

JFCKB: Japanese Feature Change Knowledge Base

Tetsuaki Nakamura[†], Daisuke Kawahara^{†‡}

[†]Kyoto University, [‡]JST PRESTO

Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan

tnakamura@nlp.ist.i.kyoto-u.ac.jp, dk@i.kyoto-u.ac.jp

Abstract

Commonsense knowledge plays an essential role in our language activities. Although many projects have aimed to develop language resources for commonsense knowledge, there is little work focusing on connotational meanings. This is because constructing commonsense knowledge including connotational meanings is challenging. In this paper, we present a Japanese knowledge base where arguments in event sentences are associated with various feature changes caused by the events. For example, “my child” in “my wife hits my child” is associated with some feature changes, such as *increase in pain*, *increase in anger*, *increase in disgust*, and *decrease in joy*. We constructed this knowledge base through crowdsourcing tasks by gathering feature changes of arguments in event sentences. After the construction of the knowledge base, we conducted an experiment in anaphora resolution using the knowledge base. We regarded anaphora resolution as an antecedent candidate ranking task and used Ranking SVM as the solver. Experimental results demonstrated the usefulness of our feature change knowledge base.

Keywords: emotion, commonsense knowledge, knowledge base, crowdsourcing, anaphora resolution

1. Introduction

Commonsense knowledge plays an essential role in our language activities. Such knowledge also plays an important role for computers to understand texts. In the area of language resource development, most studies on commonsense knowledge focus on denotational meanings such as predicate-argument structures (Baker et al., 1998; Palmer et al., 2005). Meanwhile, there are few studies focusing on connotational meanings (Nakamura and Kawahara, 2016; Rashkin et al., 2016). Even in such studies, abstract features such as polarities are used. However, concrete knowledge (that is, fine-grained knowledge) is better for computers than abstract knowledge because events actually cause various concrete feature changes to participants in the events or those who know of the events.

There are two approaches to acquiring commonsense knowledge. One is the automatic acquisition approach; the other is the manual acquisition approach. The automatic acquisition approach uses machine learning techniques and pattern matching methods. Although this approach is useful when the amount of data to be acquired is extremely large, the quality of acquired knowledge might be low. Moreover, the existence of *reporting bias* in texts (Gordon et al., 2010; Gordon and Durme, 2013) shows that infrequent events tend to appear more in texts than quotidian events. The manual approach is useful for gathering subjective information (such as emotional information) and quotidian commonsense knowledge, which are difficult to acquire automatically. This approach may use a fully manual technique (at a very early stage of artificial intelligence studies) or collective intelligence (e.g., crowdsourcing and games with a purpose).

In this study, we aim to solve some problems in previous work on connotational meanings by constructing a large-scale knowledge base of concrete feature changes of arguments in event sentences with a controlled gran-

ularity. For example, “my child” in “my wife hits my child” is associated with some feature changes such as *increase in pain*, *increase in anger*, *increase in disgust*, and *decrease in joy*. The proposed knowledge base is for Japanese. We combined automatic and manual approaches (collective intelligence) to exploit the merits of both techniques. After the construction of our feature change knowledge base, we conducted an experiment on anaphora resolution using the knowledge base. In this experiment, we used a Japanese translated version of the Winograd Schema Challenge (WSC) dataset (JWSC) (Levesque, 2011; Shibata et al., 2015).

2. Related Work

There are many studies focusing on events themselves. FrameNet (Baker et al., 1998) is a corpus in which deep cases (semantic roles) of predicates are defined. In FrameNet, each argument in an example sentence is labeled with a deep case. ConceptNet (Speer and Havasi, 2012) is a large semantic network constructed in the Open Mind Common Sense project (Singh et al., 2012). It is composed of relationships between concepts, where concepts are noun, verb or adjectival phrases.

To date, some knowledge bases have been developed to capture connotational meanings of predicates. Lexical Conceptual Structure (LCS) is a useful way to describe behavior of arguments in event sentences (Jackendoff, 1983). In LCS, not only behaviors of arguments but also relationships between arguments in event sentences are described, for each verb. However, LCS focuses only on direct and explicit information. It is necessary to develop a method to process indirect and implicit information because such knowledge is frequently used. Moreover, there is a possibility that many predicates are excessively generalized because the information used in LCS is abstract. WordNet-Affect (Strapparava and Valitutti, 2004) is an extended version of WordNet (Fellbaum, 1998). Although synsets of WordNet are associated with some emotions

in WordNet-Affect, it is hard to know who/what is associated with the emotions in events. *Connotation Frames* are a knowledge base of emotional implications of events (Rashkin et al., 2016). In *Connotation Frames*, implications of an event are represented by a set of polarities categorized into five types (writers’ perspective, entities’ perspective, effect, value, and mental state). To construct *Connotation Frames*, crowdsourcing and a news corpus were used. Although this knowledge base is useful for understanding information conveyed by event descriptions, it is hard to know details of the emotions associated with events because emotions themselves and associated behaviors are abstracted by polarities in the knowledge base.

As for connotational meanings, especially for mental states, there are many studies that exploit traditional emotion models proposed by psychologists (Tokuhisa et al., 2008; Tokuhisa et al., 2009; Hasegawa et al., 2013; Vu et al., 2014). In these studies, Ekman’s Big Six Model (Ekman, 1992) or Plutchik’s wheel of emotions (Plutchik, 1980) are used to automatically extract emotion knowledge from large corpora or Web documents. However, as pointed out in Gui et al. (2017), most natural language processing studies on emotion focus on emotion classification (Tokuhisa et al., 2008; Tokuhisa et al., 2009; Li et al., 2015; Zhou et al., 2016) and emotion information extraction (Vu et al., 2014). There are few studies on emotion cause extraction based on relationships between events and emotions caused by the events (Ghazi et al., 2015; Gui et al., 2016; Gui et al., 2017). These studies use explicit keywords in texts to recognize relationships between emotions and events.

There are many studies on automatic acquisition of relationships between events (Chambers and Jurafsky, 2009; Chambers and Jurafsky, 2010; Vanderwende, 2005; Shibata et al., 2014). Knowledge bases constructed in these studies are useful to determine what events happen after other events. However, these studies do not focus on motivations for events or effects caused by the events. As described in the previous section, this approach cannot guarantee quotidian connotational meanings from texts because of the reporting bias. Therefore, it is important to construct knowledge bases of such commonsense knowledge by manual intervention as well as through automatic acquisition.

3. Design of JFCKB

Our final goal is to construct a knowledge base of various feature changes of arguments in event sentences with the controlled granularity and to develop various methods for deep understanding of texts. To achieve the goal, we propose a knowledge base structure shown in Table 1. In this structure, each argument in an event sentence is associated with various feature changes caused by the event. For example, in the case of “my wife hits my child,” “my child” is associated with changes of some features such as *increase in pain*, *increase in anger*, *increase in disgust*, *decrease in joy*, and *decrease in trust*. We gathered such information through collective intelligence (i.e., crowdsourcing). Event sentences to be stored in our knowledge base are created

Sentence	Case (word)	Probability
My wife hits my child	<i>ga</i> (nominative) (wife)	<i>joy</i> (+, −, UNC) = (0, 1, 0) :
	<i>wo</i> (accusative) (child)	<i>anger</i> (+, −, UNC) = (1, 0, 0) :
	<i>ni</i> (dative) (NULL)	N/A
	<i>reader</i> (NULL)	<i>disgust</i> (+, −, UNC) = (0.99, 0, 0.01) :

Table 1: Example of the proposed knowledge base structure (JFCKB). This knowledge base is for Japanese. In JFCKB, each sentence has various feature changes for three cases (*ga* case, *wo* case, and *ni* case) and those of *readers*. The three cases are Japanese language specific syntactic roles. The *ga*, *wo*, and *ni* cases roughly correspond to nominative, accusative, dative, respectively. *Readers* are not arguments in the sentence. In the “Probability” column, symbols +, −, and UNC denote *increased*, *decreased*, and *unchanged*, respectively.

based on some knowledge bases automatically constructed from large-scale Web documents. Hereafter, we call our knowledge base *JFCKB* (Japanese Feature Change Knowledge Base). *JFCKB* is a Japanese knowledge base.

We pay attention to the controlled granularity for feature definition. Many studies suggest that it is important to focus on the types of changes caused by events to recognize the events. In infant cognitive development studies, there are several reports that even infants use information about the basic features of participants in events to understand the events (Massey and Gelman, 1988; Baillargeon et al., 1989; Spelke et al., 1995). In cognitive linguistics, *basic level category* is an extremely important concept (Rosch et al., 1976; Taylor, 1995). Basic level categories minimize attributes shared with other categories. Members belonging to a basic level category maximize attributes shared with other members of that same category. For example, “dog” is a basic level category while “Great Pyrenees” and “German shepherd” are not. Many cognitive linguists believe that our languages are built around basic level categories. Therefore, the granularity of knowledge is an important aspect of our language activity that needs to be captured.

Although there are many features altered by events, such features can be roughly classified into at least four categories: (1) physical features (such as size and temperature); (2) mental features (such as anger and joy); (3) sensory features (such as pain and sleepiness), and (4) relational features (such as position and social relation). According to these categories, we assumed 47 features shown in Table 3. These features represent the *basic level categories*

Verb: case frame ID	Case	Word
<i>yaku</i> : yaku1 (bake)	<i>ga</i> (nominative) <i>wo</i> (accusative) <i>de</i> (tools/ingredients)	<i>watashi</i> (I): 114, <i>haha</i> (mom): 75, <i>musume</i> (daughter): 74, ... <i>pan</i> (bread): 54076, <i>ke-ki</i> (cake): 31693, <i>niku</i> (meat): 14059, ... <i>koubo</i> (yeast): 888, <i>be-kari-</i> (bakery): 768, <i>o-bun</i> (oven): 515, ...
<i>yaku</i> : yaku2 (have difficulty)	<i>ga</i> (nominative) <i>wo</i> (accusative) <i>ni</i> (dative)	<i>mina</i> (all persons): 23, <i>sensei</i> (teacher): 11, <i>hito</i> (person): 8, ... <i>te</i> (hand): 26449 <i>kodomo</i> (child): 168, <i>musuko</i> (son):108, ...
<i>yaku</i> : yaku3 (burn)	<i>ga</i> (nominative) <i>ni</i> (dative)	<i>daitouryou</i> (president): 1, <i>shidousya</i> (mentor): 1, ... CD:13812, DVD:12200, ...

Table 2: Case frame examples. Each row denotes one case frame. In the “Word” column, each number denotes the frequency of the noun in the Web corpus.

as closely as possible. This decision was based on a traditional emotion study (Plutchik, 1980), Japanese thesauri (Ikehara, 1997; NINJAL, 2004), sentiment analysis studies (Tokuhisa et al., 2008; Tokuhisa et al., 2009), and features used in the VerbCorner project (Hartshorne et al., 2014). Although our final version of the proposed knowledge base will have all the features in Table 3, at present, emotional and sensory features in the table have been mainly investigated through crowdsourcing tasks described in the following section.

4. Construction of JFCKB

In our previous work (Nakamura and Kawahara, 2016), we constructed a trial version of a knowledge base in which each argument of an event sentence was associated with various feature changes caused by the events. As a result of a subjective evaluation experiment, it was shown that such feature changes can be appropriately acquired using crowdsourcing. In this trial version, 857 sentences (including 144 verbs (types) and 1,788 arguments (tokens) (742 types), 391 case frames (types)) were used.

In this study, we constructed a full version of the knowledge base and extended the dataset as follows: (1) for the representation manner of behaviors, we used only triples (*increased*, *decreased*, and *unchanged*) whose values are probabilities. In the trial version, some feature changes were represented by the triples while the other feature changes were represented by pairs (*changed* and *unchanged*) whose values are probabilities. (2) In addition to event sentences based on the 200 most frequent verbs in the *Kyoto University Web Document Leads Corpus* (KWDL) (Hangyo et al., 2012)¹ used in our previous work, we used sentences from other Japanese language resources. The resources are *Kyoto University Case Frames* (KUCF) (Kawahara and Kurohashi, 2006)² and the Japanese translated version of the WSC (JWSC) dataset.

KWDL is a Japanese text corpus that comprises 5,000 documents (15,000 sentences) with annotations of morphology, named entities, dependencies, predicate-argument structures including zero anaphora and coreferences. KUCF is a database of case frames automatically constructed from a corpus of 10 billion Japanese sentences taken from the Web. Case frames describe what kinds of nouns are related to each verb. Many Japanese verbs have

¹<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KWDL>

²<http://www.gsk.or.jp/en/catalog/gsk2008-b/>

Category	Sub category	Feature
physical	form	length, size , width, thickness (around) thickness (depth)
	color	redness, orangeness yellowness, greenness blueness, purpleness brownness, whiteness blackness, brightness
	touch	temperature , rigidity roughness, stickiness
	smell	goodness, badness
	sound	silence
	taste	sweetness, sourness bitterness, astringency hotness (not temperature)
	density	denseness
	amount	quantity
mental	emotion	joy, trust, surprise disgust, fear, sadness anger, anticipation
	evaluation	polarity
sensory	sensory	pain, sleepiness tiredness
relation	relation	interaction, possession physical contact physical force existence social relationship
	position	closeness

Table 3: Features assumed in this study. These features were decided by considering various studies such as traditional psychological studies, studies on cognitive development of infants, Japanese thesauri, sentiment analysis studies, and VerbCorner project. Features with bold fonts have been investigated so far.

some meanings. Examples are shown in Table 2. In KUCF, each case frame is composed of case frame ID, verb, cases, nouns filled in the cases, frequencies of the nouns in the Web corpus. KUCF has about 110,000 predicates with 5.4 case frames on average for each predicate. WSC (JWSC) dataset is basically composed of two sentences including one anaphor, two antecedent candidates, and a correct antecedent.

The construction of JFCKB was achieved using crowdsourcing. The details are described in the next subsections.

4.1. Phase 1: Event Sentence Creation

Event sentences in JFCKB are representative sentences of case frames of verbs in KUCF. The sentence creation procedure was as follows. Step 1: Selection of verbs. We used the 200 most frequent verbs in KWDLC, the 1,000 most frequent verbs in KUCF, and all verbs in the JWSC dataset. Step 2: Representative argument selection. For each case (syntactic role) of each case frame of each verb, the most frequent argument was used as the representative argument of the case. In the case of KWDLC, the two most frequent arguments were used. We used *ga* case, *wo* case, and *ni* case as arguments. These cases are Japanese grammatical syntactic roles and roughly correspond to nominative, accusative, and dative, respectively. Step 3: Representative sentence creation. Sentences were created by combining the verbs with the three types of representative arguments. Step 4: Incomprehensible sentence exclusion. Sentences difficult to understand were pruned based on a crowdsourcing task. The crowdsourcing workers were asked whether they could understand the sentences presented. The comprehensibility of each sentence was judged by ten workers. In total, 1,559 people participated in the task.

After the crowdsourcing task, we estimated the probability that each sentence was judged *comprehensible* by the participants based on the aggregation method proposed by Whitehill et al. (2009). Sentences with probabilities of *comprehensible* less than 0.9 were discarded. Unlike majority voting, this method calculates the probability based on worker agreements. As a result, 9,073 sentences remained out of 11,189 representative sentences. These 9,073 sentences included 19,052 arguments (tokens) (4,882 types).

4.2. Phase 2: Animacy Investigation

When a word does not denote a concrete object, it is not necessary to ask about feature changes of the word (e.g., “discussion”). Even if the word denotes a concrete object, it is not necessary to investigate feature changes of emotions and senses of the word when the word denotes an object without emotions and senses (e.g., “stone”). To avoid such redundant asking in the feature change investigation task, we conducted a crowdsourcing task to investigate the animacy of words (arguments) in event sentences in advance.

In this task, crowdsourcing workers were asked whether presented words denote concrete objects or may have emotions or senses. That is, we divided this phase into two sub phases. The former (sub phase 1) is a task to investigate the concreteness of arguments. The other (sub phase 2) is a task to investigate whether arguments have emotions. In total, 1,011 participants completed these two sub tasks. The method for calculating the probability of answers was the same as in the event sentence creation phase. We discarded words whose probabilities of *concrete object* or *object with emotion* were less than 0.5. As a result, as far as we investigated, 1,575 types of words were judged as those representing concrete objects out of 3,457 types of words. 677 types of words were judged as those with emotions out of

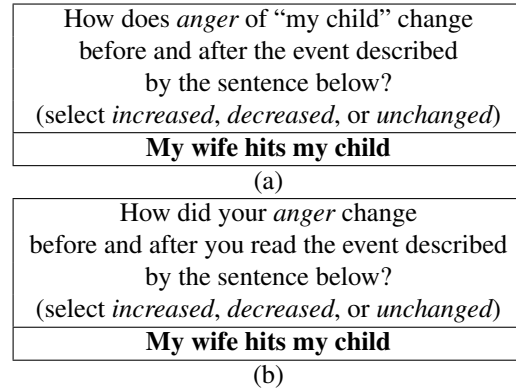


Figure 1: Questions used for crowdsourcing tasks. These tasks are to acquire concrete feature changes caused by events described by the sentences.

3,457 types of words. Note that the number of words whose animacies were investigated (3,457 types) is different from the number of words in JFCKB (4,882 types) because this phase was applied from the middle of constructing JFCKB.

4.3. Phase 3: Feature Change Investigation

According to results of the previous two phases (i.e., event sentence creation and animacy investigation), we conducted a crowdsourcing task to gather feature changes of arguments in event sentences. We also attempted to gather those of sentence readers. That is, crowdsourcing workers were asked to answer one of feature changes of presented arguments in sentences (Figure 1 (a)) or those of the workers themselves (Figure 1 (b)). In total, 33,683 people participated in this task. Probability calculations were carried out in the same way as the previous phases. The resulting knowledge base includes 5,647 case frames (types). These case frames are for 975 verbs (types). Note that one verb has one or more case frames because of the verbal polysemy as shown in Table 2.

5. Validating Usefulness of JFCKB: Anaphora Resolution

5.1. Evaluation Settings

To validate the usefulness of JFCKB, we conducted an experiment of anaphora resolution using JFCKB. We adopted JWSC as an anaphora resolution problem. Each problem in JWSC is basically composed of two sentences including one anaphor, two antecedent candidates, and a correct antecedent. This anaphora resolution task can be regarded as an antecedent candidate ranking task because it is a task to select the correct antecedent from two candidates. We used Ranking SVMs³ (Joachims, 2002) as the solver and used two-degree polynomial kernel as the kernel function. We set the cost parameter (trade-off between training error and margin) to 0.01.

In this experiment, three conditions were compared: (1) the case using only feature change information (hereafter, *FC*); (2) the case using only word embeddings of word2vec (hereafter, *W2V*), and (3) the case using both feature change information and word embeddings (hereafter, *BOTH*).

³https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html

S1 = The bee landed on the flower. S2 = Because it wanted pollen.
 antecedent 1 = bee, antecedent 2 = flower, anaphora = it, correct antecedent = bee
 diff1 = it – bee, diff2 = it – flower

$$\begin{aligned}
 1st = & \begin{bmatrix} \text{bee} & \text{flower} & \text{it} \\ \text{oo..o} & \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{land} & \text{land} & \text{want} \\ \text{oo..o} & \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{S1} & \text{S2} \\ \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{diff1} & \text{diff2} \\ \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{bee} \\ \text{oo..o} \end{bmatrix} \\
 2nd = & \begin{bmatrix} \text{bee} & \text{flower} & \text{it} \\ \text{oo..o} & \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{land} & \text{land} & \text{want} \\ \text{oo..o} & \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{s1} & \text{s2} \\ \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{diff1} & \text{diff2} \\ \text{oo..o} & \text{oo..o} \end{bmatrix} + \begin{bmatrix} \text{flower} \\ \text{oo..o} \end{bmatrix}
 \end{aligned}$$

Figure 2: Example of anaphora resolution problem and vectors used in the anaphora resolution task. S1 and S2 denote the first sentence and the second sentence in the problem. Diff denotes the difference between an anaphor and an antecedent candidate. The vectors given to Ranking SVMs were composed of vectors expressing anaphors and antecedent candidates, vectors expressing the nearest (in parse trees) predicates of anaphoras and antecedent candidates, vectors expressing the first and second sentences, vectors expressing differences between anaphoras and antecedent candidates, and vectors expressing either of two antecedent candidates. As for the order in Ranking SVMs, the vector with the last component expressing the correct antecedent was regarded the first one.

The feature change vector of each argument is a 33 dimensional vector because we used eight emotion features (joy, trust, surprise, disgust, fear, sadness, anger, and anticipation) and three sensory feaues (pain, sleepiness, and tiredness). Each feature change is composed of three probabilities (increased, decreased, and unchanged). The feature change vector of each predicate is a 99 dimensional vector because each predicate vector is composed of vectors for three cases (*ga* case (nominative), *wo* case (accusative), and *ni* case (dative)). Vectors in word2vec are 500 dimensional vectors learned from 100 million Web sentences analyzed by JUMAN++⁴.

Vectors composed of the following feature vectors were given to Ranking SVM in all three conditions. (1) vectors expressing anaphors and antecedent candidates, (2) vectors expressing the nearest (in parse trees) predicates of anaphoras and antecedent candidates, (3) vectors expressing the first and second sentences, (4) vectors expressing differences between anaphoras and antecedent candidates, and (5) vectors expressing either of two antecedent candidates. We used KNP⁵ for dependency parsing.

To solve JWSC problems, we created each vector of antecedent candidates based on five types of vectors described above and compared them. For each antecedent candidate, the five types of vectors were concatenated. Therefore, according to the dimension numbers of feature change vectors and word2vec vectors described above, vectors in *FC*, *W2V*, and *BOTH* are 693 dimensional, 5,500 dimensional, and 6,193 dimensional vectors respectively. As for the order in Ranking SVM, the vector with the last component expressing the correct antecedent was regarded the first one. For example, when the first sentence “the bee landed on the flower,” the second sentence “Because it wanted pollen,” two antecedent candidates “bee, flower,” the anaphor “it,” and the correct antecedent “bee” were given, the vector in which the last component represents “bee” is the first one (Figure 2).

We conducted ten-fold cross validation to evaluate system performance. Although JWSC is composed of 1,321 problems, we used 441 problems in this evaluation. These prob-

	FC	W2V	BOTH
Accuracy	52.15%	47.39%	47.62%

Table 4: Anaphora resolution result.

lems satisfied the following conditions: (1) each was composed of two sentences; (2) all the predicates of anaphors and those of antecedent candidates have information about their feature changes and word embeddings.

5.2. Results

The result of the cross validation is shown in Table 4. Table 4 shows that the accuracy of *FC* outperformed those of the others. An example that feature changes worked well was the case “James always gives orders to Owen. Because he is very bossy.” Figure 3 (a) shows feature changes of this example. In this case, *FC* and *BOTH* estimated the correct answer while *W2V* failed. An example that feature changes did not work was the case “Bill punched Larry. And he was hurt.” Figure 3 (b) shows feature changes of this example. In this case, *FC* and *BOTH* failed to estimate the correct answer while *W2V* succeeded. Note that Figure 3 shows representative feature change values. The representative value of each feature is the weighted average of the feature, where weights of *increased*, *decreased*, and *unchanged* are +1, -1, and 0, respectively. For example, when the feature change probabilities are 0.8, 0.15, and 0.05 for *increased*, *decreased*, and *unchanged*, respectively, the representative value of the feature is 0.65.

In the case that feature changes worked well, as can be seen in Figure 3 (a), the correct antecedent has some feature changes in common with the anaphors but the incorrect antecedent has a conflicting feature change (e.g., anticipation). In the case that feature changes did not work, as can be seen in Figure 3 (b), there are no conflicting feature changes. Considering these results, we speculate that such common feature changes and conflicting feature changes influence the estimation of the correct antecedent candidates in anaphora resolution.

6. Conclusion

In this study, we constructed a knowledge base where arguments in event sentences are associated with various feature

⁴<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN++>

⁵<http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP>

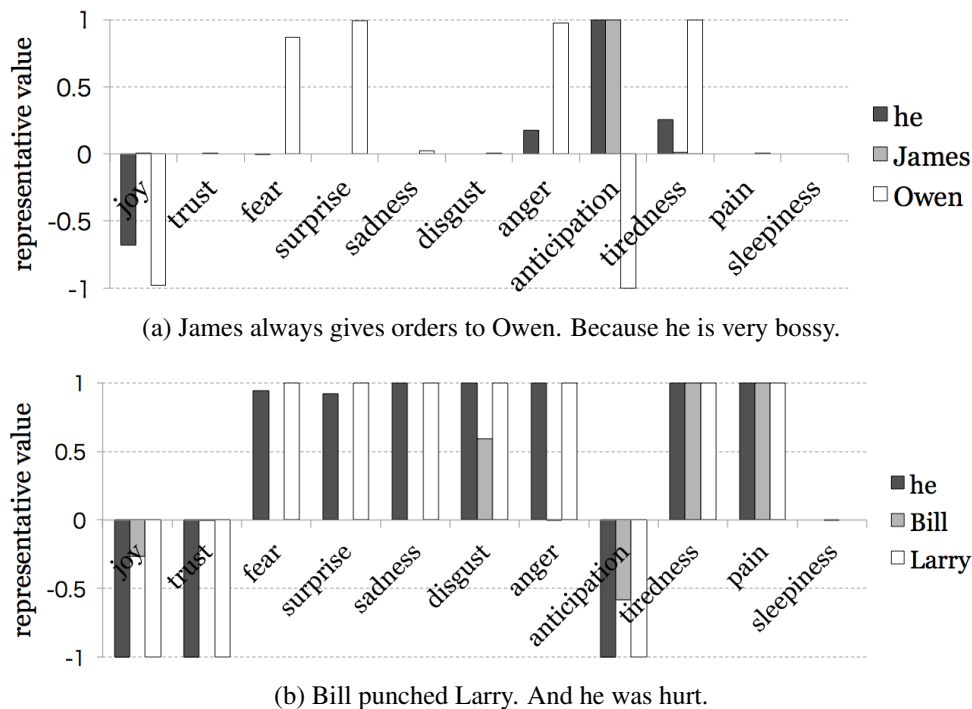


Figure 3: Examples of feature changes. (a) is an example where feature changes worked well. (b) is an example where feature changes did not work. The value of each feature is the representative feature change value. The representative value of each feature is the weighted average of the feature, where weights of *increased*, *decreased*, and *unchanged* are +1, -1, and 0, respectively.

changes caused by the events. Features used in the proposed knowledge base are determined by considering various studies such as traditional psychological studies, studies on cognitive development of infants, sentiment analysis studies. This knowledge base was constructed through crowdsourcing tasks. After the construction of our feature change knowledge base, we conducted an experiment of anaphora resolution using the knowledge base. In this experiment, we regarded anaphora resolution as an antecedent candidate ranking task and compare three conditions (the case using only the feature change information, the case using only the word2vec information, and the case using both of the feature change information and the word2vec information). As a result, it was shown that the condition where only the feature change information was used outperformed the other conditions. This result suggests the usefulness of our feature change knowledge base.

7. Acknowledgements

This work was supported by JST PRESTO Grant Number JPMJPR1402, Japan.

8. Bibliographical References

- Baillargeon, R., Devos, J., and Graber, M. (1989). Location memory in 8-month-old infants in a non-search ab task: Further evidence. *Cognitive Development*, 4(4):345–367.
- Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of ACL-COLING 1998*, pages 86–90.
- Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. 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*, pages 602–610.
- Chambers, N. and Jurafsky, D. (2010). A database of narrative schemas. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC2010)*.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6:169–200.
- Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. The MIT Press.
- Ghazi, D., Inkpen, D., and Szpakowicz, S., (2015). *Detecting Emotion Stimuli in Emotion-Bearing Sentences*, pages 152–165. Springer.
- Gordon, J. and Durme, B. V. (2013). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, pages 25–30.
- Gordon, J., Durme, B. V., and Schubert, L. (2010). Learning from the web: Extracting general world knowledge from noisy text. In *Proceedings of the AAI 2010 Workshop on Collaboratively-built Knowledge Sources and Artificial Intelligence (WikiAI 2010)*, pages 8–10.
- Gui, L., Wu, D., Xu, R., Lu, Q., and Zhou, Y. (2016). Event-driven emotion cause extraction with corpus construction. In *Proceedings of EMNLP 2016*, pages 1639–1649.
- Gui, L., Hu, J., He, Y., Xu, R., Lu, Q., and Du, J. (2017). A question answering approach to emotion cause extraction. In *arXiv:1708.05482v1*.
- Hangyo, M., Kawahara, D., and Kurohashi, S. (2012).

- Building a diverse document leads corpus annotated with semantic relations. In *Proceedings of the 26th Pacific Asia Conference on Language Information and Computing*, pages 535–544.
- Hartshorne, J. K., Bonial, C., and Palmer, M. (2014). The verbcorner project: Findings from phase 1 of crowdsourcing a semantic decomposition of verbs. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 397–402.
- Hasegawa, T., Kaji, N., Yoshinaga, N., and Toyoda, M. (2013). Predicting and eliciting addressee’s emotion in online dialogue. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 964–972.
- Ikehara, S. (1997). *Nihongo Goi Taikai*. Iwanami Syoten, Tokyo.
- Jackendoff, R. (1983). *Semantics and Cognition*. MIT Press.
- Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2002*.
- Kawahara, D. and Kurohashi, S. (2006). Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 1344–1347.
- Levesque, H. J. (2011). The Winograd Schema Challenge. In *Proceedings of AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.
- Li, S., Huang, L., Wang, R., and Zhou, G. (2015). Sentence-level emotion classification with label and context dependence. In *Proceedings of ACL 2015*, pages 1045–1053.
- Massey, C. M. and Gelman, R. (1988). Preschoolers’ ability to decide whether a photographed unfamiliar object can move itself. *Developmental Psychology*, pages 307–317.
- Nakamura, T. and Kawahara, D. (2016). Constructing a dictionary describing feature changes of arguments in event sentences. In *Proceedings of the 4th Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 46–50.
- NINJAL. (2004). *Bunrui Goihyo*. Dainippon Tosho, Tokyo.
- Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
- Plutchik, R., (1980). *A General Psychoevolutionary Theory of Emotion, 1*, pages 3–33. Academic Press.
- Rashkin, H., Singh, S., and Choi, Y. (2016). Connotation frames: A data-driven investigation. In *Proceedings of ACL 2016*, pages 264–269.
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.
- Shibata, T., Kohama, S., and Kurohashi, S. (2014). A large scale database of strongly-related events in Japanese. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014)*, pages 3283–3288.
- Shibata, T., Kohama, S., and Kurohashi, S. (2015). Nihongo Winograd Schema Challenge no kouchiku to bunseki (in Japanese). In *Proceedings of NLP2015*, pages 493–496.
- Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2012). Open mind common sense: Knowledge acquisition from the general public. In *Proceedings of On the Move to Meaningful Internet Systems, 2002*, pages 1223–1237.
- Speer, R. and Havasi, C. (2012). Representing general relational knowledge in conceptnet 5. In *Proceedings of LREC 2012*, pages 86–90.
- Spelke, E. S., Phillips, A., and Woodward, A. L., (1995). *Infants’ knowledge of object motion and human action*. Oxford University Press.
- Strapparava, C. and Valitutti, A. (2004). Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1083–1086.
- Taylor, J. R. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory*. Clarendon Press.
- Tokuhisa, R., Inui, K., and Matsumoto, Y. (2008). Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 881–888.
- Tokuhisa, R., Inui, K., and Matsumoto, Y. (2009). Emotion classification using massive examples extracted from the web. *Journal of Information Processing (in Japanese)*, 50(4):1365–1374.
- Vanderwende, L. (2005). Volunteers created the web. In *Proceedings of the 2005 AAAI Spring Symposium, Knowledge Collection from Volunteer Contributors*.
- Vu, H. T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2014). Acquiring a dictionary of emotion-provoking events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 128–132.
- Whitehill, J., Ruvolo, P. L., Ting-fan Wu, J. B., and Movellan, J. R. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, et al., editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc.
- Zhou, D., Zhou, Y., Zhang, X., Zhao, Q., and Geng, X. (2016). Emotion distribution learning from texts. In *Proceedings of EMNLP 2016*, pages 638–647.