クラウドソーシングを用いた談話関係アノテーションの改良

岸本 裕大　澤田 晋之介　村脇 有吾　河原 大輔　黒橋 禎夫
京都大学大学院情報学研究科
{kishimoto, sawada, murawaki, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

我々は文章を読解する際、文単位や節単位で読解するのではなく、周辺の文や節との意味的なつながりを理解しながら読み進めている。このようなつながりを談話関係と呼び、談話関係を解析するタスクを談話関係解析と呼ぶ。次に例を示す。

(1) (i) あ〜、やっぱ屋根と側板に隙間ができた。
(ii) あのイケズなゲートのせいだぞ。
(iii) 頑張って埋めるか。

(1) の i と ii は「ゲートが原因で隙間ができた」という意味的なつながりを持つので談話関係がある。その関係は後述の談話関係タグセット (表 1) に従うと、「原因・理由」に分類できる。同様に、i と iii についても「原因・理由」の関係を持つ。一方、ii と iii は意味的なつながりを持たないので、談話関係はない。

談話関係は談話標識を持つ場合 (以降、Explicit と呼ぶ) と談話標識を持たない場合 (以降、Implicit と呼ぶ) に分類される。(1) では、i と ii は「〜のせいだ」という談話標識を持つので Explicit に、i と iii は談話標識を持たないので Implicit に分類される。

このタスクは自然言語処理の基盤的な解析の一つであるが、談話関係がアノテーションされたコーパスは言語によっては存在しない、もしくは少量しかない。また、エキスパートがアノテーションしているため、大規模化には長い時間と莫大なコストがかかる。

この問題を解決するために、河原らはクラウドソーシングを用いた [1]。クラウドソーシングの懸念点はエキスパートによるアノテーションと比較して品質が劣ることである。しかし、河原らは得られたアノテーションの品質を評価していない。

そこで本論文では、コーパスの一部にエキスパートがアノテーションをすることで、河原らのアノテーションの品質を検証する。次に、エラー分析に基づき、クラウドワーカーに提示するアノテーション基準を見直すなどの改良を施す。改良版を用いてクラウドソーシングを再実行したところ、品質向上が確認された。

2　関連研究

エキスパートが構築した談話関係コーパスとしては、Penn Discourse Treebank (PDTB) [2] や RST Discoures Treebank (RST-DT) [3] が有名である。PDTB は英語の新聞記事（2,159 文書）を対象として、スパン単位のペアに談話関係タグを付与したコーパスである。RST-DT は、英語の新聞記事（385 文書）に対して談話構造を 1 つの木構造で表現している。英語以外の言語では、PDTB のフォーマットで構築された中国語 [4]、RST で構築されたドイツ語 [5] などが存在するが、いずれも小規模である。

クラウドソーシングを用いて言語リソースの構築を行っている研究は、本論文で扱う河原ら [1] の他にも、Snow ら [6] や Guillaume ら [7] が行っている。Snow らは単語の類似度判定や RTE など 5 つのアノテーションタスクを実施した。また Guillaume らはゲーミフィケーションを用いて係り受け関係のアノテーションを実施した。

3　クラウドソーシングを用いた談話関係アノテーション

3.1　コーパスの概要

河原らがクラウドソーシングを用いて構築した談話関係コーパス（以降、クラウドコーパスと呼ぶ）[1] について述べる。

このコーパスは、新聞記事を対象とした既存のコーパスと異なり、Web ページの冒頭 3 文を収集した京都大学ウェブ文書リードコーパスを対象としている。これは Web テキストは新聞記事と異なり様々なドメインの文書を含んでいることと、3 文程度の長さであれば簡単にアノテーションが可能であるからである。

談話関係のアノテーション方針は以下の通りである。まず談話関係を解析する単位は節とする。節は文を比

上位タイプ	下位タイプ	例文
根拠・条件	原因・理由	1. 花が陰暦五月に咲くため 2.「皐月」とよばれている。
	目的	1.「泉響」に耳を傾けるには、 2. やはり泊まるしかない。
	条件	1. あなたが本気ならば、 2. 真実を教えましょう。
	その他根拠	1. ここにカバンがあるから 2. まだ社内にいるだろう。
転換	対比	1. 雪は降らないけれど、 2. コットンが舞う。
	譲歩	1. あのレストランは確かにおいしいが 2. 値段は高い。
談話関係なし・その他弱い関係		ー

表 1: 談話関係タグセットと例文

図 1: クラウドソーシングのタスク画面（１段階目）

較的強い区切りによって分割したもので、日本語構文・格解析器 KNP[1]がルールを用いて自動認定する。節にした理由は、文単位の場合は談話関係を捉えるには荒すぎることと、PDTB で採用されているスパン単位はクラウドソーシングでアノテーションするには難易度が高いからである。アノテーションはすべての節間（以降、節ペアと呼ぶ）に対して行う。談話関係タグセットは、PDTB を参考に表１に示した上位タイプ３種類と下位タイプ７種類と定義した。

3.2 アノテーション手法

河原らは、大規模なアノテーションをする際に長い時間と莫大なコストがかかるという問題を解決するためにクラウドソーシングを用いた。クラウドソーシングとは、ネットを通じて不特定多数の人 (以降、ワーカーと呼ぶ) にタスクを発注するシステムである。クラウドソーシングでは、短時間かつ安価で大規模なアノテーションが実施できる。一方、エキスパートによるものと比較して品質が劣るという欠点がある。その欠点を解決するため、河原らは 10 人に対して同じ設

[1]http://nlp.ist.i.kyoto-u.ac.jp/?KNP

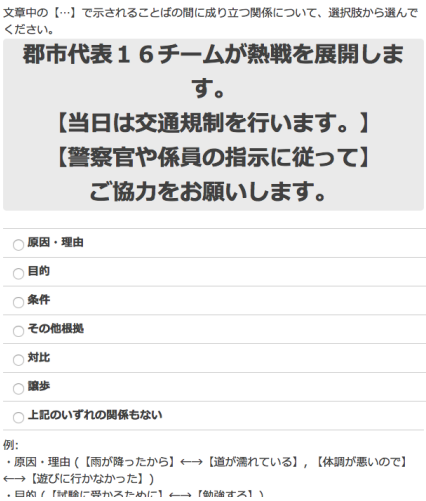

図 2: クラウドソーシングのタスク画面（２段階目）

問を実施し、重み付き投票を行った。さらに、タスクを分割した問題に簡単化した。具体的には、「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」の２段階に分けた。

１段階目では、各文書に含まれる節ペアごとに談話関係を持つか否かの２値分類を行う。まずワーカーに、ある文書に含まれるすべての節ペアを提示する (図 1)。ワーカーは、談話関係を持つ節ペアをすべて選択する。その後、集まった回答に対して Whitehill らの手法 [8] を用いて談話関係を持つ確率を計算する。計算の結果、確率が 0.01 を超えた節ペアは談話関係を持つ可能性があると判定する。

２段階目では、１段階目で談話関係を持つ可能性があると判定された節ペアを対象とする。まずワーカーに、ある節ペアとその前後の文を提示する（図 2）。ワーカーは、提示された節ペアに対して表１の下位タイプの中から最も適切な関係を選択する。その後、集まった回答に対して Whitehill らの手法を用いて各関係の確率を計算し、最も確率が高い関係を出力する。

河原らは Yahoo!クラウドソーシング[2]を用いてクラウドソーシングを実施し、約８時間（１段階目：約３時間、２段階目：約５時間）で１万文書（節ペア数：59,426 個）のアノテーションを行った。しかし、アノテーションの品質は検証していない。

4 エキスパートによるアノテーション結果との比較

クラウドコーパスの品質を検証するために、クラウドコーパスから文書を無作為に抽出し、節ペア全てに

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

	All			Explicit			Implicit		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
原因・理由	0.645 (89/138)	0.674 (89/132)	0.659	0.774 (48/62)	0.889 (48/54)	0.828	0.539 (41/76)	0.526 (41/78)	0.532
目的	0.630 (17/27)	0.354 (17/48)	0.453	0.909 (10/11)	0.667 (10/15)	0.769	0.438 (7/16)	0.212 (7/33)	0.286
条件	0.684 (39/57)	0.600 (39/65)	0.639	0.833 (35/42)	0.648 (35/54)	0.729	0.267 (4/15)	0.364 (4/11)	0.308
その他根拠	0.048 (1/21)	0.167 (1/6)	0.074	-	-	-	0.000 (0/15)	0.000 (0/5)	0.000
対比	0.516 (16/31)	0.457 (16/35)	0.485	-	-	-	0.533 (16/30)	0.457 (16/35)	0.492
譲歩	0.071 (1/14)	0.167 (1/6)	0.100	-	-	-	0.077 (1/13)	0.167 (1/6)	0.105
[談話関係あり]	0.566 (163/288)	0.558 (163/292)	0.562	0.769 (93/115)	0.756 (93/123)	0.762	0.418 (69/165)	0.411 (69/168)	0.414
談話関係なし	0.835 (396/474)	0.843 (396/470)	**0.839**	0.698 (37/53)	0.712 (37/52)	**0.705**	0.853 (359/421)	0.859 (359/418)	**0.856**

表 2: 評価セットを正解とした時のクラウドコーパスの適合率、再現率、F 値

	All			Explicit			Implicit		
	適合率	再現率	F 値	適合率	再現率	F 値	適合率	再現率	F 値
原因・理由	0.655 (95/145)	0.720 (95/132)	**0.686**	0.821 (46/56)	0.852 (46/54)	**0.836**	0.551 (49/89)	0.628 (49/78)	**0.587**
目的	0.601 (25/41)	0.521 (25/48)	**0.562**	0.929 (13/14)	0.867 (13/15)	**0.897**	0.444 (12/27)	0.364 (12/33)	**0.400**
条件	0.774 (48/62)	0.738 (48/65)	**0.756**	0.824 (42/51)	0.778 (42/54)	**0.800**	0.545 (6/11)	0.545 (6/11)	**0.545**
その他根拠	0.111 (3/27)	0.500 (3/6)	**0.182**	-	-	-	0.105 (2/19)	0.400(2/5)	**0.167**
対比	0.400 (22/55)	0.629 (22/35)	**0.489**	-	-	-	0.431 (22/51)	0.629 (22/35)	**0.512**
譲歩	0.143 (1/7)	0.167 (1/6)	**0.154**	-	-	-	0.143 (1/7)	0.167 (1/6)	**0.154**
[談話関係あり]	0.576 (194/337)	0.664(194/292)	**0.617**	0.834 (101/121)	0.821 (101/123)	**0.828**	0.446 (91/204)	0.542 (91/168)	**0.489**
談話関係なし	0.861 (366/425)	0.779 (366/470)	0.818	0.767 (33/43)	0.635 (33/52)	0.695	0.872 (333/382)	0.797 (333/418)	0.833

表 3: 評価セットを正解とした時の新コーパスの適合率、再現率、F 値

対して談話関係のアノテーションを行った（以降、評価セットと呼ぶ)。評価セットは 676 文書からなり、2 名のエキスパートが実施した。この際、両者の意見が一致しなかった場合は、合議により 1 種類の関係を決定した。

この評価セットを正解としたときのクラウドコーパスの適合率、再現率、F 値を表 2 に示す。All はクラウドコーパス全体の適合率、再現率、F 値である。談話関係を示す 6 種類のマイクロ平均（表 2 の [談話関係あり]）を求めると全体の F 値は 0.562 となり、クラウドコーパスの品質はあまり良くない。エラー分析を行ったところ、大きく 2 つの問題点が判明した。

1 つ目の問題は、Explicit であるのに 1 段階目で切り捨てられた節ペアが多数含まれている点である。

(2) (i) 特に自分は料金を知っていたので
(ii) 驚くほどでした。

(2) は i の節末に談話標識「〜ので」が含まれているため、評価セットでは「原因・理由」と判定している。しかし、クラウドコーパスでは 1 段階目で談話関係を持たないと判定された。このように、節が隣り合っており、かつ Explicit であるが、1 段階目で切り捨てられた節ペアが 34 個あった。そこで、節が隣り合っており、かつ「原因・理由」「目的」「条件」のいずれかを示す談話標識を持つ節ペアについては、1 段階目の結果にかかわらず 2 段階目を実施することとした。

2 つ目の問題はワーカーに対してアノテーション基準を例文形式で提示していた点である。難易度が高い談話関係のアノテーションタスクを簡単化するためには、シンプルな方法でアノテーション基準を効果的に伝える必要がある。河原らは例文をワーカーに提示することでアノテーション基準を示していた。具体例として、「原因・理由」のアノテーション基準を示す。

原因・理由
(例:【雨が降ったから】←→【道が濡れている】
【体調が悪いので】←→【遊びに行かなかった】)

この形式では、ワーカーが例文からアノテーション基準を推測することを期待していたが、これだけでは不十分であったと考えられる。そのため、より明示的な方法でアノテーション基準を示す必要がある。

そこで、ワーカーに提示するアノテーション基準に言語テストを追加して提示することとした。具体例として、「原因・理由」の新アノテーション基準を示す。

原因・理由
（「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。）
例：【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）

この形式は、ワーカーに提示した表現「〜したがって〜」を利用して言語テストを行ってもらうことを目的としている。もし、1 と 2 の間に「したがって」を挿入しても問題ない場合は「原因・理由」を判定するようにワーカーを誘導する。

また、「原因・理由」と「対比」のアノテーション基準には「〜と言えない」という否定の言語テストを追加している。これは予備実験を行ったところ、以下

の (3) のように談話関係を持たないのに、談話関係を持つと判定された節ペアが「原因・理由」と「対比」で多く見られたためである。

(3) (i) 神に捧げるといわれる伝統芸能バリ島・ダンスやガムラン音楽は観光客にも人気が高い。
(ii) 神々の島、神秘の楽園といわれ、
(iii) 美しいビーチもあわせもつリゾート地であります。

i と iii の談話関係：「談話関係なし」（情報の追加）

5 アノテーションの改良実験

4 節の対策によってアノテーションの品質が向上するかどうかを検証するために、クラウドソーシングを用いてアノテーションを再実施した。新しい談話関係コーパス（以降、新コーパスと呼ぶ）は表 2 で用いた節ペア 762 個を対象としている。

評価セットを正解としたときの新コーパスの適合率、再現率、F 値を表 3 に示す。表 2 と比べ、「談話関係なし」以外のすべてで F 値が向上した。Explicit を比較すると「目的」「条件」の再現率が大幅に向上していることがわかる。これは、1 段階目で除外されていた Explicit に正しい関係が付与されたためと考えられる。また、全体的に再現率が大幅に改善している。これは言語テストを行う表現を追加したことが効果的であったと考えられる。次に改善した例と悪化した例を示す。

(4) (i) 本日、自由奔放な我ら夫婦の生活は 5 年目の節目を迎えました。
(ii) 日ごろの感謝を込めて、
(iii) 妻には薔薇を贈ります。

i と ii の談話関係：
- 評価セット：「原因・理由」
- クラウドコーパス「談話関係なし」
- 新コーパス：「原因・理由」

(5) (i) 千里に一目惚れし、
(ii) 積極的にアプローチをするが、
(iii) 全く相手にされない。

ii と iii の談話関係：
- 評価セット：「談話関係なし」
- クラウドコーパス「談話関係なし」
- 新コーパス：「対比」

(4) は談話標識が含まれていなかったため、クラウドコーパスでは「談話関係なし」となっていた。しかし、新コーパスではワーカーに提示した談話関係「原因・理由」を示す表現（「i したがって ii」と言え、「i さらに ii」と言えない）を挿入しても問題がないため改善した。

対して (5) は本来談話関係を持たない節ペアであるが、誤った関係が付与されていた。これは「対比」を示す表現（「ii けれども iii」と言え、「ii ただし iii」と言えない）を挿入しても問題が無いとワーカーが判断したからだと推測される。改善策としては、言語テストを行うための表現をさらに増やすことが考えられる。しかし、複数の表現を提示しても、ワーカーが言語テストを実施するかは保証がなく、また (5) のように悪化する例が更に増える恐れもある。そのため、今以上に過不足なく、かつ出来る限り短い表現を探す必要がある。

一番の問題は、Implicit の精度が依然として低いことである。この問題への対策として、Implicit である節ペア間に挿入しうる談話標識をアノテーションさせるというタスクの追加を検討している。この追加タスクを通じてワーカーに明示的に言語テストを行わせることで、アノテーションの精度向上が期待できる。

6 おわりに

本研究では、エキスパートがアノテーションした評価セットを用いて、河原らが構築した談話関係コーパスの品質を検証した。検証結果をもとにクラウドソーシングの実施方法を改良し、品質が良くなることを示した。今後は、5 節で述べた改善策を実施し、Implicit のさらなる精度の向上を目指す。

参考文献

[1] Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. Rapid Development of a Corpus with Discourse Annotations using Two-stage Crowdsourcing. In *Proceedings of COLING2014*, pp. 269–278, 2014.

[2] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In *LREC 2008*, 2008.

[3] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of SIGDIAL2001*, 2001.

[4] Yuping Zhou and Nianwen Xue. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of ACL2012*, Vol. 1, pp. 69–77, 2012.

[5] Manfred Stede and Arne Neumann. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of LREC2014*, 2014.

[6] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of EMNLP2008*, pp. 254–263, 2008.

[7] Bruno Guillaume, Karën Fort, and Nicolas Lefebvre. Crowdsourcing Complex Language Resources: Playing to Annotate Dependency Syntax. In *Proceedings of COLING2016*, pp. 3041–3052, 2016.

[8] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Proceeding of NIPS22*, pp. 2035–2043, 2009.