Vol.2009-NL-192 No.9
2009/7/23

様々な文書を対象とした段落一貫性の解析

板倉由知[†1]　白井治彦[†2]　黒岩丈介[†1]
小高知宏[†1]　小倉久和[†1]

本報告では，段落の一貫性を数値的に評価する尺度である段落一貫度を定義し，段落一貫性を評価する．このとき，評価対象として，章，段落の構成がより一貫した形で構成されると考えられる科学技術文書を対象にし段落一貫性について論じ，段落一貫性の評価として扱う．実験では，数々の科学技術文書を対象に，章，段落それぞれの構成についての段落一貫度の結果をまとめ，比較を行い，それらの特徴を考察する．

Analysis of coherency of paragraph in several documents

Yoshitomo Itakura,[†1] Haruhiko Shirai,[†2]
Josuke Kuroiwa,[†1] Tomohiro Odaka[†1]
and Hisakazu Ogura[†1]

This paper defined the coherent value of paragraph that numerically evaluate the coherency of paragraph and evaluated the coherency of paragraph in the document. In this evaluated method, we have discussed the coherency of paragraph that intended scientific and technical documentation that is more coherent construct of chapter and paragraph and have dealt with evaluation of the coherency of paragraph. In the experiment, we targeted several scientific and technical documentation, and combined results from coherent value of paragraph about the construct of chapter and paragraph. we compared the result, and considered its features.

†1 福井大学大学院工学研究科
graduate school,University of Fukui
†2 福井大学工学部
Faculty of Engineering,University of Fukui

1. はじめに

文書を評価する基準として，段落の内容一貫性を評価する段落一貫度を定義し文書評価を行う[1)]．特に，科学技術論文などの学術文書については章，段落ごとに内容が適切にまとめられていると考えられ，定義した段落一貫度は，技術論文において適切な評価基準となることが期待できる．

論文執筆に慣れていない学生における執筆技術は章，段落構成を意識せず執筆することがしばしばあり，散文的に文書を構成してしまうことがある．このような文書構成では筆者が伝えたい内容が読者に明確に伝わってこない．そのような場合，筆者が容易に評価することができる基準として段落一貫度を提案し，文書校正支援への利用が期待できる．

文書校正の研究分野では，すでに誤字脱字の指摘における文書の表面的な誤りの指摘を行う文書校正支援ツールが存在する[5),6)]．本報告の段落一貫度は，段落の一貫性という意味内容に関して評価を行うことで，段落の一貫性についての文書校正への支援ができると考えられ，この可能性についても論ずる．

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，概念シソーラス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，段落を構成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．

本報告では，対象とする文書として，科学技術文書を対象にした様々な文書を用い，段落一貫度における比較を行い文書における段落一貫性を解析する．

実験として，学術論文，操作マニュアル，新聞等の科学記事などに代表される科学技術文書を対象とし，段落一貫度の比較を行う．これらの結果から段落一貫度の特徴を論じる．

2. 段落の一貫性

文書を評価する１つの基準として，段落構成の一貫性が挙げられる．章，段落では，その構成要素に従い，文書内容の一貫性を意識して文書を記述することが考えられる．このような文書において内容に一貫性が読者にとって見出せない場合，文書内容を相手に十分伝えることができないこととなり，文書としての役割を果たせなくなる．

段落一貫度は，段落一貫性についての１つの評価尺度であり，文書校正に利用が期待でき

る評価基準であるといえる．

本研究では，文書における要素としての段落に注目している．段落とは，読者に伝えたい内容を適切な単語を用い，複数の文で表現した文集合であるといえる．このとき段落内容は，その伝えたい内容について一貫した記述となっており，特に科学技術文書であればこの特徴がより顕著に見られる．科学技術文書では，１段落内に極めて範囲の狭い主張を記述しており，同一段落内に複数の主張が含まれることはない．もし段落内に複数の主張が混在している場合，段落構成として主張が散漫となり，そのような段落は適切な段落構成とは言えない．仮に技術論文内に複数の主張が混在している段落が存在する場合，段落が主題とする内容がゆれてしまい，その段落だけでなく，章や文書全体について，読者に対し理解されにくくなる．この特徴によって，段落内で用いられる単語は伝えたい段落内容に強く関連した単語が多く含まれていると考えられる．

本研究では，この特徴を段落内容の一貫性として解釈し，単語間の意味類似度を用いて段落の内容の一貫性を評価するための段落一貫度という評価尺度を検討した．段落一貫度とは段落内容がある主張についてどれだけ一貫した記述をしているかを示す評価基準である．

2.1 単語間意味類似度

段落一貫度は，概念シソーラス上における単語間意味類似度を用いて算出が行われる．単語間意味類似度とは，概念シソーラス上で 2 単語間の意味的距離を算出される尺度の１つである．図 1 は，概念シソーラスにおける構造の一部を示している．本報告では概念シソーラスとして EDR 概念辞書を用いた．

図は，ルートノードを“ 概念 ”とする木構造を作っている．各ノードは概念を示しており，それぞれのリンクが概念関係を示している．図では，“ 食べる ”,“ 飲む ”といった概念は，“ 飲食する ”という上位概念に包含されているという関係を示している．このとき，各ノードの左肩に記してある数字はルートノードからの距離を示している．

本報告では以下の式に基づき，単語間の意味類似度の計算を行っている[10]．

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

α, β は定数であり，ともに $0 > (\alpha, \beta) > 1$ の範囲の値をとる．このとき α, β はそれぞれ，$\alpha = 0.1, \beta = 0.6$ とした．また，Sim は $(0 > Sim > 1)$ の値をとる．

Sim は単語間意味類似度であり，w_1, w_2 はそれぞれ単語 1，単語 2 を示す．d は，w_1, w_2 の概念シソーラス上におけるノード間距離である．図を例にすると，この場合，“ 食べる ”,“ 飲む ”の二単語間の距離は２となる．h は，w_1, w_2 の単語の持つ概念をどちらも包含する上位概念におけるルートノードからの距離となる．図中では，“ 食べる ”,“ 飲む ”の上位概念は“ 飲食する ”となり，ルートノードからの距離は１０となる．

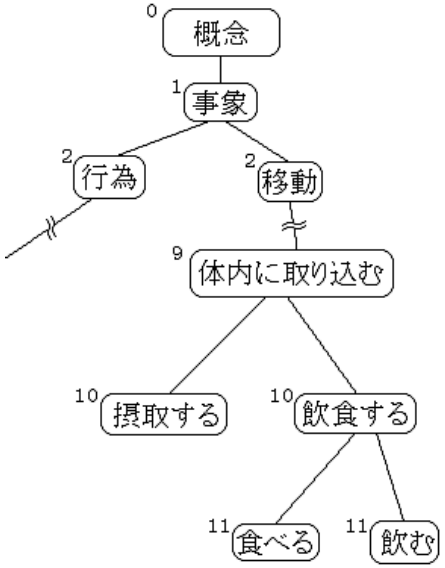

図 1 EDR 概念辞書の構造 (一部)
Fig. 1 A part of structure of EDR dictionary.

d における単語間のノード距離とは，２単語における意味が近ければ，概念シソーラス上では近くに配置されることが分かる．また h におけるルートノードから２単語の上位概念までの距離は，その距離が深くなれば深くなるほど，具体的な概念で共有される．これらの特徴により，単語間意味類似度は意味的に近しい内容の単語であればその値は高く示されることになる．

2.2 段落一貫度と 1 文の関連度

本報告で定義する段落一貫度は，単語間意味類似度を用いた文章間の意味的な距離といえる．

段落の意味内容の一貫性を評価する指標である段落一貫度を算出し文書評価に用いる．段

落一貫度を求める際，段落内容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 を算出するまでの手順を以下に示す[7)–9)]．

2.2.1 段落一貫度の算出

段落内容を構成するすべての文について，段落と 1 文との関連度 R_i を計算することでその平均値を段落一貫度 C と定義する．

$$C = 1/n \cdot \sum R_i$$

このときの n は段落を構成するすべての文の数である．また i は段落内の文番号を示す．

2.2.2 文の関連度の算出

文の関連度 R_i は，先に示した単語間意味類似度 Sim を用いて算出を行う．文の関連度 R_i とは，段落における文 S_i と，段落を構成するその他の文集合との間の意味的な距離を示している．このとき，文の関連度算出のため単語間意味類似を用いることが必要となり，段落文における単語抽出が不可欠となる．段落文から単語を抽出するために，形態素解析を行い，単語の中でも段落内容を特徴づけると考えられる名詞，動詞の抽出を行う．このときの形態素解析には MeCab を用いた．

文の関連度の算出は以下の手順によって求められる．

(1) 単語集合の抽出
段落を構成する 1 文とその他の文集合から，名詞，動詞を形態素解析によって抽出する．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集合から抽出した単語の集合 $W_{P_{(s_i)}} = \{w_1(P_{(s_i)}), w_2(P_{(s_i)}) \cdots\}$ を生成する．このときの m は文 S_i に含まれる名詞，動詞の単語数である．

(2) 1 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，$w_a(s_i)$ に対して最大となる意味類似度 $\max(\mathrm{Sim}(w_a(s_i), w_b(p))$ を求める．$W(s_i)$ 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．

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

以上の手順によって段落一貫度 C は算出される．この段落一貫度は，段落内で使われる単語を基準とした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，段落内容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．

3. 実　　験

本報告では様々な科学技術文書を対象に，文書，章ごとにおける段落一貫性を比較検討するためにそれぞれの段落一貫度を求める．ここでは 25 編の論文を対象にした章構成における段落一貫度についての実験を述べる．

25 編の論文はインターネット上に公開されているもので，検索ワードとして「自然言語」で検索される論文を対象とした．このとき，論文ごとに章構成，章数が独自の形で記述されており，章ごとにおける段落一貫度の単純な比較を行うことができない．本実験では章構成を独自に，導入部，展開部，結末部の 3 つの部に分け，段落一貫度の比較を行った．このときの分類を行った主な内訳として，導入部は 1 章，展開部は 2 章以降，結末部は終章として分類を行っている．以上の内容に基づく実験結果は図 2 に示す．

この結果の具体的な値を表 1 に示す．このとき同時に導入部，展開部，結末部における段落数も示す．

以上の結果では，各論文における章構成の段落一貫度の比較を示している．この結果から，多くの論文における段落一貫度の導入部，展開部，結末部での推移はほとんどなく，平坦な変化を示していることがわかる．その一方で，一部に特徴的な傾向として導入部，展開部に比べ，結末部における段落一貫度の極端な減少が数例確認することができる．

これらの導入部，展開部，結末部における結果について，各部の段落数を交えて比較を行う．図 3,4,5 はそれぞれ導入部，展開部，結末部における段落一貫度とそれらを構成する段落数の相関図である．

図より段落数と段落一貫度の関係は，段落数の多少に関わらず，一定の平均段落一貫度を示すことが分かる．これは導入部，展開部，結末部それぞれにおいて同様の傾向が見られる．

4. 考　　察

本報告における段落一貫度とは，単語の関連性を元に段落の一貫性について評価する指標である．本指標は，科学技術文書のような章，段落構成が厳密に行われ，段落内に 1 つの主張だけが体系的に記述される文書にとって，有益な指標になると考えられる．しかし，同一内容の繰り返した段落構成の場合，段落として適切な体裁ではなく，この際の段落一貫度の値は非常に高い値を算出してしまうことが考えられる．極端な例を示すと，同じ文が繰り返される段落の場合，本提案の段落一貫度は最大値を示すことになる．このような場合，段

Vol.2009-NL-192 No.9
2009/7/23

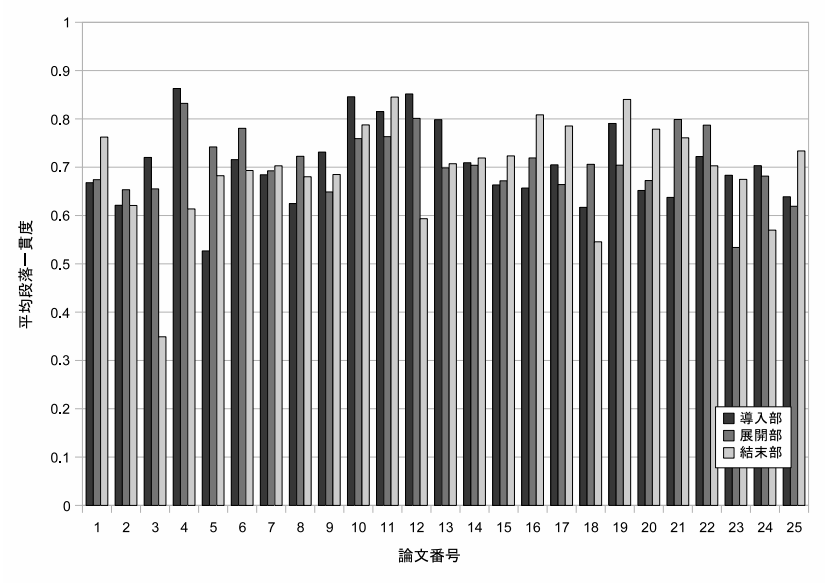

図 2 実験：章ごとにおける段落一貫度
Fig. 2 Coherent value of paragraph in chapters.

落一貫度は有益な指標といえないため，段落一貫度を元にした評価関数の必要性が考えられる．

本報告では，実験として論文を構成する章ごとの段落一貫度を算出し，比較を行った．論文執筆の際，1 章を構成する導入部においては，研究の背景，先行研究などの話題が含まれ，比較的多岐に渡る内容を含むことが通常であると考えられる．同時に，導入部で段落一貫度を算出した場合，導入部ではこの特徴により段落一貫度は展開部よりも低く算出されると考えられる．また，展開部の場合，論文中において最も伝えたい内容であると言え，その他の話題を含む余地がない．この特徴から段落一貫度は，展開部では，他の導入部，結末部における段落一貫度よりも高く算出されると言える．終章を含む結末部では，まとめ，今後の話題などの主張が含まれることで，展開部の段落関連度よりも若干低くなると考えられる．

本報告で示した実験の結果，多くの論文ではそれぞれの導入部，展開部，結末部において上記で示した段落一貫度の特徴が部分的に確認することが出来た．中には上記の特徴を持たない段落一貫度の変化を示すものもあるが，それぞれ平坦な変化を示すものが多く確認できた．しかし，その一方で，導入部，展開部と比べ結末部での段落一貫度が極端な低下を示す論文が数件確認された．

表 1 実験：章ごとにおける段落一貫度と段落数
Table 1 Coherent value of paragraph and the number of paragraph in chapters.

論文番号	導入部	展開部	結末部	導入部	展開部	結末部
1	0.67	0.67	0.75	4	10	3
2	0.62	0.65	0.62	3	27	2
3	0.73	0.66	0.35	4	17	1
4	0.86	0.83	0.61	1	14	2
5	0.53	0.74	0.68	2	13	1
6	0.72	0.78	0.69	2	24	4
7	0.68	0.69	0.70	1	10	1
8	0.62	0.72	0.68	2	28	1
9	0.73	0.65	0.68	2	10	3
10	0.85	0.76	0.79	2	14	1
11	0.82	0.76	0.85	1	7	1
12	0.85	0.80	0.59	1	16	2
13	0.80	0.70	0.71	5	22	2
14	0.71	0.70	0.72	15	60	2
15	0.66	0.67	0.72	5	55	1
16	0.66	0.72	0.81	4	60	1
17	0.70	0.66	0.79	5	37	1
18	0.62	0.71	0.55	7	46	2
19	0.79	0.70	0.84	4	38	1
20	0.65	0.67	0.78	5	37	1
21	0.64	0.80	0.76	6	49	2
22	0.72	0.79	0.70	7	12	2
23	0.68	0.53	0.67	1	77	1
24	0.70	0.68	0.57	3	24	3
25	0.64	0.62	0.73	4	20	1

この結果を算出した論文を具体的に挙げると論文番号 3，4，12，18 の 4 編の論文が相当する．この 4 編の論文は，結末部である終章の段落数を確認すると 1，2 段落だけで構成されていることが分かる．結末部である終章では，論文内容の総括を行うとともに，今後の課題について記述することがある．しかしそれらの話題は別々の主題であり，同一の段落に含むことは，段落内容の一貫性の観点から，問題を生じさせることがある．

今回，特徴的な結果を示した 4 編の論文はそれぞれ 1，2 段落内に上記で示した総括の話

Vol.2009-NL-192 No.9
2009/7/23

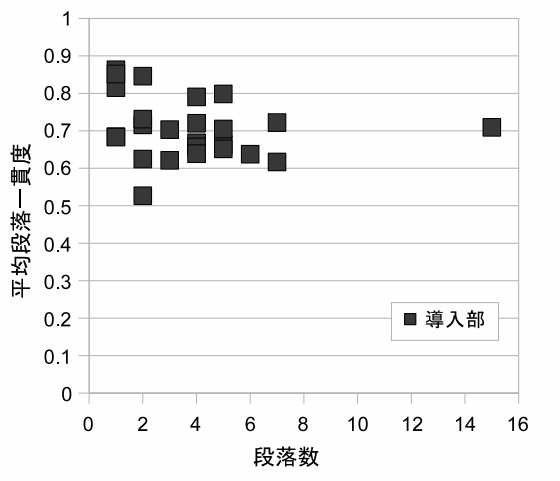

図 3 実験：段落数と段落一貫度 (導入部)
Fig. 3 the number of paragraph and coherent value of paragraph in introduction

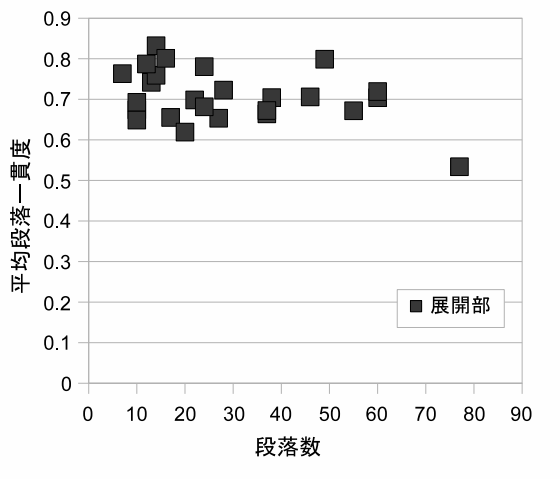

図 4 実験：段落数と段落一貫度 (展開部)
Fig. 4 the number of paragraph and coherent value of paragraph in middle chapters

題，今後の課題についてなどの 2，3 の主張を記述していた可能性が考えられる．実際に論文本文を確認したところ，1 段落中にまとめと今後の課題について記載されているものが確認出来た．この特徴以外にも，同一のものを指し示す名詞の誤字による，いわゆる表記の揺れにより，本手法が処理しきれていない事例が確認された．

この場合，名詞の表記の揺れによって文全体の整合性がとれなくなり，文書の段落一貫性を論ずる以前に文書の信頼性として大きな問題となる場合がある．しかし，そのような場合，実験結果で示したように，段落一貫度の異常な減少が確認出来る．このとき，該当する章の段落ではなんらかの異常を抱えていると考えられる．段落一貫度は，そのような異常を検出することができる文書校正支援のための一手法として利用可能性が考えられる．

本報告の実験では，25 編の論文における各章ごとの段落一貫度と段落数も示した．実験結果からは，段落数の多少に関わらず，段落一貫度は一定の値以上に集中することが判明した．これは通常の論文における段落での段落一貫度は，その内容的に一定の完成度が満たされていると考えられる．段落一貫度とは，前述の通り，単語の関連性を元に算出した段落の一貫性評価指標であり，科学技術文書における段落にとって有益な指標であるといえる．実験結果から，段落数の多少に関わらず段落一貫度が一定以上の値を示すことにより，一般的な科学技術文書における適切な段落一貫度の範囲を検討する．

実験結果を確認すると導入部，展開部，結末部について，段落数の多少に関わらず，どれも 0.6 の値を超えるところに値が集中していることがわかる．一部に 0.6 を下回る論文も存在するが，導入部での段落数が極端に少ない状態で多くの主張を記述し，結末部でも同様に１つの段落に複数の話題を記述するなどの段落構成に難のある論文が確認できた．つまり，段落一貫度の値がある一定の閾値を下回る段落は，何かしらの問題を抱えている可能性があることを示す文書校正支援のための判断材料になり得ることができると思われる．

導入部，結末部は本実験を行う際にあたり，それぞれ 1 章，終章を割り当てているが，展開部である 2 章以降の構成に比べ，章数に大きな差が生まれている．これにより実際の段落構成は展開部における段落数が極端に多くなっている．その一方で，導入部，結末部はひとつの章だけであり，その段落数も限定されてしまう．しかし，それらの章における段落数はこのような背景があるにも関わらず，極端に少ない段落数で構成される章をもつ論文が確認できる．1 章や終章については，少なくとも複数の話題を記述する必要があり，1 段落だけで構成される章はその段落内容の一貫性に問題を抱えている可能性が考えられる．本実験では，すでに考察したようにある論文の結末部で 1 段落内に複数の話題が記述され，段落一貫

Vol.2009-NL-192 No.9
2009/7/23

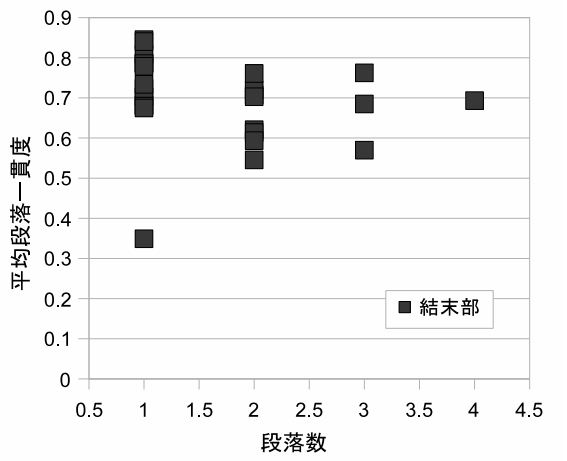

図 5 実験：段落数と段落一貫度 (結末部)
Fig. 5 the number of paragraph and coherent value of paragraph in last chapter

度の極端な低下を示すものが確認できた．それ以外にも，導入部において同一段落内に複数の主張を記述するものが確認できた．これらの結果から，本報告の段落一貫度はその章を構成する段落数を元に，章の構成についても評価できる可能性が考えられる．

5. おわりに

本報告では，段落内容の一貫性を評価するための尺度として段落一貫度の定義し，段落一貫性を評価する指標としての有効性について論じた．本報告における段落一貫度とは，文書の段落内における単語間の関連性を元に算出した段落の一貫性を評価する１つの指標であるといえる．段落一貫度を用いることで文書をある１つの側面からではあるが，有効に評価することができるといえる．

今後の課題として，段落一貫度を文書校正支援のための指標となり得るための検討を行い，段落一貫度を元にした評価関数の必要性についても検討を行う．

参 考 文 献

1) 板倉 由知, 白井 治彦, 黒岩 丈介, 小高 知宏, 小倉 久和,”単語の概念関係を用いた段落の一貫性評価手法”, 電子情報通信学会論文誌 D, vol.J91-D, No.2, pp.1672-1675, June. (2008).
2) 藤沢晃治,“「分かりやすい文章」の技術”, 講談社 (2004).
3) 木下是雄,“理科系の作文技術”, 中公新書 (1981).
4) 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).
5) Microsoft Office Word,“http://www.microsoft.com/japan/office/word/prodinfo/default.mspx”.
6) Justsystem JustRight!2,“http://www.justsystem.co.jp/justright/”.
7) 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).
8) 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).
9) 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).
10) 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.
11) 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).
12) 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).

(平成？年？月？日受付)
(平成？年？月？日採録)