6Q-4

文書校正における単語の概念関係の利用

板倉 由知 (福井大学大学院工学研究科)
白井 治彦・高橋 勇・黒岩 丈介・小高 知宏・小倉 久和 (福井大学大学院)

1 はじめに

本研究では、文書を書きなれていない学生を対象とした、単語の概念関係を用いることで計算する段落と1文との間の意味的距離を用いた文書校正支援ツールを提案する。

文書を書きなれていない学生の書く論文やレポートは、段落構成の熟考が行われてなく、散文的に文章を記述してしまうため、その段落で何を主張しているかを読者が読解することが困難となる場合がある。

本研究では、文書中に使われている単語に注目し、既存の単語概念辞書から求められる単語間の意味類似度を用いることで文の関連性を示す距離を算出する。求められた文の関連性から段落内において校正すべき文を利用者に提示することで文書校正を促すための校正支援ツールを提案、実装して有効性を確認した。

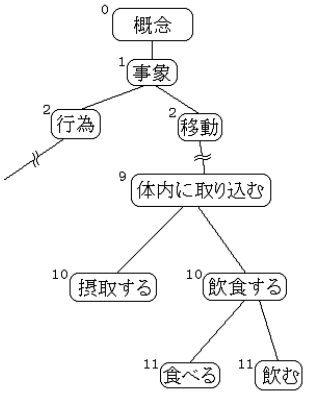

図 1: EDR 概念シソーラス（一部）

2 不適切文判別方法

本研究において文書校正の必要な学生の書く論文に存在する問題を、次の3つに分類して考える。

- 文書作成上の問題
- 文書構成上の問題
- 文書内容上の問題

文書作成上の問題として挙げられる問題として、誤字、脱字、係り受けの誤りなど様々な問題があるが、これらの問題を校正するための技術や製品は既に存在している。[1][2] 本研究では文書内の段落に不適切な文が挿入される問題を文書構成上の問題として扱い、単語間の意味類似度を利用することで、段落における不適切文の抽出、利用者への提示をするための手法を提案する。

2.1 単語間の意味類似度

本手法では、単語間の意味的な類似性を定量化した単語間の意味類似度を用いる。この単語間意味類似度は、EDR 概念辞書を用い概念シソーラスにおける単語間の距離や、共通概念情報を用い、Yuhua Li 氏の提案する手法 [3] により計算する。図 1 は、EDR 概念辞書における概念シソーラスの一部である。この概念シソーラスは、ルートに”概念”ノードを持ったツリー構造となっており、このシソーラスに存在する全ての単語は”概念”から派生している。図に示されている数字は、ルートノード”概念”からの距離、つまり深さを表している。この数値が高いほど、その単語はより具体的な概念を持っているということができ、共通概念情報はより有効な情報となる。

2.2 不適切文抽出法の検討

本研究では、ある主張する内容について複数の文で表現される集まりを段落として扱う。一方、この段落内に主張する内容とは違った別の主張をする文が挿入されるとき、段落にとっては不適切な文を持ち、本来主張する意味内容がゆれてしまう。その結果、読者は著者が意図する内容を理解しにくくなり、誤解したまま読み終えることも考えられる。本稿で提案する手法は、文中で使われる単語に注目し、単語間の意味類似度から文と段落との関連度を算出することで不適切な文かどうかの判断を行う。もし、段落にそぐわない不適切な文が挿入されている場合、その文を構成する単語と、段落で主張している内容に関する単語との間の意味類似度は低く算出されると考えられる。そこで提案する本手法は、ある段落を構成する 1 文が、段落の主張している内容について適切な記述をしているかどうかを判断するために 1 文を構成する単語集合と、段落を構成する単語集合との間で、単語間の意味類似度を算出し、それを元に 1 文が段落にとって適切がどうかを判断する。このとき、段落で使われる単語間の意味類似度を算出したときに、高い値を示すのであれば、主題に関する単語が用いられ、その段落は表現したい主題を実現しているといえる。

以下に、文の関連度算出に注目した具体的な処理を記述する。

段落を構成する文を文集合 S とし、各文 s_i(文番号：i=1,2,3,...) に対し、形態素解析を行い、文ごとに構成している単語集合 W_{ix}(x=1,2,3,...) を抽出する。この際、抽出する単語は名詞、動詞のみである。ある 1 文の単語集合 W_{ix}(x=1,2,3,...) に注目し、その 1 文以外の文集合 S の単語集合 W_{j_y} との全組み合わせにおける単語間意味類似度を算出する。1 文と段落との単語組み合わせの中から、各単語は最も高い関連を持つ単語との意味類似度を平均することで 1 文の関連度とする。例えば、段落に対し不適切な文の関連度は、主張する内容に関係しない単語が使われていると考えられるの

で、関連度は低く算出され、段落に適切な文の関連度は、高く算出されると考えられる。以上の処理を、段落を構成するすべての文に対して行う。すべての文の関連度が求められたら、それらの平均関連度を求め、その値を段落全体の一貫度とする。段落にとって不適切な文の関連度は、その平均関連度を大きく下回っていると考えられる。よって、平均関連度を大幅に下回る関連度を持つ文は、段落にとって不適切な文だと見なす。図 2 に、以上の処理を図に示す。

これら各文の関連度と一貫度は、段落における話題の中心から、文がその内容に対し意味的に近いかどうかを判断する指標になると考えられる。

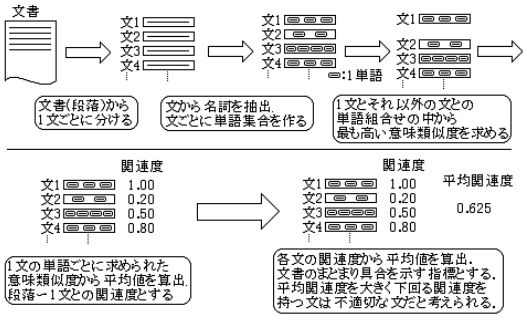

図 2: 段落－１文の関連度算出

これら各文の関連度は、一貫度よりも高い値を示せば、段落を構成する文は主題に沿っており意味的にまとまっていると考え、一貫度よりも低い値を示すのであれば、文の主題が散漫していて意味的にまとまっていないと考える。

3 実験

本手法の評価実験として、研究室に所属する学生の書いた卒業論文を用いて実験を行う。対象とした卒業論文は、指導教員の校正前の内容がまとまってなく、本論文で挙げた不適切な文の挿入された段落を含む論文と、指導教員の校正が行われた内容がまとまった論文とで比較を行い、その違いを確認する。

本手法では 1 文と段落との関連度を、1 文を構成する各単語組み合わせの単語間最高意味類似度を平均することで算出していたが、本実験では、更に、1 文と段落との単語組み合わせから、単語間最高意味類似度、最低意味類似度をそれぞれ算出し、それらを 1 文と段落との関連度として扱う。従って、1 文と段落との関連度を、各単語組み合わせから求められる単語間意味類似度の平均値、最高値、最低値の 3 種類算出し、それぞれ、校正前後の論文がどのように評価されるかを確認し、有効性を確かめる。

図 3 は、校正前後のある段落内における文の関連度を示す。校正を行ったことで文の数が 4 文から 6 文へ変化しているが、文の関連度は校正前よりも校正を行ったことで向上しているといえる。さらに詳しい実験結果は、発表当日に述べる。

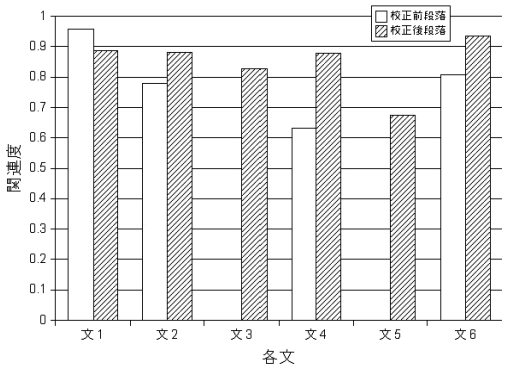

図 3: ある段落における実験結果

4 考察とまとめ

本稿では、EDR 概念辞書を利用することで単語の概念関係を用いた文書校正のための手法を提案した。本手法は、何らかの主題をもっているはずの段落において、意味的にふさわしくない文を発見、抽出するものである。この手法により、普段文書を書きなれていないような学生に対し、主題とすべき内容に沿った文書を書くための支援ツールとしての利用を考える。さらに本手法は機械的に判断を行うため、客観的に自分の執筆した論文を見直すきっかけともなると考えられる。

本システムでは、EDR 概念辞書に基づいて単語間意味類似度を算出しているため、辞書に存在する単語しか処理することができず、新出語については対応しきれない問題が存在する。また、形態素解析に用いているソフトの解析能力にも依存しているため、改良の余地が伺える。本手法では、不適切文判断のため、一貫度を判断基準としたが、誤判断の可能性も考えられる。よって、別の判断基準などを考える必要がある。ただし本稿で提案した手法による文書校正だけでも、利用者にとって、文書推敲の機会となり、結果として読者にとって、その文書内容を理解しやすくなると考えられる。

参考文献

[1] Microsoft Office Word,“http://www.microsoft.com/japan/office/word/prodinfo/default.mspx”.

[2] Justsystem JustRight!2,“http://www.justsystem.co.jp/justright/”.

[3] Yuhua Li et al.”An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources” IEEE Transactions on Knowledge and Data Engineering Vol.15 No.4 pp.871-882

[4] 板倉由知, 白井治彦, 高橋勇, 黒岩丈介, 小高知宏, 小倉久和 ”単語の概念関係を用いた文書校正ツールの開発” 情報処理学会第 68 回全国大会講演論文集 4N-7(2006.3)