2段階のクラウドソーシングによる
談話関係タグ付きコーパスの構築

河原 大輔1,2,a) 町田 雄一郎1,b) 柴田 知秀1,2,c) 黒橋 篤夫1,2,d) 小林 隼人3,e) 颱倉 学3,f)

概要：本稿では、クラウドソーシングを活用することによって、談話関係タグ付きコーパスを構築する手法を提案する。談話関係のタグ付けは従来非常に長い時間とコストがかかりものであったが、提案手法ではこれを2段階のクラウドソーシングのタスクに分けることによって短時間で構築する。1文書が3文からなる短い日本語テキストを対象としてクラウドソーシングを行い、1万文書（3万文）からなるコーパスが8時間で構築できることを示す。また、構築したコーパスを利用して談話関係解析器を開発し、コーパスの有用性を示す。

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

Daisuke Kawahara1,2,a) Yuichiro Machida1,b) Tomohide Shibata1,2,c) Sadao Kurohashi1,2,d) Hayato Kobayashi3,e) Manabu Sassano3,f)

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. はじめに

我々はテキストを読んで解釈するとき、節や文などのそれぞれを単独で解釈するわけではなく、他の節や文などの関係を理解しながらテキスト全体を解釈する。計算機によるテキスト理解を実現するためには、節・文間の関係を高精度に解析するシステムを構築する必要がある。このようなシステムは談話関係解析や談話構造解析と呼ばれ、自然言語処理の重要かつ基盤的な解析の一つであるが、タグ付きコーパスの欠如のため、英語のようなメジャーな言語以外には開発されていない。

英語に対しては、Penn Discourse Treebank (PDTB) [1]、RST Discourse Treebank [2]、Discourse Graphbank [3]などの談話関係タグ付きコーパスが長い月日をかけて構築されており、それらを用いて機械学習した談話関係解析システムがいくつか公開され実際に使われている。他の言語について同じ枠組みで開発するにはやはり長い時間、莫大なコストがかかり、大きな問題となる。短時間、低コストでこれを実現する一つの手はクラウドソーシングを使うことである。しかし、これまでの典型的な談話関係アノテーションは、テキスト中で関係をもつスパンのベアをまず見つけ、次にスパン間の関係の種類を明らかにするという難
267
Table 1 Discourse relation tagset.

| 上位タイプ | 下位タイプ | 例 |
|------------|------------|----|
| 根拠・条件 | 原因・理由 | 【ボタンを押したので】【お湯が出た。】 |
| | 目的 | 【試験に受かるために】【必死に勉強した。】 |
| | 条件 | 【ボタンを押せば】【お湯が出る。】 |
| | その他根拠 | 【ここにカバンがあるから】【まだ社内にいるだろう。】 |
| 転換 | 対比 | 【あのレストランは寿司はおいしいが】【ラーメンは普通だ。】 |
| | 演化 | 【あのレストランは確かにおいしいが】【価段が高い。】 |
| その他 | (その他) | 【家にいてもから】【雨が降ってきた。】 |

3.1.2 談話関係タグセット

談話関係解析のひとつアプリケーションは、賛否両論があるトピックに対する俯瞰的なマップを生成する論理マップ [5]やDispute Finder [6]である。これらのシステムにおいて、対立関係を含む、言語間のさまざまな関係の自動同定技術が使われている。この目的には、詳細化や換言のような付加的な談話関係や時間関係は重要ではないと考えた。この設定は、因果関係と独立に時間関係を付与している [28]と似ている。この研究のように、時間が必要とするアプリケーションのために、時間関係を別途付与することも考えられる。

このような考えのもとに、談話関係タグセットはPDTBを参考にして決めた。タグセットは、表1に示すように、2階層、下位7つの分類からなる。なお、談話関係の方向性は、クラウドソーシングのためにタスクを簡単化するために扱わないものとした。

我々のタグ付けタスクは、文書中の任意の節ペア上記のいずれかの談話関係が成り立つかどうかをアノテーションすることである。以下にタグ付け例を示す。ここでは、節区切りは“::”で示している。また、明示的に談話関係を示していない節ペアには“その他”関係が与えられている。

原因・理由 気がつけば：梅雨も明けました。：毎日暑い日が続きますね。：【父の手術も無事に終わりました。】：【少しだけはとってます。】

対比 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。】

3.2 2段階クラウドソーシングの手続き

本研究では、2段階のクラウドソーシングによって、談話関係タグ付けをコーパスを構築する。この2つのステージは、節ペアに対する談話関係の有無の判定と、談話関係があると判定された節ペアに対する談話関係タイプの判定からなる。以下ではそれぞれについて詳細に述べる。
表 2 談話関係有無判定の結果

| 確率 | 頻度 |
|------|------|
| = 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 |

3.2.1 ステージ1：談話関係有無判定

まず、各文書3文中に含まれる節ペアごとに、“原因結果”、“目的”、“条件”、“その他根拠”、“対比”、“説明”の談話関係（すなわち、“その他”以外）があるかどうかをクラウドソーシングで判定する。

ワークーネに提示される1設問は、1文書に含まれるすべての節ペアに対する判定からなる。こうすることによって、文書全体を考慮して、談話関係の有無を判断することが可能になる。

3.2.2 ステージ2：談話関係タイプ判定

次に、第1段階において談話関係があると判断された節ペアに対して、談話関係のタイプを判定する。タイプ判定は、談話関係タグセットの下位の7つの関係から一つを選択することによって行う。もし、ワークーネが“その他”を選択すれば、それは関係ありと判定された第1段階の結果をキャンセルすることに相当する。

ワークーネに提示される1設問は、1つの節ペアの判定である。すなわち、1文書中に判断すべき複数の節ペアがある場合には、複数の設問に分割されるが、このような場合は稀である。

3.3 結果

Yahoo!クラウドソーシングを用いて、3.2の2段階クラウドソーシングを実行した。構築するコーパスの信頼性を高めるために、それぞれの判定における各設問を判定するワークーネの数を10人とした。Snowら[7]は、専門家レベルの質のアセスメントをクラウドソーシングで実現するには、ワークーネの数を4人以上の設定すればよいと結論付けているが、ワークーネの数を10人まで増やすに連れていくつかのタスクにおける質が向上しているため、本タスクにおいては10人と設定した。また、ワークーネの質を担保するために10設問ごとに1つの臆測ゴールド問題を設定した。これは正解をあらかじめ与えてあらゆる問題であり、ワークーネがこの問題を連続して間違えればテスト問題が課され、タスクを継続するためにはこれに正解する必要がある。

第1段階のクラウドソーシング、つまり談話関係有無判定タスクを実行し、10,000文書中の59,426節ペアそれぞれについて10個の判定を得た。この結果から、Whitehillらによって提案されたGLAD[29]を用いて、それぞれの節ペアに対するラベル（談話関係の有無）の確率を計算した。この手法は、単純な多数決を用いるよりも高い精度でラベルを推定することができるのと報告されている。表2に結果を示す。第2段階のクラウドソーシングで関係タイプを判定する対象となる節ペアを選択するためには、この機能に依頼を設定した。本論文では、関値を0.1と設定し、この関値を上回る9,068個の節ペア（全体の15.3%）を選択した。この関値はかなり低いが、これは確率の低い節ペアについては第2段階のクラウドソーシングでも一度判定するためである。

第2段階のクラウドソーシングにおいては、9,068個の節ペアについて談話関係タイプの判定を行い、それぞれの節ペアについて10個の判定を得た。第1段階と同様に、GLADを多値に拡張して、節ペアごとに各ラベル（談話関係タイプ）の確率を求めた。最終的に、節ペアごとに、もっととも高い確率をもつ談話関係タイプに付与した。表3に談話関係タグセット下位クラスに対する分布を示す。表において、2つ以上の列は、それぞれのタイプの頻度を示し、3つ以上の列は確率が0.8以上の頻度を示している。表4には、

http://crowdsourcing.yahoo.co.jp/
表 5 アノテーション結果の例。第 1 列は確率、第 2 列は第 3 列のタイプを付与したワーカーの数を示す。第 4 列において、第 3 列のタイプをもつ節ペアを【】で示す。

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 【】.

| 確率 | #ワーカー | タイプ | 文書 |
|------|----------|-------|------|
| 1.00 | 6/10 | 原因-理由 | ツツジ-ツツジ属。【花が臨時五月に咲くため】【「早月」と呼ばれている。】市制 20 年を記念して、1979年11月1日に制定された。 |
| 0.99 | 4/10 | 条件 | 【マッサ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。市町村案をクリックすると「見どころ・体験・食」の情報がご覧になれます。緑色の実線は各スポットの写真がご覧になれます。 |
| 0.81 | 3/10 | 目的 | 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。 |
| 0.61 | 2/10 | 原因-理由 | スケール（一部を除き）1／32とされている。これは単形乾電池2本が入るやすいようにしたサイズである。動力は単形乾電池2本と F－A－130サイズのモーター1個で、【ギャガ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軸を駆動する。 |
| 0.54 | 3/10 | 対比 | 来年春には、阪急百貨店が新宿駅に乗り換ハンドと共に目を奪えます。そうなると【百貨店による顧客の奪い合いが激しくなる。】そこに浮上するのが、三越福岡の閉鎖の可能性である。 |

下位クラスを上位クラスにマージした結果の分布を示す。
表 5 に結果のアノテーションの例を示す。

クラウドソーシングによる上記 2 つのタスクの実行には、それぞれ約 3 時間、5 時間かたり、1,458 人、1,100 人のワーカーが作業を行った。もしこのタスクを 1 つのタスクとして実行ならば、約 33 時間（5 時間 / 0.153）かたり、これは、我々の提案する 2 段階の手法と比べて 4 倍長くなることになる。また、このような 1 段階の手法は、2 段階の手法がもっているようなダブルチェックの機構がないため、頑健さに欠けると思われる。金額的なコストとしては、それぞれのタスクに 11.1 万円、11.3 万円かたり、従来の大规模な談話関係アノテーションプロジェクトと比べて、非常に安価にコーパスを構築できたと考えられる。

表 5 のアノテーション例において、上から 4 つ目までの例の談話関係タイプは正しいが、一番下の例の“対比”は正しくないと思われる。この例では、2 つの括弧で示された節は、1 つ目の例示であると考えられるので、正しいタイプは「その他」である。このような間違は、特に確率 0.8 以下の節ペアについて見つかっている。

4. 談話関係解析器の構築

構築した談話関係付きコーパスの有用性を検証するために、このコーパスを教師データに用いた機械学習に基づく談話関係解析器を生成し、評価した。解析器としては、談話関係タグセットの上位タイプおよび下位タイプのそれぞれを対象とした 2 種類を作成した。

第 1 段階の談話関係有無判定の結果において、関係ありの確率が 0.01 より低い節ペアに “その他” タイプを与えられた。それ以外の節ペアについては、第 2 段階の談話関係タイプ判定の結果において、確率 0.8 を越える談話関係タイプが含まれる。したがって、節ペアを実験に用いないことはなかった。この結果の結果、下位タイプについて 58,135 個 (50,358 + 7,777) の節ペア、上位タイプについて 58,521 個 (50,358 + 8,163) の節ペアが得られた。それぞれのうち、4,024 個 (6.9%) と 4,410 個 (7.5%) の節ペアは “その他” 以外のタイプが付与されていた。この 2 種類のデータを用いて、5 分割交差検定を行った。

機械学習ツールとしては、opal[30]と JUMAN[2]と日本語文解析器クノップ[3]を上記コーパスに適用した。用いた素性を表 6 に示す。これらの他言語の談話関係解析器で標準的に使われているものである。

機械学習ツールとしては、opal[30]と JUMAN[2]と日本語文解析器クノップ[3]を用いた。このツールでは、多数式カーネルを用いたオンライン学習が可能である。学習には、2 次の多項式カーネルを用いた。

Table 6 Features for our discourse parsers.

| 名前 | 説明 |
|------|------|
| 節距離 | 節間の距離（節単位） |
| 文距離 | 節間の距離（文単位） |
| 特語 | 特に含まれる単語（原形） |
| 適語 | 節の適語（原形） |
| 評語活用型 | 節の適語の活用型 |
| 接続詞 | 節の単語が接続詞またはその単語（原形） |
| 単語重複率 | 2 節間における単語重複率 |
| 節タイプ | KNP によって出力される節のタイプ（約 100 種類） |
| 「は」有無 | 節中に助辞「は」があるかどうか |
| 「は」共起 | 節ともに助辞「は」があるかどうか |

表 6. 談話関係解析器の素性。
| タイプ          | 適合率 | 再現率 | F1   |
|----------------|--------|--------|------|
| 原因・理由     | 0.623  | 0.240  | 0.346|
| 目的           | 0.489  | 0.075  | 0.131|
| 条件           | 0.581  | 0.277  | 0.375|
| その他根拠    | 0.000  | 0.000  | 0.000|
| 対比          | 0.857  | 0.017  | 0.033|
| 議論         | 0.000  | 0.000  | 0.000|
| その他        | 0.944  | 0.992  | 0.968|

| タイプ          | 適合率 | 再現率 | F1   |
|----------------|--------|--------|------|
| 根拠・条件     | 0.625  | 0.272  | 0.379|
| 議論          | 0.412  | 0.017  | 0.032|
| その他        | 0.942  | 0.988  | 0.964|

表7 下位タイプに対する談話関係解析器の精度

Table 7 Performance of our lower-type discourse parser.

表8 上位タイプに対する談話関係解析器の精度

Table 8 Performance of our upper-type discourse parser.

5. おわりに

本稿では、2段階のクラウドソーシングを用いて、談話関係タグ付きコーパスを構築する手法を述べた。談話構造のタグ付け数は非常に長い時間とコストがかかりるものであったが、提案手法ではこれを2段階のクラウドソーシングのタスクに分けないことによって、8時間弱で構築することができた。完成したコーパスは、一般に公開する予定であり、これによって、談話関係解析の研究が進展することが期待される。

参考文献
[1] Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, B.: The Penn Discourse TreeBank 2.0, Proceedings of the 6th International Conference on Language Resources and Evaluation, pp. 2961–2968 (2008).
[2] Carlson, L., Marcu, D. and Okurowski, M. E.: Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory, Proceedings of the Second SIGdial Workshop on Discourse and Dialogue (2001).
[3] Wolf, F. and Gibson, E.: Representing Discourse Coherence: A Corpus-Based Study, Computational Linguistics, Vol. 31, No. 2, pp. 219–287 (2005).
[4] Zirn, C., Niepert, M., Stuckenschmidt, H. and Strube, M.: Fine-Grained Sentiment Analysis with Structural Features, Proceedings of 5th International Joint Conference on Natural Language Processing, pp. 336–344 (2011).
[5] Murakami, K., Nichols, E., Matsu yo, S., Sumida, A., Masuda, S., Imui, K. and Matsumoto, Y.: Statement Map: Assisting Information Credibility Analysis by Visualizing Arguments, Proceedings of the 3rd Workshop on Information Credibility on the Web, pp. 43–50 (2009).
[6] Emmals, R., Trushkowsky, B. and Agosta, J. M.: High-

© 2014 Information Processing Society of Japan
lighting Disputed Claims on the Web, Proceedings of the 19th international conference on World Wide Web, pp. 341–350 (2010).

[7] Snow, R., O’Connor, B., Jurafsky, D. and Ng, A.: Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks, Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 254–263 (2008).

[8] Negri, M., Bentivogli, L., Mehldau, Y., Giampiccolo, D. and Marchetti, A.: Divide and Conquer: Crowdsourcing the Creation of Cross-Lingual Textual Entailment Corpora, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 670–679 (2011).

[9] Fossati, M., Giuliano, C. and Tonelli, S.: Outsourcing FrameNet to the Crowd, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 742–747 (2013).

[10] Hong, J. and Baker, C. F.: How Good is the Crowd at “real” WSD?, Proceedings of the 5th Linguistic Annotation Workshop, pp. 30–37 (2011).

[11] Steede, M.: The Potsdam commentary corpus, Proceedings of the 2004 ACL Workshop on Discourse Annotation, pp. 96–102 (2004).

[12] Pardo, T. A. S., Nunes, M. d. G. V. and Rino, L. H. M.: Dizer: An automatic discourse analyzer for Brazilian Portuguese, Advances in Artificial Intelligence-SBI 2004, Springer, pp. 224–234 (2004).

[13] da Cunha, I., Torres-Moreno, J.-M. and Sierra, G.: On the development of the RST Brazilian treebank, Proceedings of the 5th Linguistic Annotation Workshop (LAW V), pp. 1–10 (2011).

[14] Pitler, E., Louis, A. and Nenkova, A.: Automatic sense prediction for implicit discourse relations in text, 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, pp. 683–691 (2009).

[15] Pitler, E. and Nenkova, A.: Using Syntax to Disambiguate Explicit Discourse Connectives in Text, Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 13–16 (2009).

[16] Subba, R. and Di Eugenio, B.: An effective Discourse Parser that uses Rich Linguistic Information, Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 566–574 (2009).

[17] Hernalst, H., Prendinger, H., duVerle, D. and Ishizuka, M.: HILDA: A discourse parser using support vector machine classification, Dialogue & Discourse, Vol. 1, No. 3, pp. 1–33 (2010).

[18] Ghosh, S., Tonelli, S., Riccardi, G. and Johansson, R.: End-to-End Discourse Parser Evaluation, Fifth IEEE International Conference on Semantic Computing (ICSC), pp. 169–172 (2011).

[19] Feng, V. W. and Hirst, G.: Text-level Discourse Parsing with Rich Linguistic Features, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, pp. 60–68 (2012).

[20] Joty, S., Carenini, G. and Ng, R.: A Novel Discriminative Framework for Sentence-Level Discourse Analysis, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 904–915 (2012).

[21] Joty, S., Carenini, G., Ng, R. and Mehldau, Y.: Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 486–496 (2013).

[22] Biran, O. and McKeown, K.: Aggregated Word Pair Features for Implicit Discourse Relation Disambiguation, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 69–73 (2013).

[23] Lan, M., Xu, Y. and Niu, Z.: Leveraging Synthetic Discourse Data via Multi-task Learning for Implicit Discourse Relation Recognition, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 476–485 (2013).

[24] Lin, Z., Ng, H. T. and Kan, M.-Y.: A PDTB-styled end-to-end discourse parser, Natural Language Engineering, Vol. 20, No. 2, pp. 151–181 (2014).

[25] Hangyo, M., Kawahara, D. and Kurohashi, S.: Building a Diverse Document Leads Corpus Annotated with Semantic Relations, Proceedings of 26th Pacific Asia Conference on Language Information and Computing, pp. 535–544 (2012).

[26] Shiibata, T. and Kurohashi, S.: Automatic Slide Generation Based on Discourse Structure Analysis, Proceedings of Second International Joint Conference on Natural Language Processing, pp. 754–766 (2005).

[27] 齊田隆，徳永健伸：文内に出現する談話関係を認定するための接続表現の調査，言語処理学会第20回年次大会，pp. 173–176 (2014).

[28] Bethard, S., Corvey, W., Klingenstein, S. and Martin, J. H.: Building a Corpus of Temporal-Causal Structure, Proceedings of the 6th International Conference on Language Resources and Evaluation, pp. 908–915 (2008).

[29] Whitehill, J., Ruvolo, P., fan Wu, T., Bergsma, J. and Movellan, J.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, Advances in Neural Information Processing Systems 22 (Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I. and Culotta, A., eds.), pp. 2035–2043 (2009).

[30] Yoshinaga, N. and Kitsuregawa, M.: Kernel slicing: Scalable online training with conjunctive features, Proceedings of the 23rd International Conference on Computational Linguistics (COLING2010), pp. 1245–1253 (2010).

[31] Matsushima, S., Shimizu, N., Yoshida, K., Ninomiya, T. and Nakagawa, H.: Exact Passive-Aggressive Algorithm for Multiclass Classification Using Support Class, Proceedings of 2010 SIAM International Conference on Data Mining (SDM2010), pp. 303–314 (2010).