クラウドソーシングによる関係知識のアノテーション

塙 一晃　佐々木 彬　岡崎 直観　乾 健太郎
東北大学 情報科学研究科
{hanawa, aki-s, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

自然言語処理の研究を進めるうえで，品詞や係り受けなどの言語知識や，エンティティとその関係などの世界知識を記述した言語資源は欠かせない．以前は，専門家に作業を依頼して言語資源を構築することが多かったが，近年はクラウドソーシングを活用し，大規模な言語資源を低コストで構築できるようになった [1]．クラウドソーシングで構築された言語資源のタスクは，品詞タグ付け [2]，統語情報 [3]，固有表現抽出 [4, 5]，類似度判定 [6]，評判抽出 [7]，関係インスタンス [8]，談話関係 [9] など，多岐にわたる．

しかし，自然言語処理のすべてのタスクにクラウドソーシングが向いている訳ではない．クラウドソーシングの作業者は専門家ではないので，明快で，気軽にできて，単純な作業を設計する必要がある．また，クラウドソーシングでの作業は，選択肢への回答や自由記述などに限定されることが多い．このため，テキスト中の任意の単語を作業者が選び，その単語にラベルを付与したり，別の単語との関係を付与するようなアノテーションには向かない．先行研究では，付与する単語の場所とラベルの候補を予め抽出しておき，選択式の問題に落とし込むことが多い．しかし，付与する単位（単語なのか句なのか等）や付与する箇所の候補（体言のみか用言も含むか等）を前もって決めておくのは難しい．

本論文では，コーパスに関係知識を付与する作業をクラウドソーシングで完結させるため，アノテーションツールである brat [10] を改変し，Yahoo!クラウドソーシング[*1]の外部作業サイトとして自由に利用する方法を紹介する．この方法を利用し，Wikipedia の概要文に対して（促進と抑制の）因果関係の事例を付与する実験を行い，付与対象の単位や正解の優先順位を明確に与えなくても，クラウドソーシングで比較的質の高いコーパスを構築できることを示す．アノテーションの一致度や極性反転表現などの分析，構築したコーパスを学習データとした因果関係抽出器の実験などを報告し，本研究で構築したコーパスの有効性を示す．なお，作成したコーパスはウェブサイト[*2]上で公開している．

2 Wikipedia 記事への促進・抑制関係付与

2.1 促進・抑制関係

本研究では，促進・抑制関係 [11, 12] のアノテーションに取り組む．ここで，「X が Y を促進する」とは X が活性化したときに，Y も活性化するような関係であり，同義関係なども含む．「X が Y を抑制する」とは X が活性化したときに，Y は不活性化される関係である．このような促進・抑制関係による知識は，病気や失敗などの要因の分析や，質問応答 [13]，賛否分類 [14] などのタスク等で有効である．

本研究では，記事のタイトルが促進するもの (Pro)，タイトルが抑制するもの (Sup)，タイトルを促進するもの (Pro_by)，タイトルを抑制するもの (Sup_by) を，記事の概要文中の表現に対してアノテーションすることを考える．各関係の片方の引き数を記事のタイトルに固定しておくことで，アノテーション作業を簡略化するだけでなく，Wikipedia 記事からの知識獲得として現実的なタスクを設定している．付与対象の記事は，社会問題，災害，病気，技術革新，政策の 5 つのカテゴリと，そのサブカテゴリ，サブサブカテゴリに収録されている記事の中から，ランダムに 1,000 件を選んだ．これらのカテゴリを採用したのは，記事中に促進・抑制関係の事例が多く含まれると予測したからである．

2.2 アノテーション方針

促進・抑制といった因果関係をアノテーションする際に問題になるのが，付与対象の表現をどのように規定するかである．本研究では，体言にアノテーションする場合と，用言にアノテーションする場合の 2 通りを検討したが，いずれの場合でも不満が残ることが分かった．

例として，「柑皮症」の Wikipedia 記事中の 1 文，「柑皮症とは、β－クリプトキサンチンやβ－カロテンといったカロテノイド色素の過剰な摂取で皮膚が黄色くなることをいう。」を考える．この 1 文から，⟨Pro, 柑皮症, 皮膚が黄色くなる⟩ という関係事例を取り出したくなるが，体言にのみアノテーションするという方針を採用してしまうと「皮膚が黄色くなる」という箇所に付与することはできない．代わりに，用言のみにアノテーションするという方針を取った場合は，⟨Pro_by, 柑皮症, β－クリプトキサンチン⟩ という関係事例にアノテーションできない．さらに，体言と用言のどちらを採用しても，体言もしくは用言の単位をどのように規定するかという問題が残る．先程の例では，「カロテノイド色素の過剰な摂取」と「カロテノイド色素」のいずれも Pro_by 関係にあると解釈できる．このように，正解が複数あり得る状況では，どれか一つに決めるための基準を作っても，アノテーションの一貫性が保証されない．そこで，1 つの記事に対して複数人のアノテーションを収集することで，付与箇所ごとに異なる確信度（一致度）を持ったコーパスが作れるのではないかと考えた．

3 クラウドソーシングにおける brat の活用

1 節で説明したように，一般的なクラウドソーシング・サービスでは選択式や自由記述などの決められた形式の作業しか行えない．この場合，付与すべき単位を予め決めておき，付与すべき箇所の候補を作業者に提示する必要があるが，これ

[*1] http://crowdsourcing.yahoo.co.jp/
[*2] http://www.cl.ecei.tohoku.ac.jp/

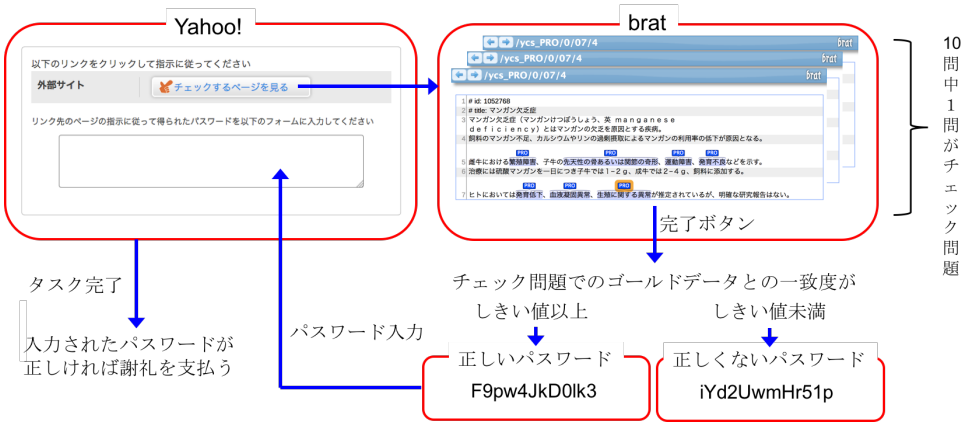

図 1　Yahoo!クラウドソーシングと brat によるアノテーションの概略

は 2 節で述べたように現実的ではない．

他に問題となるのが，作成されるデータの質を保証するためのチェック設問の取扱いである．Yahoo!クラウドソーシングでもチェック設問での完全一致による正解判定を行うことは可能である．しかし，今回のアノテーション作業では付与すべき単位を明確に決めておらず，複数の正解があり得るため，完全一致による正解判定を行ってしまうと，ほぼ全ての作業者が不正解と判定されてしまう．そこで，本研究では Yahoo!クラウドソーシングから（本研究で構築した）外部サイトに誘導し，作業者には brat によるアノテーションを依頼することにした．また，チェック設問の正解判定も brat 側で行えるように，システムを改変した．

図 1 に，提案システムの概要を示す．このシステムは，以下の流れでアノテーション作業を進めていく．

1. Yahoo!クラウドソーシングの作業画面に外部サイトへのリンクを貼り，brat で構築したアノテーション・ツールへ誘導する．
2. 作業者は brat 上でアノテーション作業を行う．
3. 1 セットの作業が完了したら，その作業の中に紛れ込ませておいたチェック設問を使い，作業の正確度を測定する．作業の正確度は，こちらが用意した正解と作業者のアノテーションの一致度を文字レベルでの F スコアで測定したものを採用する．
4. 作業者にパスワードを発行する．このとき，作業の正確度が閾値（0.3）を超えていたら，報酬が支払われるパスワード，閾値未満ならば報酬を支払わないパスワードを発行する．
5. 作業者は Yahoo!クラウドソーシングの画面に戻り，パスワードを入力する．正確度が閾値を超えていた場合はそのアノテーションを採用し，作業者には謝礼が支払われる．

4　アノテーション結果

前節で説明したシステムを用い，1 つの記事につき 10 人のアノテーションが採用されるように収集した．促進・抑制に関する 4 つの関係は，それぞれ独立のタスクとして作業を発注することで，作業を単純化するとともに，他の 3 つの関係を意識しない時のアノテーション結果を得ることにした．実際に得られたアノテーションの例を図 4 に示す．本文の下にある色は付与された関係を表し，その濃淡は作業者の一致度

PRO	SUP	PRO_BY	SUP_BY
0.345	0.289	0.334	0.354

表 1　関係ごとのアノテーションの一致度（F 値）

記事数	1,000
文数	5,680
PRO ラベル数	5,937
SUP ラベル数	2,337
PRO_BY ラベル数	3,937
SUP_BY ラベル数	933

表 2　正解データの統計量（2 人以上一致）

を表している．脳膿瘍を引き起こすのは，「バクテリア」という判定が一番多く，次いで「バクテリアなどが侵入」「感染」など判定に迷う事例が続いているのが興味深い．また，脳膿瘍は「脳の組織の一部が壊死」を促進するが，その部分表現である「脳の組織の一部」を抑制するという入れ子が確認できる．ここから，促進から抑制へ極性を反転させる表現（ここでは「壊死」）を抽出することができる（4.2 節参照）．

4.1　アノテーションの一致度

このように構築した因果関係コーパスの質はどの程度なのか？　表 1 は，各記事に付与された 10 件のアノテーションの一致度の平均を計算し，関係の種類毎に示したものである．ここでは，2 つのアノテーション間の一致度として文字単位の F 値を採用し，アノテーションの全て（${}_{10}C_2 = 45$ 個）のペアの一致度をマイクロ平均で算出し，ある記事に付与されたアノテーションの一致度を算出している．アノテーションの一致度は 0.3 くらいであるが，タスクの難しさを考えると，妥当な数字である．

10 人の作業者の全ての作業結果を使うのではなく，n 人以上が一致している箇所のみを採用することで，アノテーションの一致度を高め，データの質を高めることができる．図 4 は，n 人以上のアノテーションが一致している箇所のみを取り出して「正解データ」を作成したとき，その正解データと元々の 10 件のアノテーション間の一致度のマイクロ平均を求めたものである．この図が示しているように，$n = 2$，すなわち 2 人以上のアノテーションが一致している箇所を取り出して正解データとした場合に，一致度が最も高くなった．そこで，以降の実験では $n = 2$ として得られたアノテーションを正解データとして使用する．表 2 に，この正解データの

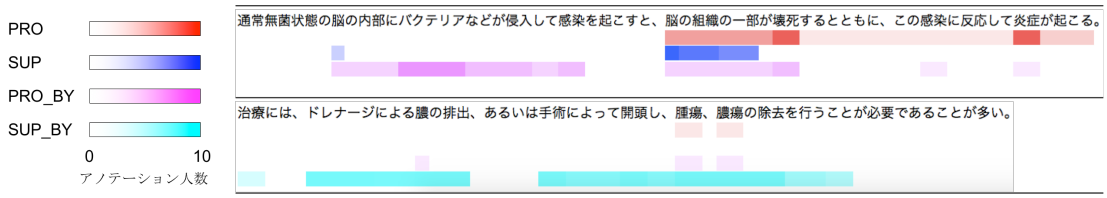

図 2　「脳膿瘍」の Wikipedia 記事に対するアノテーション結果の抜粋

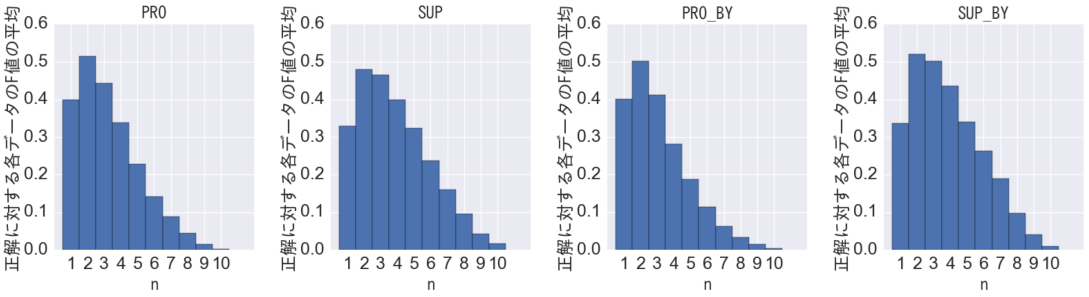

図 3　n 人以上が一致する箇所を正解データとした時のアノテーションの一致度

記事数，文数，各ラベル数を示す．

4.2　促進・抑制の入れ子構造

図 4 に示した例のように，句全体では促進関係にあるものの，句の一部である A には抑制関係があるといった入れ子の構造がしばしば出現する．$n = 2$ として正解データを作成したとき，促進と抑制の重なりは 5 つに場合分けができ，その内訳は表 3 のとおりであった．

圧倒的に多いのは，PRO が SUP を完全に含む事例で，「A の減少」などの極性反転がよく使われることを示している．その逆である SUP が PRO を完全に含む事例は，極性表現を二重に使う場合などに見られる．このような，促進と抑制の極性を反転させているパターンを抽出し，その出現頻度を測定したのが表 4 である．「A の低下」「A を防止」など，一見すると人手で作れそうなパターンが多いが，「A 炎」（胃腸炎）や「A 被害」（健康被害）など，文節内の極性反転などの興味深い事例も観察される．

4.3　アノテーション間違い

表 3 の完全一致の 115 件は，全て作業者のアノテーション間違いによるものであった．では，作業者はどの関係のアノテーションを取り違えやすいのだろうか？　ここでは，正解データをとその他の全てのデータを比較することで，アノテーション間違いの傾向を分析する．

正解データのアノテーション結果を事象 X，10 人全てのアノテーション結果を事象 Y とみなし，その事象間の独立性を分析する．分析には，χ^2 検定で用いられる観測値と期待値のずれを計算する式において，分子を二乗しないものを採用する．すなわち，以下の式を用いる．

$$\frac{\text{観測値} - \text{期待値}}{\text{期待値}} = \frac{\text{観測値}}{\text{期待値}} - 1 \qquad (1)$$

例えば，10 人全てのアノテーション（Y）において，PRO, SUP, PRO_BY, SUP_BY のラベルが付与された割合が 0.4, 0.3, 0.2, 0.1 で，正解データで PRO で付与された数が 600 件とする．アノテーションの間違いが，事象 Y の生起確率分布に従うと仮定すると，SUP, PRO_BY, SUP_BY のラベルが付与される期待値は，それぞれ 300, 200, 100 となる．ここで，間違いのみに注目しているため，PRO を除く 3 つの関係の比，すなわち $Sup : Pro\_by : Sup\_by = 0.3 : 0.2 : 0.1$ を用いて計算している点に注意する．このとき，正解データでは PRO になっているものが 10 人全てのアノテーションでは SUP になっていた事例が 200 件だった場合，式 1 の値は $200/300 - 1 = -0.333$ となる．これは，アノテーションの間違いが 10 人全てのアノテーションのラベルの分布の通りに発生すると仮定した場合と比べて，33.3% 少なかったことを表している．

このようにしてアノテーションの間違いを定量化したものが表 4.3 である．この結果から，PRO と PRO_FROM などの因果関係の向きの取り違えが多いこと，PRO と SUP のような因果関係の極性の取り違えは少ないことが分かる．

5　因果関係の自動認識

本研究で構築した正解データは，Wikipedia 記事からの因果関係知識獲得にどのくらい貢献するのか？本研究で構築した正解データを学習データとみなし，概要文中の単語に対して促進・抑制に関するラベルを予測するタスクを系列ラベリング問題として定式化した．4.2 節で説明したように，促進・抑制関係が重なって付与される箇所があるため，ラベルを予測するモデルを各関係ごとに構築した．系列ラベリングの手法として，双方向 LSTM を採用した．入力単語ベクトルと中間層の次元数はいずれも 300 に設定し，順方向と逆方向の LSTM を 1 層ずつ用いた．また，単語ベクトルは Wikipedia で訓練された単語ベクトル[*3]を用いて初期化した．因果関係に IOB2 記法を適用し，B-PRO, I-PRO, B-SUP, I-SUP などの 8 種類のラベルに展開した．概要文中の中に出てくるタイトルの単語は，すべて__TITLE__に置換し，括弧表現を削除した[*4]．本研究でアノテーションした 1,000 記事のうち，800 記事を学習データ，100 記事を開発

[*3] https://github.com/overlast/word-vector-web-api
[*4] Wikipedia の概要文では読み仮名を表すことが多い．

パターン	出現回数	例		
		文	Pro	Sup
Pro が Sup を完全に含む	1,467	血小板の減少を呈する	血小板の減少	血小板
Sup が Pro を完全に含む	45	不本意な結果を防ぐことに失敗	不本意な結果	不本意な結果を防ぐこと
Pro の左側に Sup の右側が重なる	68	鶏、兎、猫等の家畜が大量死	家畜が大量死	鶏、兎、猫等の家畜
Sup の左側に Pro の右側が重なる	0	-	-	-
Pro と Sup が完全に一致	115	四肢の麻痺が生じる	四肢の麻痺	四肢の麻痺

表 3　Pro，Sup のオーバーラップの統計

A 障害 (51), A の低下 (24), A 低下 (16), A 異常 (12),
A が低下 (9), A 減少 (8), A が障害される (6), A の減少 (6),
A 炎 (6), A を防止 (5), A の軽減 (5), A が困難 (5),
A 防止 (5), A 失調 (5), A 制限 (5), A が損なわれる (4),
A 欠損症 (4), A を補償 (4), A を無視 (4), A の解消 (4),
A が萎縮 (4), A 被害 (4), A 汚染 (4), A 放棄 (4),
A 対策 (4), A 困難 (4), A 不全 (4), 防 A(3),
A を最小限に抑える (3), A を防いだ (3), A を減らす (3),
A を最小限 (3), A への影響 (3), A に悪影響 (3),
A 欠乏症 (3), A を軽減 (3), A を排除 (3), A を拒否 (3),
A の麻痺 (3), A の防止 (3), A の悪化 (3), A の復興 (3),
A の変性 (3), A の代替 (3), A の不全 (3), A に障害 (3),
A が阻害 (3), A 阻害 (3), A 遅滞 (3), A 疾患 (3),

表 4　極性反転表現の出現回数上位 50 件

		間違って付けられた関係			
		Pro	Sup	Pro_by	Sup_by
正解の関係	Pro	-	-0.510	0.425	0.019
	Sup	-0.612	-	-0.405	1.037
	Pro_by	0.556	-0.198	-	-0.567
	Sup_by	-0.222	0.969	-0.670	-

表 5　関係ごとのアノテーション間違い数の期待値からのずれ (割合)

データ，100 記事をテストデータとして用いた．

ラベル毎の F スコア（括弧内数字）は，Pro (0.424), Sup (0.310), Pro_by (0.397), Sup_by (0.211) であった．図 4 に示した通り，人間がアノテーションをしても一致度（F スコア）は 0.5 程度であったことから，現状の自動認識性能は比較的高いと考えている．

6　おわりに

本論文では，Yahoo!クラウドソーシングと brat の連携により，コーパスに関係知識を付与する作業をクラウドソーシングで完結させる方法を提案した．この手法を利用し，Wikipedia の概要文に対して促進・抑制の関係事例を付与する作業を依頼し，コーパスを構築した．促進・抑制の関係事例を付与する場合は，付与対象の単位や複数の正解を絞り込む基準を明確に与えることができないが，そのようなタスクでもクラウドソーシングを活用し，比較的質の高いコーパスを構築することができた．構築したコーパスを用いて，促進と抑制の入れ子現象，極性反転表現，双方向 LSTM による自動認識の性能など，有用な知見を得ることができた．今後は，アノテーションの一致度を高めるための基準を検討しながら，コーパスの規模を大きくしたいと考えている．

謝辞

本研究は，文部科学省科研費 15H01702，15H05318，および JST，CREST の支援を受けたものである．

参考文献

[1] K. Fort, G. Adda, and K. B. Cohen, “Amazon Mechanical Turk: Gold mine or coal mine?” *Computational Linguistics*, pp. 413–420, 2011.

[2] D. Hovy, B. Plank, and A. Søgaard, “Experiments with crowdsourced re-annotation of a POS tagging data set,” in *Proc. of ACL 2014*, 2014, pp. 377–382.

[3] M. Jha, J. Andreas, K. Thadani, S. Rosenthal, and K. McKeown, “Corpus creation for new genres: A crowdsourced approach to PP attachment,” in *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 13–20.

[4] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, “Annotating named entities in Twitter data with crowdsourcing,” in *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 80–88.

[5] N. Lawson, K. Eustice, M. Perkowitz, and M. Yetisgen-Yildiz, “Annotating large email datasets for named entity recognition with mechanical turk,” in *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 71–79.

[6] S. Takase, N. Okazaki, and K. Inui, “Composing distributed representations of relational patterns,” in *Proc. of ACL 2016*, 2016, pp. 2276–2286.

[7] A. Brew, D. Greene, and P. Cunningham, “Using crowdsourcing and active learning to track sentiment in online media,” in *Proc. of ECAI 2010*, 2010, pp. 145–150.

[8] M. R. Gormley, A. Gerber, M. Harper, and M. Dredze, “Non-expert correction of automatically generated relation annotations,” in *Proc. of NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 204–207.

[9] D. Kawahara, Y. Machida, T. Shibata, S. Kurohashi, H. Kobayashi, and M. Sassano, “Rapid development of a corpus with discourse annotations using two-stage crowdsourcing,” in *Proc. of COLING 2014*, 2014, pp. 269–278.

[10] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, “brat: a web-based tool for NLP-assisted text annotation,” in *Proc. of EACL 2012 (demonstrations)*, 2012, pp. 102–107.

[11] J. Fluck, S. Madan, T. R. Ellendorff, T. Mevissen, S. Clematide, A. van der Lek, and F. Rinaldi, “Track 4 overview: Extraction of causal network information in biological expression language (BEL),” in *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, 2015, pp. 333–346.

[12] C. Hashimoto, K. Torisawa, S. De Saeger, J.-H. Oh, and J. Kazama, “Excitatory or inhibitory: a new semantic orientation extracts contradiction and causality from the web,” in *Proc. of EMNLP-CoNLL 2012*, 2012, pp. 619–630.

[13] J.-H. Oh, K. Torisawa, C. Hashimoto, R. Iida, M. Tanaka, and J. Kloetzer, “A semi-supervised learning approach to why-question answering,” in *Proc. of AAAI-16*, 2016, pp. 3022–3029.

[14] A. Sasaki, J. Mizuno, N. Okazaki, and K. Inui, “Stance classification by recognizing related events about targets,” 2016, pp. 582–587.