段落の一貫性評価手法の提案

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

1 はじめに

学術論文を含めた技術論文では、段落ごとに主張が一貫していることが必要である [1][2]。しかし、卒業論文に見られるように、論文記述に不慣れな学生が書いた論文では、1つの段落に様々な観点からの主張が書かれ、一貫性に乏しい。既存の自動文書校正ツールでは、誤字脱字の指摘などの表面的な誤りを指摘するものは存在するが、段落の主張の一貫性を検査するものはない [3][4][5]。そこで本研究では、段落を構成する複数の文が担う概念が一貫していることを自動的に評価する方法を提案する。

本研究では段落一貫度という尺度によって段落の概念一貫性を評価する．段落一貫度は，ある段落に含まれる文と，段落を構成するその他の文との間の概念距離から，ある1文とその段落との関連度を求め，段落を構成する全文の関連度の平均値を段落内容の一貫性評価の指標としたものである．文の関連度は，EDR 概念辞書によって定義される単語間概念距離を用いることで，ある段落に含まれる文と，段落を校正するその他の文との間から単語の関係性を対象とした関連性を示す尺度である．

段落一貫度は，単語の概念距離に基づいた段落の概念一貫性を示す指標である．段落一貫度が段落の主張の一貫性を評価できることを示すために，実験を行った．実験から、段落一貫度の有効性を示した．

2 単語間意味類似度

本研究では段落一貫度を求める際、EDR 概念辞書における単語間の意味類似度を利用している。単語間意味類似度とは、EDR 概念辞書に示された単語間の距離から求められる概念距離 d やルートノードから単語間で共通する概念まで距離を示す共通概念距離 h を利用し、単語間の意味的な類似性を数値で示した尺度である。

ここでは図1に示した EDR 概念辞書の一部を例に説明する。図では食べる、飲むといった単語の概念関係が木構造に示されている。食べるという単語に最も近い概念として “食べる” が存在し、同様に飲むという単語には “飲む” という概念が存在する。これら1つ1つの概念をノードとして考える。

“食べる”、“飲む” という単語間から、そのノード間の距離を求めると、EDR 概念辞書における単語間距離は図1から2となる。また、“食べる”、“摂取する” という単語間の距離は3である。これらの単語間距離を $d(w_1, w_2)$ と定義する。このとき、w_1,w_2 はそれぞれ単語を示す。この距離が近ければ近いほど、単語間の類似性は高くなる。

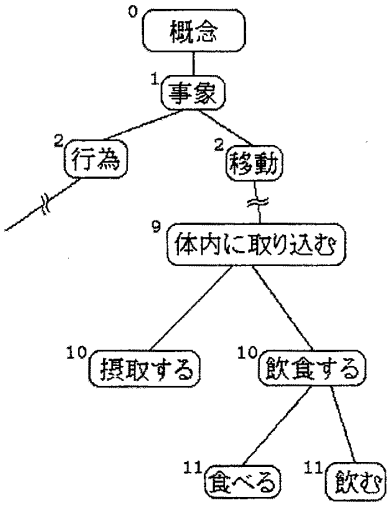

図 1: EDR 概念辞書シソーラス (一部)(数字はルートノードからの深さを示す)

さらに、双方の概念の間で共通した上位概念として “飲食する” という概念があり、さらに上位概念を見ていくと最終的には “概念” という概念に達する。このときの “概念” をルートノードとし、ルートノードからある単語間が共有する概念までの距離を概念の深さ、つまり共通概念距離 h とする。“食べる”、“飲む” という単語間の共通する概念 “飲食する” までの共通概念距離 h は図1中の各ノードの左肩に記してある 10 である。また、“食べる”、“摂取する” という単語間における共通する概念 “体内に取り込む” までの共通概念距離は 9 である。これらの共通概念距離を $h(w_1, w_2)$ と定義する。この距離はルートノードからの距離であり、ルートノードから遠ければ遠いほど、単語間の類似性は高くなる。

従って、“食べる”、“摂取する” の単語間よりも、“食べる”、“飲む” の単語間のほうが意味類似性が高いといえる。

本研究では、単語間意味類似度の計算のために以下の式を使用した [9]。

$$\mathrm{Sim}(w_1, w_2) = e^{-\alpha d(w_1,w_2)} \frac{e^{\beta h(w_1,w_2)} - e^{-\beta h(w_1,w_2)}}{e^{\beta h(w_1,w_2)} + e^{-\beta h(w_1,w_2)}}$$

$\alpha(>0)$, $\beta(>0)$ はそれぞれ任意の定数である．

3 関連度と段落一貫度

本論文で提案する手法は，段落の主張の一貫性を評価する指標である段落一貫度 C を求めることで，段落を評価する．段落一貫度 C を求める際，段落

を構成する各文について段落内容との関連性を示す文の関連度 R_i を利用する．段落一貫度 C は以下の手順によって求める [6][7][8].

- 段落一貫度
  段落を構成するすべての文について関連度 R_i を計算し，その平均値を段落一貫度 C と定義する．

$$C = \frac{1}{n} \sum R_i$$

n は段落における文の数.

このとき，段落におけるそれぞれの文の関連度 R_i は，1 文を構成する単語を利用し前章で示した単語間意味類似度 Sim から求める．

この単語間意味類似度 Sim を用い，段落におけるある文の関連度 R_i を計算する．関連度 R_i とは，ある段落に含まれる文 S_i と，段落を構成するその他の文との間の概念距離である．関連度 R_i を求める際，文 S_i と段落を構成するその他の文から単語を抽出する必要があるため，形態素解析ツール MeCab を用いて名詞と動詞を抽出する．関連度 R_i は，次の手順により計算する．

- 単語集合の抽出
  段落を構成する文から，名詞と動詞を形態素解析ツール MeCab を用いて抽出する．文 S_i から抽出した単語の集合 $W(s_i) = \{w_1(s_i), w_2(s_i)...w_m(s_i)\}$ と，段落を構成するその他の文から抽出した単語の集合 $W_{P_{(s_i)}} = \{w_1(P_{(s_i)}), w_2(P_{(s_i)})...\}$ を作成する．

- 文の関連度
  単語集合 $W(s_i)$ のひとつの要素 $w_a(s_i)$ と，$W_{P_{(s_i)}}$ の要素 $w_b(p)$ との間の単語間意味類似度 $\mathrm{Sim}(w_a(s_i), w_b(p))$ を計算し，$w_a(s_i)$ に対する最大意味類似度 $\max(\mathrm{Sim}(w_a(s_i), w_b(p))$ とする．$W(s_i)$ のすべての要素について最大となる単語間意味類似度を計算し，その平均値を文 S_i の関連度 R_i と定義する．

  $R_i = \frac{1}{m} \sum \max(\mathrm{Sim}(w_a(s_i), w_b(p)))$

m は文 S_i を構成する単語 (名詞，動詞) の数.

以上の手順によって算出された段落一貫度 C は段落の内容一貫性を評価する指標である．段落一貫度とは，段落を構成する単語の概念関係における関係性の強さを示している．

4 実験

本手法の評価実験として、段落一貫度の有効性を示すための実験を行った。詳しい実験結果については、発表当日に発表する。

5 考察とまとめ

本稿では、段落の一貫性を評価するための指標として、段落一貫度という指標を提案した。この指標は、段落が同一する概念で一貫して記述されているかを判断するための評価基準になり得ると考えられる。

また段落一貫度を求める手法を応用することで、段落一貫度を著しく下回る関連度を持つ文を段落内容に不適切な文であると判断する文書校正のための支援ツールとしての活用が考えられる。

References

[1] 藤沢晃治,“「分かりやすい文章」の技術”, 講談社,2004.

[2] 木下是雄,“理科系の作文技術”, 中公新書,1981.

[3] 鈴木 恵美子,“日本語文書校正支援システムの設計と評価” 情報処理学会論文誌, vol.30, pp.1402-1412, Nov. 1989.

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

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

[6] 板倉 由知,”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大会講演論文集, F-70, 9. 2005.

[7] 板倉 由知,”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, 3. 2006.

[8] 板倉 由知,”文書校正における単語の概念関係の利用”, 情報処理学会第 69 回全国大会講演論文集, 6Q-4, 3. 2007.

[9] Yuhua Li, “An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources” IEEE Transactions on Knowledge and Data Engineering, vol.15, pp.871–882.

[10] 深谷 亮,“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誌 D-2, vol.J87-D-2, No.2, pp.661-672, Feb. 2004.

[11] 岡本 潤,“連想概念辞書の距離情報を用いた重要文の抽出”, 自然言語処理, vol.10, No.5, pp.139-151, Nov. 2003.

[12] MeCab“http://mecab.sourceforge.jp/”.