論文間参照情報のアノテーションにおけるクラウドソーシングの利用検討

井上 絢翔　　韓 東力
日本大学大学院 総合基礎科学研究科

1. はじめに

論文を執筆する上で論文サーベイは重要なタスクの 1 つである。論文サーベイによく利用されているツールとして、CiNii[1]や Google Scholar[2]、CiteSeerX[3]などの電子図書館が挙げられるが、キーワード検索がいまだに主流となっており、論文間の関連性に基づく検索がほとんど実現されていない。

論文サーベイを行う際によく利用されている手法の一つに、起点となる論文(以下「起点論文」と呼ぶ)を 1 つ選定し、その論文が参照している論文や、さらに参照論文が参照している論文という順にサーベイの対象を広げていくという方法がある。我々はこの方法に注目し、起点論文とそれが参照している論文(以下「被参照論文」と呼ぶ)間の参照関係を利用した検索が行える電子図書館を新たに構築することを考えた。

論文間の参照関係の付与には機械的に行うものと手動によるものがある。前者に関するものとして難波ら[4]、小出ら[5]と Teufel ら[6]の研究がある。難波らは論文間の参照タイプを 3 種類に分類しているが、効率的な論文サーベイを行うのに分類数が不十分と思われる。小出らと Teufel らの研究では、それぞれ 9 種類と 12 種類の参照理由を定義し機械学習を用いて参照理由を付与しているが、精度は最大で 60%台に留まっている。論文間関係の解明を最終目標とするような研究では上記の精度でも一定の有効性があるかもしれないが、自動付与された参照理由を異なる目的で再利用する場合には、連鎖的誤りを回避するためにはより正確な分類結果が必要であろう。

手動付与ではより高精度のアノテーションを行うことができるが、時間がかかることや論文サーベイに精通している専門家を雇うのに多大なコストがかかることなどが問題点としてあげられる。

そこで、クラウドソーシングを利用することによりコストの問題に対処できるのではないかと考えた。以下第 2 章では本研究における基本的考え方、第 3 章ではアノテーションの基本方針、第 4 章では評価実験の結果について述べる。

2. 研究手法

クラウドソーシングとはインターネット上の不特定多数の作業者に仕事を依頼する雇用形式で、低コストで迅速な作業が可能である。本研究では、論文間参照情報のアノテーションにおけるクラウドソーシングの利用可能性について検討する。

論文間参照情報のアノテーションは比較的難易度の高いタスクであるため、不特定多数の作業者がどの程度遂行できるのか、また専門家と比べるとどのような差があるのかなど大きな不安がある。

本研究では上記のような検討課題を念頭に、まずは論文間参照情報をアノテーションするためのプロトタイプを構築した。次に、論文サーベイに精通している大学教員を専門家に、大学生をクラウドソーシングで働く一般作業者に見立て、構築されたプロトタイプを利用してアノテーションしてもらった結果を比較した。この過程を通じて論文間参照情報のアノテーションにクラウドソーシングを利用する可能性を検討していく。

3. アノテーション

小出らや Teufel らの研究[5][6]では、タグの種類はどちらでも単一階層で 10 種類前後と定められている。これを本研究で採用すると一般作業者によるアノテーションの難易度が高くなってしまうため、構築されたプロトタイプでは図 1 のようにタグの階層および種類を採用している。

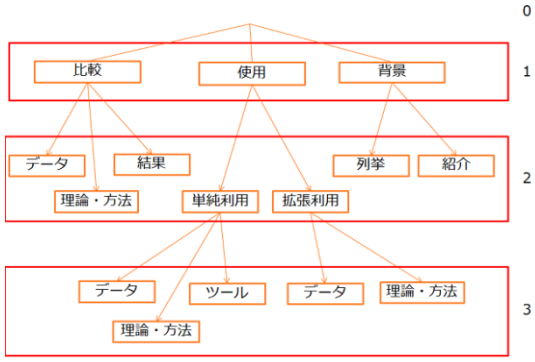

図 1 論文間参照関係の分類

まず第 1 階層に「背景」、「利用」と「比較」の三つのタグがあり、どれかを選択するとそれに応じた第 2 階層、そして第 3 階層の候補タグが画面に順次表示されていく。このように少ない選択肢を段階的に与えることでタグの網羅性を保ちつつ、アノテーションの難易度低下を狙う。

論文を引用するにあたり、「利用」や「比較」のために参照を行う場合、事前に被参照論文に関する簡単な紹介が行われたケースが多い。このような場合、たとえ「利用」や「比較」を行っていたとしても同時に「背景」も当てはまってしまうため、「背景」を選択する優先度を下げ、被参照論文を紹介する程度にとどまっている場合のみ選択できるようにする。

また、1 つの起点論文と 1 つの被参照論文の間に複数の参照が存在するケースがある。このような複数回参照を行っている場合は、「被参照論文の理論・方法を拡張利用して、最終的に起点論文の提案手法と比較した」といったように複数の参照理由を付与できるケースが起こり得る。そのため、複数回参照を行っている論文のみ、複数のタグを選択できるものとする。

4. 評価実験

論文サーベイに精通している教員 3 人と情報科学を専門とする大学生 13 人を対象に論文間参照情報のアノテーションを行ってもらった。

言語処理学会年次大会発表論文集に掲載されているものから異なる難易度で 3 編の起点論文と 10 編の被参照論文を選んだ。クラウドソーシングに近い条件で実施するため、実験用プロトタイプはウェブシステムとして構築されている。作業の時間や場所を選ぶ必要がなく、インターネットにアクセスできる端末があればどこでも作業可能であった。

表 1 ルートの深さによる類似度

	論文A-1	論文A-2	論文A-3	論文A-4	論文A-5	論文B-1	論文B-2	論文B-3	論文C-1	論文C-2	平均	平均	中央値
学生1	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.243	0.500	0.174		
学生2	0.500	0.000	0.000	0.000	0.000	0.000	0.500	0.000	0.444	0.000	0.144		
学生3	0.250	0.667	1.000	0.667	1.000	0.000	0.500	0.833	0.000	1.000	0.592		
学生4	0.417	0.000	1.000	0.000	0.000	0.000	0.500	0.300	0.500	0.000	0.272		
学生5	0.500	1.000	1.000	1.000	1.000	0.750	0.500	0.833	0.444	1.000	0.803		
学生6	0.500	1.000	1.000	1.000	1.000	0.475	0.500	0.000	0.233	0.000	0.571		
学生7	0.500	1.000	1.000	1.000	1.000	0.750	1.000	0.000	0.256	1.000	0.751	0.570	0.628
学生8	0.500	1.000	1.000	1.000	1.000	0.500	1.000	0.000	0.444	0.500	0.694		
学生9	0.333	0.333	1.000	1.000	0.500	0.500	0.500	0.667	0.444	1.000	0.628		
学生10	0.500	1.000	1.000	1.000	1.000	0.750	0.500	0.889	0.394	0.500	0.753		
学生11	0.350	0.667	1.000	0.667	1.000	0.750	0.500	0.000	0.333	1.000	0.627		
学生12	0.500	1.000	1.000	1.000	1.000	0.750	1.000	0.000	0.444	0.500	0.719		
学生13	0.000	1.000	1.000	1.000	1.000	0.000	1.000	1.000	0.314	0.500	0.681		
教員1	0.500	1.000	1.000	1.000	1.000	0.750	1.000	0.833	0.394	0.500	0.798		
教員2	0.000	1.000	1.000	1.000	1.000	0.750	0.500	0.833	0.000	1.000	0.708	0.742	0.719
教員3	0.250	1.000	1.000	1.000	1.000	0.500	1.000	0.000	0.444	1.000	0.719		
平均	0.350	0.729	0.875	0.771	0.781	0.452	0.719	0.387	0.333	0.625			

アノテーションの結果は事前に作成された模範解答との類似度を計算することにより評価された。模範解答については、著者らと実験に参加していなかった数名の大学生が検討を重ねて作成した。結果の評価方法は以下の 2 つである。

4.1 ルートの深さによる類似度

選択されたタグを図 1 のように樹系図に見立て、その深さにより類似度を計算する。計算式は次式で示す。

$$\frac{2 * D_{ab} + F}{D_a + D_b}$$

D_a、D_b はそれぞれアノテータ(a)が付与したタグの深さと模範解答(b)のタグの深さで、D_{ab} は a と b の共通の深さである。また D_a と D_b のルートが完全に一致していた場合は類似度は 1 とする。F は a と b のラベル名が一致していた場合に 1、そうでない場合は 0 で計算する。これは「比較」と「利用」もしくは「単純利用」と「拡張利用」を間違えた場合でも、ラベル名が一致していたらより近い回答として見ることができるという考えに基づいたためである。

複数回参照を行っている論文はアノテータの選択した複数のタグと模範解答にある複数のタグに対してそれぞれ、最も類似度の高いタグ対を求め、それらの平均をその論文の類似度とする。

4.2 タグの数による一致率

4.1 のルートの深さによる類似度計算は複数回参照の場合に最も模範解答に近いタグとの類似度を求めているため、公平性に欠けている。ここでは次式のように、選ばれた全てのタグに対して評価を行う。

$$\frac{2 * N(a \cap b)}{N(a) + N(b)}$$

N(a)はアノテータ(a)が選択したタグの数、N(b)は模範解答(b)のタグの数、N(a∩b)はアノテータ(a)と模範解答(b)が共通して選択したタグの数を表す。4.1 の評価方法と違い、N(a∩b)はタグが完全一致しているものの数である。

4.3 結果

表 1 は 4.1 のルートの深さによる類似度計算の結果である。表の「A」、「B」と「C」はそれぞれ起点論文を表しており、それぞれが参照している論文が「論文 A-1」、「論文 A-2」・・・となっている。「論文 A-1」、「論文 B-1」、「論文 B-3」と「論文 C-1」は同じ起点論文から複数回参照された論文である。大学生の平均は教員の平均に及ばないものの、最高得点は学生が出しているこ

とがわかる。また、教員の中央値である 0.719 を超えた学生は 4 人もいることから、論文サーベイにそれほど精通していない学部生でも良質なアノテーションを行うことが可能ではないかと考えられる。

しかし、複数回参照の結果では、教員・学生ともに一回参照の半分程度に留まっている。これは選択の幅が大きく広がったことによる難易度の上昇が原因であるとみられる。複数回参照を除いて平均を計算したところ、学生が 0.705、教員が 0.944 であった。

さらに、類似度が明らかに低い学生(1・2・4)については、選択したタグの内容にシステムの初期状態で設定されている「背景」がほとんどを占めていることがわかる。

表 2 は 4.2 で述べたタグの数による一致率の計算結果である。

表 2 タグの数による一致率

		平均	中央値
学生 1	0.143		
学生 2	0.231		
学生 3	0.320		
学生 4	0.375		
学生 5	0.769		
学生 6	0.444		
学生 7	0.643	0.488	0.462
学生 8	0.615		
学生 9	0.385		
学生 10	0.643		
学生 11	0.462		
学生 12	0.692		
学生 13	0.621		
教員 1	0.690		
教員 2	0.538	0.632	0.667
教員 3	0.667		

教員の平均値である 0.632 を超えた学生は 4 人であり、最高得点はやはり学生の方が出している。また、ほぼシステムの初期状態でアノテーションを終えた学生を外して計算し直したところ学生の平均値は 0.611、中央値は 0.632 まで向上した。

5. 終わりに

本研究では論文間参照情報のアノテーションをクラウドソーシングにより実施する可能性を検討するため、独自で開発したアノテーションツールをクラウドソーシングに近い状態で利用してもらい、その評価を行った。

評価の結果により、アノテーション方法の簡易化や悪質アノテーターの事前排除、複数回参照の単一化などの点おいて工夫すれば、論文サーベイにそれほど精通していない一般作業者でも専門家に近い、良質なアノテーションを行う可能性が十分あることが示唆された。

また、今回の評価実験では情報系の大学生を対象にしているが、今後は他分野の作業者にもアノテーションしてもらい、作業者の分野がもたらした影響を分析する。

参考文献

[1] http://ci.nii.ac.jp/
[2] http://scholar.google.co.jp
[3] http://citeseerx.ist.psu.edu/
[4] 難波英嗣，神門典子，奥村学，「論文間の参照情報を考慮した関連論文の組織化」，情報通信学会論文誌，42(11)，pp.2640-2649. (2001)
[5] 小出寛史，韓東力「論文間参照情報のデータベース化に基づく参照タイプの同定」，自然言語処理研究会報告 2012-NL-209(2), 1-7(2012)
[6] Teufel, S. The Structure of Scientific Articles –Applications to Citation Indexing and Summarization. CSLI Publications. (2010)