頻出する同義語の提示による論文の単語推敲支援

-A Tool to sport revision of technical papers by proposing synonymous words-

*1 熊本　聖, *2 竹内　章
Satoru Kumamoto and Akira Takeuchi
*1 九州工業大学大学院情報工学府情報科学専攻
Computer Science and Systems Engineering, Kyushu Institute of Technology
*2 九州工業大学大学院情報工学研究院
*2 Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
Email: kumamoto@minnie.ai.kyutech.ac.jp

あらまし：本研究では、論文執筆の初心者に対して推敲作業における言葉遣いの適切性に着目した支援システムの実現を目指している．現在、執筆者が書いた論文内の内容語に対し、一般の論文に頻出する同義語を提示する機能を実装している．この支援機能を用いて執筆初心者に自身の研究概要を修正してもらい、その内容を基に評価を行った結果、システムが指摘する単語の中で修正すべき単語の半数が修正されていることがわかった．さらに実験結果から分かった改善点について述べる．
キーワード：論文推敲支援，出現頻度付シソーラス，形態素解析，語釈文

1. はじめに

論文執筆における初心者が書いた論文の文章は，論文ではあまり使われない書き方となっている場合がある．そのため執筆者は自身の論文の完成度を高めるために推敲をする必要があるが、論文を読み書きしてきた経験の少なさから、推敲がうまく行えていないのが現状である．

執筆された日本語文章の推敲を支援する研究は多くあり、推敲する際にチェックする要点ごとに着目した研究が進められている．そのほとんどは論文執筆にも適用できるが、「用語は適切か，よりよい表現はないか，言い回しは効果的か」といった言葉遣いの推敲は、論文特有の表現があるために、論文で書かれている文章に絞った手法をとる必要がある．そこで我々は論文に適切な言葉遣いの支援に着目した支援システムの実現を目指している．本報告では、論文に頻出する単語の提示による支援方法と、その支援の効果について述べる．

2. 支援方法の概要

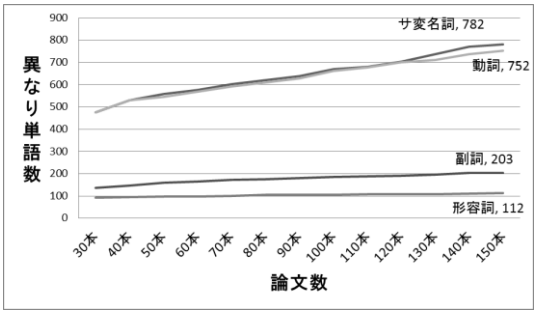
論文の中で実際によく使われる単語を執筆者に提示することで、執筆した論文の言葉遣いをより適切な用語に書き換えるための推敲支援を考えている．そこでまず実際に論文でよく使われる単語を収集するために、論文集の論文本文を形態素解析[(1)]して、使われた単語を取り出す．そして執筆した論文内の単語に対し、論文集に出現した同義語を執筆者に提示することで推敲支援を実現する．

3. 論文出現単語の収集

論文で実際に用いられた単語を調査するために、150 本の情報系の論文[(2)]（以降「サンプル論文」と呼ぶ）から、本文中での普通名詞を除く内容語（サ変名詞，動詞，形容詞，副詞）を取り出した．サンプル論文に出現した各内容語がシソーラスである日本語 WordNet[(3)]に登録されていれば、その単語にサンプル論文中での出現頻度を追加し、出現頻度付シソーラスへと拡張した．出現頻度付シソーラスを用いることで、執筆者が使った単語に対して、サンプル論文で実際に使われている同義語を提示することができる．日本語 WordNet には単語が 93,834 語登録されており、サ変名詞と動詞が合わせて 15437 語、形容詞が 8533 語、副詞が 4076 語である．

調査に用いたサンプル論文数が十分であるか確認するために、日本語 WordNet に登録されておりサンプル論文に出現する単語について、論文数の増加に伴う異なり単語数の変化を品詞ごとに調査した．調査結果を表 1 に示す．形容詞、副詞の単語は論文数に対して横ばいになっているが、サ変名詞、動詞の単語は異なり単語数が増え続けていることがわかる．異なり単語数が増加している品詞に関しては、未収集の単語がまだ存在するため、論文をさらに収集する必要がある．

表 1：各論文数における異なり単語数



4. 同義語の提示による支援

執筆した論文で用いられた単語の内、書き換え候補単語が存在する内容語（普通名詞は除く）を支援対象単語とする．書き換え候補単語とは出現頻度付シソーラスにおいて支援対象単語と同じ意味を持つ単語であり、サンプル論文において出現頻度が２以

上のものである．支援対象単語が多義性をもつ場合も存在するため、書き換え候補単語を取り出す際は支援対象単語が持つすべての意味に対して行う．支援情報を提示する際には、書き換え候補単語がどの意味から取り出されたかを表すために、単語だけでなく意味の説明である語釈文も合わせて提示する．

本支援システムのフィードバック例を図１、図２に示す．図１では執筆した論文の本文が掲載され、支援対象単語に下線が引かれている．利用者は下線が引かれた単語を選択することで、図２に示す同義語の提示画面へ遷移する．書き換え候補単語がサンプル論文では実際にどの程度使われているのかを利用者に示すために、書き換え候補単語を、支援対象単語のサンプル論文中での出現頻度を基準に２つに分けている．支援対象単語より出現頻度が高い書き換え候補単語を提案単語とし、利用者に「この単語の方が多く利用されている」として提示する．支援対象単語より出現頻度は低いが、論文集にいくつか出現した書き換え候補単語を参考単語とし、利用者に「この単語も論文で使われている」として提示する．図１で [!] がついているのが提案単語の存在する単語であり、参考単語しか存在しない場合には [*] がつけられる．

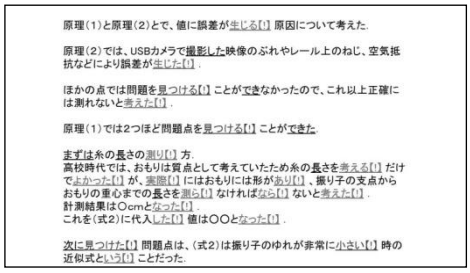


図１：リンク付された論文テキスト画面

図２：同義語の提示画面

5.　評価実験

5.1　実験方法と結果

本支援システムを用いて推敲を行うことの有効性を確認するために、2011 年度の本学の学部 4 年生 8 名を対象に、支援システムを利用して完成途中の自身の卒業研究概要の修正を行ってもらった．卒業研究概要は A4 用紙 1 枚 1200 字程度の文章である．卒業研究概要を支援システムに与え、そのフィードバックにより利用者は提示された書き換え候補を自由に確認し、提示されている単語のほうが適していると判断すれば、修正を行ってもらう．その後各修正の妥当性と、修正されなかった不適切な単語がないかを教員がチェックした．書き換え候補単語の内、すくなくとも提案単語は利用者がすべて確認すると期待されたため、チェックは提案単語がある支援対象単語に絞っている．8 つの卒業研究概要に対し、提案単語が存在する支援対象単語は 340 か所であり、それらを利用者の修正の有無と修正内容の適切性により 6 つに分類し集計を行った．各分類とそれらに属す数を以下に示す．

・修正された箇所：61
① 修正が適切な箇所：43
② 修正が不適切であり、提案単語に適切な単語が存在した箇所：8
③ 修正が不適切であるが、適切な単語が提示単語に存在しない箇所：10
・修正されなかった箇所：279
④ 修正する必要がない箇所：245
⑤ 修正する必要があり、提案単語に適切な単語が存在している箇所：6
⑥ 修正する必要があるが、提案単語に適切な単語が存在しない箇所：28

5.2　評価実験の分析

提案単語が存在する単語の内、修正が必要な箇所は④以外の 95 か所であり、このうち正しく修正されたのが①の 43 か所である．つまり支援システムの指摘によって修正した 45.3%が適切であった．

実験を通して判明した課題について述べる．支援システムが十分に単語を提示できていないため③と⑥が存在している．これは 3 章で述べたように、収集論文数が足りていないことが原因だと考えられる．

②と⑤のように利用者が書き換え候補単語から適切な単語を選択できなかった原因は、システムが提示する内容が同義語とその意味だけであり、実際の論文ではその単語がどのように使われているか分からないためと考えている．そこで提案する単語と意味に加えて、サンプル論文で使われた実例も提示する機能を追加中である．

6.　おわりに

本支援システムが修正候補を提示した単語のうち、修正を要する部分の約半数が適切に修正されており、初心者の論文の完成に貢献ができたことがわかった．現在、サンプル論文中の実例提示については、書き換え候補単語を含むすべての例文をそのまま提示する機能を完成させている．今後はこの実例の中で利用者にとって有用な情報が何であるかを調査し、それを取り出す機能の実装を行う予定である．

参考文献

(1) 形態素解析エンジン JUMAN http://nlp.ist.i.kyoto-u.ac.jp/
(2) 電子情報通信学会論文誌 D 分野　2000 年，2002 年
(3) 日本語 WordNet　http://nlpwww.nict.go.jp/wn-ja/