

単語の概念関係を用いた段落一貫性評価指標の有効性

板倉 由知, 白井 治彦, 黒岩 丈介, 小高 知宏, 小倉 久和
福井大学大学院工学研究科

本研究では文書における段落の一貫性に注目し、その段落一貫性を評価するための段落一貫度という指標を提案し、その有効性を確かめた。特に技術文書における段落は、段落ごとにその主張が一貫していることが必要になる。本研究で提案する段落一貫度は、EDR 概念辞書を用い、単語間の意味類似度から求められる値である。本論文の実験では、論文の抄録を段落に見立て段落一貫度を算出したものと、抄録内の1文を他抄録の1文に変更したものととの比較を行い、段落一貫度の変化を示した。また、学生8名の執筆した卒業論文を対象に指導教員の論文指導前後から、各段落、各章の段落一貫度の比較を行った。それらの実験により、段落一貫度の有効性を示した。

キーワード：EDR 概念辞書、概念距離、文書評価

Availability of evaluation method of paragraph consistency with words semantic relation.

Yoshitomo ITAKURA, Haruhiko SHIRAI, Josuke KUROIWA, Tomohiro ODAKA,
Hisakazu OGURA
Graduate School of Engineering, Fukui University

We proposed new index ,Paragraph Consistency Degree(PCD),that expressed coherency of paragraph. PCD is index that shows whether content of paragraph composed of one insistence. In this paper, we had performed experiments for availability of evaluation method of paragraph consistency.

Keyword:EDR,concept dictionary,textual,document evaluation

1 まえがき

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、従来用いられる自動文書校正ツールでは、誤字脱字の指摘などの文書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]. また、卒業論文に見られるように論文記述に不慣れた学生が書いた論文は、思いついた文をそのまま記述してしまうような散文的な文構成になっているものがあり、結果的に様々な観点からの主張が記述されてしまう場合がある。そのような文面は一貫性に乏しく、読者にとって理解し難い内容になってしまう。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、2つの実験を行った。ひとつは、完成された学術文書に対して意図的に一貫性を阻害する文を挿入し、

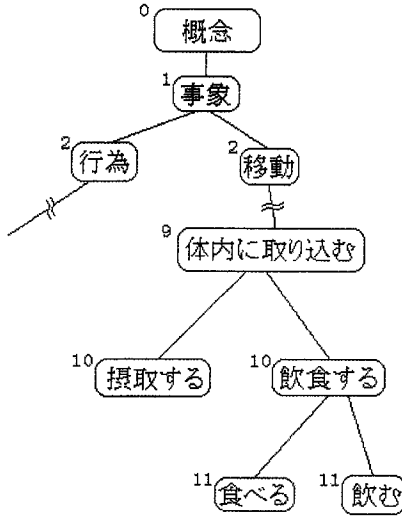


図 1: EDR 概念辞書シソーラス (一部)(数字はルートノードからの深さを示す)

挿入前後の段落一貫度を比較することで、挿入により一貫度が低下することを示した。もうひとつの実験では、卒業論文の推敲過程において、推敲後に一貫度が向上することを示した。

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。

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]。

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

$\alpha (> 0)$, $\beta (> 0)$ 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 $W(s_i) = \{w_1(s_i), w_2(s_i) \cdots w_m(s_i)\}$ と、段落を構成するその他の文から抽出した単語の集合 $W_{P(s_i)} = \{w_1(P(s_i)), w_2(P(s_i)) \cdots\}$ 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 $\text{Sim}(w_a(s_i), w_b(p))$ を計算し、 $w_a(s_i)$ に対する最大意味類似度 $\max(\text{Sim}(w_a(s_i), w_b(p)))$ 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。

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

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

本手法の実験として、3つの実験を行った。実験1では完成された学術論文に対し、内容に全く関連性の無い文を挿入し、挿入前後の段落一貫度を比較した。実験2は実験1で対象とした学術論文について、それぞれの論文を構成する文から他の論文を構成する文を挿入し、挿入前後の段落一貫度を比較した。実験3は、学術論文の推敲過程において、推敲の前後で段落一貫度の変化を調べた。

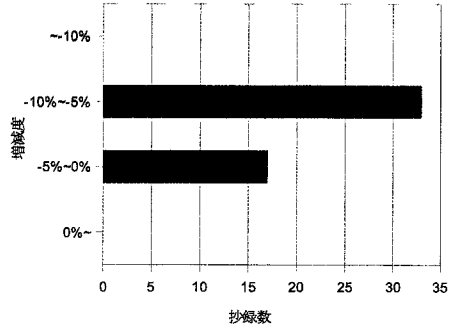


図2: 実験1における変更前後による段落一貫度の増減の割合

実験1では、文献検索サービス J-DreamII より検索フレーズ “情報処理 自然言語” から検索される50篇の論文の抄録 (A) を段落として扱う。扱った抄録は4文から10文で構成されている文章である。このような学術論文における抄録は、研究内容という主題について端的に表現している文集合であるため、抄録を段落として扱い、本提案手法による評価を行った。

このとき50篇の論文抄録 (A) の末尾に、内容に全く関連性が無いと考えられる以下の文を追加したもの抄録 (B) とする。

- 本日の天気は晴れです。

抄録 (A) と抄録 (B) の段落一貫度の増減割合をまとめた結果を図2に示す。

また、抄録 (B) において、抄録内の段落一貫度と挿入した関連性の無い文の関連度の増減割合を示した結果を図3に示す。

抄録 (A) と抄録 (B) を比較したとき、抄録 (B) について全抄録において段落一貫度の減少を確認することができた。段落の一貫性を阻害するために挿入した文は、内容理解を妨げることを段落一貫度から確かめることができた。また、図3で示した挿入文と段落一貫度の増減割合からは、各々の段落一貫度に比べ、挿入文の関連度は著しい減少が確認できた。挿入文は明らかに内容と関連性の無い文であるということが数値から確認できたと考える。

実験2対象となる文書は実験1でも扱った抄録 (A) である。このとき、50篇の論文の抄録 (A) に

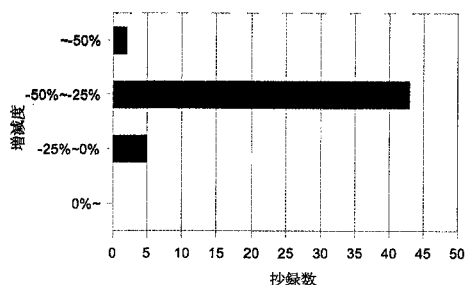


図 3: 実験 1 における挿入文と段落一貫度の増減の度合

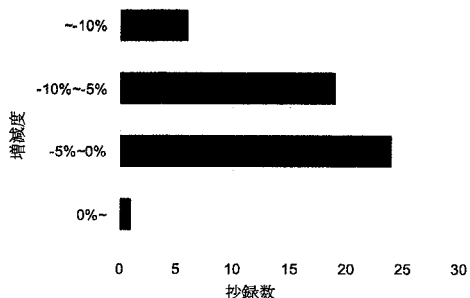


図 4: 実験 2 における変更前後による段落一貫度の増減の度合

対し、それぞれの抄録内の適当な 1 文を別の抄録内からランダムに選択した 1 文に変更した別の抄録 (C) を作成し、抄録 (A) と抄録 (C) の段落一貫度の変化を確かめた。

変更前抄録の段落一貫度と変更後抄録の段落一貫度の増減度合をまとめた結果を図 4 に示す。

実験 2 において、50 篇の論文抄録のうち、ほぼ全てである 49 篇の論文抄録において本提案手法による段落内容の一貫度が減少することで、本手法の指標の効果が確認できた。1 篇のみ比較した段落一貫度が増加しているものがあった。これは似通ったテーマの論文抄録を対象としたことで偶然にも変更後の抄録内容の意味が通る状態になってしまったためであると考えられる。

実験 3 では、工学部学生 8 名が執筆した卒業論文を対象として、本提案手法による段落一貫度の変化を調べた。このとき、指導教員は学生に対して

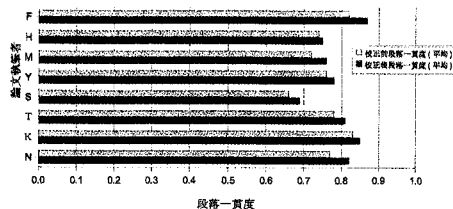


図 5: 卒業論文の段落一貫度

要点が定まらない段落内容に対し 1 つの主題について記述するように論文指導を行った。対象とした卒業論文は、論文指導を行う前の初稿が執筆時間 1ヶ月程度で書き、20 から 30 の段落で構成される論文である。それらの段落はそれぞれ 4 文から 5 文で構成されており、全体として 80 文から 120 文で構成される論文である。一方、論文指導が行われた完成稿は、執筆時間 3ヶ月程度で書き上げられ、60 程度の段落、総文数にして 260 文程度で構成された論文である。各学生の執筆した卒業論文は、全段落に対し本手法による段落一貫度で評価し、それぞれの平均を求め、校正前後の卒業論文について段落一貫度 C を比較した結果を図 5 に示す。

実験 3 においては、学生の執筆した卒業論文で指導教員の校正が行われたことによって、全ての論文で段落一貫度が向上した結果となった。未完成卒業論文は、論文として体裁は整っているものの推敲不足、内容の不備などの問題点により、論文の完成度としては低く要点の定まらない段落が多く存在していた。完成した卒業論文は、未完成卒業論文で見られた要点の定まらない段落は減少し、内容も熟考されて読みやすい文章で記述されていた。

また、実験 3 で対象とした卒業論文について、各章ごとにおける段落一貫度の平均を表 1 に示す。

執筆者 Y,S,T,K は卒論を 5 章構成で作成し、それ以外の執筆者は 6 章構成で卒論を作成した。表中の-は、その時点で段落、章が記述されてなく、段落一貫度を算出できなかったことを示し、表中の/は、該当する章が存在しないものである。

さらに各執筆者の章ごとにおける初稿と最終稿との平均段落一貫度の差を示したものを図 6 に示す。

この結果から、1 章を初め、ほとんど全ての章に

表 1: 各章の平均段落一貫度

執筆者		1章	2章	3章	4章	5章	6章
F	初稿	0.80	0.83	0.79	-	-	-
	終稿	0.81	0.87	0.89	0.89	0.84	0.87
H	初稿	0.77	0.64	0.67	0.79	0.75	0.80
	終稿	0.64	0.75	0.73	0.87	0.92	0.76
M	初稿	0.71	0.84	0.52	-	-	0.66
	終稿	0.78	0.78	0.84	0.73	0.82	0.65
Y	初稿	0.82	0.75	0.74	-	-	/
	終稿	0.80	0.77	0.78	0.76	0.71	/
S	初稿	-	0.66	-	-	-	/
	終稿	0.83	0.77	0.65	0.62	0.74	/
T	初稿	0.69	0.76	0.82	0.72	0.84	/
	終稿	0.85	0.77	0.82	0.82	0.85	/
K	初稿	-	0.87	0.67	0.80	-	/
	終稿	0.83	0.87	0.80	0.85	0.86	/
N	初稿	0.69	0.77	0.78	-	-	-
	終稿	0.78	0.86	0.78	0.83	0.84	0.85

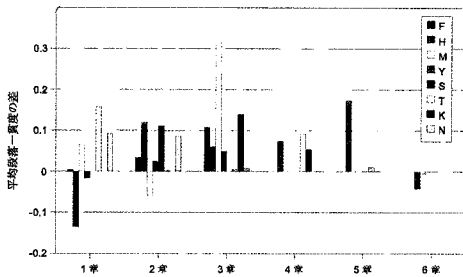


図 6: 各章の平均段落一貫度の差

において初稿よりも最終稿のほうが段落一貫度の向上が確認できる。

5 考察

本手法では、文書の重要な構成要素の1つである段落に注目し、段落はある1つの主題について複数の文で表現したものであることを利用している。それらの文で構成される段落には、それぞれ関連性をもった単語が用いられ、読者に伝えたい主題について記述しようとしていると考えられる。例えば、“自然言語”という主題の段落に、“言語”

に関する名詞が頻出していると考えられる。もしそれら内容のまとまった段落内の文内に段落が主張している内容にそぐわない別の主張をしている不適切文が挿入されている場合、段落内容はその内容に一貫性が保てなくなり、読者にとっても読みにくい文書となる。

本手法では、EDR概念辞書を用い、単語の概念関係から、一文と段落との間の関連性を数値として計算し算出している。また、段落内におけるすべての文の平均関連度を段落一貫度とした。この段落一貫度は、段落内容が同一の主題に関して記述されていた場合は高い値を示し、逆に、段落内容がそれぞれの文で別の主題に関して記述されていた場合は低い値を示すと考えられる。このとき、ある程度内容がまとまった段落内に不適切文が存在していた場合、段落内容に対し不適切であるということから、段落内で用いられる単語と不適切文内で用いられる単語には関連性を見出すことができず、段落内容の一貫性は認められない。

実験1では明らかに違う話題を扱う文を抄録に挿入した上で、段落一貫度の評価を行った。結果として、全ての抄録において段落一貫度の減少が確認できた。段落として評価した同一の抄録内に本来の話題とは違う文の挿入が、内容の一貫性を低

下させ、挿入した文にいたっては段落一貫度をはるかに下回る値が数多く確認できた。これは、数値上では段落に相応しくない不適切な値となっており、挿入した文が抄録内容を阻害していたと判断することができる。

実験2では実験1とは違い、似たような論文抄録の1文を別抄録の1文に変更するという条件のもとで行った実験結果として、50篇の抄録のうち、49篇の抄録の一貫度の減少が確認できた。実験1、2より本手法の提案する指標である段落一貫度が正しく評価し得る有用な指標であるといえる。

実験3の結果は、学生が執筆したそれぞれの卒業論文の段落から求められた段落一貫度の平均値をグラフにして表示している。これは、各学生の卒業論文の段落数はそれぞれ異なっており整合性がとれないため、段落一貫度の平均値として結果を示した。また、論文指導前後の段落を比較したものの、段落構成が変更されている箇所が数多くあり、単純な比較ができないため、段落一貫度の平均値として結果を示した。

そこで卒業論文の各章の平均段落一貫度を比較した場合、初稿の段階では全ての章について執筆している者は少なく、初稿と最終稿との比較は一部の章でしか行えなかった。しかし、比較できる章だけで判断すると、それぞれの執筆者における多くの章の平均段落一貫度が向上していることが確認できる。初稿の段階で比較的高い平均段落一貫度を示した章は、最終稿の段階において大きな変化は無く、高い平均段落一貫度を保ったままだった。逆に、初稿の段階で、他の章の平均段落一貫度と大きな差のある章は、最終稿の段階で、その章の平均段落一貫度が大きく向上しているものが多く存在した。図3から確認できることとして、比較を行うことができた章については、全ての執筆者のほとんどの章において、段落一貫度の向上が判断できる。これは、指導教員が内容のまとまらない卒業論文の1段落に対し、1つの主題について記述するように論文指導した内容が的確に反映されたものだと考えられる。

一方では、1、2章の平均段落一貫度が初稿の段階に比べ、最終稿の段階の方が低いものが存在した。これは、1章や2章の章構成にもよるが、序章にあたる1章では、研究の目的や背景を記述するため、多様な話題を含んでいることが多い。そのため、1章の平均段落一貫度は、比較的低く示され

る場合があると考えられる。実験結果からも、それが確認できる。2章以降は、その研究内容について詳しく記述するため、1章であったような多様な話題を含むことは少ないため、それらの段落一貫度は高く示されると考えられる。このように本論文で提案した段落一貫度は、章によって段落一貫度の値に傾向が見ることができる。

3つの実験結果から、本論文で主張する単語間の関係性によって、段落一貫度は技術文書の段落の一貫性を評価できるものと考えられる。

6 まとめ

本論文では、段落の一貫性評価のために段落一貫度という新しい指標を提案し、実験によって、学術論文を含む技術文書の段落一貫性を評価できることを示した。本提案手法によって算出される段落一貫度は、技術文書の段落一貫性を評価することができ、有益な概念といえる。

本論文で提案する段落一貫度は、章、文書単位で適用することが可能だが、文集合として最小単位である段落を対象とすることが適切だと考えた。仮に、適用範囲を章、文書単位まで拡張した場合、その文書構成法によっては、研究の応用的事例や関連研究を記した内容で記述される文面があるため、本提案手法を用いて章全体から一貫度を求めた場合、その章内容が複数の話題を記してしまい、一貫性を正しく評価できない可能性が考えられる。このようなことから段落という文集合の最小単位の一貫性を評価することが、文書内容の評価に有効だと考えた。

本稿では、段落一貫度という指標によって段落の一貫性を評価しているが、本手法を拡張することによって章や文書全体の一貫性評価を行うことができる。また、段落内の一貫性評価の際、段落を構成する1文ごとに求められるある文の関連度が他の文の関連度に比べ、極端に低い値を示している場合、その文は段落内容において不適切な記述をした文であると判断できる。したがって段落内における不適切な文を判断するための文書校正支援ツールとしての利用が期待できる。

参考文献

- [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/>”.