sentence2vec を用いた学術論文の引用意図分類手法の検討

Examination of a citation intention classification method for academic papers using sentence2vec

吉次 優
Yu Yoshitsugu

岡山大学 太田研究室
Ohta Laboratory, Okayama University

概要 学術論文が参照する論文は一般に複数あるが，それらすべてを確認することは困難である．そこで学術論文閲覧支援を目的として，閲覧論文中の引用箇所から手掛かり語により引用意図を推定し，被引用論文中の適切な被引用箇所を特定する手法が提案されている．しかし，引用意図の分類精度が十分でなかったため，本研究では sentence2vec により引用箇所をベクトル化し，これらを用いた引用意図分類について検討する．

1 はじめに

学術論文では多くの場合，その研究の根拠や用いた手法に関する論文を引用する．そのため，被引用論文を読むことで閲覧中の論文をより深く理解することができる．しかし，被引用論文は 1 つの論文に対して複数あることが多く，それらすべてを引用意図に応じて閲覧するのは論文の読者にとって大きな負担である．

この支援に関する研究として，被引用論文から適切な被引用箇所を特定し，論文閲覧者に提示する方法が提案されている．石井らの方法[1]では，まず閲覧論文中から引用箇所を特定し，人手で収集した手掛かり語を用いて引用意図を分類する．次に，分類した引用意図に適すると考えられる被引用論文中の節を限定する．限定した節の各文に含まれる語と，引用箇所に含まれる語や人手で収集した手掛かり語とを照合し，一致した語数の多かった文とその周辺を適切な被引用箇所として提示している．

本研究では，引用箇所の引用意図分類について，石井らの手法の課題であった汎用性の問題について改善を検討する．

2 石井らの引用意図分類

2.1 引用意図の分類クラス

石井ら[1]は，引用意図の分類クラスを“Group”，“Method”，“Result”，“Data”，“Equation”，“Other”の 6 種類と定義した．これらは NTCIR-9[2]の論文を分析して定められたもので，詳細は以下の通りである．

(1) Group
タスクやフォーラム，ワークショップの詳細が引用したい内容である場合．

(2) Method
著者が研究で使用する手法，または著者の手法の比較対象となる既存手法等が引用したい内容である場合．

(3) Result
既存研究の実験結果が引用したい内容である場合．

(4) Data
既存研究で用いられた実験データなどが引用したい内容である場合．ただし，論文中に使用したデータの件数が述べられているもの．

(5) Equation
計算式の詳細が引用したい内容である場合．ただし，計算式そのものが引用されているもの．

(6) Other
上記 5 つのどのクラスにも当てはまらない場合．

2.2 分類方法

石井らの分類方法では，NTCIR-9 の論文から人手で収集した手掛かり語を用いて，以下のような手順で引用箇所を分類する．ここで，引用箇所とは論文中で引用を表す文字列を持つ文と定義する．

(1) Equation の手掛かり語が 1 つ以上あれば Equation に分類する．

(2) (1)の条件を満たさず，かつ Data の手掛かり語が 1 つ以上あれば Data に分類する．

(3) (1)と(2)の両方を満たさず，かつ Group，Method，Result のいずれかの手掛かり語が 1 つ以上あれば，最も手掛かり語の数が多いクラスに分類する．

(4) (1)～(3)を全て満たさなかった場合 Other に分類する．

2.3 石井らの手法の問題点

石井らの手法では，引用箇所を 2.1 節で挙げた各クラスに分類する際，人手で収集した手掛かり語を用いている．また，引用箇所の分類クラスごとの件数の偏りを考慮して定められており，汎用性が高いとは言えない．この方法では分類対象の論文が変わった場合の対応が難しいため，極力人為的な作業を減らす方が望ましい．

3 引用意図分類方法

3.1 分類に用いるデータ

本稿では実験のため，NTCIR-9 に投稿された論文の内，GeoTime タスクの論文 9 件から 76 文，INTENT タスクの論文 12 件から 119 文，SpokenDoc タスクの論文 9 件から 78 文の合計 273 文の引用箇所を抽出した．まず論文 PDF ファイルを pdftotext[3]を用いて TXT ファイル化する．この TXT ファイルから「[1]」のような表現を含む文を抽出して引用箇所とし，それぞれ 2.1 節のクラスに手動で分類し，学習とテストに用いるデータとする．この中で，Other 以外の 5 種類のいずれかに分類された 236 文を sentence2vec[4]を用いて

ベクトル化する．各引用箇所の引用意図の内訳を表 1 に示す．また，今回 sentence2vec の学習に用いたデータは以下の 3 種類である．

① NTCIR-9 の論文 102 件
② ①の 102 件及び NTCIR-10[5]の論文 104 件の計 206 件
③ ②の 206 件及び NTCIR-8[6]の論文 65 件の計 271 件

ベクトルの次元は 10～100 まで 10 刻みと 200，300，500 の 13 種類とした．

3.2 重心ベクトルを用いた分類

引用箇所の引用意図を分類する．まず，分類対象とする引用箇所 1 件を除いた 235 件の各引用箇所のベクトルから引用意図ごとに重心ベクトルを算出する．この重心ベクトルと，分類する引用箇所のベクトルとのコサイン類似度を計算し，最も値の大きい引用意図クラスに対象の引用箇所を分類する．

4 引用意図の分類実験

3.2 節で示した方法を用いて引用意図の分類実験を行った．3.1 節で用意した 236 文の引用箇所を分類対象として分類した．学習データごとの正解率を図 1 に示す．sentence2vec ではベクトル化の際に乱数が用いられており，学習の度に出力結果が異なるため，今回は次元ごとに 10 回ずつベクトルを出力して平均し，そのベクトルを引用箇所のベクトルとした．図 1 より，どの次元においても NTCIR-9 のみを学習データとした①よりも，他の論文を加えた②や③の方が正解率は高くなっている．図 1 で②と③を比べると大きな差は無いが，正解率の平均では 2 ポイントほど③が上回っている．

次に，次元数による結果の変化を確認するため，学習データを③とした場合の引用意図の分類クラスごとの分類の F 値を図 2 に示す．図 2 より，次元数による F 値の大きな変化は見られないが，低次元（10～30 次元）よりはある程度高次元（80～100 次元）の方が F 値は概ね高くなっており，これ以上に次元を上げても F 値の向上には繋がらないという結果になった．また，各引用意図クラスに属する引用箇所の件数と F 値に相関が見られる．実際の分類クラスを確認するため，F 値の平均が最も高かった 80 次元の場合の分類結果の内訳を表 2 に示す．表 2 より，Equation 以外の 4 クラスにおいては過半数が正しく分類されており再現率は高い．しかし，件数の多い Method の引用箇所を誤って分類していることが多く，適合率が下がっている．今後は重心ベクトルの扱い方を検討し Method の再現率を上げるとともに，1 件も正しく分類されていなかった Equation についての改善を検討する．

5 まとめ

本稿では，学術論文における引用箇所の引用意図分類について述べた．今後は引用箇所の少ない引用意図を持つ引用箇所の追加や他の方法での分類を検討し，精度の向上を図る．

表 1 各引用意図の引用箇所の内訳

引用意図	Group	Method	Result	Data	Equation
件数	24 件	179 件	22 件	6 件	5 件

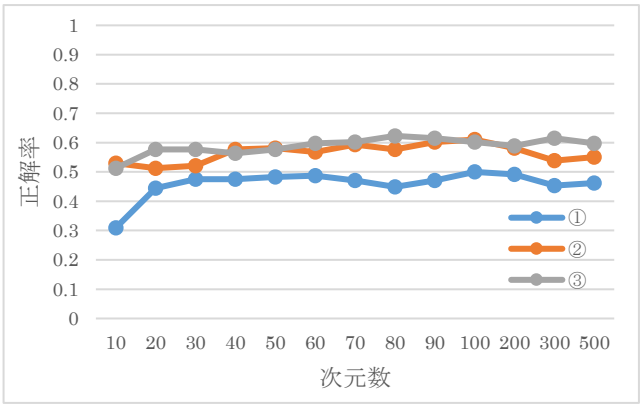

図 1 学習データごとの正解率

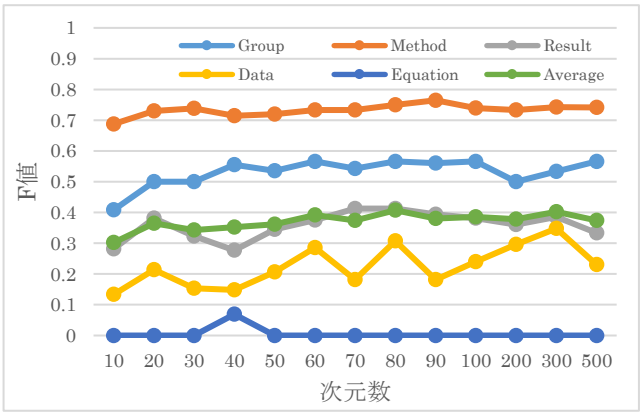

図 2 引用意図ごとの分類の F 値

表 2 分類された引用意図と件数（80 次元）

		分類されたクラス				
		Group	Method	Result	Data	Equation
正しいクラス	Group	15 件	3 件	1 件	2 件	3 件
	Method	12 件	115 件	26 件	12 件	14 件
	Result	1 件	6 件	13 件	2 件	0 件
	Data	0 件	1 件	0 件	4 件	1 件
	Equation	1 件	3 件	1 件	0 件	0 件

参考文献

[1] 石井仁子，太田学，高須淳宏，“引用意図を利用した学術論文閲覧支援のための適切な被引用箇所の特定”，第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2015)，F3-5，2015.
[2] NTCIR-9 : http://reserch.nii.ac.jp/ntcir/ntcir-9/
[3] Xpdf : http://www.foolabs.com/xpdf/index.html
[4] klb3713/sentence2vec · GitHub : https://github.com/klb3713/sentence2vec
[5] NTCIR-10 : http://reserch.nii.ac.jp/ntcir/ntcir-10/
[6] NTCIR-8 : http://reserch.nii.ac.jp/ntcir/ntcir-8/