

段落間関係の評価に基づく文章構造推敲支援

Text Structure Polish Support based on Paragraph Relationship Evaluation

山手 砂都美* 砂山 渡
Satomi Yamate Wataru Sunayama

広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

Abstract: We have many opportunities to write a text. However, relationship among paragraphs are hard to be grasped. Therefore, we focused on top-down structures and bottom-up structures between paragraphs. Top-down means texts that a paragraph including a conclusion or what the writer wants to say comes at first and bottom-up means texts that such a paragraph comes lastly. In this paper, a system that supports polish of text structure is proposed. Relationships between paragraphs are expressed as a tree structure and writers can confirm whether a text is top-down or bottom-up. Users of the system can polish their texts to be top-down or bottom-up by seeing output of the system.

1 はじめに

文章を書く機会はさまざまなところである。文章を書いた本人が分かりやすく書いたつもりでも、他人が読むと分かりにくかったり、間違った解釈をしてしまったりすることもある。自分の伝えたい意図を正しく伝えるためには、現状の文章の構造を把握するのはもちろんのこと、文章の構造を推敲する必要がある。しかし、自分で書いた文章の構造を正確に理解し、推敲するのは難しく、自分で推敲するには限界がある。他人に文章構造の推敲を依頼せず、自分で簡単に文章構造を推敲していきたい。

そこで本研究では、文章を段落間のつながりに着目した段落間関係の評価に基づく文章構造推敲支援システムの構築を目的とする。

2 関連研究

2.1 段落間関係の評価に関する研究

本節では、段落間関係の評価に関する研究について述べる。

単語の概念関係を用いて段落の一貫性を解析する研究 [1][2] がある。これらの研究は、あらかじめ一貫した形で構成されていると考えられる技術文章を対象とし、

単語間の意味類似度を用いて提案する段落一貫度が有効であるかどうかを検証したものである。本研究では単語間の意味には着目せず、条件付き確率で段落間の関係性を評価していく。

また、文章のセグメント間関係解析に基づく文章構造解析をする研究 [3] がある。この研究では、小さな意味段落内を修辞構造で扱いつつ、意味段落間の関係づけを行っている。本研究では、意味段落間の関係づけを行っていない。また、前者の研究同様に、条件付き確率で段落間の関係性を評価していく。

2.2 文章の構造化に関する研究

本節では、文章の構造化に関する研究について述べる。

文章構造解析に基づく小論文の理論性についての自動採点を行う研究 [4] がある。この研究では、文章の構造化を文に関して行っている。本研究では、文章の構造化を段落間に関して行っている点、理論性について着目している点と着目していない点が異なっている。

また、情報理論による、文章の理論的構造の解析をする研究 [5] がある。この研究では、文章の理論構造を分析して可視化を行っている。本研究では、理論構造の分析を行わず、条件付き確率の大きい段落間をつなぐことで構造化を行っている。

また、理工系学生を対象とした技術文書作成支援システムを作成した研究 [6] がある。この研究では、あらかじめ、論文のルールに沿って書かれていない文、意

*連絡先：広島市立大学大学院情報科学研究科システム工学専攻
〒731-3194 広島県広島市安佐南区大塚東三丁目4番1号
E-mail: {yamate,sunayama}@sys.info.hiroshima-cu.ac.jp

図が分かりにくい文を検出する機能があり、分かりにくい文をクラス図で可視化している。文章の構造を可視化している点は同じだが、本研究では、段落間を対象とした構造を可視化している。

更に、以上の関連研究では、話の分岐として、話を広げている段落、話をまとめている段落に着目していない点で異なる。

2.3 文章の推敲に関する研究

本節では、文章の推敲に関する研究について述べる。

文の理論構造デザインツールの試作と校正・推敲支援ツールを作成した研究 [7][8] がある。これらの研究では、推敲は文章の理解向上を指しており、句読点の統一チェック、長文チェック等、細かい点に着目している。本研究では、推敲に関して細かい点に着目せず、文章の構造に着目し、文章の枠組みに着目している点異なる。

また、文を分かりにくくする要因の分析と改善支援手法の提案をした研究 [9] がある。この研究では、ユーザである学生が陥りやすい誤りを分析し、その結果より、修正すべき箇所の優先順位と改善方法を指摘している。本研究では、ユーザの誤りから改善方法を指摘するのではなく、現状の文章構造から、より良い文章構造を提案して行く。

3 本研究で扱う文章構造

本章では、本研究で扱う文章構造について述べる。

本研究では、トップダウン構造とボトムアップ構造に着目する。トップダウン構造とは、結論や話の全体の構造に関する説明を先の段落で述べた後、その詳細を後の段落で述べている構造のことを指し(図1左)、ボトムアップ構造は、トップダウン構造とは逆に、先の段落で部分的な話の詳細を述べ、結論や全体のまとめを後の段落で述べている構造のことを指す(図1右)。

図2にトップダウン構造とボトムアップ構造の具体例を示し、図3に図2の各段落に使われている単語を示す。

まず、トップダウン構造に着目する。第1段落とつながっている第2段落では単語Aのみ共通に使われており、それぞれの段落の単語の使用割合は、第1段落では、3単語中に1単語、第2段落では7段落中に1単語である。第1段落よりも第2段落では使用単語が増えていることから、第1段落から第2段落に単語Aについて話が展開していると捉えることができ、矢印の向きを付けることが出来る。第3段落、第4段落にも着目すると、第3段落では単語Bについて、

第4段落では単語Cについて話が展開していると捉えられる。

次に、ボトムアップ構造に着目する。第5段落とつながっている第2段落間をトップダウン構造と同様に考えると、単語Aを共通に使われ第2段落から第5段落へ単語Aについて話がまとまっていると捉えることが出来る。同様に、矢印の向きを付けることが出来る。ただし、文章は小さい段落番号から大きい段落番号へ話が進むため、この場合、第5段落から第2段落へ話が広がっているのではなく、第2段落から第5段落へ話がまとまっていると考える。

結論が先に書かれている場合、第5段落を取り除くとトップダウン構造の文章、逆に結論が最後に書かれている場合、第1段落を取り除くとボトムアップ構造の文章になる。

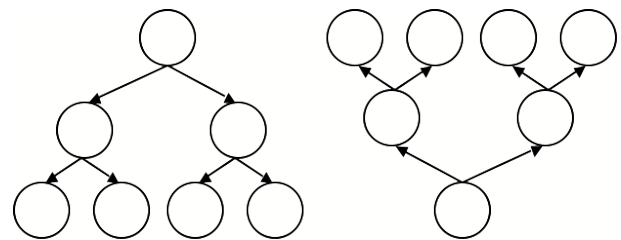


図1: トップダウン構造とボトムアップ構造

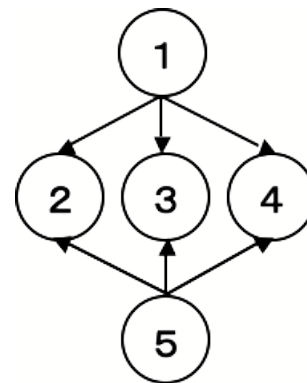


図2: トップダウン構造とボトムアップ構造の例

1段落	2段落	3段落	4段落	5段落
ABC	A DEFGHI	BJKLMNO	CPQRSTU	ABC

図3: 図2の各段落で使われている単語

4 段落間関係の評価に基づく文章構造推敲支援システム (SAT)

4.1 段落間関係の評価に基づく文章構造推敲支援システム (SAT) の構成

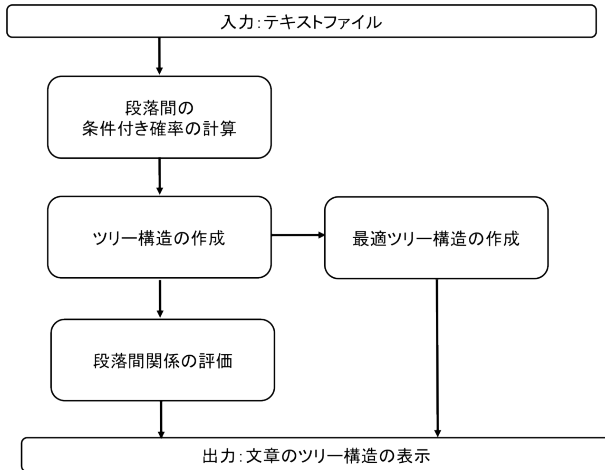


図 4: 段落間関係の評価に基づく文章構造推敲支援システム (SAT) の構成

本節では段落間関係の評価に基づく文章構造推敲支援システム (SAT: Segment Association Tree) の構成について述べる。以下, SAT システムと呼ぶ。

図 4 に SAT システムの全体の構成を示す。SAT システムにテキストファイルを与え、それを元に段落間の条件付き確率を計算を行い、リンクをつなぐ段落間を決定する。次に、段落間の条件付き確率を元にリンクをつなぎ、文章のツリー構造を作成し段落間関係の評価を行う。また、ツリー構造を推敲するために、段落並べ替えによる最適ツリー構造を作成する。最後に、それぞれの結果をインタフェース上へ出力し、ツリー構造と最適ツリー構造の比較を行い、文章構造の推敲をする。

以下、各処理について述べる。

4.2 段落間関係の評価

本節では、トップダウン構造とボトムアップ構造の 2 つで表される段落間関係の評価方法について述べる。

4.2.1 段落間の条件付き確率の計算

本項では、段落間の条件付き確率の計算方法について述べる。以下の式 (1) で全ての段落間の条件付き確率 $Relation(A, B)$ を計算をする。段落 A で使われてい

る単語集合を W_A 、段落 B で使われている単語集合を W_B とし、それらを数える関数 n へ与える。段落 A で使われている単語集合中、段落 A と段落 B で共通して使われている割合を求めることによって、段落間のリンクに向きをつけることが出来ると考えられる。

$$Relation(A, B) = \frac{n(W_A \cap W_B)}{n(A)} \quad (1)$$

4.2.2 段落間関係の評価

本項では、段落関係の評価方法について述べる。

以下、後述する段落間の関係を表すツリー構造 (段落をノードとし、枝には条件付き確率 $Relation(A_i, A_j)$ が与えられる) が作成されたとき、 n 段落構成の文章のトップダウン構造の評価値を与える際のアルゴリズムを示す。

1. 各段落 A_i について、枝分かれしている枝数 $Branch_i$ をカウントする。
2. 各段落 A_i の、 $Branch_i$ が 2 以上の段落 A_i の各枝 j について、最も長い葉ノードまでの枝に与えられている、条件付き確率 $Relation(A_i, A_j)$ を加算し、 $BValue_{ij}$ とする。
3. 各段落 A_i について、すべての $BValue_{ij}$ を掛け合わせた $NValue_i$ を求める。
4. $i < n/2$ の段落の $NValue_i$ を加算、 $i \geq n/2$ の段落の $NValue_i$ を減算した合計値を、トップダウン構造の評価値 $TopValue$ とする。

なおステップ 4 で、話を広げるのは前半段落、話をまとめるのは後半段落、の考えから前半の段落にトップダウンがある場合は加算し、後半の段落にある場合は減点する。

またボトムアップ構造の評価値は、トップダウン構造の評価値の計算アルゴリズムと同様で、作成されるツリー構造の天地逆転をしてできるツリーのトップダウン構造の評価値と同じとする。

4.3 文章の強リンク構造の作成

本項では、文章の強リンク構造を作成する方法について述べる。強リンク構造とは、 n 段落構成の文章について、各段落をノード、ノード間のリンクを条件付き確率 $Relation(A_i, A_j)$ として、最少の強いリンクのみで作成したツリー構造を指す。以下に、強リンク構造を作成するアルゴリズムを示す。

1. 全ての段落 $A_i, A_j (0 \leq i, j \leq n, i \neq j)$ 間の条件付き確率 $Relation(A_i, A_j)$ を計算する。

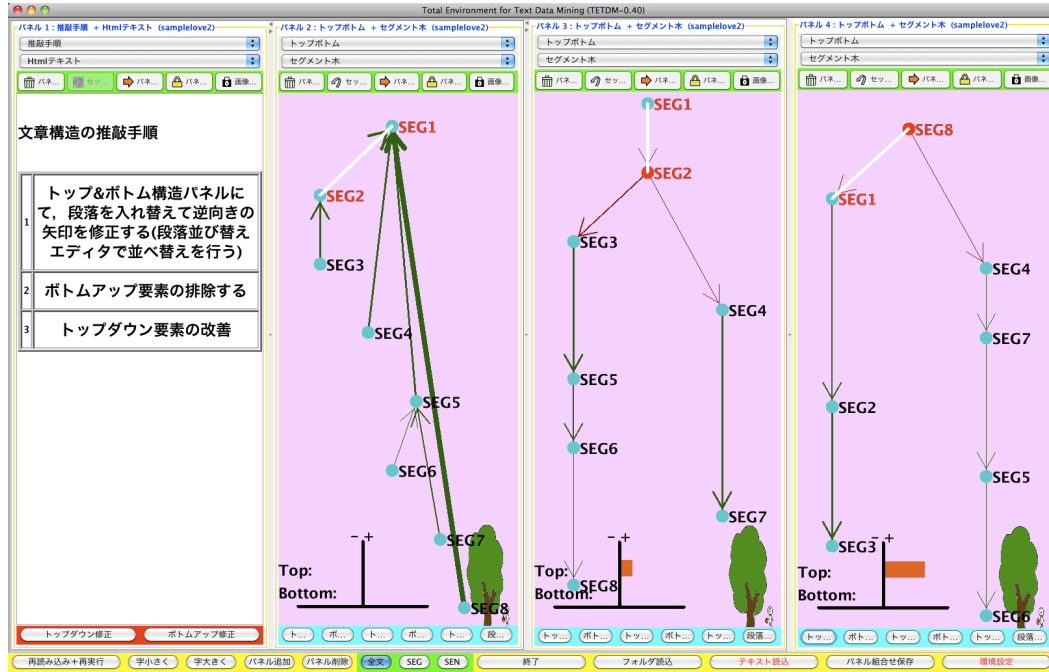


図 5: 文章構造推敲支援システムの使用例

2. 各段落 A_i に関して、段落 A_i との条件付き確率 $Relation(A_i, A_j)$ が最も高い段落 A_j との間にリンクを生成する。(この時点ではリンクの向きは考えない)。
3. n 個の段落のすべてがリンクでつながっていれば (リンクによる段落間のパスが存在すれば) 5.へ。そうでなければ、全段落間の条件付き確率 $Relation(A_i, A_j)$ を大きい順にソートする。
4. ソートされた条件付き確率 $Relation(A_i, A_j)$ の大きい順に、段落 A_i と段落 A_j との間にリンクを生成し、すべての段落がリンクでつながるまでリンクの生成を繰り返す。
5. 第1段落を根ノード (親ノード) とし、つながっているノードを順に子ノードとしたツリー構造を生成する。
6. リンク生成に用いた条件付き確率 $Relation(A_i, A_j)$ をもとに、リンクに $A_i \rightarrow A_j$ の向きを与える。

まず、全ての段落 A_i, A_j 間の条件付き確率 $Relation(A_i, A_j)$ の計算を行い、各段落 A_i に関して、条件付き確率 $Relation(A_i, A_j)$ が最も高い A_j との間にリンクを引く。次に、 n 個の段落の全てがリンクでつながれていなければ、全段落間の条件付き確率 $Relation(A_i, A_j)$ を計算し、まだつながれていない段落間で一番大きい段落間にリンクをつなぐ。一番条件付き確率が高い段落間にリンクを引くことによって、つながりの強いツリー構造が出来ると考えられる。

4.3.1 トップダウン構造の作成

本項は、トップダウン構造の作成について述べる。トップダウン構造のアルゴリズムは、4.3 のアルゴリズムのステップ2で段落 A_j に着目し、下向きの矢印が入ってくる段落 $A_i (i < j)$ の中で、条件付き確率 $Relation(A_i, A_j)$ が一番高い段落間にリンクを生成する。また、他のステップは同様である。下向きの矢印の条件付き確率 $Relation(A_i, A_j)$ に着目することによって、トップダウンの段落からつながっている段落元を探すことができる。

4.3.2 ボトムアップ構造の作成

本項はボトムアップ構造の作成について述べる。ボトムアップ構造のアルゴリズムは、トップダウン構造と同様に4.3のアルゴリズムのステップ2が異なる。段落 A_i に着目し、上向きの矢印入ってくる段落 $A_j (i < j)$ の中で、条件付き確率 $Relation(A_i, A_j)$ が一番高い段落間にリンクを生成、他のステップは同様である。上向きの矢印の条件付き確率 $Relation(A_i, A_j)$ に着目することによって、ボトムアップ構造からつながっている段落元を探すことができる。

4.4 最適ツリー構造の作成

本節では、最適ツリー構造の作成について述べる。最適ツリー構造とは、文章内の段落の順番を任意に変更

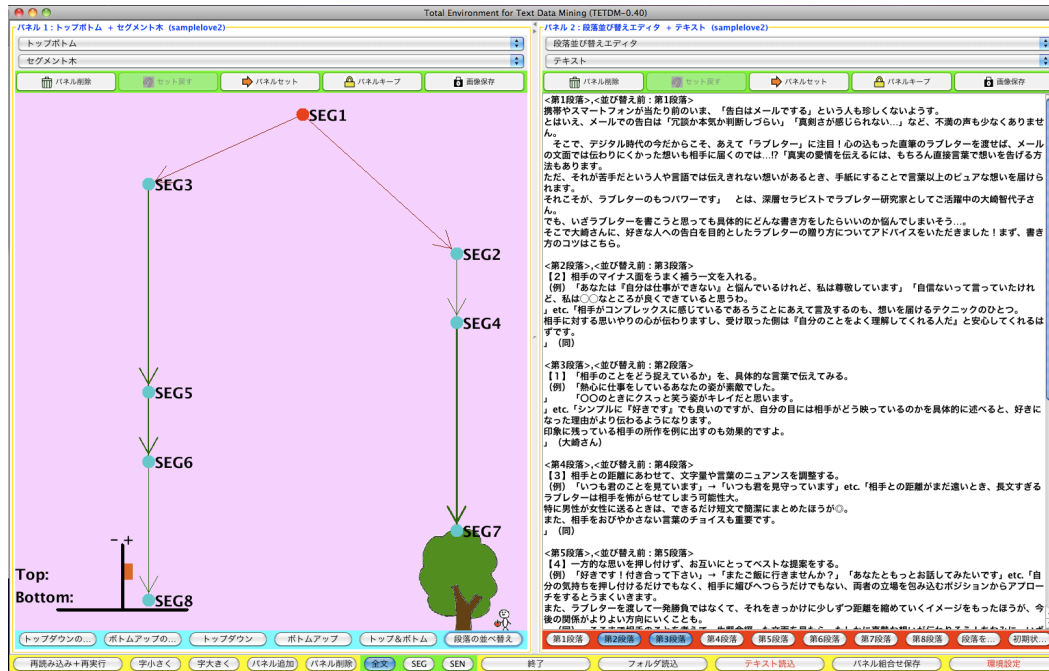


図 6: 段落並び替え可能のインタフェース

する中で、トップダウン構造、またはボトムアップ構造の評価値が最も高くなる構造のことを指す。

1. 文章内の段落 (段落数 n) を並べ替えて出来る $n!$ 通りの全てのパターンを作成する。
2. 作成した各パターンについてツリー構造を作成し、評価値を計算する。
3. 評価値が最も高くなった段落の並びについて、先頭の段落を根ノード (親ノード) つながっているノードを順に子ノードとしたツリー構造を生成する。

これにより現時点のツリー構造と最適ツリー構造との比較を促し、文章構造の推敲を支援する。

4.5 出力：文章のツリー構造の表示

本節では、システムの出力：文章のツリー構造の表示について述べる。4.2 段落間関係の評価、4.4 最適ツリー構造で述べたものを TETDM[11] へ実装している。

SATシステムの出力例を図5へ示す。文章は、mixi[12]のコラムに掲載されていた「ラブレターの書き方のアドバイス」を用いている。段落間をつなぐリンクを線で表し、条件付き確率の大きさによってリンクの太さを変更し pixel で表す。トップダウン構造、ボトムアップ構造の場合は、該当する段落とそのリンク先の色を変える。また、左下にトップダウン、ボトムアップの評価値の値を表示している。

4.5.1 各パネルの表示

図5より、左パネルから順に、文章構造の推敲手順、4.3 節の強リンク構造、4.3.1 項トップダウン構造、4.4 節トップダウンの最適構造を表示しており、これらを用いて文章構造の推敲を行う。強リンク構造では、矢印の向きが全て上を向いており、あとの段落から第1段落をまとめていると解釈が出来る。次に、真ん中のパネルのトップダウン構造では、第2段落は下向きに複数の段落に話を広げていることが分かり、第2段落は話を展開していると読み取ることができる。最後に、右パネルのトップダウンの最適構造では、段落の並び替えによる最適ツリー構造を表示している。

4.5.2 並び替え可能のインタフェース

本項は、並び替え可能のインタフェースについて述べる。

図6に段落並び替え可能のインタフェースを表示する。右側のパネルは文章を示しており、背景が赤いボタンに段落番号を表示している。並び替えを行いたい2つの段落を選んで「段落を並び替える」ボタンを押すことによって段落を入れ替えることが出来る。次に、左のツリー構造パネルで「並べ替えの結果」ボタンを押すと、右パネルで並び替えた結果を反映し、最適ツリー構造を表示を行う。このインタフェースを使用することによって、適宜、段落の順番を並び替えてツリー構造を確認することができる。

4.5.3 文章構造推敲支援システム (SAT) の使用の流れ

本項は、文章構造をトップダウン構造に修正したい場合の SAT システムの使用の流れについて述べる。なお、ボトムアップ構造に修正した場合はトップダウン構造の場合と逆になる。

1. 強リンク構造のパネルで、段落を入れ替えて逆向きの矢印を修正する (段落並び替え可能のインタフェースで並べ替えを行う)
2. 同パネルで、つながっている段落間で共通に使っている単語を減らすことでボトムアップ要素の排除を行う。
3. 現状のトップダウン構造より更にトップダウン構造にするため、トップダウンの最適ツリー構造と比較を行い、よりトップダウン構造になる段落間に共通単語を多く使う。また、段落の順序入れ替えを行う。

以下、各手順について述べる。

1 について、段落の順番を入れ替えて文章の修正を行う。トップダウン構造よりに修正を行うため、矢印は全て下向きにする。段落の入れ替えが出来ない場合、この手順を飛ばす。

2 について、ボトムアップ要素の排除を行う。文章をトップダウン構造にしたい場合、ボトムアップ構造を排除する必要があるため、複数の段落と話がつながっている段落を一緒にする。または、共通に使う単語数を減らして複数段落をつなぐのではなく、1つの段落のみをつなぐ修正する。

3 について、現状の構造よりも更にトップダウン構造になる修正を行う。主に、該当する段落間に共通単語を使用する。

5 結論

本研究では、文章を段落間のつながりに着目した段落感関係の評価に基づく文章構造推敲支援システム (SAT システム) を提案した。文章構造の推敲のために、ツリー構造表示と最適ツリー構造表示を行い、それぞれのツリー構造を比較することにより、文章構造を推敲を行うことを示した。

今後は、実装した SAT システムの精度を計る実験を行うために、実験で使う文章の収集、準備をおこなっていく。

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一貫性評価指標の有効性, 情報処理学会研究報告, NL-183, pp.107-113, (2008)
- [2] 板倉由知, 白井治彦, 黒岩丈介, 小高知宏, 小倉久和:様々な文書を対象とした段落一貫性の解析:情報処理学会研究報告, NL-192, pp.1-6, (2009)
- [3] 春日隆緒, 田村直良:文章のセグメント間関係解析に基づく文章構造解析:情報処理学会研究報告, NL-155, pp.59-64, (2003)
- [4] 藤田彬, 田村直良:文章構造解析に基づく小論文の理論性についての自動採点, 第9回情報科学技術フォーラム講演論文集 7(2), pp.21-22, (2008)
- [5] 奥出 信一郎:情報理論による, 文章の理論的構造の解析, 電子情報通信学会技術研究報告, pp.1-5, (2008)
- [6] 松本章代, 山田未央佳, 山田翔, 鈴木雅人:理工系学生を対象とした技術文書作成支援システム, 情報処理学会研究報告, CE-98, pp.91-96, (2009)
- [7] 大野博之, 稲積宏誠:文の構造デザインツールの試作と校正・推敲支援ツールとの連携, 電子情報通信学会技術研究報告, ET, 教育工学 108(146), p.73-78, (2008)
- [8] 奥村有希, 大野博之, 稲積宏誠:技術文章作成支援ツールの推敲支援機能の拡張-長い修飾節に起因する悪文の検出手法の提案 (次世代情報教育の構築に向けて/一般), 電子情報通信学会技術研究報告, ET, 教育工学 107(536), (2008)
- [9] 須藤崇志, 丸山広, 中村太一:文を分かりにくくする要因の分析と改善支援手法の提案, 電子情報通信学会技術研究報告, KBSE, 知能ソフトウェア工学 108(65), p41-46 (2008)
- [10] 松本裕治, 山下達雄, 平野義隆, 松田寛, 高岡一馬, 浅原正幸:形態素解析システム『茶筌』, Ver.2.4.0, 使用説明書, (2007)
- [11] 砂山渡, 高間康史, ダヌシカ ボレガラ, 西原陽子, 徳永秀和, 串間宗夫, 松下光範:テキストデータマイニングのための統合環境, 人工知能学会論文誌, Vol.26, No4, p483-493, (2011)
- [12] mixi:<http://mixi.jp/>