ISS-SP-048

学術論文の被引用文章生成の一手法

田邊　俊介　　太田　学
岡山大学大学院自然科学研究科

1. はじめに

学術論文では研究の根拠などが書かれた文献を参考文献として引用する．そのためこれらを読むことは論文理解の助けとなるが，読者にとって負担でもある．そこで本稿では，引用箇所との類似度に基づき被引用論文中の重要な文を被引用箇所として特定し，それをもとに被引用文章を生成する手法を提案する．

2. 引用箇所の特定

本稿では，論文中の引用を示す“[12]”のような表現を含む文を引用文とする．そして，文の長さや含まれている単語から引用文とその前後の文が関係しているかを推定し，条件を満たす文を引用文と結合して引用箇所として特定する[1]．

3. 被引用箇所の特定

本稿では，[1]と同様に doc2vec[2]により文の特徴ベクトルを生成し，引用箇所と被引用論文の各文のコサイン類似度を算出する．そして，引用箇所との類似度が高い方から合計で100単語を超えるまで被引用論文中の文を選択し，それを被引用箇所とする．

本稿ではまた，要約アルゴリズムである LexRank[3]に，引用箇所が重要であるという情報を加えた手法により被引用箇所を特定する．まず，引用箇所と被引用論文中の各文をノードとし，それらの間の類似度をエッジの値とした無向グラフを作成する．そして，引用箇所と被引用論文の各文を繋ぐエッジの閾値を0.1，その他のエッジの閾値を0.2として，閾値以下のエッジを枝刈りする．その後，そのグラフで PageRank[4]を計算して，被引用論文中から高い固有値を持つ文を選択し，合計で100単語程度の被引用箇所を決定する．

4. 被引用文章生成手法

特定した被引用箇所の文から以下の方法で被引用文章を生成する．

まず，Stanford CoreNLP[5]の共参照解析の結果に基づき，被引用箇所中の参照表現(mention)を代表参照表現(representative mention)に置換する．

そして，学術論文は明確な順序で文章が展開されるという考えに基づき被引用論文中の出現順に文を並び替える．ただし，論文の要約を表す梗概やまとめから得た被引用箇所は文章の先頭に移動し，複数の文に共参照の関係にある表現が出現した場合は，後に出現した文を先に出現した文と連なるように並び替える．

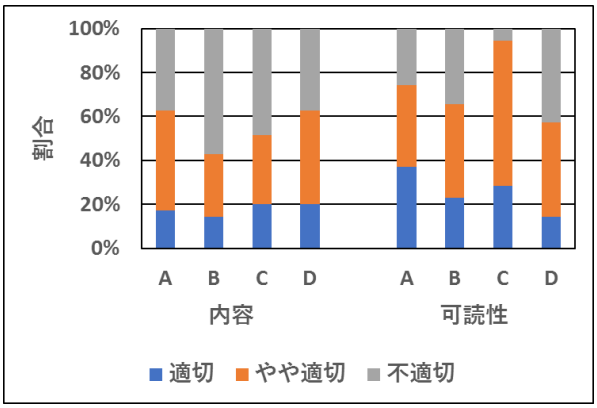

図1：各文章の内容および可読性の評価

5. 被験者アンケートによる評価

NTCIR-13[6]の5件の論文から1箇所ずつ選んだ合計5箇所の引用箇所に対して，被引用論文のアブストラクトから人手で選んだ3文の文章(A)，[1]の特定手法により得た文章(B)，提案手法である，文ベクトルのコサイン類似度，またはLexRankにより特定した被引用箇所をもとに生成した被引用文章(それぞれ(C)，(D))の4種類の被引用文章を生成し，大学院生5名と大学生2名の計7名を対象とした被験者アンケートにより評価した．

図1は5箇所の引用箇所に対して被引用文章として提示した各文章の内容および可読性を被験者が3段階で評価したアンケートの集計結果である．内容に関しては D と A がほぼ同じ評価，可読性については C が最も「不適切」が少なくなった．

また，各被引用文章を1文ごとに分割し，各文に対して内容を○，△，×の3段階で被験者に評価させ，○を1，△を0.5，×を0 とし，その平均を各被引用文章のスコアとして算出した．その結果，A が0.33，B が0.21，C が0.39，D が0.37となり，提案手法である C と D が他と比較してわずかに高くなった．

6. まとめ

本稿では，学術論文の引用箇所に対する被引用文章の生成手法を提案した．

参考文献

[1] 田邊ほか，DEIM Forum 2018，G4-5，2018.
[2] Q. Le，*et.al.*，CoRR，abs/1405. 4053，pp. 1-9，2014.
[3] G. Erkan，*et.al.*，JAIR，no. 1，pp. 457-479，2004.
[4] L. Page，*et.al.*，Technical report，Stanford InfoLab，1999.
[5] Stanford CoreNLP: https://stanfordnlp.github.io/CoreNLP
[6] NTCIR-13: http://research.nii.ac.jp/ntcir/ntcir-13