Improving Crowdsourcing-Based Annotation of Japanese Discourse Relations

Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{kishimoto, sawada, murawaki, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Although discourse parsing is an important and fundamental task in natural language processing, few languages have corpora annotated with discourse relations and if any, they are small in size. Creating a new corpus of discourse relations by hand is costly and time-consuming. To cope with this problem, Kawahara et al. (2014) constructed a Japanese corpus with discourse annotations through crowdsourcing. However, they did not evaluate the quality of the annotation. In this paper, we evaluate the quality of the annotation using expert annotations. We find out that crowdsourcing-based annotation still leaves much room for improvement. Based on the error analysis, we propose improvement techniques based on language tests. We re-annotated the corpus with discourse annotations using the improvement techniques, and achieved approximately 3% improvement in F-measure. We will make re-annotated data publicly available.

Keywords: discourse annotation, crowdsourcing

1. Introduction

Humans understand text not by interpreting clauses or sentences individually, but by linking such a text fragment with another in a particular context. To allow computers to understand text, it is essential to capture the precise relations between these text fragments. The task of analyzing these relations is called discourse parsing. Discourse relations are conventionally divided into two types: explicit and implicit. Explicit relations are overtly marked with discourse connectives such as "and" and "however." By contrast, implicit relations lack discourse connectives.

Discourse parsing is an important and fundamental task in natural language processing. Systems for discourse parsing are, however, available only for major languages, such as English and Chinese. This is because few languages have corpora annotated with discourse relations and if any, they are small in size. Moreover, creating a new corpus of discourse relations by hand is costly and time-consuming.

Kawahara et al. (2014) addressed this problem by using crowdsourcing. Crowdsourcing allows for cheap and speedy annotation. However, crowdsourcing-based annotation tends to be of poorer quality than expert annotation.

In this paper, we evaluate the quality of the annotations produced by Kawahara et al. (2014). By asking experts to annotate a part of the corpus, we evaluate the quality of the annotation. Next, we analyze the corpus and find two problems. Then we propose solutions to each of these problems. Finally, we re-annotate the corpus with discourse relations through crowdsourcing. Experimental results show that the accuracy is improved by an F-measure of 3%.

2. Related Work

There are several corpora with discourse annotations for English, such as the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) and the RST Discourse Treebank (RST-DT) (Carlson et al., 2001). PDTB's and RST-DT's annotations are done as another layer on the Wall Street Journal section of the Penn Treebank. The PDTB consists 2,159 articles and each discourse relation consists of two text spans (arguments) and a relation label. The RST-DT consists 385 articles and discourse relations can be represented as a tree structure. Discourse corpora for Chinese (Zhou and Xue, 2012) and Turkish (Zeyrek et al., 2013) have also been developed based on the PDTB. In Japanese, Kaneko and Bekki (2014) built a Japanese corpus with temporal and causal relations using the Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003). They annotated only 66 sentences.

In recent years, various language resources were created through crowdsourcing. Snow et al. (2008) ran five crowdsourcing tasks including word similarity and RTE. Guillaume et al. (2016) produced a French corpus with dependency syntax annotation by using gamification. Kawahara et al. (2014) produced a Japanese corpus with discourse annotations, which we will review in section 3.

3. Annotating Discourse Relations Using Crowdsourcing

3.1. Corpus Specifications

We overview the corpus with discourse annotations produced by Kawahara et al. (2014). The target documents are web pages extracted from the Kyoto University Web Leads Corpus (Hangyo et al., 2012). Each document consists of the first three sentences of a Japanese web page. The web pages cover a variety of domains and the first three sentences are long enough to annotate with discourse relations through crowdsourcing.

They adopted a clause as the discourse unit. The clause is a span delimited by relatively strong boundaries in a sentence. They are automatically identified with hand-written rules by the KNP parser (Kurohashi and Nagao, 1994). Kawahara et al. (2014) annotated all possible combinations of clauses with discourse annotations.

Table 1 shows the discourse relation tagset. This tagset consists of two layers, where the upper layer contains three classes and the lower layer contains seven classes.

Upper type	Lower type	Example
CONTINGENCY	Cause/Reason	【ボタンを押したので】【お湯が出た】 [Since (I) pushed the button] [hot water came out]
	Purpose	【試験に受かるために】【必死に勉強した】 [To pass the exam] [(I) desperately studied]
	Condition	【ボタンを押せば】【お湯が出る。】 [If (you) push the button] [hot water will be turned on]
	Ground	【ここにカバンがあるから】【まだ社内にいるだろう。】 [Here is (his) bag] [(he) would be still in the company]
COMPARISON	Contrast	【京都は雨だが、】【宮崎は晴れだ。】 [It is raining in Kyoto] [however it is sunny in Miyazaki]
	Concession	【あのレストランは確かにおいしいが】【値段は高い。】 [That restaurant is surely good] [but the price is high]
Other or None (hereafter referred to as "OTHER")		—

Table 1: Discourse relation tagset with examples.

3.2. Crowdsourcing-Based Annotation

Annotating a corpus with discourse relations consumes a great deal of time and cost. Kawahara et al. (2014) addressed this problem by using crowdsourcing.

Crowdsourcing is a mechanism for ordering tasks to internet users (hereafter referred to as "workers"). Using crowdsourcing, we can produce language resources cheaply and speedily. However, the quality of the resources is often questionable. To mitigate this problem, Kawahara et al. (2014) asked 10 workers to answer each question and aggregated the answers.

Another technique for quality control is to simplify the annotation task by dividing it into subtasks. Kawahara et al. (2014) proposed a two-step annotation. The first subtask was to determine whether a clause pair has a discourse relation other than "OTHER." In this subtask, workers were given a document and asked to choose between "OTHER" or non-"OTHER" for every clause pair. Once 10 answers were corrected, Kawahara et al. (2014) calculated the probability that the clause pair has a non-"OTHER" discourse relation using GLAD (Whitehill et al., 2009), which proved to be more reliable than majority voting. If the probability was larger than 0.01, the clause pair was passed to the next subtask. Otherwise, the clause pair was labeled as "OTHER." This task can be seen as a filtering step because the vast majority of clause pairs are to be labeled as "OTHER."

The second subtask was to classify the discourse relation of a clause pair that passed the first subtask. In this subtask, workers were given a clause pair and its context, and asked to select one of the 7 relations. Once 10 answers were collected, Kawahara et al. (2014) calculated the probability of each discourse relation type using GLAD and assigned the discourse relation type with the highest probability to each clause pair.

Kawahara et al. (2014) conducted the two-stage crowdsourcing experiment using Yahoo! Crowdsourcing[1] and created the annotation comprising 10,000 Japanese web pages in less than eight hours (the first subtask ran for three hours, and the second one ran for five hours). In the first subtask, 9,068 clause pairs (15.3% of all the clause pairs) were passed to the next subtask, and 4,927 clauses pairs (54.3% of the 9,068 clause pairs) were annotated with discourse relations other than "OTHER" in the second subtask.

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

4. Evaluation and Improvement of Annotation Quality

One major question left unanswered by Kawahara et al. (2014) is how good the quality of the crowdsourcing-based annotation is. In this section, we compare a part of the crowdsourcing-based annotation with the annotation given by experts. The crowdsourcing-based annotation is hereafter referred to as the "Old Annotation." We also report two problems found in the Old Annotation and propose a solution to each of these problems.

We annotated 500 documents as gold data. Three professional annotators with background in linguistics annotated these documents. We assigned a discourse relation type with majority vote. If all annotators disagreed with each other, one consensus label was chosen through discussion. In evaluation, we randomly sampled 313 documents from gold data. This annotation is hereafter referred to as the "Expert Annotation."

Table 2 shows the accuracy of the Old Annotation if the Expert Annotation is used as gold data. We calculated a micro average of discourse relations excluding "OTHER" ($[MicroAve.]$ in Table 2), and the F-measure was 49.5%. This result indicates that the Old Annotation had much room for improvement.

We analyzed the Old Annotation and found two problems. The first problem is that some explicit clause pairs were labeled as "OTHER" in the first subtask.

(1) (i) 特に自分は料金を知っていた**ので**
[Especially because I knew the fee]
(ii) 驚くほどでした。
[it is amazing]

Example (1) has a discourse connective "ので (because)", and consequently example (1) was given Cause/Reason in the Expert Annotation. In the Old Annotation, however,

	All relations (Explicit and Implicit)			Explicit relations			Implicit relations		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Cause/Reason	0.574 (35/61)	0.565 (35/62)	0.569	0.500 (2/4)	1.000 (2/2)	0.667	0.579 (33/57)	0.550 (33/60)	0.564
Purpose	0.417 (5/12)	0.385 (5/13)	0.400	-	-	-	0.417 (5/12)	0.385 (5/13)	0.400
Condition	0.654 (17/26)	0.515 (17/33)	0.576	1.000 (7/7)	0.438 (7/16)	0.609	0.526 (10/19)	0.588 (10/17)	0.556
Ground	0.333 (3/9)	0.273 (3/11)	0.300	0.000 (0/1)	-	-	0.375 (3/8)	0.273 (3/11)	0.316
Contrast	0.167 (2/12)	0.500 (2/4)	0.250	0.000 (0/2)	0.000 (0/0)	0.000	0.200 (2/10)	0.500 (2/4)	0.286
Concession	0.636 (7/11)	0.280 (7/25)	0.389	0.500 (1/2)	0.200 (1/5)	0.286	0.667 (6/9)	0.300 (6/20)	0.414
$[MicroAve.]$	0.527 (69/131)	0.466 (69/148)	0.495	0.625 (10/16)	0.435 (10/23)	0.513	0.513 (59/115)	0.472 (59/125)	0.492

Table 2: Accuracy of the Old Annotation.

	All relations (Explicit and Implicit)			Explicit relations			Implicit relations		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
Cause/Reason	0.588 (40/68)	0.645 (40/62)	0.615	0.500 (2/4)	1.000 (2/2)	0.667	0.594 (38/64)	0.633 (38/60)	0.613
Purpose	0.444 (8/18)	0.615 (8/13)	0.516	-	-	-	0.444 (8/18)	0.615 (8/13)	0.516
Condition	0.821 (23/28)	0.697 (23/33)	0.754	1.000 (10/10)	0.625 (10/16)	0.769	0.722 (13/18)	0.765 (13/17)	0.743
Ground	0.364 (4/11)	0.364 (4/11)	0.364	-	-	-	0.364 (4/11)	0.364 (4/11)	0.364
Contrast	0.083 (2/24)	0.500 (2/4)	0.143	0.000 (0/5)	0.000 (0/0)	0.000	0.105 (2/19)	0.500 (2/4)	0.174
Concession	0.500 (3/6)	0.120 (3/25)	0.194	1.000 (1/1)	0.200 (1/5)	0.333	0.400 (2/5)	0.100 (2/20)	0.160
$[MicroAve.]$	0.516 (80/155)	0.541 (80/148)	0.528	0.650 (13/20)	0.565 (13/23)	0.605	0.496 (67/135)	0.536 (67/125)	0.515

Table 3: Accuracy of the New Annotation.

example (1) was classified as "OTHER" at the first subtask. It turned out that 40.2% (384/955) of the explicit clause pairs in the Old Annotation were annotated with "OTHER." This is probably because workers had not paid attention to discourse connectives.

Moreover, about 47% (182/384) of the explicit clause pairs annotated with "OTHER" at the first subtask have discourse connectives related to Condition. Let us consider the following examples:

(2) (i) 送信が行われる**と**、
[If you send (data),]
(ii) 送信終了のメッセージが表示されます。
[the display shows the message "send completely".]

(3) (i) ミルクで粉をこねて、
[(She) kneaded a powder with milk,]
(ii) おナベで焼く**と**、
[baked it in a pot,]
(iii) パンケーキがフワッとふくらみました。
[then, a pancake was swollen.]

In example (2), the clause pair (i) and (ii) has a Condition relation, because the discourse connective "と" means "if" in English. However, in example (3), the clause pair (ii) and (iii) does not have a discourse relation, because the discourse connective "と" means "then" in English. These examples illustrate the fact that some discourse connectives have ambiguity, and it is difficult to classify such discourse connectives, especially related to Condition.

Our solution to this problem is to skip the first subtask if adjacent clause pairs have discourse connectives without ambiguity. To detect such discourse connectives, we use the KNP parser, which identifies them by hand-crafted rules. These pairs are directly passed to the second subtask. Note that adjacent clause pairs with discourse connectives are sometimes to be labeled as "OTHER" because discourse connectives can connect non-adjacent pairs. Moreover, clause pairs which have discourse connectives with ambiguity are treated as implicit relations.

The second problem is that the instruction was not instructive enough for workers to understand the criteria to determine the discourse relations. Because the discourse annotation task is a bit complicated for crowd workers, we have to simplify the annotation task and to make workers understand the criteria with a simple instruction. Kawahara et al. (2014) showed workers the following instruction:

Condition
e.g.)【明日、晴れれば】←→【ゴルフに行く】
[If it is fine tomorrow,] [I will play golf.]

Kawahara et al. (2014) expected workers to understand the criteria from the instruction. However, the result shown in Table 2 suggests that workers could not understand the criteria. To alleviate this problem, we add the description of discourse connective phrases to the instruction. This additional explanation aims to force workers to do positive language tests using discourse connectives.

Condition
(「１すれば２」と言える関係。１・２が逆でも可。)
[we can insert "if" between 1 and 2.]
e.g.)【1 明日晴れれば、】午前は買い物に行こう。【2 午後は映画に行こう。】(「１すれば２」と言える)
[【1 If it is fine tomorrow, 】let's go to the shopping in the morning.【2 Let's go to the movies in the afternoon. 】(we can insert "if" between 1 and 2)]

In the example above, if "すれば (if)" can be inserted between a clause pair, workers are expected to choose Condition. In the instructions of Cause/Reason and Contract, positive language tests are accompanied by negative language tests:

Cause/Reason
(「１したがって２」と言えるが、「１さらに２」と言えない関係。１・２が逆でも可。)
[we can insert "therefore" between 1 and 2, but cannot insert "moreover" between 1 and 2.]
e.g.)【1 雨が降った。】よく見ると【2 道が濡れている。】(「１したがって２」と言える)
[【1 It rained. 】Look closely,【2 the road is wet. 】(we can insert "therefore" between 1 and 2)]

In the example above, if "したがって (therefore)" can be inserted between the clause pair, workers are expected to choose Cause/Reason. Meanwhile, if "さらに (moreover)" can be inserted between the clause pair, workers can rule out Cause/Reason. We need the negative tests because positive tests sometimes applied to non-target clause pairs in our preliminary experiment.

5. Experiment

With the two improvement techniques described in section 4. we re-annotated the corpus with discourse annotations (hereafter referred to as "New Annotation") and evaluated the quality of the annotation.

Table 3 shows the accuracy of the New Annotation. The New Annotation achieved about 3% higher accuracy than the Old Annotation. The micro-averaged accuracy of explicit relations in the New Annotation was about 9% higher than the Old Annotation. The improvement can mainly be attributed to the first subtask where clause pairs wrongly classified as "OTHER" in the Old Annotation were now given correct labels. Table 3 also shows that the recall was 7.5% higher than in the Old Annotation. This indicates the effectiveness of the new instruction with language tests.

Let us consider the following examples:

(4) (i) どなたにも飲みやすいおいしいワインです。
[This wine is delicious and easy to drink for everyone.]
(ii) おどや販売部長の土谷が山形のワイナリーにお願いして
[Tuchiya, who is a sales manager of Odoya, requested (a brand new product) for winery in Yamagata,]
(iii) おいしいワインを作ってもらいました。
[(and) They made delicious wine.]

The discourse relation between (ii) and (iii)：
- Expert Annotation: OTHER
- Old Annotation: Purpose
- New Annotation: OTHER

(5) (i) ２泊で香川県に旅行に行ってきた。
[(I) went on a trip to Kagawa Prefecture for two nights (and three days).]
(ii) 一応おそめの夏休みである。
[It was a late summer vacation.]
(iii) 奥さんと２歳のきょうこを連れての家族旅行である。
[It was a family trip with my wife and a 2 year old (doughter), Kyoko.]

The discourse relation between (ii) and (iii)：
- Expert Annotation: OTHER
- Old Annotation: OTHER
- New Annotation: Cause/Reason

In example (4), the Old Annotation disagrees with the Expert Annotation, but the New Annotation agrees with the Expert Annotation presumably because this pair failed the language tests (we can insert "in order to"[2] between (ii) and (iii)). However, the New Annotation disagrees with the Expert Annotation in example (5). We conjecture that workers misjudged that this pair passed the language tests. To cope with this problem, we plan to add more expressions for language tests in the instruction. However, new instructions are not guaranteed to force workers to run the language tests because the instruction will become complicated for crowdsourcing. We need to craft concise instructions.

[2]This discourse connective is used for the positive language test of Purpose.

6. Conclusion

In this paper, we evaluated the quality of a Japanese corpus with discourse annotations and proposed improvement techniques based on language tests. The experiment showed that the quality of re-annotation data using our methods was improved by an F-measure of 3%. We will make the re-annotated data publicly available.

The experimental results indicate that the annotation of implicit discourse pairs remains an issue. In the future, we would like to follow in a more direct manner the workflow of the PDTB's annotation procedure for implicit pairs: (1) identifying a discourse connective that could be inserted between arguments without changing the discourse relation between them, and then (2) specifying the discourse relation (Prasad et al., 2014). To adapt this procedure to crowdsourcing, we plan to implement discourse connective identification in the form of a cloze test.

7. Bibliographical References

Asher, N. and Lascarides, A. (2003). *Logics of Conversation*. Cambridge University Press.

Carlson, L., Marcu, D., and Okurowski, M. E. (2001). Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue - Volume 16*, SIGDIAL '01, pages 1–10. Association for Computational Linguistics.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3041–3052. The COLING 2016 Organizing Committee.

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 Computation*, pages 535–544, Bali,Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

Kaneko, K. and Bekki, D. (2014). Building a japanese corpus of temporal-causal-discourse structures based on sdrt for extracting causal relations. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 33–39.

Kawahara, D., Machida, Y., Shibata, T., Kurohashi, S., Kobayashi, H., and Sassano, M. (2014). Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 269–278.

Dublin City University and Association for Computational Linguistics.

Kurohashi, S. and Nagao, M. (1994). A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4).

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC2008)*, pages 2961–2968.

Prasad, R., Webber, B., and Joshi, A. (2014). Reflections on the penn discourse treebank, comparable corpora, and complementary annotation. *Computational Linguistics*, 40(4):921–950.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. (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. Association for Computational Linguistics.

Whitehill, J., Wu, T., Bergsma, J., Movellan, J. R., and Ruvolo, P. L. (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.

Zeyrek, D., Demirşahin, I., Çallı, A. B. S., and Çakıcı, R. (2013). Turkish discourse bank: Porting a discourse annotation style to a morphologically rich language. *Dialogue & Discourse*, 4(2):174–184.

Zhou, Y. and Xue, N. (2012). PDTB-style discourse annotation of Chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 69–77. Association for Computational Linguistics.