overrightarrow{q}_4$ $\overrightarrow{q}_5$
$\overleftarrow{h}_1$ $\overleftarrow{h}_2$ $\overleftarrow{h}_3$ $\overleftarrow{h}_4$ $\overleftarrow{h}_5$
$\overrightarrow{h}_1$ $\overrightarrow{h}_2$ $\overrightarrow{h}_3$ $\overrightarrow{h}_4$ $\overrightarrow{h}_5$
We
should
adopt
free
trade
$\overleftarrow{v_1}$
...
$\overleftarrow{v_i}$
...
$\overleftarrow{v_6}$
$\overleftarrow{h_1}$ $\overleftarrow{h_2}$ $\overleftarrow{h_3}$
genetically
modified
food
embeddings
NONE
PRO
PRO
SUP
SUP
PROBY
SUPBY
×
0.97
0.00
0.03
0.00
0.00
0.00
0.00
softmax
κ
0.1
6.5
0.8
−0.4
0.1
0.2
(= 10)
dot product
$\overrightarrow{v_1}$
...
$\overrightarrow{v_i}$
...
$\overrightarrow{v_6}$
$\overrightarrow{h_1}$ $\overrightarrow{h_2}$ $\overrightarrow{h_3}$
genetically
modified
food
D_z(extracted form D for the given topic, TPP)
1. ("genetically modified food", PRO)
2. ("free trade", PRO)
3. ("customs duties", SUP)
4. ("Japanese agriculture", SUP)
5. ("Ministry of Economy, Trade and Industry", PROBY)
6. ("Japan Agricultural Cooperatives", SUPBY)

Figure 3: Incorporating promote and suppress relations into stance detection via attention-based matching.

As an example, let us consider predicting a stance for the sentence "We should adopt free trade," with respect to the TPP topic, i.e., z = TPP and s = ("we", "should", "adopt", "free", "trade"). In addition, suppose the database D_z consists of just {("free trade", PRO)}. Then, the values of p_1, p_2, and p_3 would be NONE and those of p_4 and p_5 would be PRO. Next, the model would embed the relation of the word w_t to the topic z by concatenating the word vector x'_t and the relation embedding:

$$x''_t = x'_t \oplus \mathrm{emb}(p_t). \tag{7}$$

Here, $\mathrm{emb}(p) : p \longmapsto \mathbb{R}^{d_r}$ is a function that looks up the embedding vector for a given relation p, and d_r is the dimensionality of the relation embeddings. This model incorporates these relation embeddings into the baseline model by using x''_t instead of x'_t in Equations 3 and 4.

4.3 Attention-based matching (with KB)

One issue with the above exact matching approach is that we cannot guarantee that a given text will use the exact phrases included in the KB $\mathcal{D}$. Thus, we now propose a neural network model that carries out more flexible matching against the KB, as illustrated in Figure 3. This method essentially computes, for each word w_t in the input text, an attention score for the i-th record in the KB by comparing the hidden vectors of the word w_t and the mention $\mathfrak{m}_i$, as encoded by LSTMs sharing the same parameters.

First, we obtain the database D_z for the given topic z using Equation 6. Let $\mathfrak{m}_i$ and $\mathfrak{r}_i$ denote the mention and relation associated with the i-th record $(\mathfrak{m}_i, \mathfrak{r}_i)$ in D_z $(i \in \{1, ..., |D_z|\})$. The records in D_z are stored as key–value pairs, where the keys are the mentions ($\mathfrak{m}$) and the values are the relations ($\mathfrak{r}$). We use single-layer bidirectional LSTMs to encode the keys (mentions) as sequences of words, for example, encoding the phrase “free trade” for the record (“free trade”, PRO). In addition, let $\overrightarrow{v}_i \in \mathbb{R}^{d_h}$ and $\overleftarrow{v}_i \in \mathbb{R}^{d_h}$ be vectors representing the mention $\mathfrak{m}_i$ in D_z encoded in the forward and backward directions, respectively.

We also construct the vectors $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ for the word w_t in the input text by using LSTMs sharing the same parameters as were used to encode $\overrightarrow{v}_i$ and $\overleftarrow{v}_i$. Then, we compute attention scores $\overrightarrow{a}_{t,0}, \overrightarrow{a}_{t,1}, \cdots, \overrightarrow{a}_{t,|D_z|}$, where $\overrightarrow{a}_{t,0}$ is the score when the word w_t is not matched with any record in D_z and $\overrightarrow{a}_{t,i}$ $(i \in \{1, ..., |D_z|\})$ are the scores when the word w_t is matched with the i-th record of D_z. Specifically, the score $\overrightarrow{a}_{t,i}$ is computed between the forward hidden states of $\overrightarrow{h}_t$ and $\overrightarrow{v}_i$, as follows:

$$\overrightarrow{a}_{t,i} = \frac{\exp(\mathrm{sim}(\overrightarrow{h}_t, \overrightarrow{v}_i))}{\sum_{i'=0}^{|D_z|} \exp(\mathrm{sim}(\overrightarrow{h}_t, \overrightarrow{v}_{i'}))}. \tag{8}$$

Here, the function sim essentially computes the dot product of the two vectors:

$$\mathrm{sim}(\overrightarrow{h}_t, \overrightarrow{v}_i) = \begin{cases} \overrightarrow{h}_t \cdot \overrightarrow{v}_i & (\text{if } 0 < i) \\ \kappa & (\text{if } i = 0) \end{cases}. \tag{9}$$

Note, however, that this yields the constant value κ (a hyper-parameter) when $i = 0$, corresponding to the case where w_t cannot be matched with any record in D_z.

Let r_i represent the relation $\mathfrak{r}_i$ for the i-th record $(\mathfrak{m}_i, \mathfrak{r}_i) \in D_z$ when $i \in \{1, ..., |D_z|\}$ and NONE otherwise ($i = 0$). We also introduce the function $\mathrm{emb}(r) : r \longmapsto \mathbb{R}^{d_r}$, which looks up the embedding vector for a given relation r. With these, we compute the relation embedding for the word w_t as follows:

$$\overrightarrow{q}_t = \sum_{i=0}^{|D_z|} \overrightarrow{a}_{t,i} \cdot \mathrm{emb}(r_i). \tag{10}$$

We also compute a relation embedding $\overleftarrow{q}_t$ for the backward direction similarly. Finally, we incorporate these embeddings into the baseline model by feeding in $[\overrightarrow{h}_t; \overleftarrow{h}_t; \overrightarrow{q}_t; \overleftarrow{q}_t]$ instead of $[\overrightarrow{h}_t; \overleftarrow{h}_t]$ at the boundary between the baseline model's first and second LSTM layers.

5 Experiments

5.1 Setting

We evaluated the contribution of the external knowledge (promote and suppress relations) via topic-wise seven-fold cross-validation. Each run used data on five topics to train the models, one topic's data as the test set, and the remaining topic's data as the validation set. The fact that the test set (target topic) and validation set were excluded from the training data means that the relation extraction model was required to make predictions about unseen topics, making this a difficult task.

With regard to the parameters used, we set $d_w = d_h = d_e = 300$, $d_r = 100$, and $\kappa = 10$. In addition, we decided on a probability threshold of $\alpha = 0.85$ after conducting a search using the validation sets. As noted above, the word embeddings were initialized to the results of training them on Japanese Wikipedia articles[6]. The parameters in the neural network models (i.e., those of the LSTMs, fully connected layers, word and relation embeddings) were optimized by Adam. Following Mohammad et al. (2016), we used the macro-average F_1 score F_{avg} as an evaluation metric, calculated as $F_{avg} = (F_{for} + F_{against})/2$, where F_{for} and $F_{against}$ are the F_1 scores for the *for* and *against* stance predictions.

We also explored an ensemble approach that involved training 10 models, each initialized randomly, and then considering their majority vote. If this resulted in a tie, we broke the tie in favor of the most common label in the training data (resulting in a priority order of *against*, *neutral*, then *for*).

5.2 Results

Model	Knowledge	Ensemble	TPP	PreFri	AntiCons	NPP	OsakaMetro	JapanMil	SelfDef	Overall
Majority baseline		no	0.375	**0.400**	0.431	0.425	0.228	0.310	0.363	0.374
Baseline		no	0.490	0.258	0.489	0.430	0.332	0.533	0.470	0.466
		yes	0.498	0.267	0.490	0.439	0.348	0.543	0.488	0.478
Exact match	topic	no	0.488	0.257	0.502	0.431	0.331	0.528	0.475	0.470
		yes	0.499	0.260	0.502	0.448	0.352	0.549	0.494	0.475
	automatic	no	0.484	0.255	0.488	0.447	0.359	0.544	0.477	0.475
		yes	0.489	0.258	0.514	0.448	0.360	0.561	0.482	0.486
	gold	no	0.501	0.264	0.525	0.444	0.368	0.548	0.477	0.483
		yes	0.516	0.262	0.529	0.452	**0.369**	0.566	0.489	0.490
Attention match	topic	no	0.491	0.250	0.488	0.425	0.347	0.539	0.475	0.468
		yes	0.489	0.251	0.495	0.436	0.352	0.546	0.479	0.476
	automatic	no	0.491	0.252	0.508	0.430	0.352	0.564	0.481	0.473
		yes	0.509	0.263	0.527	0.442	0.360	0.570	0.499	0.491
	gold	no	0.500	0.249	0.515	0.448	0.361	0.566	0.486	0.490
		yes	**0.523**	0.258	**0.539**	**0.465**	0.362	**0.582**	**0.500**	**0.507**

Table 4: Stance detection performance (F_1 score) of each model for each of the seven topics.

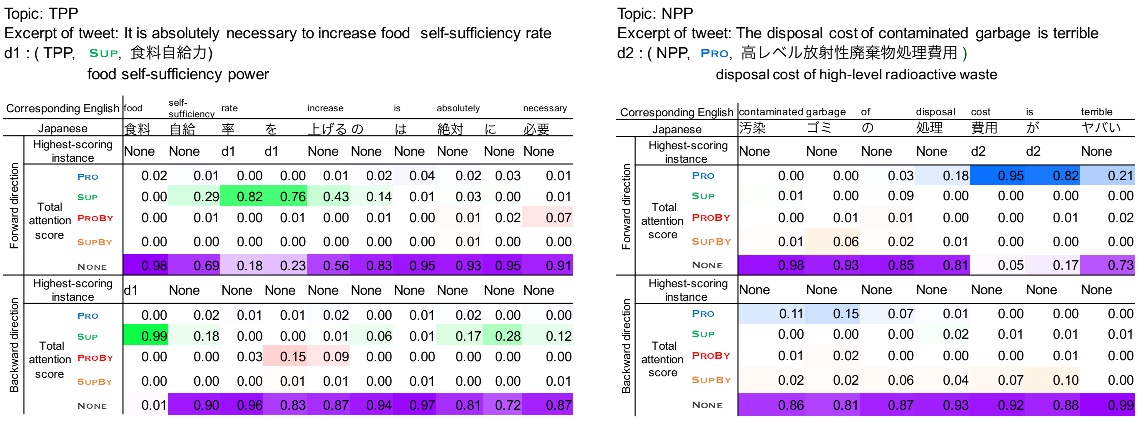
Table 4 shows the different models' stance classification performance. The overall F_1 scores are micro-averages of the F_{avg} scores for each topic. The discussion below focuses on the ensemble approach, as this always produced better results than using a single model.

The baseline method obtained an overall F_1 score of 0.478. Using a naïve topic-knowledge-only method that associates occurrences of the topic name in the text with PRO and PROBY relations did not improve performance, yielding scores of 0.475 (exact matching) and 0.476 (attention-based matching). Here, we should again emphasize that we measured the models' performance via cross-topic validation, meaning that, during the test phase, they had to predict stances for a topic on which they had not been trained. This is the main reason why these performance results are so low and there is no strong baseline: most of the existing methods were not designed for this cross-topic setting.

We then explored two other ways of incorporating knowledge into stance detection, which are listed as the *automatic* and *gold* knowledge results in Table 4. The *automatic* approach used the relations extracted by the method proposed in Section 3.2, gaining topic knowledge from Wikipedia articles. In other words, this setting corresponds to the process where a computer reads a Wikipedia article to learn knowledge about the topic and predicts stances with the knowledge. The *gold* method instead used the gold standard relations created by crowdsourcing annotations to the Wikipedia articles, and thus, represents an upper bound on the performance of our approach.

As we can see from Table 4, using the promote and suppress relations improved performance, with the best results obtained by applying attention-based matching with the gold-standard relations (0.507). Using the automatically extracted relation information reduced performance (to 0.491), but still yielded better results than that obtained without including this knowledge. In consequence, we believe that manually curating the promote and suppress relations can be a reasonable strategy for realizing intelligent systems, but these results also demonstrate that even less accurate topic knowledge can improve the stance classification performance.

Table 4 also shows the superiority of attention-based matching over exact matching, especially when using the automatically derived and gold standard promote and suppress relations. Visualization of the attention scores (Figure 4) demonstrates that attention-based matching could find topic-related concepts/events even when the phrases



Topic: TPP

Excerpt of tweet: It is absolutely necessary to increase food self-sufficiency rate

d1 : (TPP, SUP, 食料自給力)

food self-sufficiency power

		Corresponding English	food	self-sufficiency	rate		increase		is	absolutely		necessary
		Japanese	食料	自給	率	を	上げる	の	は	絶対	に	必要
Forward direction	Highest-scoring instance		None	None	d1	d1	None	None	None	None	None	None
	Total attention score	PRO	0.02	0.01	0.00	0.00	0.01	0.02	0.04	0.02	0.03	0.01
		SUP	0.00	0.29	0.82	0.76	0.43	0.14	0.01	0.03	0.00	0.01
		PROBY	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.02	0.07
		SUPBY	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
		NONE	0.98	0.69	0.18	0.23	0.56	0.83	0.95	0.93	0.95	0.91
Backward direction	Highest-scoring instance		d1	None	None	None	None	None	None	None	None	None
	Total attention score	PRO	0.00	0.02	0.01	0.01	0.02	0.00	0.01	0.02	0.00	0.00
		SUP	0.99	0.18	0.00	0.00	0.01	0.06	0.01	0.17	0.28	0.12
		PROBY	0.00	0.00	0.03	0.15	0.09	0.00	0.00	0.00	0.00	0.00
		SUPBY	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.01
		NONE	0.01	0.90	0.96	0.83	0.87	0.94	0.97	0.81	0.72	0.87

Topic: NPP

Excerpt of tweet: The disposal cost of contaminated garbage is terrible

d2 : (NPP, PRO, 高レベル放射性廃棄物処理費用)

disposal cost of high-level radioactive waste

		Corresponding English	contaminated	garbage	of	disposal	cost	is	terrible
		Japanese	汚染	ゴミ	の	処理	費用	が	ヤバい
Forward direction	Highest-scoring instance		None	None	None	None	d2	d2	None
	Total attention score	PRO	0.00	0.00	0.03	0.18	0.95	0.82	0.21
		SUP	0.01	0.00	0.09	0.00	0.00	0.00	0.00
		PROBY	0.00	0.01	0.01	0.00	0.00	0.01	0.02
		SUPBY	0.01	0.06	0.02	0.01	0.00	0.00	0.00
		NONE	0.98	0.93	0.85	0.81	0.05	0.17	0.73
Backward direction	Highest-scoring instance		None	None	None	None	None	None	None
	Total attention score	PRO	0.11	0.15	0.07	0.01	0.00	0.00	0.00
		SUP	0.00	0.00	0.00	0.02	0.01	0.01	0.01
		PROBY	0.01	0.02	0.00	0.00	0.00	0.01	0.00
		SUPBY	0.02	0.02	0.06	0.04	0.07	0.10	0.00
		NONE	0.86	0.81	0.87	0.93	0.92	0.88	0.99

Figure 4: Examples of instances with sums of attention scores at each word visualized.

	TPP		PreFri		AntiCons		NPP		OsakaMetro		JapanMil		SelfDef	
	p	r	p	r	p	r	p	r	p	r	p	r	p	r
PRO	0.41	0.12	0.48	0.37	0.26	0.27	0.29	0.30	0.38	0.15	0.48	0.18	0.43	0.10
SUP	0.67	0.09	1.00	1.00	0.58	0.06	0.34	0.18	1.00	0.09	1.00	0.15	0.59	0.29
PROBY	0.18	0.17	0.38	0.29	0.14	0.18	0.23	0.10	0.19	0.13	0.16	0.13	0.50	0.06
SUPBY	0.32	0.08	0.0	0.0	0.29	0.15	0.19	0.05	0.25	0.18	0.30	0.08	0.85	0.10

Table 5: Character-level precision (p) and recall (r) values for the automatically acquired promote and suppress relations.

in the text were not identical to those in the relations obtained from Wikipedia.

Table 5 shows the performance of the automatic promote and suppress relation extraction process in terms of the character-level precision (p) and recall (r) of the recognition results with respect to the manually annotated data. Because we chose a relatively high threshold for extracting relation instances ($\alpha = 0.85$) after considering the validation sets, the model had a tendency to produce a limited number of highly confident relations, resulting in relatively high precision but low recall.

6 Related work

Many researchers have recently addressed stance detection (Thomas et al., 2006; Mohammad et al., 2016), but stance detection across different topics (*cross-topic* stance detection) remains a challenge: although these methods can achieve fairly high performance for known topics, their performance drops substantially for unseen topics (absent from the training data) (Anand et al., 2011; Zarrella and Marsh, 2016; Du et al., 2017).

The previous studies on cross-topic stance detection can be divided into two groups. Methods in the first group create pseudo-training data for unseen topics by using certain clues, e.g., hashtags (Wei et al., 2016) or user profiles (Ebrahimi et al., 2016), taken from an SNS service. However, these approaches suffer from two disadvantages: they require large amounts of unlabeled data, which may not be available for infrequently discussed topics, and they rely on the existence of clues specific to the topic and SNS service, making them difficult to generalize to arbitrary topics and SNS services.

The other group of methods explores the use of external knowledge. Previous studies have considered various kinds of relational knowledge about topics, such as paraphrases (Ferreira and Vlachos, 2016), comparisons (Jindal and Liu, 2006), and relation aspects (Somasundaran and Wiebe, 2009), as well as entailment (Cabrio and Villata, 2013) and cause–effect (Sasaki et al., 2016) relations. Boltuzic and Šnajder (2017) defined eight types of relations (promote, suppress, allow, entail, contradict, purpose, equal, and have) to analyze claims in terms of microstructures. While using such fine-grained relations is a reasonable approach, it can be challenging to discriminate them.

To deal with this issue, we reduced these fine-grained relations to just promote and suppress relations, considering only those between a topic and related concepts/events that affect stance polarity. Bar-Haim et al. (2017) presented a similar approach, employing consistent and contrastive relations for stance detection. However, they relied on linguistic patterns such as “A vs B” and “A versus B” to extract contrastive relation instances from query logs and Wikipedia titles/headers. This method is suitable for extracting competing concepts, e.g., TPP and NAFTA, but not for causally related concepts, e.g., TPP and customs duties: we cannot expect queries such as “TPP vs. customs duties.” In addition, they only found an improvement in stance detection performance when they evaluated a high-confidence subset of the test data; the effect of using consistent and contrastive relations disappeared when they evaluated the whole test dataset.

7 Conclusion

This paper presents a novel approach to stance detection that utilizes external knowledge about the topics involved. To evaluate it, we built a stance detection dataset consisting of 6,701 tweets on seven topics. Our analysis of this dataset showed that detecting the stances of 40.2% of the tweets required knowledge of the topics' promote and suppress relations, which we obtained from Wikipedia articles. We also propose a neural network model that attends to the relation instances based on an input sentence. The experimental results demonstrate that including promote and suppress relation instances can have a positive impact on stance detection (yielding F_1 score improvements of 0.013 and 0.029 when the knowledge is automatically extracted and manually annotated, respectively).

In the future, we plan to expand the range of external knowledge sources to include newspaper articles and SNS data in order to collect more relation instances for each topic. Another interesting direction would be to explore an end-to-end architecture covering both knowledge acquisition and stance detection, tasks that are currently handled by two separate models.

Acknowledgements

I am deeply grateful to Dr. Kentaro Inui and Dr. Jun Suzuki for able guidance and generous support. I would like to show my greatest appreciation to Dr. Naoaki Okazaki and Dr. Akira Sasaki for insightful comments and constructive suggestions. Discussions with my academic colleagues in our laboratory have been quite illuminating.

References

Abu-Jbara, A., P. Dasigi, M. Diab, and D. Radev
2012. Subgroup detection in ideological discussions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Pp. 399–409.

Anand, P., M. Walker, R. Abbott, J. E. F. Tree, R. Bowmani, and M. Minor
2011. 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)*, Pp. 1–9.

Bar-Haim, R., I. Bhattacharya, F. Dinuzzo, A. Saha, and N. Slonim
2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Pp. 251–261.

Bermingham, A. and A. Smeaton
2011. On using twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP)*, Pp. 2–10.

Boltuzic, F. and J. Šnajder
2017. Toward stance classification based on claim microstructures. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, Pp. 74–80.

Cabrio, E. and S. Villata
2013. A natural language bipolar argumentation approach to support users in online debate interactions. *Argument and Computation*, 4(3):209–230.

Du, J., R. Xu, Y. He, and L. Gui
2017. Stance classification with target-specific neural attention networks. In *Proceedings of the 26th International Joint Conferences on Artificial Intelligence (IJCAI)*, Pp. 3988–3994.

Ebrahimi, J., D. Dou, and D. Lowd
2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Pp. 1012–1017.

Ferreira, W. and A. Vlachos
2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Pp. 1163–1168.

Fluck, J., S. Madan, T. R. Ellendorff, T. Mevissen, S. Clematide, A. van der Lek, and F. Rinaldi
2015. Track 4 overview: Extraction of causal network information in biological expression language (BEL). In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, Pp. 333–346.

Graefe, A.
2016. Guide to automated journalism. Technical report, The Tow Center for Digital Journalism.

Gray, J., L. Chambers, and L. Bounegru
2012. *The Data Journalism Handbook*. O'Reilly Media.

Hanawa, K., A. Sasaki, N. Okazaki, and K. Inui
2017. A crowdsourcing approach for annotating causal relation instances in Wikipedia. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC)*.

Hasan, K. S. and V. Ng
2013. Extra-linguistic constraints on stance recognition in ideological debates. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Pp. 816–821.

Hashimoto, C., K. Torisawa, S. De Saeger, J.-H. Oh, and J. Kazama
2012. Excitatory or inhibitory: a new semantic orientation extracts contradiction and causality from the web. In *Proceedings of the 2012 Conference on Empirical*

Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Pp. 619–630.

Huang, Z., W. Xu, and K. Yu
2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991.*

Jindal, N. and B. Liu
2006. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, Pp. 244–251.

Kim, S.-M. and E. Hovy
2007. Crystal: Analyzing predictive opinions on the web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Pp. 1056–1064.

Mohammad, S., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry
2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, Pp. 31–41.

Murakami, A. and R. Raymond
2010. 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)*, Pp. 869–875.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer
2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, Pp. 2227–2237.

Qiu, M., L. Yang, and J. Jiang
2013. Modeling interaction features for debate side clustering. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management (CIKM)*, Pp. 873–878.

Sasaki, A., J. Mizuno, N. Okazaki, and K. Inui

2016. Stance classification by recognizing related events about targets. In *Proceedings of the 2017 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Pp. 582–587.

Somasundaran, S. and J. Wiebe

2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, Pp. 226–234.

Somasundaran, S. and J. Wiebe

2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, Pp. 116–124.

Thomas, M., B. Pang, and L. Lee

2006. 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)*, Pp. 327–335.

Wei, W., X. Zhang, X. Liu, W. Chen, and T. Wang

2016. 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)*, Pp. 384–388.

Zarrella, G. and A. Marsh

2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, Pp. 458–463.