Vol.2010-NL-199 No.17
2010/11/19

テキストの断片に対する局所的一貫性モデル

横野 光[†1] 奥村 学[†1]

本論文では，テキスト中の各文について，その先行文脈における局所的一貫性を判定する，テキストの断片に対する局所的一貫性モデルを提案する．このモデルはテキストに出現する要素の構文的な役割の遷移の傾向に基づく局所的一貫性モデルである Entity Grid モデルを基にしている．先行文脈内の一貫性は保たれていると仮定を置き，先行文脈中に一貫性が悪い箇所が存在した場合，その箇所より前の文を先行文脈から除外する．人工的に一貫性の悪い箇所を作成したデータを用いた実験により，提案モデルがベースラインよりも性能が上回ることを示した．

Local Coherence Model for Text Segments

Hikaru Yokono [†1] and Manabu Okumura[†1]

In this paper, we propose the local coherence model for text segments which consist of a target sentence and its precedent context. Our model is based on the Entity Grid model, which is the local coherence model considering tendency of grammatical roles of entities in text. Assuming that the precedent context is coherent, we exclude an incoherent fragment inside it. The experiment using automatically created incoherent texts shows that our model outperforms the baseline model.

1. はじめに

テキストの自動評価は自動要約や機械翻訳のようなテキストを生成するタスクにおけるモデルの評価だけでなく，小論文の校正支援などにも用いられる．テキストの評価は大きく分けると，内容に対する評価とテキストの質に対する評価の 2 種類が考えられる．前者はそのテキストに必要とされている情報が正しく含まれているかということに対する評価であり，後者は読みやすさや文法性，一貫性などテキストの容認可能性に対する評価である．テキストの性質の一つであるテキスト一貫性[1)] は文章の意味的なまとまりの良さであり，例えば因果関係や文章構造などによって示される文同士の繋がりである．一般的に，テキストの評価では内容に関する評価に主眼が置かれるが，正しく内容が理解されるためにはテキストが読む人にとって理解しやすいものでなければならない．これはテキストの質の評価が高くなければならないということを意味する．

本研究ではテキスト一貫性の中でも局所的一貫性と呼ばれる性質に焦点を当てる．局所的一貫性とは相前後する 2 文間における一貫性である．これに対して，大域的一貫性と呼ばれる性質もある．これはテキストにおける話題の遷移に関する繋がりの良さのことを示す．

また，評価の単位に関しては，一般的なテキストの評価はテキスト全体を対象として行われる．しかし，小論文の添削支援や校正支援などを目的としたテキスト評価においては，テキスト全体の良さを評価するだけでなく，より具体的にテキスト中のどこに問題があるかを指摘する必要があると考えられる．既存のテキスト評価に関する研究では，その多くがテキスト全体を対象にしたものである．

以上のことから，本研究では複数文からなる断片に対する局所的一貫性モデルを提案する．提案モデルはテキストの断片中の最後の一文に着目し，先行文脈に対してその文が一貫的かどうかを評価する．これによって具体的に一貫性の悪い箇所を推定することが可能になると考えられる．

2. 関連研究

局所的一貫性のモデルとして Barzilay ら[2)] は Entity Grid を利用したモデルを提案している．Entity Grid とはテキストの各文に出現する要素を行列で表現したものであり，行列の成分には単なる出現の有無や出現した要素の構文役割が示される．このモデルはテキスト中で述べられている要素の遷移に着目している．これは，センタリング理論[3)] で示されているように，一貫性のあるテキストではその文中の要素の出現に規則性があるという考えに基づいている．

この Entity Grid を用いたモデルを基にした様々な一貫性モデルが提案されている．Elsner ら[4)] は，Entity Grid モデルがテキストの一貫性において要素の遷移にのみ着目しているということに言及し，例えば参照表現や対象の要素がこれまでに既に述べられている要素かどうか，などといった他の要素をモデルに組み込んでいる．Filippova ら[5)] は Entity Grid モ

†1 東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology

デルに要素間の関係を考慮したモデルを提案し，このモデルをドイツ語の新聞記事に対して適用した結果を報告している．Cheung ら[6] はドイツ語のテキストを対象にして，抽象化した要素の位置を考慮したモデルを提案している．横野ら[7] はテキスト結束性が一貫性に影響を与えるという仮定を置き，結束性に関係する要素を組み込んだモデルを提案している．

一方，大域的な一貫性に関しては，隠れマルコフモデル (HMM) を採用したモデルを Barzilay ら[8] が提案している．このモデルでは文章中の話題を HMM における隠れ状態と見なし，話題の一貫性を隠れ状態の遷移確率によって表現している．また，Soricut ら[9] や Elsner ら[10] は局所的な一貫性と大域的な一貫性を同時に考慮するモデルをそれぞれ提案している．これらのモデルは Entity Grid を用いたモデルと HMM を組み合わせたものである．

一般的にテキストにおいて話題が変わる箇所では局所的一貫性が悪くなると考えられるため，本研究で提案するモデルは局所的一貫性という観点からのテキストセグメンテーションとみることができる．テキストセグメンテーションに関する先行研究には，文間の類似度を基にした Hearst による TextTiling 法[11] や，この手法の問題点を解決するために語の結束性などを考慮した平尾らのモデル[12] がある．また，Sun ら[13] は，まず Latent Dirichlet Allocation を用いてテキストを複数のブロックに分割し，それらの意味的類似度などに基づいて最終的な実際の分割点を動的計画法によって求めるモデルを提案している．

3. 断片に対する局所的一貫性モデル

本研究ではテキスト中の各文について，その先行文脈における局所的一貫性を判定する，テキストの断片に対する局所的一貫性モデルを提案する．対象の文とその先行文脈からなるテキストの断片を 1 つの事例とし，それに対する素性ベクトルを作成して，先行文脈に対する対象の文の一貫性の有無を判定する．既存の Entity Grid を用いたモデルは 2 個のテキストの一貫性を比較するモデルであるが，提案モデルは 2 値分類モデルである．なお，先行文脈のサイズはあらかじめ人手で設定する．

提案モデルでは先行文脈内の一貫性は保たれているものと仮定している．従って，テキストの最初の文から順番に一貫性の判定を行い，先行文脈中に一貫性の悪い箇所が存在した場合，それより前の文を先行文脈から除外する．

以降，断片から素性ベクトルを作成する方法と先行文脈の決定についてそれぞれ述べる．

3.1 断片の素性ベクトル

素性ベクトルの作成方法に関して本研究では 2 種類のモデルを提案する．1 つは筆者らが以前に提案した Entity Grid の拡張を利用したモデルであり，もう 1 つは対象の文と先行文脈中の各文との関係を考慮したモデルである．

3.1.1 Entity Grid の拡張を利用したモデル

テキストの断片を一つのテキストと見なして，Barzilay らの Entity Grid を用いたモデル[2] を拡張したモデル[7] における素性ベクトルの作成手法を用いて素性ベクトルを作成する．

Entity Grid を用いたモデルは，一貫性のあるテキストではその中に出現する要素の出現に傾向があるという考えに基づいて，要素の構文役割の遷移確率を素性としたベクトルをモデルに使用している．Barzilay らが用いた構文役割は主語，目的語，それら以外での出現である．

これに対し，拡張したモデルでは一貫性と同様にテキストの性質の一つであるテキスト結束性に着目し，結束性は一貫性と密接に関係しているという仮定に基づき，Entity Grid モデルに対して結束性に寄与する要素である接続表現や語彙的結束性を考慮に入れるといったことなどで拡張している．

接続表現の考慮では，要素の遷移を接続表現から導かれる接続関係毎に別々に考え，各関係毎に遷移確率を計算している．このモデルでは 3 種類の接続関係を考える．この関係は市川による文脈展開に基づいて 3 種類に分けられたものである[14]．接続関係を考慮した遷移確率の計算の例を以下に示す．**表 1** 中の “S”，“O”，“X” はそれぞれ主語，目的語，それら以外を表す．

s_1. $[e_1]_S$ ... $[e_2]_O$ ...
s_2. **そして，** $[e_2]_S$ ... $[e_3]_O$ ...
s_3. $[e_3]_S$ ... $[e_1]_X$ ...
s_4. $[e_4]_S$ ...

図 1 テキスト例

表 1 **図 1** に対する Entity Grid

	e_1	e_2	e_3	e_4
s_1	S	O	-	-
s_2	-	S	O	-
s_3	X	-	S	-
s_4	-	-	-	S

図 1 のベクトルを**図 2** に示す．SS_{G_i} はある接続関係の分類 G_i における構文役割の遷移 [SS] の遷移確率を表す．この例において拡張モデルでは接続詞 “そして” によって隣接している文 s_1 と文 s_2 の遷移とそれ以外の遷移は別に扱われる．

語彙的結束性に関しては語彙的連鎖[15] を利用して要素のクラスタリングを行い，同じクラスタに属する要素を同一のものとして遷移を数えている．本研究ではクラスタ間の遷移確

Vol.2010-NL-199 No.17
2010/11/19

SS_{G_1}	$\cdots$	$--_{G_1}$	SS_{G_2}	$\cdots$	$S\text{-}_{G_2}$	$\cdots$	$--_{G_2}$	SS_{G_3}	$\cdots$	$S\text{-}_{G_3}$	$\cdots$	$--_{G_3}$
0	$\cdots$	0	0	$\cdots$	.25	$\cdots$	.25	0	$\cdots$	.06	$\cdots$	.13

図 2 接続関係を考慮した文書ベクトルの例 (一部)

率に関する素性に加えて，要素間の遷移確率も素性に加えている．

これらの他にも，文の主題を表す取り立て語の副助詞 “は” を別の構文役割とするなどの構文役割の拡張などを行っている．

3.1.2 先行文脈中の各文との関係を考慮したモデル

このモデルでは対象の文と先行文脈中の各文との構文役割の遷移を考える．対象の文を s_i，その先行文脈を $s_{i-k}, \ldots, s_{i-1}$ としたとき，各 $j(1 \leq j \leq k)$ に対して文 s_{i-k} から文 s_i への構文役割の遷移に関してその遷移確率の計算を行い，それらを結合させたものを対象の断片の素性ベクトルとする．

素性ベクトルについては Entity Grid を拡張したモデルと同様の接続関係，語彙的結束性，構文役割の拡張を考慮した要素間の構文役割を利用して作成する．

3.2 先行文脈の決定

提案モデルでは先行する文脈に対して対象の文の一貫性を判定する．このとき先行文脈中の一貫性は保たれているものと仮定している．そこで，先行文脈が一貫性のない箇所を含む場合は，一貫性がないとされている箇所より前の文を先行文脈から除外する．従って，先行文脈の文数はそれまでの一貫性判定の結果によって変化し，最大であらかじめ指定した文数となる．

例として，**図 3** のテキストの各文に対する先行文脈を**表 2** に示す．なお，先行文脈として考慮する文数を 2 とする．ここで，文 s_3 に関しての一貫性を判定する場合は，先行文脈として文 s_1 と文 s_2 が用いられる．その結果，一貫性が無いと判定されると，その次の文 s_4 の一貫性の判定の際の先行文脈には，文 s_2 と文 s_3 ではなく，文 s_3 のみを考慮するということになる．

4. 実　　験

本研究で提案したモデルの性能を評価するために，新聞記事を用いて一貫性が悪い箇所を推定するという実験を行った．実際の新聞記事は出版される前に人手による校正が入るため，我々が目にする記事の一貫性は保証されており，その状態から文の順番を操作すると一貫性は悪くなると考えることができる．従って，実験で使用するテキストに対して，文の順序を操作することによって人工的に一貫性が悪い箇所を作成し，その箇所をどの程度推定することができるかによってモデルを評価する．

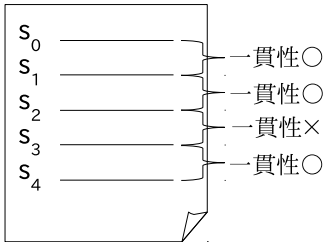

図 3 テキスト例

表 2 各文に対する先行文脈

対象の文	先行文脈
s_1	s_0
s_2	s_0, s_1
s_3	s_1, s_2
s_4	s_3

実験では一貫性の欠落箇所を作成するための文の操作として，テキスト中の文の削除を行った．**図 4** に例を示す．この例では文 s_2 の削除によって，本来は連続していない文 s_1 と文 s_3 が隣接することになり，この間の一貫性は悪くなっていると考える．

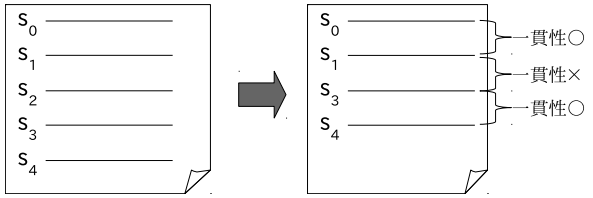

図 4 一貫性の欠落箇所の作成の例

また，削除する文の数が多くなるほど，その間の一貫性はさらに悪くなると考えられる．従って，2 文を削除することで作られた一貫性の欠落を発見するタスクは 1 文の削除による一貫性の欠落の発見のタスクに比べて容易であると考えることができる．そこで本実験では，タスクの難易度による性能の違いについて調査する．実験では，1 文の削除という操作をテキスト中の 2 箇所で行った場合 (以下，skip.1x2 と呼ぶ) と，連続する 2 文の削除をテキストの 2 箇所で行った場合 (以下，skip.2x2 と呼ぶ) の 2 種類のデータセットに対して実験を行った．

実験ではモデルの学習に 2000 年版の毎日新聞コーパスの記事を，テストに 2001 年版の毎日新聞コーパスの記事 400 件を用いた．識別モデルには線形カーネルの Support Vector

Vol.2010-NL-199 No.17
2010/11/19

Machine(SVM) を用いた．SVM の実装には SVM$^{light\star 1}$ を利用している．

Entity Grid からの素性ベクトルの作成においては，遷移確率の計算には文 2-gram と文 3-gram の遷移を利用した．また，語彙的結束性で使用する閾値，SVM のパラメータなどは実験で用いたものとは別のデータによって同様の実験を行い，最も性能が良かったものを使用している．

本実験は一貫性の悪い箇所を推定するというタスクであるため，一貫性の悪い箇所を正例，そうでない箇所を負例として扱う．skip.1x2，skip.2x2 ともに 1 テキスト中には正例となる事例が最大でも 2 箇所しか存在せず，その他は全て負例であるという偏りのあるデータになっている．そのため，正例と負例のバランスを取るために，学習で使用するデータはテキスト単位ではなく，事例単位で正例と負例の数が等しくなるように設定したものを用いている．これに対して，テストではテキスト単位で行っている．これはこのモデルが実際に用いられる場合，その入力はテキストであることを想定している．

評価には各テキストに対して求めた F 値の平均を用いている．ベースラインには Barzilay らの Entity Grid モデルを断片に対して適用したモデルを用いた．先行文脈の決定に関しては提案モデルと同じ方法を採用している．

先行文脈を 2 としたときの skip.1x2 の実験結果の表とグラフを**表 3**，**図 5** に，skip.2x2 の実験結果の表とグラフを**表 4**，**図 6** に示す．表中の “each” は先行文脈中の各文から対象の文への遷移を考慮するモデル，“ext-EG” は Entity Grid の拡張を利用したモデル，“org-EG” はベースラインのモデルを表している．横軸は学習に使用した事例の数である．表中の斜体はベースラインのモデルよりも F 値が低いことを表している．

skip.1x2，skip.2x2 の両方において，提案したモデルはベースラインの結果を上回っている．skip.1x2 では Entity Grid の拡張を利用したモデルが，skip.2x2 では先行文脈中の各文との遷移を見たモデルが比較的良い結果を得ている．また，タスクの難易度に関しては，仮定通り skip.2x2 の方が全体的に良い結果を得ている．

以降の実験では Entity Grid の拡張を利用したモデルで skip.1x2 のデータセットに対する結果のみに関して考察を行う．

本研究で提案したモデルでは先行文脈の数を人手によって設定する．そこで，先行文脈の文数を変化させたときの性能の比較を行った．先行文脈の文数を 2 から 4 まで変化させた時の実験結果の表とグラフを**表 5**，**図 7** に示す． この結果から先行文脈を 2 とした時に比較的良好な結果を得ることができるということが分かった．

表 3 実験結果 (skip.1x2)

モデル	200	400	600	800	1000	1200	1400	1600	1800	2000
each	0.197	0.265	0.241	0.240	0.233	0.226	0.227	0.228	0.233	0.234
ext-EG	0.248	0.266	0.281	0.295	0.274	0.273	0.286	0.279	0.286	0.284
org-EG	0.132	0.100	0.086	0.108	0.112	0.139	0.154	0.179	0.183	0.184

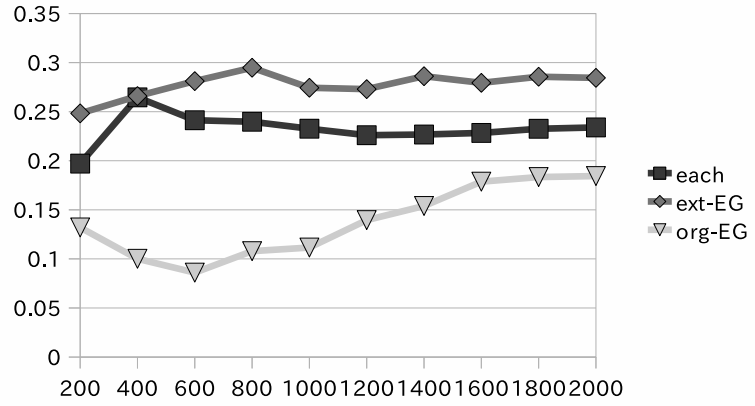

図 5 実験結果 (skip.1x2)

5. 形式段落との関係

一般的に，必ずしも一致するわけではないが，ある程度形式段落が話題の転換を示すと考えられる．従って，段落境界は話題が変化するという意味で一貫性が悪くなっている箇所とみなすことができる．

局所的一貫性モデルは話題の転換を特に考慮していないため，段落境界を誤って一貫性が悪い箇所と判定してしまうことが考えられる．そこで，各モデルによって一貫性が悪いと推定した箇所と，段落境界との関係を調査した．具体的には誤って一貫性が悪いと判定した箇所のうち，段落境界であった割合を求めている．結果を**表 6**，**図 8** に示す． Entity Grid の拡張を利用したモデルに比べて，先行文脈中の各文との遷移確率を計算するモデルが段落境界を一貫性が悪いと判定している割合が高く，これが skip.1x2 において Entity Grid の拡張を利用したモデルよりも性能が低い原因の一つだと考えられる．

⋆1 http://www.cs.cornell.edu/people/tj/svm_light/

表 4 実験結果 (skip.2x2)

モデル	200	400	600	800	1000	1200	1400	1600	1800	2000
each	0.347	0.352	0.348	0.347	0.348	0.348	0.352	0.350	0.348	0.345
ext-EG	0.304	0.305	*0.272*	0.297	0.330	0.326	0.335	0.334	0.331	0.323
org-EG	0.298	0.297	0.297	0.297	0.297	0.297	0.297	0.297	0.297	0.297

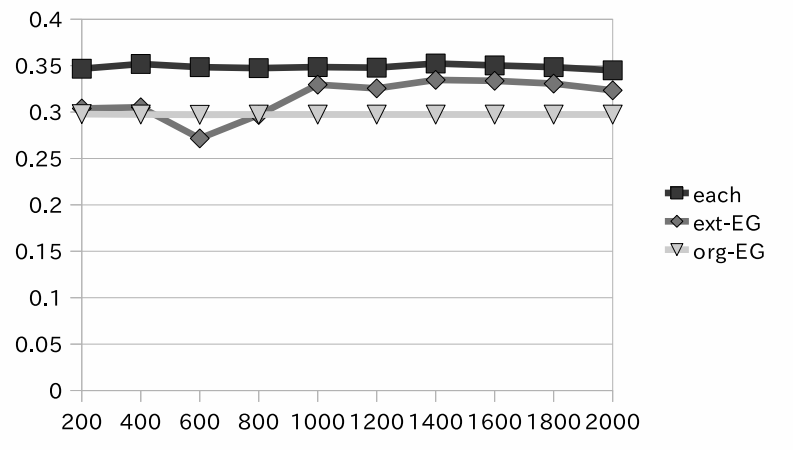

図 6 実験結果 (skip.2x2)

6. おわりに

本研究では，テキスト中で具体的に一貫性の悪い箇所を推定する，テキストの断片に対する局所的一貫性モデルを提案した．このモデルは推定対象の文とその先行文脈から素性ベクトルを作成する．また，先行文脈の一貫性は保たれているという仮定の元に，対象の文以前の推定結果を先行文脈の決定に利用している．新聞記事を用いた実験ではベースラインのモデルよりも良い性能であることを示したが，全体的には F 値は低い結果となった．

実験結果の分析で示したように，本研究のモデルでは段落境界を誤って一貫性が悪いと判定してしまうことがある．これは本モデルが局所的一貫性のみを考慮しているからであると考えられる．この問題を解決するためには，トピックセグメンテーションを用い，これによって同定されたトピック内でのみ本論文で提案したモデルを適用するといったようにテキストのトピックを考慮した手法などが考えられる．この場合，トピック間の一貫性は考慮されないので，大域的な一貫性モデルを組み込んでトピックの一貫性を別に評価する必要があると考えられる．

表 5 先行文脈の文数毎の実験結果

先行文脈	200	400	600	800	1000	1200	1400	1600	1800	2000
2	0.248	0.266	0.281	0.295	0.274	0.273	0.286	0.279	0.286	0.284
3	0.206	0.286	0.271	0.255	0.232	0.229	0.227	0.228	0.228	0.231
4	0.190	0.271	0.291	0.283	0.244	0.242	0.243	0.244	0.238	0.234

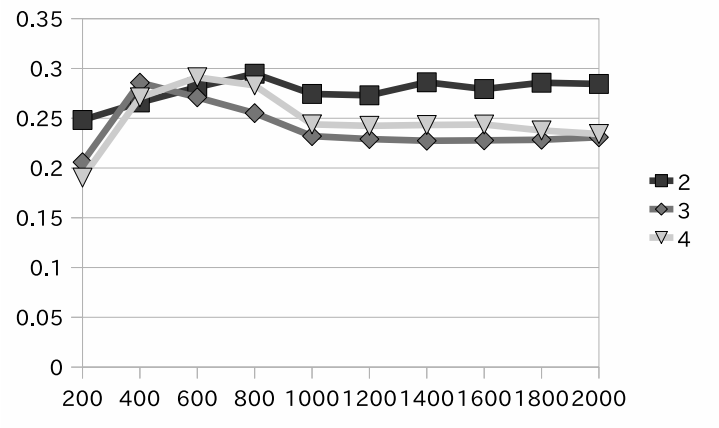

図 7 先行文脈の文数毎の実験結果

また，本研究の実験で用いたデータは人工的に一貫性の欠落箇所を作成したものであるため，必ずしも実際に存在する一貫性の欠落を反映しているとは言えない．実際に人手によって作られた一貫性が悪い箇所を持つテキストをデータとして用いるのが理想的ではあるが，そのようなデータを大量に作成することは非常に困難である．そのため，一貫性の欠落の要因の分析，それに基づいた様々な操作による多様な人工的データの作成など，実験に用いるテキストをより実際のデータに近づける必要がある．

参 考 文 献

1) 田窪行則，西山佑司，三藤　博，亀山　恵，片桐恭弘: 談話と文脈，岩波書店 (2004).
2) Barzilay, R. and Lapata, M.: Modeling Local Coherence: An Entity-Based Approach, *Computational Linguistics*, Vol.34, No.1, pp.1–34 (2008).

表 6 段落境界の影響

	200	400	600	800	1000	1200	1400	1600	1800	2000
each	0.334	0.344	0.356	0.348	0.356	0.357	0.357	0.357	0.357	0.381
ext-EG	0.342	0.311	0.320	0.337	0.328	0.326	0.327	0.323	0.332	0.335
org-EG	0.287	0.224	0.159	0.238	0.250	0.276	0.297	0.314	0.331	0.333

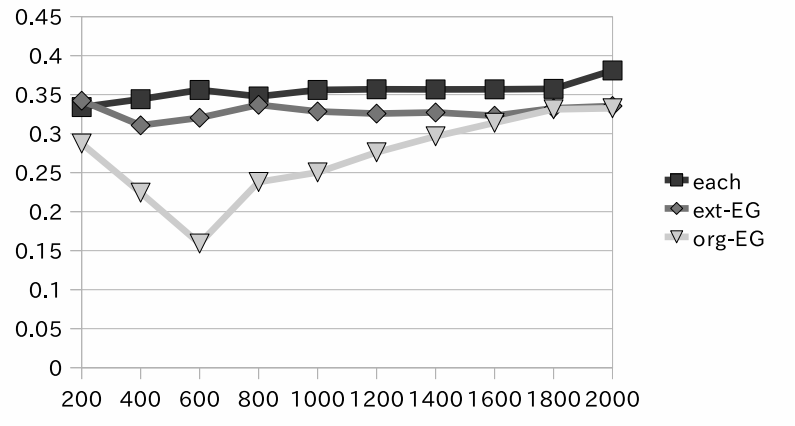

図 8 段落境界の影響

3) Grosz, B.J., Weinstein, S. and Joshi, A.K.: Centering: a Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics*, Vol. 21, No. 2, pp. 203–225 (1995).
4) Elsner, M. and Charniak, E.: Coreference-inspired Coherence Modeling, *Proceedings of ACL-08: HLT, Short Papers*, pp.41–44 (2008).
5) Filippova, K. and Strube, M.: Extending the Entity-grid Coherence Model to Semantically Related Entities, *Proceedings of the 11th European Workshop on Natural Language Generation* (2007).
6) Cheung, J. C.K. and Penn, G.: Entity-Based Local Coherence Modelling Using Topological Fields, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp.186–195 (2010).
7) 横野　光，奥村　学: テキスト結束性を考慮した entity grid に基づく局所的一貫性モデル，自然言語処理，Vol.17, No.1, pp.161–182 (2010).
8) Barzilay, R. and Lee, L.: Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization, *HLT-NAACL 2004: Main Proceedings*, pp.113–120 (2004).
9) Soricut, R. and Marcu, D.: Discourse Generation Using Utility-Trained Coherence Models, *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pp.803–810 (2006).
10) Elsner, M., Austerweil, J. and Charniak, E.: A Unified Local and Global Model for Discourse Coherence, *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp.436–443 (2007).
11) Hearst, M. A.: TextTiling: Segmenting Text into Multi-paragraph SubtopicPassages, *Association for Computational Linguistics*, Vol.23, No.1 (1997).
12) 平尾努，北内啓，木谷強: 語彙的結束性と単語重要度に基づくテキストセグメンテーション，情報処理学会論文誌，Vol.41, No.SIG 3(TOD 6), pp.24–36 (2000).
13) Sun, Q., Li, R., Luo, D. and Wu, X.: Text Segmentation with LDA-Based Fisher Kernel, *Proceedings of ACL-08: HLT, Short Papers*, pp.269–272 (2008).
14) 市川　孝: 国語教育のための文章論概説，教育出版 (1978).
15) Mochizuki, H., Iwayama, M. and Okumura, M.: Passage-Level Document Retrieval Using Lexical Chains, *Proceedings of RIAO 2000*, pp.491–506 (2000).