Rapid Development of a Corpus with Discourse Annotations using Two-stage Crowdsourcing

Daisuke Kawahara[†‡] **Yuichiro Machida**[†] **Tomohide Shibata**[†‡] **Sadao Kurohashi**[†‡]
Hayato Kobayashi[§] **Manabu Sassano**[§]

[†]Graduate School of Informatics, Kyoto University
[‡]CREST, Japan Science and Technology Agency
[§]Yahoo Japan Corporation

{dk, shibata, kuro}@i.kyoto-u.ac.jp, machida@nlp.ist.i.kyoto-u.ac.jp,
{hakobaya, msassano}@yahoo-corp.jp

Abstract

We present a novel approach for rapidly developing a corpus with discourse annotations using crowdsourcing. Although discourse annotations typically require much time and cost owing to their complex nature, we realize discourse annotations in an extremely short time while retaining good quality of the annotations by crowdsourcing two annotation subtasks. In fact, our experiment to create a corpus comprising 30,000 Japanese sentences took less than eight hours to run. Based on this corpus, we also develop a supervised discourse parser and evaluate its performance to verify the usefulness of the acquired corpus.

1 Introduction

Humans understand text not by individually interpreting clauses or sentences, 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. This kind of analysis is called discourse parsing or discourse structure analysis, and is an important and fundamental task in natural language processing (NLP). Systems for discourse parsing are, however, available only for major languages, such as English, owing to the lack of corpora with discourse annotations.

For English, several corpora with discourse annotations have been developed manually, consuming a great deal of time and cost in the process. These include the Penn Discourse Treebank (Prasad et al., 2008), RST Discourse Treebank (Carlson et al., 2001), and Discourse Graphbank (Wolf and Gibson, 2005). Discourse parsers trained on these corpora have also been developed and practically used. To create the same resource-rich environment for another language, a quicker method than the conventional time-consuming framework should be sought. One possible approach is to use crowdsourcing, which has actively been used to produce various language resources in recent years (e.g., (Snow et al., 2008; Negri et al., 2011; Hong and Baker, 2011; Fossati et al., 2013)). It is, however, difficult to crowdsource the difficult judgments for discourse annotations, which typically consists of two steps: finding a pair of spans with a certain relation and identifying the relation between the pair.

In this paper, we propose a method for crowdsourcing discourse annotations that simplifies the procedure by dividing it into two steps. The point is that by simplifying the annotation task it is suitable for crowdsourcing, but does not skew the annotations for use in practical discourse parsing. First, finding a discourse unit for the span is a costly process, and thus we adopt a clause as the discourse unit, since this is reliable enough to be automatically detected. We also limit the length of each target document to three sentences and at most five clauses to facilitate the annotation task. Secondly, we detect and annotate clause pairs in a document that hold logical discourse relations. However, since this is too complicated to assign as one task using crowdsourcing, we divide the task into two steps: determining the existence of logical discourse relations and annotating the type of relation. Our two-stage approach is a robust method in that it confirms the existence of the discourse relations twice. We also designed the tagset of discourse relations for crowdsourcing, which consists of two layers, where the upper layer contains the following three classes: "CONTINGENCY," "COMPARISON" and "OTHER." Although the task

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 269–278, Dublin, Ireland, August 23-29 2014.

settings are simplified for crowdsourcing, the obtained corpus and knowledge of discourse parsing could be still useful in general discourse parsing.

In our experiments, we crowdsourced discourse annotations for Japanese, for which there are no publicly available corpora with discourse annotations. The resulting corpus consists of 10,000 documents, each of which comprises three sentences extracted from the web. Carrying out this two-stage crowdsourcing task took less than eight hours. The time elapsed was significantly shorter than the conventional corpus building method.

We also developed a discourse parser by exploiting the acquired corpus with discourse annotations. We learned a machine learning-based model for discourse parsing based on this corpus and evaluated its performance. An F1 value of 37.9% was achieved for contingency relations, which would be roughly comparable with state-of-the-art discourse parsers on English. This result indicates the usefulness of the acquired corpus. The resulting discourse parser would be effectively exploited in NLP applications, such as sentiment analysis (Zirn et al., 2011) and contradiction detection (Murakami et al., 2009; Ennals et al., 2010).

The novel contributions of this study are summarized below:

- We propose a framework for developing a corpus with discourse annotations using two-stage crowdsourcing, which is both cheap and quick to execute, but still retains good quality of the annotations.
- We construct a Japanese discourse corpus in an extremely short time.
- We develop a discourse parser based on the acquired corpus.

The remainder of this paper is organized as follows. Section 2 introduces related work, while Section 3 describes our proposed framework and reports the experimental results for the creation of a corpus with discourse annotations. Section 4 presents a method for discourse parsing based on the corpus as well as some experimental results. Section 5 concludes the paper.

2 Related Work

Snow et al. (2008) applied crowdsourcing to five NLP annotation tasks, but the settings of these tasks are very simple. There have also been several attempts to construct language resources with complex annotations using crowdsourcing. Negri et al. (2011) proposed a method for developing a cross-lingual textual entailment (CLTE) corpus using crowdsourcing. They tackled this complex data creation task by dividing it into several simple subtasks: sentence modification, type annotation and sentence translation. The creative CLTE task and subtasks are quite different from our non-creative task and subtasks of discourse annotations. Fossati et al. (2013) proposed FrameNet annotations using crowdsourcing. Their method is a single-step approach to only detect frame elements. They verified the usefulness of their approach through an experiment on a small set of verbs with only two frame ambiguities per verb. Although they seem to be running a larger-scale experiment, its result has not been revealed yet. Hong and Baker (2011) presented a crowdsourcing method for selecting FrameNet frames, which is a part of the FrameNet annotation process. Since their task is equivalent to word sense disambiguation, it is not very complex compared to the whole FrameNet annotation process. These FrameNet annotations are still different from discourse annotations, which are our target. To the best of our knowledge, there have been no attempts to crowdsource discourse annotations.

There are several manually-crafted corpora with discourse annotation for English, such as the Penn Discourse Treebank (Prasad et al., 2008), RST Discourse Treebank (Carlson et al., 2001), and Discourse Graphbank (Wolf and Gibson, 2005). These corpora were developed from English newspaper articles. Several attempts have been made to manually create corpora with discourse annotations for languages other than English. These include the Potsdam Commentary Corpus (Stede, 2004) for German (newspaper; 2,900 sentences), Rhetalho (Pardo et al., 2004) for Portuguese (scientific papers; 100 documents; 1,350 sentences), and the RST Spanish Treebank for Spanish (da Cunha et al., 2011) (several genres; 267 documents; 2,256 sentences). All of these consist of relatively small numbers of sentences compared with the English corpora containing several tens of thousands sentences.

In recent years, there have been many studies on discourse parsing on the basis of the above hand-annotated corpora (e.g., (Pitler et al., 2009; Pitler and Nenkova, 2009; Subba and Di Eugenio, 2009; Hernault et al., 2010; Ghosh et al., 2011; Lin et al., 2012; Feng and Hirst, 2012; Joty et al., 2012; Joty et al., 2013; Biran and McKeown, 2013; Lan et al., 2013)). This surge of research on discourse parsing can be attributed to the existence of corpora with discourse annotations. However, the target language is mostly English since English is the only language that has large-scale discourse corpora. To develop and improve discourse parsers for languages other than English, it is necessary to build large-scale annotated corpora, especially in a short period if possible.

3 Development of Corpus with Discourse Annotations using Crowdsourcing

3.1 Corpus Specifications

We develop a tagged corpus in which pairs of discourse units are annotated with discourse relations. To achieve this, it is necessary to determine target documents, discourse units, and a discourse relation tagset. The following subsections explain the details of these three aspects.

3.1.1 Target Text and Discourse Unit

In previous studies on constructing discourse corpora, the target documents were mainly newspaper texts, such as the Wall Street Journal for English. However, discourse parsers trained on such newspaper corpora usually have a problem of domain adaptation. That is to say, while discourse parsers trained on newspaper corpora are good at analyzing newspaper texts, they generally cannot perform well on texts of other domains.

To address this problem, we set out to create an annotated corpus covering a variety of domains. Since the web contains many documents across a variety of domains, we use the Diverse Document Leads Corpus (Hangyo et al., 2012), which was extracted from the web. Each document in this corpus consists of the first three sentences of a Japanese web page, making these short documents suitable for our discourse annotation method based on crowdsourcing.

We adopt the clause as a discourse unit, since spans are too fine-grained to annotate using crowdsourcing and sentences are too coarse-grained to capture discourse relations. Clauses, which are automatically identified, do not need to be manually modified since they are thought to be reliable enough. Clause identification is performed using the rules of Shibata and Kurohashi (2005). For example, the following rules are used to identify clauses as our discourse units:

- clauses that function as a relatively strong boundary in a sentence are adopted,
- relative clauses are excluded.

Since workers involved in our crowdsourcing task need to judge whether clause pairs have discourse relations, the load of these workers increases combinatorially as the number of clauses in a sentence increases. To alleviate this problem, we limit the number of clauses in a document to five. This limitation excludes only about 5% of the documents in the original corpus.

Our corpus consists of 10,000 documents corresponding to 30,000 sentences. The total number of clauses in this corpus is 39,032, and thus the average number of clauses in a document is 3.9. The total number of clause pairs is 59,426.

3.1.2 Discourse Relation Tagset

One of our supposed applications of discourse parsing is to automatically generate a bird's eye view of a controversial topic as in Statement Map (Murakami et al., 2009) and Dispute Finder (Ennals et al., 2010), which identify various relations between statements, including contradictory relations. We assume that expansion relations, such as elaboration and restatement, and temporal relations are not important for this purpose. This setting is similar to the work of Bethard et al. (2008), which annotated temporal relations independently of causal relations. We also suppose that temporal relations can be annotated separately for NLP applications that require temporal information. We determined the tagset of discourse relations

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 was turned on]
	Purpose	【試験に受かるために】【必死に勉強した。】 [to pass the exam] [(I) studied a lot]
	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/her bag] [he/she would be still in the company]
COMPARISON	Contrast	【あのレストランは寿司はおいしいが】【ラーメンは普通だ。】 [at that restaurant, sushi is good] [ramen is so-so]
	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] [the price is high]
OTHER	(Other)	【家に着いてから】【雨が降ってきた。】 [After being back home] [it began to rain]

Table 1: Discourse relation tagset with examples.

by referring to the Penn Discourse Treebank. This tagset consists of two layers, where the upper layer contains three classes and the lower layer seven classes as follows:

- CONTINGENCY
  - Cause/Reason (causal relations and not conditional relations)
  - Purpose (purpose-action relations where the purpose is not necessarily accomplished)
  - Condition (conditional relations)
  - Ground (other contingency relations including pragmatic cause/condition)
- COMPARISON (same as the Penn Discourse Treebank)
  - Contrast
  - Concession
- OTHER (other weak relation or no relation)

Note that we do not consider the direction of relations to simplify the annotation task for crowdsourcing. Table 1 shows examples of our tagset.

Therefore, our task is to annotate clause pairs in a document with one of the discourse relations given above. Sample annotations of a document are shown below. Here, clause boundaries are shown by "::" and clause pairs that are not explicitly marked are allocated the "OTHER" relation.

Cause/Reason 気がつけば::梅雨も明けてました。::毎日暑い日が続きますね。::【父の手術も無事に終わり、】::【少しだけほっとしてます。】
... [the surgery of my father ended safely] [(I) am relieved a little bit]

Contrast 今日とある企業のトップの話を聞くことが出来た。::経営者として何事も全てビジネスチャンスに変えるマインドが大切だと感じた。::【生きていく上で追い風もあれば、】::【逆風もある。】
... [There is tailwind to live,] [there is also headwind.]

3.2 Two-stage Crowdsourcing for Discourse Annotations

We create a corpus with discourse annotations using two-stage crowdsourcing. We divide the annotation task into the following two subtasks: determining whether a clause pair has a discourse relation excluding "OTHER," and then, ascertaining the type of discourse relation for a clause pair that passes the first stage.

Probability	Number
$= 1.0$	64
> 0.99	554
> 0.9	1,065
> 0.8	1,379
> 0.5	2,655
> 0.2	4,827
> 0.1	5,895
> 0.01	9,068
> 0.001	12,277
> 0.0001	15,554

Table 2: Number of clause pairs resulting from the judgments of discourse relation existence.

3.2.1 Stage 1: Judgment of Discourse Relation Existence

This subtask determines whether each clause pair in a document has one of the following discourse relations: Cause/Reason, Purpose, Condition, Ground, Contrast, and Concession (that is, all the relations except "OTHER"). Workers are shown examples of these relations and asked to determine only the existence thereof.

In this subtask, an item presented to a worker at a particular time consists of all the judgments of clause pairs in a document. By adopting this approach, each worker considers the entire document when making his/her judgments.

3.2.2 Stage 2: Judgment of Discourse Relation Type

This subtask involves ascertaining the discourse relation type for a clause pair that passes the first stage. The result of this subtask is one of the seven lower types in our discourse relation tagset. Workers are shown examples of these types and then asked to select one of the relations. If a worker chooses "OTHER," this corresponds to canceling the positive determination of the existence of the discourse relation in stage one.

In this subtask, an item is the judgment of a clause pair. That is, if a document contains more than one clause pair that must be judged, the judgments for this document are divided into multiple items, although this is rare.

3.3 Experiment and Discussion

We conducted an experiment of the two-stage crowdsourcing approach using Yahoo! Crowdsourcing.[1] To increase the reliability of the produced corpus, we set the number of workers for each item for each task to 10. The reason why we chose this value is as follows. While Snow et al. (2008) claimed that an average of 4 non-expert labels per item in order to emulate expert-level label quality, the quality of some tasks increased by increasing the number of workers to 10. We also tested hidden gold-standard items once every 10 items to examine worker's quality. If a worker failed these items in serial, he/she would have to take a test to continue the task.

We obtained judgments for the 59,426 clause pairs in the 10,000 documents of our corpus in the first stage of crowdsourcing, i.e., the subtask of determining the existence of discourse relations. We calculated the probability of each label using GLAD[2] (Whitehill et al., 2009), which was proved to be more reliable than the majority voting. This probability corresponds to the probability of discourse relation existence of each clause pair. Table 2 lists the results. We set a probability threshold to select those clause pairs whose types were to be judged in the second stage of crowdsourcing. With this threshold set to 0.01, 9,068 clause pairs (15.3% of all the clause pairs) were selected. The threshold was set fairly low to allow low-probability judgments to be re-examined in the second stage.

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

[2] http://mplab.ucsd.edu/~jake/OptimalLabelingRelease1.0.3.tar.gz

Lower type	All	prob > 0.8
Cause/Reason	2,104	1,839 (87.4%)
Purpose	755	584 (77.4%)
Condition	1,109	925 (83.4%)
Ground	442	273 (61.8%)
Contrast	437	354 (81.0%)
Concession	80	49 (61.3%)
Sum of the above discourse relations	4,927	4,024 (81.7%)
Other	4,141	3,753 (90.6%)
Total	9,068	7,777 (85.8%)

Table 3: Results of the judgments of lower discourse relation types.

Upper type	All	prob > 0.8
CONTINGENCY	4,439	3,993 (90.0%)
COMPARISON	516	417 (80.8%)
Sum of the above discourse relations	4,955	4,410 (89.0%)
OTHER	4,113	3,753 (91.2%)
Total	9,068	8,163 (90.0%)

Table 4: Results of the judgments of upper discourse relation types.

The discourse relation types of the 9,068 clause pairs were determined in the second stage of crowdsourcing. We extended GLAD (Whitehill et al., 2009) for application to multi-class tasks, and calculated the probability of the labels of each clause pair. We assigned the label (discourse relation type) with the highest probability to each clause pair. Table 3 gives some statistics of the results. The second column in this table denotes the numbers of each discourse relation type, while the third column gives the numbers of each type of clause pair with a probability higher than 0.80. Table 4 gives statistics of the results when the lower discourse relation types are merged into the upper types. Table 5 shows some examples of the resulting annotations.

Carrying out the two separate subtasks using crowdsourcing took approximately three hours and five hours with 1,458 and 1,100 workers, respectively. If we conduct this task at a single stage, it would take approximately 33 (5 hours / 0.153) hours. It would be four times longer than our two-stage approach. Such single-stage approach is also not robust since it does not have a double check mechanism, with which the two-stage approach is equipped. We spent 111 thousand yen and 113 thousand yen (approximately 1,100 USD, respectively) for these subtasks, which would be extremely less expensive than the projects of conventional discourse annotations.

For the examples in Table 5, we confirmed that the discourse relation types of the top four examples were surely correct. However, we judged the type (Contrast) of the bottom example as incorrect. Since the second clause is an instantiation of the first clause, the correct type should be "Other." We found such errors especially in the clause pairs with a probability lower than 0.80.

4 Development of Discourse Parser based on Acquired Discourse Corpus

To verify the usefulness of the acquired corpus with discourse annotations, we developed a supervised discourse parser based on the corpus, and evaluated its performance. We built two discourse parsers using the annotations of the lower and upper discourse relation types, respectively. From the annotations in the first stage of crowdsourcing (i.e., judging the existence of discourse relations), we assigned annotations with a probability less than 0.01 as "OTHER." Of the annotations acquired in the second stage (i.e., judging discourse relation types), we adopted those with a probability greater than 0.80 and discarded the rest. After this preprocessing, we obtained 58,135 (50,358 + 7,777) instances of clause pairs for the lower-type discourse parser and 58,521 (50,358 + 8,163) instances of clause pairs for the upper-type

Prob	# W	Type	Document
1.00	6/10	Cause/Reason	ツツジ科・ツツジ属。【花が陰暦五月に咲くため】【「皐月」と呼ばれている。】市制２０年を記念して、１９７９年１１月３日に制定された。 ... [Since the flower blooms in the fifth lunar month] [it is called "Satsuki."] ...
0.99	4/10	Condition	【↓マップ上の吹き出しをクリックすると】【おすすめルートがご覧になれます。】市町村名をクリックすると「見どころ・体験・食」の情報がご覧になれます。緑色の表記は各スポットの写真がご覧になれます。 [If you click the balloon on the map] [you can see the recommended route] ...
0.81	3/10	Purpose	ダイランティアはマナによって支えられた世界。しかし、人類の繁栄と共に世界樹が３年に一度結実させる「大いなる実り」だけでは人類の繁栄を支えることができなくなってしまった。【そして「大いなる実り」を求めて】【各国が戦争を繰り広げていく。】 ... [And seeking "Great harvest"] [each country is engaged in a war]
0.61	2/10	Cause/Reason	スケールは（一部を除き）１／３２とされている。これは単３形乾電池２本が入りやすいようにしたサイズである。動力は単３形乾電池２本とＦＡ－１３０サイズのモーター１個で、【ギヤーとシャフトの組み合わせにより動力を前後の車軸に伝達し、】【４輪を駆動する。】 ... [by transmitting power to the front and rear axle with the combination of gears and shafts] [(it) drives the four wheels.]
0.54	3/10	Contrast	来年春には、阪急百貨店が新博多駅に東急ハンズと共にお目見えする。そうなると【百貨店による顧客の奪い合いが厳しくなる。】【そこに浮上するのが、三越福岡の閉鎖の可能性である。】 ... [a scramble for customers by department stores would be severe.] [What comes out is the possibility of the closure of Fukuoka Mitsukoshi.]

Table 5: Examples of Annotations. The first column denotes the estimated label probability and the second column denotes the number of workers that assigned the designated type. In the fourth column, the clause pair annotated with the type is marked with 【】 ([ ] in English translations).

discourse parser. Of these, 4,024 (6.9%) and 4,410 (7.5%) instances, respectively, had one of the types besides "OTHER." We conducted experiments using five-fold cross validation on these instances.

To extract features of machine learning, we applied the Japanese morphological analyzer, JUMAN,[3] and the Japanese dependency parser, KNP,[4] to the corpus. We used the features listed in Table 6, which are usually used for discourse parsing.

We adopted Opal (Yoshinaga and Kitsuregawa, 2010)[5] for the machine learning implementation. This tool enables online learning using a polynomial kernel. As parameters for Opal, we used the passive-aggressive algorithm (PA-I) with a polynomial kernel of degree two as a learner and the extension to multi-class classification (Matsushima et al., 2010). The numbers of classes were seven and three for the lower- and upper-type discourse parsers, respectively. We set the aggressiveness parameter C to 0.001, which generally achieves good performance for many classification tasks. Other parameters were set to the default values of Opal.

To measure the performance of the discourse parsers, we adopted precision, recall and their harmonic mean (F1). These metrics were calculated as the proportion of the number of correct clause pairs to the

[3]http://nlp.ist.i.kyoto-u.ac.jp/EN/?JUMAN
[4]http://nlp.ist.i.kyoto-u.ac.jp/EN/?KNP
[5]http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/

Name	Description
clause distance	clause distance between two clauses
sentence distance	sentence distance between two clauses
bag of words	bag of words (lemmas) for each clause
predicate	a content word (lemma) of the predicate of each clause
conjugation form of predicate	a conjugation form of the predicate of each clause
conjunction	a conjunction if it is located at the beginning of a clause
word overlapping ratio	an overlapping ratio of words between the two clauses
clause type	a lexical type output by KNP for each clause (about 100 types)
topic marker existence	existence of a topic marker in each clause
topic marker cooccurrence	existence of a topic marker in both clauses

Table 6: Features for our discourse parsers.

Type	Precision	Recall	F1
Cause/Reason	0.623 (441/708)	0.240 (441/1,839)	0.346
Purpose	0.489 (44/90)	0.075 (44/584)	0.131
Condition	0.581 (256/441)	0.277 (256/925)	0.375
Ground	0.000 (0/12)	0.000 (0/273)	0.000
Contrast	0.857 (6/7)	0.017 (6/354)	0.033
Concession	0.000 (0/0)	0.000 (0/49)	0.000
Other	0.944 (53,702/56,877)	0.992 (53,702/54,111)	0.968

Table 7: Performance of our lower-type discourse parser.

Type	Precision	Recall	F1
CONTINGENCY	0.625 (1,084/1,735)	0.272 (1,084/3,993)	0.379
COMPARISON	0.412 (7/17)	0.017 (7/417)	0.032
OTHER	0.942 (53,454/56,769)	0.988 (53,454/54,111)	0.964

Table 8: Performance of our upper-type discourse parser.

number of all recognized or gold-standard ones for each discourse relation type. Tables 7 and 8 give the accuracies for the lower- and upper-type discourse parsers, respectively.

From Table 8, we can see that our upper-type discourse parser achieved an F1 of 37.9% for contingency relations. It is difficult to compare our results with those in previous work due to the use of different data set and different languages. We, however, anticipate that our results would be comparable with those of state-of-the-art English discourse parsers. For example, the end-to-end discourse parser of Lin et al. (2012) achieved an F1 of 20.6% – 46.8% on the Penn Discourse Treebank.

We also obtained a low F1 for comparison relations. This tendency is similar to the previous results on the Penn Discourse Treebank. The biggest cause of this low F1 is the lack of unambiguous explicit discourse connectives for these relations. Although there are explicit discourse connectives in Japanese, many of them have multiple meanings and cannot be used as a direct clue for discourse relation detection (e.g., as described in Kaneko and Bekki (2014)). As reported in Pitler et al. (2009) and other studies, the identification of implicit discourse relations are notoriously difficult. To improve its performance, we need to incorporate external knowledge sources other than the training data into the discourse parsers. A promising way is to use large-scale knowledge resources that are automatically acquired from raw corpora.

5 Conclusion

We presented a rapid approach for building a corpus with discourse annotations and a discourse parser using two-stage crowdsourcing. The acquired corpus is made publicly available and can be used for research purposes.[6] This corpus can be used not only to build a discourse parser but also to evaluate its performance. The availability of the corpus with discourse annotations will accelerate the development and improvement of discourse parsing. In the future, we intend integrating automatically acquired knowledge from corpora into the discourse parsers to further enhance their performance. We also aim to apply our framework to other languages without available corpora with discourse annotations.

References

Steven Bethard, William Corvey, Sara Klingenstein, and James H. Martin. 2008. Building a corpus of temporal-causal structure. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 908–915.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. 2011. On the development of the RST Spanish treebank. In *Proceedings of the 5th Linguistic Annotation Workshop (LAW V)*, pages 1–10.

Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World Wide Web*, pages 341–350.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 60–68. Association for Computational Linguistics.

Marco Fossati, Claudio Giuliano, and Sara Tonelli. 2013. Outsourcing FrameNet to the crowd. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 742–747.

Sucheta Ghosh, Sara Tonelli, Giuseppe Riccardi, and Richard Johansson. 2011. End-to-end discourse parser evaluation. In *Fifth IEEE International Conference on Semantic Computing (ICSC)*, pages 169–172.

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

Hugo Hernault, Helmut Prendinger, David duVerle, and Mitsuru Ishizuka. 2010. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33.

Jisup Hong and Collin F. Baker. 2011. How good is the crowd at "real" WSD? In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 30–37.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 486–496.

Kimi Kaneko and Daisuke Bekki. 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.

[6] http://nlp.ist.i.kyoto-u.ac.jp/EN/?DDLC

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 476–485.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2012. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, pages 1–34.

Shin Matsushima, Nobuyuki Shimizu, Kazuhiro Yoshida, Takashi Ninomiya, and Hiroshi Nakagawa. 2010. Exact passive-aggressive algorithm for multiclass classification using support class. In *Proceedings of 2010 SIAM International Conference on Data Mining (SDM2010)*, pages 303–314.

Koji Murakami, Eric Nichols, Suguru Matsuyoshi, Asuka Sumida, Shouko Masuda, Kentaro Inui, and Yuji Matsumoto. 2009. Statement map: Assisting information credibility analysis by visualizing arguments. In *Proceedings of the 3rd Workshop on Information Credibility on the Web*, pages 43–50.

Matteo Negri, Luisa Bentivogli, Yashar Mehdad, Danilo Giampiccolo, and Alessandro Marchetti. 2011. Divide and conquer: Crowdsourcing the creation of cross-lingual textual entailment corpora. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 670–679.

Thiago Alexandre Salgueiro Pardo, Maria das Graças Volpe Nunes, and Lucia Helena Machado Rino. 2004. Dizer: An automatic discourse analyzer for Brazilian Portuguese. In *Advances in Artificial Intelligence–SBIA 2004*, pages 224–234. Springer.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. 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 683–691.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 2961–2968.

Tomohide Shibata and Sadao Kurohashi. 2005. Automatic slide generation based on discourse structure analysis. In *Proceedings of Second International Joint Conference on Natural Language Processing*, pages 754–766.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew 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.

Manfred Stede. 2004. The Potsdam commentary corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 96–102.

Rajen Subba and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 566–574.

Jacob Whitehill, Paul Ruvolo, Ting fan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043.

Florian Wolf and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 31(2):249–287.

Naoki Yoshinaga and Masaru Kitsuregawa. 2010. Kernel slicing: Scalable online training with conjunctive features. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING2010)*, pages 1245–1253.

Cäcilia Zirn, Mathias Niepert, Heiner Stuckenschmidt, and Michael Strube. 2011. Fine-grained sentiment analysis with structural features. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 336–344.