4M-1

語の共起頻度の提示と用例文の検索に基づく論文執筆支援システム

馮 思萌† 井上 慧† 松原 茂樹† 長尾 確†

†名古屋大学大学院情報科学研究科

1 はじめに

学術論文には、一般文書とは異なる特有の表現や決まった言い回しが多数存在する。論文の執筆経験が乏しい学生は、論文特有の表現の使い方を身につけておらず、執筆している論文中に、論文として適切でない表現（以後、**非論文的表現**と呼び、論文として適切な表現を**論文的表現**と呼ぶ）が含まれることが多い。

本稿では、執筆者が作成した論文に非論文的表現が含まれているときに、自らそれに気付き、対応する論文的表現を発見し修正することを支援するシステムを提案する。非論文的表現の発見を促すために、ユーザが論文を執筆中に語の共起頻度を提示する。また、論文的表現の発見を支援するために、大量の論文データから参考となる用例文を効率的に検索できる環境を提供する。

2 システム構成

提案システムは、論文執筆中に語の共起頻度を提示する機能及び、論文的表現を含む用例文を提示する機能から構成される。両機能の関係を図 1 に示す。

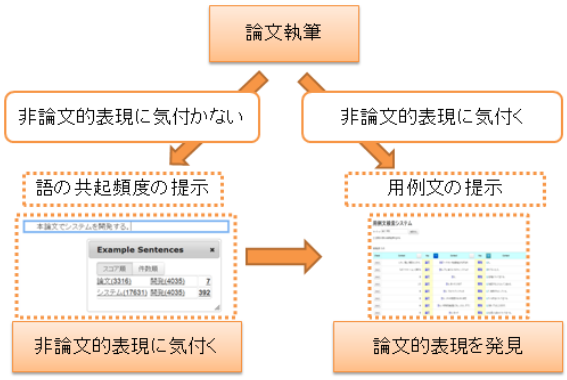

図 1 語の共起頻度提示と用例文提示の関係

語の共起頻度の提示機能は、ユーザが論文を作成中に、非論文的表現に気付けるように、執筆者が使用した表現に含まれる語の共起頻度を求め、逐次提示する。

用例文提示機能では、ユーザが必要な論文的表現およびそれを含む用例文を用例文データベースから検索し、参照することができる。

Paper writing support system based on presentation of word co-occurrence frequencies and search of example sentences

†FENG Simeng (ryou@nagao.nuie.nagoya-u.ac.jp)

†INOUE Kei (kinoue@nagao.nuie.nagoya-u.ac.jp)

†MATSUBARA Shigeki (matubara@nagoya-u.jp)

†NAGAO Katashi (nagao@nuie.nagoya-u.ac.jp)

†Graduate School of Information Science, Nagoya University

3 語の共起頻度の提示

語の共起頻度の提示は、本研究室で開発している論文執筆支援システム TDEditor[1]と連携して行う。TDEditor は、作成・蓄積されたコンテンツの検索とコンテンツの参照・引用の機能を備える。提案システムでは、ユーザが TDEditor で論文を作成時に、入力した文章に含まれる表現の共起頻度が逐次的に提示される。システムの動作を図 2 に示す。

図 2 語の共起頻度の提示

3.1 非論文的表現

非論文的表現は下記の 2 つのタイプに分類できる。

1. 単語自体が適さない
2. 単語の組み合わせが適さない

タイプ 1 の例として、「だんだん」は論文では使われず、「徐々に」の方が相応しい。タイプ 2 の例として、「本論文で～システムを開発する」という文では、「論文」と「開発する」はいずれも論文に適した単語であるものの、「論文で開発する」は組み合わせとして適さず、「論文で提案する」や「研究で開発する」の方が相応しい。

非論文的表現は、公表済みの学術論文に出現することは少ないと見込まれる。すなわち、既存の学術論文から該当単語あるいは単語の組み合わせを含む用例文の頻度を調べれば、表現の適切さを判定できる。

3.2 語の共起頻度

提案システムでは、上記の 2 つのタイプの非論文的表現の発見を促すために、文法的関係をもって共起する頻度を用いる。このために、ユーザが入力した文に対して形態素解析および構文解析を行い、単語と単語対を抽出し、それぞれの頻度を算出する。なお、単語は自立語に限定する。また、単語対とは、単語が属する文節が係り受け関係にある 2 つの自立語の対を指す。単語対に対するスコアは以下のように求める。

$$Score = \log_2 \frac{\text{単語対頻度} \times \text{コーパス総語数}}{\text{語A頻度} \times \text{語B頻度}}$$

ただし、語 A 頻度と語 B 頻度は単語対を構成する 2 つの単語のそれぞれの出現頻度である。

4 用例文の提示

ユーザは語の共起頻度の提示を通じて非論文的表現に気付く手がかりを得られる。しかし、用例文の数は表現の妥当性の目安に過ぎない。実際には、具体的な用例文を確認し、論文的表現の使い方が自らが作成しようとしている文と一致するかどうかを判断する必要がある。

表現の使い方の理解を支援する手段として、KWIC（Key Word In Context）がある[2]。提案システムでは、日本語論文執筆支援における用例文提示の観点から、KWICをベースに、複数キーワードによるソートと文の圧縮の２つのアプローチを採用する。システムの動作を図３に示す。

図３ 用例文の提示

4.1 複数列によるソート

従来のKWICの単一列ソートでは、単一キーワード前後の文脈しか一覧できない。一方、複数キーワードで検索する場合、キーワード間にも文字列が存在する。このため、本システムでは、複数列からソートキーを選択し、関連付けてソートすることを可能にする。

例えば、図に示すように、ユーザが「論文」「開発」というキーワードで検索する場合、「論文」の後方の文字列の最初のソートボタンをクリックすれば、「論文」の後方に隣接する形態素でソートされる。また、「開発」についても同様である。これにより、「論文で開発する」を含む用例文の全体に占める割合を把握できる。

4.2 文の圧縮

学術論文中の文を用例文として提示するため、用例文には専門用語や独自の記述なども含まれる。表現の使い方を理解する上で、ユーザは専門的内容まで理解する必要はない。

例えば、ユーザが「論文で開発する」という表現の妥当性を調べるため、「論文」「開発」で検索すると、検索結果の中に図４に示す文が含まれることになる。この文のキーワードとなる「開発」の前に、開発されたシステムを説明する内容が詳しく記述されている。しかし、図５のように省略してもユーザにとって不都合はない。

このように、必要のない内容を削除することによって、表現の使い方を把握するための時間を短縮できる。

文圧縮は、まず用例文に対して係り受け解析を行い、各文節に対してキーワードからの係り受け関係の距離を計算する。距離がキーワードに近いほど有用となる可能性が高いので、優先的に圧縮文に含める。一覧性を重視するために、１行で表示できる長さになるまで以上の手順を繰り返す。

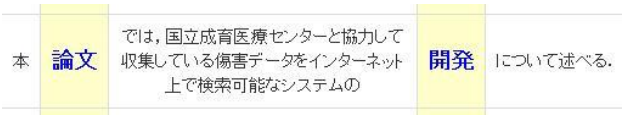

図４ 文圧縮前の用例文

図５ 文圧縮後の用例文

5 評価実験

5.1 実験の目的と方法

提案システムの有効性を検証するため、日本人学部生および大学院生11名に対して被験者実験を行った。被験者に非論文的表現が含まれる文を提示し、被験者は提案システムとKWICの２つのシステムを用いて表現を検索し、非論文的表現を論文的表現に推敲した。その際、被験者は参考となった用例文をマークした。難易度が均等な問題を７問ずつ含む問題集を３セット用意した。問題集と実験システムは被験者ごとにランダムに組み合わせた。

提案システムとKWICでの検索時間を比較した。また、参考になったとマークされた文に対して文圧縮を行い、圧縮後の文に被験者の必要な表現が含まれるかどうかを検証した。

5.2 実験結果

実験の結果、提案システムを用いた被験者の１回の検索時間の平均は39.13秒で、KWICを用いた被験者の１回の検索時間の平均は46.98秒であった。提案システムを用いることによって、効率的に必要な用例文を発見できることを確認した。

文圧縮を行った13文のすべてに必要な表現が含まれていた。平均圧縮率は51.61%であった。また、13文のうち、11文が文法的であった。ユーザのほしい情報と文法性を保持しており、文圧縮の有効性を確認した。

6 おわりに

本稿では、執筆中の文に含まれる表現の共起頻度を提示する機能と効率的に用例文を検索する機能を用いて、日本語論文執筆を支援するシステムについて述べた。今後の課題として、自立語のみではなく、誤用が多い助詞などの適切さの提示や、非論文的表現に対する代替候補の提示などが挙げられる。

参考文献

[1] 棚瀬達央，大平茂輝，長尾確，複数コンテンツの部分関連付けに基づく論文作成支援，情報科学技術フォーラム(FIT)，2013.

[2] Luhn, H.P.: Key-Word-In-Context Index for Technical Literature, IBM Corporation, 1959.