

係り受け構造と概念の近さをを用いた論文の被引用箇所の抽出方法の提案

芳賀 康敏[†]荒井 正之[†]帝京大学大学院理工学研究科[†]

1. はじめに

学生レポートの自動採点システムの開発を目指している。本研究では、システムのコアな機能の 1 つである、引用文献の整合性の判定を目的とする。本稿では、その前段階として、論文の被引用箇所の抽出方法を提案する。被引用箇所の抽出とは、ある論文（引用論文）が引用している論文（被引用論文）中のどこを引用しているかを抽出することである。提案手法は以下に示す 2 つの処理で構成される。(1)引用論文の引用箇所からキーワードの抽出、キーワードを用いた被引用論文中から候補の選出。(2)引用の際によく行われる、文の並べ替えと、単語の言い換えに対応するため、係り受け構造と単語の概念の近さをを用いた候補の絞込み。情報科学系の論文を用いて実験を行った結果、4 位累積抽出率は約 90%となった。

2. 提案手法

2.1 tf・idf によるキーワード抽出と被引用箇所候補の抽出

被引用箇所候補の抽出のため、tf・idf によるキーワード抽出を行う。

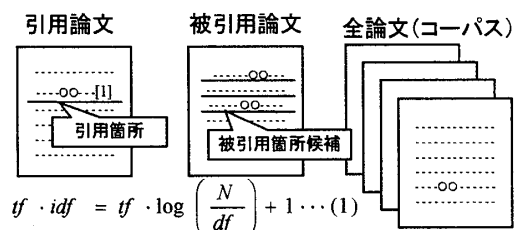


図 1 tf・idf の計算方法

図 1 中の $\circ\circ$ はキーワード、tf は被引用論文中でのキーワードの出現数、df はキーワードを含む論文数、N は全論文数を表す。

本研究では、引用番号を含む一文を引用箇所とし、引用箇所に存在する単語で、被引用論文中での tf・idf 値が最大のものをキーワードとする。決定したキーワードで被引用論文中を検索し、キーワードを含む部分を被引用箇所とする。

「An extraction method of cited sentences in technical reports using distance of dependency structure and words concept.」

[†]Yasutoshi HAGA, Masayuki ARAI, Graduate School of Science and Engineering, Teikyo University

被引用箇所の抽出範囲は、キーワードを含む「一文」、「複数文」、「一段落」、の 3 つで検討を行った。抽出した複数の被引用箇所を被引用箇所候補とし、以下の方法で候補を絞込む。

2.2 係り受け構造と単語の概念の近さをを用いた候補の絞込み

(1) 係り受け構造の近さ

文の係り受け関係を取得するため、引用箇所と被引用箇所候補を係り受け解析[1]し、解析結果を木構造として近さ求める。2 つの木の近さは Collins らの提案する Tree Kernel[2]で算出する。

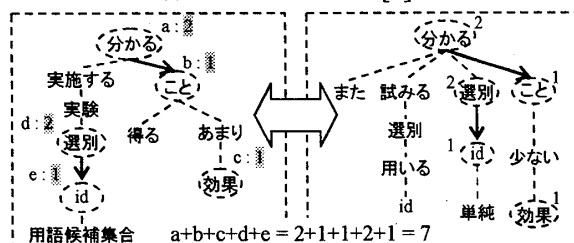


図 2 Tree Kernel の計算例

具体的には、2 つの木の一致するノードの近さを合算した値が木構造の近さとなる。

例えば、図 2 の「効果」のような子の一致が無いノードは近さが 1、「分かる」のような、自ノードと、その子「こと」の一致があるノードは、親に子の近さを足し、近さが 2 となる、このように一致するノード a-e を合算した値 7 が 2 つの木の近さとなる。

(2) 単語概念の近さ

同義語間の距離が登録されている EDR 概念辞書[3]を用い、単語の概念距離を用いて近さを求める[4]。2 つの単語 i, j の概念の近さ R_{ij} は、i, j のルートからの距離を d_{ix} , d_{jy} , i, j 共通の上位概念の深さを d_{xy} としたとき、(2)式で求まる。

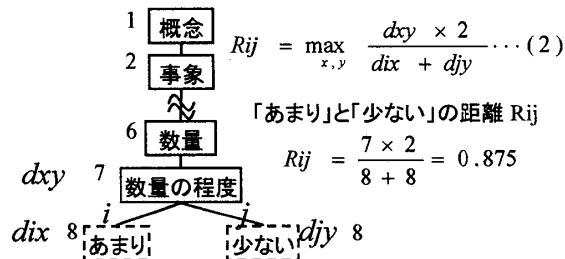


図 3 単語概念の実例と近さの計算式

(3) (1) (2)の組合せによる候補の絞り込み

本研究では、Tree Kernel に概念の近さを付加する手法を提案する。例えば、図 2 の場合、「あまり」と「少ない」は、単語が異なるため、ノードの一致が取れない。ここに図 3 で示した概念の近さを付加する。これにより、「こと」の子と「効果」の親の一致が現れ、この場合の近さは図 4 に示すように 9.75 となる。

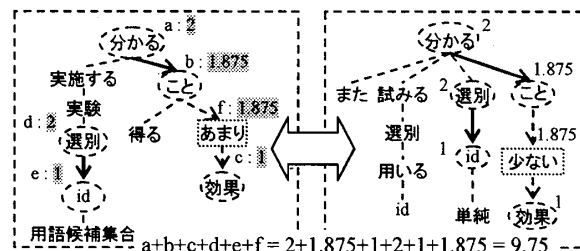


図 4 Tree Kernel と概念辞書の組合せの計算例

3. 実験結果と考察

次の条件で提案手法の評価実験を行った。

- ・ 情報科学系 11 論文を用意し、18 の引用箇所それぞれに対応する被引用論文の中から、被引用箇所候補を選出する。
- ・ 提案手法、レポートの剽窃判定などに多用されている $tf \cdot idf$ [5]、3-gram[6]の 3 手法で被引用箇所の N 位累積抽出率を求める。

なお、2 章 1 節で述べた被引用箇所候補について、検討を行ったところ、最も良い結果を得た「キーワードを含む段落」を抽出範囲とした。

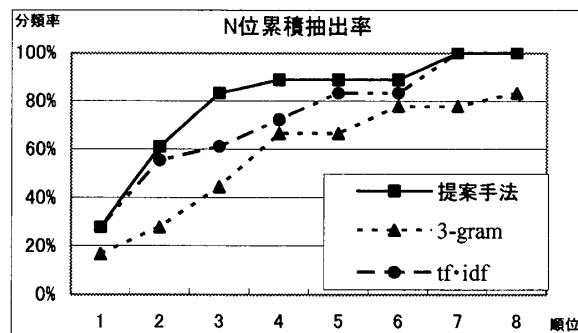


図 5 N 位累積抽出率

図 5 に示すように、1~7 位累積抽出率において提案手法が最も良い結果を得ることができた。

3-gram は引用箇所と被引用箇所から 3 文字単位で区切った文字列を作り、両者に存在する各文字列の出現頻度で 2 つの文書と比較する。これにより「です、ます調」、「だ、である調」の書き換え程度ならば対応できると考えられる。また、特徴として、比較する 2 つの文の長さに差が無いほど類似していると評価する傾向がある。しかし、一般に論文では引用箇所よりも被引用箇所が長くなるため、抽出率が下がったと考えられる。3 手法の中で抽出率が最も悪かった。

$tf \cdot idf$ は、引用箇所と被引用箇所の特徴的な単語（特徴語）を $tf \cdot idf$ により抽出し、その単語で生成されるベクトル空間の cosine 尺度で比較する。今回の実験では比較する文の長さが短いため、特徴語の tf が大きくならなかったことと、コーパスとなる論文数が不足していたため、特徴語とそうでない単語との idf にあまり差が出なかったことから、どの単語の $tf \cdot idf$ 値もほぼ同じ値になっていた。これにより、特徴語による比較を行うという $tf \cdot idf$ の良さが出ていない可能性が高い。

提案手法が最も良い結果を得たが、まだ改善すべき点も残っている。図 4 を例にするならば、右の木では「分かる」「選別」の親子関係が成立しているが、左の木では間にノードが入り、親子関係が生じないため、Tree Kernel が有効に働いていない点、概念辞書を用いる際に「少ない」「多い」などの相反する意味の単語でも、「数量の程度」という同じ上位概念を持つことから、概念的に近いと判断されてしまう点、などが挙げられる。

4. おわりに

論文の被引用箇所を抽出するため、木構造と単語概念の近さを組合せた方法を提案した。その結果、4 位累積抽出率は約 90% になった。

今後、単語の意味の近さを考慮することや、間にノードが入ったために親子関係が成立しない部分について、ノードを飛び越えることにより、類似文の木構造の近さの判定精度を向上させること、大規模実験などを行いたい。

参考文献

- [1] 工藤拓, 松本祐治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol.43, No.6, pp.1834-1842 (2002).
- [2] Collins, M. and Duffy, N: Convolution Kernels for Natural Language, In *Proceedings of NIPS 2001* (2001).
- [3] 日本電子化辞書研究所: EDR 電子化辞書仕様説明書, 日本電子化辞書研究所 (1995).
- [4] 深谷亮, 山村毅, 工藤博章, 松本哲也, 竹内義則, 大西昇: 単語の頻度統計を用いた文章の類似性の定量化—部分的類似性の考慮—, 電子情報通信学会論文, Vol.J87-D-II, No.2, pp.661-672 (2004).
- [5] 小河貴博, 岩堀祐之, 岩田彰: 情報メディア教育における類似レポート判定システムの構築, 平成 13 年度電気関係学会東海支部連合大会講演論文集, Vol.604, p304 (2001).
- [6] 田村哲也, 黒石文介, 高橋勇, 白井治彦, 小高知宏, 小倉和久: 学生レポートの n-gram による類似度評価の検討, 情報科学フォーラム(FIT) 2002 講演論文集, pp.101-102 (2002).