4N-7

単語の概念関係を用いた文書校正ツールの開発

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（福井大学工学部)
小高 知宏 (福井大学大学院)・小倉 久和 (福井大学工学部)

1 はじめに

文書を書きなれていない学生の書くレポートや卒業論文などは、文法的な誤り、文章構成上の誤りなどによる完成度の低い文書が多く存在している。

本研究では、そのような文書を校正するために、ある段落における１文が、その段落においてふさわしい主張をしているかどうかといった文章構成上の誤りを判断する手法を提案する。つまり、ある段落において主題としている何らかの内容があると考え、その段落を構成している各文が、主題に適しているのかどうかを判断する手法を提案する。そして、その手法を用い、段落内の文章が正しく構成されているかの評価を行う文書校正のためのツールを検討する。本手法では、文書を構成する単語に注目し、それらの単語が主題について関連性があるかどうかを確かめることで、一貫した主張が行われているかどうかを判断する。その際、単語間の意味類似度を利用する。

2 段落における不適切文の抽出

本稿では、文書中における段落は、複数の文がある主題を表現していると考える。つまり、それらの文で使われている単語は、表現したい主題について述べているはずであり、その単語間にも、何らかの関連性があると考えられる。そこで単語間の意味類似度を用いることで、段落で表現したい主題に関する単語と、1 文を構成する単語との間の意味類似度から、その文が不適切かどうかを判断する手法を提案する。

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 [1] 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、共通概念情報においてより有効な情報となる。例えば、”飲む”と”食べる”が共通する概念情報は”飲食する”であり、一方、”飲む”と”摂取する”が共通する概念情報は”体内へ取り込む”である。”飲食する”と”体内へ取り込む”を比べた概念の深さは、”飲食する”の方が深く、より具体的な概念を共有していることがわかり、意味的にも高い類似性が見て取れる。また、”飲む”と”食べる”の単語間距離は、図から一方のノードに至るまでの経路を数え上げると、距離 2 となり、”飲む”と”摂取する”の単語間距離は、距離 3 である。単語間の距離は近いほど、似通った意味を持つと考えられる。よって、”飲む”と”食べる”の方が、意味的な類似性が高いと考えられる。単語間の距離や共通概念情報を用いて算出される意味類似度は、単語間の意味が似通っていた場合、高い値を示し、意味が違っていた場合、0 に近い値となる。

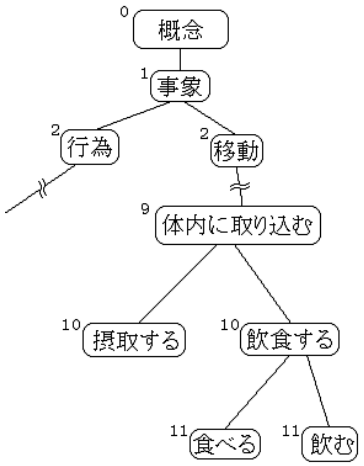

図 1: EDR 概念シソーラス（一部）

2.2 不適切文抽出法の検討

文書中における段落で、その主張にそぐわない不適切な文が挿入されている場合、読者は段落全体として、主張の揺れを感じてしまい、著者が意図する内容が伝わりにくくなってしまう。ここでいう不適切な文とは、主張したい内容に関係のない単語が使われている文とする。本稿で提案する手法は、単語間の意味類似度を算出することで不適切な文を発見、抽出を行う。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。逆に、意味類似度が低い値であれば、表現すべき主題が定まっていないといえる。

具体的な処理を以下に記述する。

まず対象とする文書から、句点を基準に文集合 S に分解する。ただし、このときの文集合 S は、ある段落を構成している文集合とする。文集合 S における各文各文 s_i(文番号：i=1,2,3,...) に対し、形態素解析を行い、各文ごとに文を構成している単語集合 W_{ix}(x=1,2,3,...) を抽出する。ただし、この際、名詞だけを単語集合として抽出する。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} との全単語組み合わせにおける単語間意味類似度を算出する。つまり、ある 1 文の単語集合から段落で使われる単語との最も関連の強い単語組を抽出することで、その単語は段落において主題となる単語と考えることができる。このように、ある 1 文を構成する単語集合において、それぞれの単語は、段落に使われている最も関連の強い単語との意味類似度を算出することによって、1 文を構成する単語集合の平均意味類似度を求めることができる。ある 1 文で使われる単語の平均意味類似度は、段落と 1 文との距離、つまり関連度を求めることだといえる。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まとまり具合の指標として扱う。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、段落ー 1 文の関連度のイメージ図を示す。図 2 では、段落のまとまり具合を示す平均関連度を点線で示し、その外側に位置する文 3 を段落の主題にそぐわない不適切な文であると判断している。

この段落ー 1 文の関連度は、段落における話題の中心から、ある 1 文がどれだけ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。

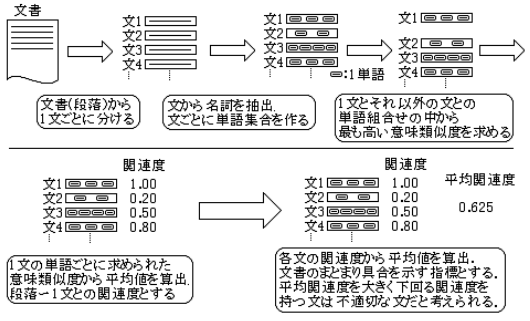

図 2: 段落ー 1 文の関連度算出

この段落ー 1 文の関連度は、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、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られる。

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。ただし、対象とした卒業論文は、指導教官の校正は無く学生自身が書き上げた文書である。普段文書を書いていない学生の書いた卒業論文は、今回対象とする文書校正上の問題が数多く存在していると考えられ、本手法によって、段落に対し不適切な文があった場合、その文を抽出できるかどうかを確かめる。

卒業論文の完成度:未完成であり、校正されていない
データ採取の対象：研究室に所属する学生 8 名

実験結果は、発表当日に詳しく述べる。

4 考察とまとめ

本稿では、EDR 概念辞書を基準とした単語が持つ概念シソーラスを用い、単語間の意味類似度を使っ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。ただし、本手法では、「例えば」などの接続詞からなる例示を表す文を含む段落には、誤抽出を行う可能性があり、その解決方法については検討中である。

文書校正としては、本稿で検討した内容以外に、様々な内容の校正が考えられる。例えば、単語の係り受けの誤りといった文法上の誤りの校正や、文書の論理的展開における主義主張のゆれといった意味内容的な誤りの校正がある。しかし、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、読者にとっても、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] 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

[2] 藤沢晃治：「「分かりやすい文章」の技術」 講談社,(2004)

[3] 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検討" 平成 17 年度電気関係学会北陸支部連合大会講演論文集 E-70(2005.9)