Constructing a Dictionary Describing Feature Changes of Arguments in Event Sentences

Tetsuaki Nakamura †
tnakamura@nlp.ist.i.kyoto-u.ac.jp

Daisuke Kawahara †‡
dk@i.kyoto-u.ac.jp

† Graduate School of Informatics,
Kyoto University

‡ JST PRESTO

Abstract

Common sense knowledge plays an essential role for natural language understanding, human-machine communication and so forth. In this paper, we acquire knowledge of events as common sense knowledge because there is a possibility that dictionaries of such knowledge are useful for recognition of implication relations in texts, inference of human activities and their planning, and so on. As for event knowledge, we focus on feature changes of arguments (hereafter, FCAs) in event sentences as knowledge of events. To construct a dictionary of FCAs, we propose a framework for acquiring such knowledge based on both of the automatic approach and the collective intelligence approach to exploit merits of both approaches. We acquired FCAs in event sentences through crowdsourcing and conducted the subjective evaluation to validate whether the FCAs are adequately acquired. As a result of the evaluation, it was shown that we were able to reasonably well capture FCAs in event sentences.

1 Introduction

Common sense knowledge plays an essential role for natural language understanding, human-machine communication and so forth. There are two approaches to acquire such 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. This approach is useful when the amount of data to be acquired is extremely large. However, the quality of acquired knowledge might be of low quality. The manual acquisition approach may use a fully manual technique (at very early stage of artificial intelligence studies) or collective intelligence (e.g., crowdsourcing and games with a purpose). This approach is useful for gathering subjective information, such as emotion information, which is difficult to acquire automatically.

In this paper, we acquire knowledge of events as common sense knowledge because there is a possibility that dictionaries of such knowledge are useful for recognition of implication relations in texts, inference of human activities and their planning, and so on. As for event knowledge, we focus on feature changes of arguments (hereafter, FCAs) in event sentences as knowledge of events because there are several studies of infant cognitive development (Massey and Gelman, 1988; Baillargeon et al., 1989; Spelke et al., 1995) that report that even infants use information about the basic features of participants in events to understand the events. There are some trials suggesting the possibility that dictionaries of FCAs are useful resources for deep understanding of texts (Rahman and Ng, 2012; Goyal et al., 2013). In (Rahman and Ng, 2012), features of relationships between event causalities and polarities (positive / negative) of participants in the events are used as features for anaphora resolution. In (Goyal et al., 2013), a dictionary of various binary effects (success(+/-), motivation(+/-), and so on) on characters caused by events (that is, verbs) are used for automatic story generation.

To construct a dictionary of FCAs, we propose a framework for acquiring such knowledge based on

both an automatic approach and a manual approach (collective intelligence) that exploits merits of both approaches. That is, we acquire seed knowledge by using collective intelligence and expand the knowledge automatically.

2 Related Work

Many studies have aimed at automatically acquiring relationships between events (Chambers and Jurafsky, 2009; Chambers and Jurafsky, 2010; Vanderwende, 2005; Shibata et al., 2014). However, these studies do not focus on the motivation of events or effects caused by the events. Although we can determine which events occur after an event by using the dictionaries constructed in such studies, we cannot understand why the events occur. For example, although we may know the event "a girl cries" happens after another event "a girl gets injured" by using the dictionary, we cannot understand why the former event occurs. To understand the motivation of the former event, we have to know that "the girl feels pain." Of course, the problem can be solved by treating the phenomenon "the girl feels pain" as an event and describing it in the dictionary. However, such a policy requires infinite descriptions. It is preferable to form and maintain the dictionary with an controlled granularity.

In the case that participants in events are animated beings, events may influence their emotions. Many studies developing software to process human emotions exploits models proposed by psychologists (Hasegawa et al., 2013; Tokuhisa et al., 2008; Tokuhisa et al., 2009; Vu et al., 2014). In these studies, Ekman's Big Six Model (Ekman, 1992) and Plutchik's wheel of emotions (Plutchik, 1980) are used to automatically extract emotion knowledge from large corpora. However, since these studies do not focus on what happens after evoking various changes of emotions, they are not able to understand complex phenomenon involving such changes.

Some dictionaries constructed manually that retain high quality are available. For example, ConceptNet[1] built in the Open Mind Common Sense project (Singh, 2002) and a large-scale knowledge database built in the CYC project (Lenat, 1995) are available. The database built in the VerbCorner project (Hartshorne et al., 2014), expanding of VerbNet (Kipper et al., 2000) through crowdsourcing, is also available. However, these studies do not focus on the granularity of knowledge.

[1] http://conceptnet5.media.mit.edu/

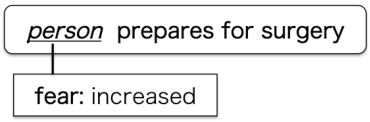

Figure 1: Example entry in the proposed dictionary. Event sentences are associated with FCAs.

For the controlled granularity, we focus on the use of basic level features of arguments in event sentences. Since both of animate beings and inanimate objects can be arguments in sentences, we assume not only physical features but also mental features. The decision criteria for these features are described in Section 3.2.

3 Proposal

3.1 Architecture

To construct a knowledge database that keeps an controlled granularity and can be used to understand the motivation of events, we focus on the feature changes of arguments (FCAs) in event sentences. Our purpose is to construct a dictionary that has the architecture shown in Figure 1. As shown in the figure, the dictionary describes relationships between events and FCAs in the event sentences.

3.2 Features

As described in Section 2, we focus on the use of basic level features of arguments in event sentences. We assume 16 features listed in Table 1 to be associated with arguments in event sentences. As for physical basic level features, we regard eight features listed in the Table as basic level features because we pay attention to the traditional sense categories ("sight", "hearing", "smell", "taste", and "touch"). As for mental basic level features, we use basic emotions (e.g., eight mental features listed in the Table) of Plutchik's model (Plutchik, 1980) because the model assumes other emotions derived from the basic emotions systematically.

Category	Features	Degrees
physical	form, color, touch, smell, sound, taste, location, temperature	changed, unchanged, irrelevant
mental	joy, fear, trust, surprise, anger, disgust, sadness, expectation	increased, decreased, unchanged, irrelevant

Table 1: Features of arguments.

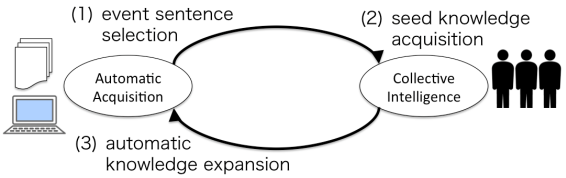

Figure 2: Expansion of our dictionary. We used automatic procedures and crowdsourcing to expand the size of the dictionary.

3.3 Construction Method

We are planning to construct the dictionary described in section 3.1 based on both of automatic procedure and collective intelligence as follows: (PHASE 1) Frequent event sentences are selected automatically. (PHASE 2) FCAs of the selected sentences are acquired as seed knowledge based on crowdsourcing because our dictionary is designed to deal with subjective information such as emotions. (PHASE 3) Knowledge is automatically expanded by using the seed knowledge.

The sequence of precedures described above is illustrated in Figure 2. We intend to expand the size of our dictionary based on the loop.

4 Experiment

We conducted an experiment to validate whether FCAs in event sentences are adequately acquired by using crowdsourcing.

4.1 Data

We created event sentences presented to crowdsourcing workers based on the "Kyoto University Web Document Leads Corpus (KWDLC) (Hangyo et al., 2012)"[2] and the "Kyoto University Case Frames (KUCF) (Kawahara and Kurohashi, 2006)."[3] The KWDLC 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. The KUCF is a database of case frames automatically constructed from 1.6 billion Japanese sentences taken from Web pages. The KUCF has about 40,000 predicates, with 13 case frames on average for each predicate.

The sentence creation procedure is as follows. (STEP 1) The 200 most frequent verbs were extracted from the KWDLC. (STEP 2) Verbs that are very abstract (e.g., "do") were omitted. As a result, 167 verbs remained. (STEP 3) The top two frequent arguments were used as the representative arguments of each case frame (meaning) of each verb. (STEP 4) Sentences were created by combining the verbs with their representative arguments. As a result, 1,935 Japanese sentences were created. (STEP 5) Sentences that were difficult to understand were pruned based on crowdsourcing[4]. The crowdsourcing workers were asked to answer whether they could understand the presented sentences. Each sentence was judged by 10 workers. In total, 244 people participated in the task. For each sentence, we estimated the probability that the "yes" was selected by using the method proposed in (Whitehill et al., 2009) and omitted sentences with probabilities less than 0.9. As a result, 857 event sentences remained and 148 verbs (types) were included in the sentences. Since each of these sentences had two or three arguments, 1,768 arguments (token) were obtained as a result.

4.2 Method

FCAs in event sentences were acquired through a crowdsourcing task. In the task, workers were asked to answer questions about feature changes of arguments in the presented event sentences as shown in Figure 3. Every worker was given a set which includes an event sentence and five questions about FCAs in the sentence. The argument was randomly selected from the 1,768 arguments described above. Although five features to be asked were selected from the 16 features in Table 1 for each task, FCAs of all the features for each argument were acquired in total. 2,910 people (token) were participated in this experiment. The each worker could be engaged in a maximum of five tasks and each feature was judged by 10 workers.

[2] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?DDLC
[3] http://www.gsk.or.jp/en/catalog/gsk2008-b/
[4] http://crowdsourcing.yahoo.co.jp/

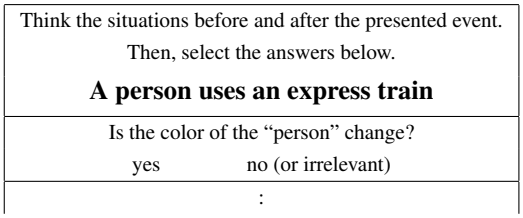

Figure 3: Layout examples of the task to acquire FCAs (physical features). Each task is composed of an event sentence and five questions about FCAs in the sentence. In the case of mental features, four options (increased, decreased, unchanged, and irrelevant) are used.

We used the method proposed by Whitehill et al. (2009) to estimate the probability of each option for each feature. These probabilities are estimated jointly incorporating the difficulty of each question and the answering skill of each worker according to the agreements of the collected judges. Therefore, the estimated probabilities are more reliable than the results of a simple counting method such as majority voting. The probabilities were used as elements in vectors expressing the FCAs (hereafter, FCA vectors). Moreover, we assigned ten workers to each question to better ensure reliability, since it is reported that annotation results provided by more than six non-expert workers are close to those provided by expert labelers (Snow et al., 2008).

4.3 Results

For example, in the case of "a child" as in "a child gets a fever," the estimated probabilities that the labels "increased" of "temperature" and "fear" were selected were 0.99. To validate whether FCAs in event sentences were adequately acquired by using crowdsourcing, we conducted a subjective evaluation as follows: (STEP 1) Ten arguments with salient elements of FCA vectors were selected (hereafter, query arguments: QAs) from 1,768 arguments. (STEP 2) For each QA, a list of arguments with impressions similar to the QA was obtained (candidate arguments: CAs). The lists were composed of the five most similar arguments. In this paper, "arguments with similar impressions" means arguments which have similar FCA vectors. We used cosine similarities between FCA vectors as similarities. (STEP 3) For each list, five judges (people engaged in studies of natural language processing) were asked to answer whether each CA was similar to the corresponding QA (similar = 2 points, a little similar = 1 point, different = 0 point).

As a result, the average point of CAs was 1.24. (For QAs with salient physical FCAs, the average point was 1.08. For QAs with salient mental FCAs, the average point was 1.40) This result suggests that FCAs were adequately acquired through crowdsourcing. Especially, it was shown that the result of QAs with salient mental FCAs was better than that of QAs with salient physical FCAs. We speculate that the reason was due to the granularities of values (that is, "whether features are changed (three options: changed, unchanged, irrelevant)" v.s. "how features are changed (four options: increased, decreased, unchanged, irrelevant)").

5 Conclusion

In this paper, we focused on feature changes of arguments (FCAs) in event sentences as knowledge of events and acquire the FCAs as common sense knowledge through crowdsourcing. As a result of the subjective evaluation to validate whether the FCAs are adequately acquired by using crowdsourcing, it was shown that we were able to reasonably well capture the FCAs.

In the next step, we must consider the automatic method that expand the size of our dictionary by using seed knowledge as described in Section 3.3. For the expansion, we are planning to analyze knowledge acquired in this study because we anticipate that there are FCAs to be shared and those not to be shared. That is, we think that there are three types of FCAs at least: (1) FCAs mainly depend on verbs themselves (e.g., disgust that "girl" feels in "A boy complains to a girl"), (2) FCAs mainly depend on combinations of arguments and verbs (e.g., color of "water" in "He pour his orange juice into the water"), and (3) FCAs mainly depend on contexts (e.g., joy that "girl" feels in "A girl see a boy"; the girl's emotion depends on whether she likes him).

For the future work, we are also planning to construct a large scale network in which structures described in Figure 1 are regarded as units and the units associated with related units are linked each other.

Note that the patent for the architecture described in this paper is pending at the time of writing this paper.

References

Renee Baillargeon, Julia Devos, and Marcia Graber. 1989. Location memory in 8-month-old infants in a non-search ab task: Further evidence. *Cognitive Development*, 4(4):345–367.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *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.

Nathanael Chambers and Dan Jurafsky. 2010. A database of narrative schemas. In *the Seventh International Conference on Language Resources and Evaluation (LREC2010)*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6:169–200.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2013. A computational model for plot units. *Computational Intelligence*, 29(3):466–488.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2012. Building a diverse document leads corpus annotated with semantic relations. In *the 26th Pacific Asia Conference on Language Information and Computing*, pages 535–544.

Joshua K. Hartshorne, Claire Bonial, and Martha Palmer. 2014. The verbcorner project: Findings from phase 1 of crowd-sourcing a semantic decomposition of verbs. In *the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 397–402.

Takayuki Hasegawa, Nobuhiro Kaji, Naoki Yoshinaga, and Masashi Toyoda. 2013. Predicting and eliciting addressee's emotion in online dialogue. In *the 51st Annual Meeting of the Association for Computational Linguistics*, pages 964–972.

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 1344–1347.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *the 7th National Conference of Artificial Intelligence*, pages 691–696.

Douglas B. Lenat. 1995. Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38.

Christine M. Massey and Rochel Gelman. 1988. Preschoolers' ability to decide whether a photographed unfamiliar object can move itself. *Developmental Psychology*, pages 307–317.

Robert Plutchik, 1980. *A General Psychoevolutionary Theory of Emotion, 1*, pages 3–33. Academic Press.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789.

Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. 2014. A large scale database of strongly-related events in japanese. In *the 9th International Conference on Language Resources and Evaluation (LREC2014)*, pages 3283–3288.

Push Singh. 2002. The open mind common sense project.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Elizabeth S. Spelke, Ann Phillips, and Amanda L. Woodward, 1995. *Infants' knowledge of object motion and human action*. Oxford University Press.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2008. Emotion classification using massive examples extracted from the web. In *the 22nd International Conference on Computational Linguistics*, pages 881–888.

Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. 2009. Emotion classification using massive examples extracted from the web. *Journal of Information Processing (in Japanese)*, 50(4):1365–1374.

Lucy Vanderwende. 2005. Volunteers created the web. In *Proceedings of the 2005 AAAI Spring Symposium, Knowledge Collection from Volunteer Contributors*.

Hoa Trong Vu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Acquiring a dictionary of emotion-provoking events. In *the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 128–132.

Jacob Whitehill, Paul L. Ruvolo, Jacob Bergsma Ting-fan Wu, and Javier R. Movellan, 2009. *Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise*, volume 22, pages 2035–2043. Curran Associates, Inc.