3K2-NFC-3-4

文章の話の組み立てと展開速度による段落間関係の評価

Valuation of Segment Relation based on Text Structure and Expansion

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

There is an opportunity to read and write a text. However, we think that we write an text to easy understand, it is difficult for the other to understand, and happen to explanation mistake. Also, we write an text to stray from subject . We want to communicate right. In this paper, I focused on the connection between segment and deployment of the story on the subject of text, a system that the valuation of segment relation based on text structure and expansion is proposed. A user input text in the system and the system output tree structure and words is the number of words is used with theme.

1. はじめに

文章を読み書きする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うことがある．また，ある事柄について文章を書いているうちに話が脱線して無駄なことを書いてしまうと，読み手に自分の意図を正確に伝えれない可能性もある．他には，話を広げるだけ広げてまとめを書いていない場合も読み手に混乱を招く恐れもある．これらを解消するためには，文章を書く本人が意識して改善できるのならいいのだが，それは難しい．文章の組み立てと構造をしっかりと理解することができ，主題と関係の無いことを述べられているところが分かることが望ましい．両者を理解することが出来れば，文章の修正も容易に行えると期待できる．

そこで本研究では，文章を段落間の繋がり，話の展開の点に着目した文章の話の組み立てと展開速度の段落間関係を評価するインタフェースの構築を目的とする．

以下，2. で関連研究，3. で文章の話の組み立てと展開速度の評価システム (SAT:Segment Association Tree)，4. で SAT の精度を計る評価実験，5. で結論を述べる．

2. 関連研究

2.1 段落間の関係を評価する研究

本節では，段落間の関係を評価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] 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った．本研究では単語間の意味には着目せず，*cos* 類似度で段落間の関係を評価していく．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 [2] 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，*cos* 類似度で段落間の関係を評価していく．

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

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．情報理論による，文章の理論的構造の解析をする研究 [3] 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，単に段落間の類似度の大きさによって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 [4] 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 文章の主題を評価する研究

本節では，文章の主題を評価する研究について述べる．文章の主題と各文の関連度を評価し，視覚的に表示する研究 [5] がある．この研究は，主題と関係のある部分と主題と関係のない部分を色で分かりやすく可視化することで，主題と関係のある箇所と全体に対する割合を確認することが出来る．本研究は，文章において主題の関係有無の可視化は行わず，主題に関する単語数を数え，段落間に数字で表現していく．単語数を見ることによって，話の展開が容易に分かると考えられる．

3. 文章の話の組み立てと展開速度の評価システム (SAT)

3.1 文章の話の組み立てと展開速度の評価システム (SAT) の構成

本節では文章の話の組み立てと展開速度の評価システム (SAT) について述べる．以下，SAT と呼ぶ．図 1 に SAT の全体の構成を示す．SAT にテキストファイルを与え，それを元に話の組み立てに関する処理，話の展開速度に関する処理をそれぞれ行う．話の組み立てに関する評価は，最初に文章の段落間の類似度の計算を行い，リンクを繋ぐ段落間を決定する．次に，段落間の類似度を元にリンクを繋ぎ文章のツリー構造を作成する．文章のツリー構造の作成の後，話の組み立ての評価を行う．話の組み立ての評価と平行して，話の展開速度の評価を行い，それぞれの結果をインタフェース上へ出力する．

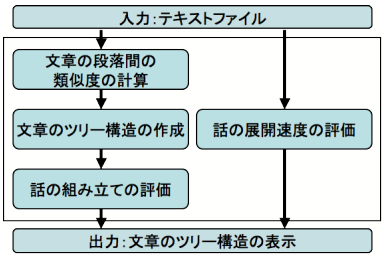

図 1: 文章の話の組み立てと展開速度の評価システム (SAT) の構成

以下，ツリー構造を表示するまでの各処理について述べる．

3.2 話の組み立ての評価

本節では，話の組み立ての評価について述べる．話の組み立ては段落の繋がりから文章を構造化することをいう．以下，評価するまでの各処理について述べる．

3.2.1 文章の段落間の類似度の計算

本項では，文章の段落間の類似度の計算方法について述べる．以下の式 (1) で全ての段落間の類似度 $Relation(A,B)$ 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，それらの数を数える関数 n へ与える．同じ単語が多く使われると類似度は高くなり，同じ単語が使われていなければ類似度は低くなる．

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

3.2.2 文章のツリー構造の作成

本項では，文章のツリー構造の作成方法について述べる．以下，アルゴリズムを示す．

1. 第 A 段落に着目し，全ての段落間の類似度 $Relation(A,B)$ を計算する．
2. 第 A 段落に関して，一番高い類似度 $Relation(A,B)$ を求める．
3. 2 で求めた段落間にリンクを繋ぐ．
4. 2〜3 を繰り返し，全ての段落について行う．
5. 4 でツリー構造が出来ない場合，全ての段落間の類似度 $Relation(A,B)$ についてソートを行う．
6. 繋がれていない段落間の中で一番高い類似度 $Relation(A,B)$ の段落間を繋ぐ．
7. ツリー構造が出来るまで 6 を繰り返す．

3.2.1 で算出した類似度を元に一番類似度の高い段落間をリンクで繋ぐが，ネットワーク化が出来ない場合がある．ネットワーク化とは，どのノードから見ても全てのノードを辿れることをいう．例えば直線的な繋がりでは，第 1 段落，第 2 段落，第 3 段落のことを指す．そこで，ネットワーク化が出来ていない場合は，全ての段落間の類似度をソートする．ソートした結果より，類似度の高い値から順に着目し，まだリンクが引かれていない段落間にリンクを追加してツリー構造を作成する．一番類似度の高い段落間にリンクを引くことによって，繋がりの強いツリー構造が出来ると考えられる．

3.2.3 話の組み立ての評価

本項では，話の組み立ての評価方法について述べる．話の組み立ての評価は，話が広がっている段落，話をまとめている段落について行い，前者を文章のトップダウン構造，後者を文章のボトムアップ構造と定義する．以下，文章のトップダウン構造のアルゴリズムを示す．なお，ボトムアップのアルゴリズムは以下のアルゴリズムの 1 の A_i, A_j の条件が逆になるとする．

1. 段落番号 A_i とリンクが繋がっている段落番号 A_j が A_i より大きい時，枝分かれしている枝数を数える関数 $Branch$ をカウントする．
2. $Branch$ が 2 以上の時，下向きの伸びている枝数を数える関数 $Length_j$ をカウントし，A_i に A_j を代入し，調べるノードを変える．
3. 2 で根までたどり着いたら，2 を $Branch$ の本数分調べる．
4. $Length_j + 1$ して $Branch$ の本数を乗算する．
5. 4 の値の対数を取る．
6. 1 から 5 を全ての段落について行う．

段落 A_i に着目し，下向きに伸びているリンク数でトップダウン，逆に，上向きに伸びているリンク数でボトムアップの割合を求める (2)．段落 A_i から伸びているリンク数 $Length_j$ を数えることによって，トップダウン，ボトムアップの割合を計算できると考え，また，$Length_j + 1$ することで，第 1 段落から全ての段落にリンクが 1 本ずつ伸びている場合も割合を求めることが出来る．なお，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点され，後半の段落にある場合は減点される．ボトムアップの場合は，トップダウンの場合の逆となる．

$$rate = log(\prod_{j=1}^{Branch(A_i)} (Length_j + 1)) \tag{2}$$

3.3 話の展開速度の評価

本節では，話の展開速度の評価方法について述べる．話の展開速度とは，ある段落で主題に関する単語と一緒に新しく使われた単語数のことをいう．話の展開速度が分かると，主題に沿った文章が書かれている段落かどうか理解することが出来る．主題の単語と一緒に使われた単語 (名詞) 数は，以下の式 (3) で計算する．ある段落 i からある段落 j へ話が進む際に，主題に関する単語と一緒に使われた単語数 K_i を計算して評価する．ただし，ここでの単語数は，形態素解析システム『茶筌』[6] で取れるものであり，主題に関する単語は，展望台システム [7] から出力されたものとする．

$$speed(i,j) = \sum_{i=0}^{j} K_i \tag{3}$$

3.4 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．3.2 話の組み立ての評価，3.3 話の展開速度で述べたものを TEDM[8] 上へ実装する．SAT システムの出力例を図 2 へ示す．段落間を繋ぐリンクを線で表し，類似度の大きさによってリンクの太さを変更し pixel で表す．段落間には主題と一緒に使われた単語数が記載され，式 (3) の値より，単語数

に応じて縦のリンクの長さが変化している．左下にトップダウン，ボトムアップの割合 (2) を表示している．出力例では，第 6 段落，第 7 段落がトップダウン構造，第 10 段落，第 13 段落がボトムアップ構造となっている．トップダウン，ボトムアップの表示には，プラスとマイナスがあるが，これは，3.2.3 で述べた通り，プラスへ表示されている．SAT を用いることにより，文章理解と文章作成の支援ができると考えている．

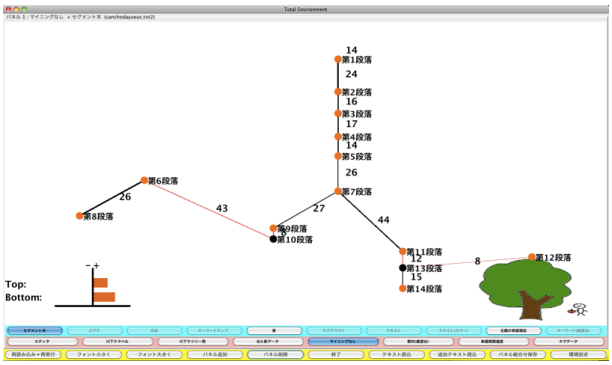

図 2: SAT の表示例

4. SAT の精度を計る評価実験

4.1 実験内容

SAT の精度を計るため，web 上で集めた 10 個の文章を情報科学を専攻する大学生・大学院生 32 名に A か B の 5 つのテキストを読んでもらい，以下の手順に沿って話の組み立てと話の展開速度に関する問いに答えてもらった．

1. 文章を通しで読んで，その主題を捉えて回答に記入する．
2. 再度文章に目を通して，以下の段落を答える．

- 複数の段落に話を広げている段落 (トップダウン)(該当する全ての段落を選択)
- 複数の段落の話をまとめている段落 (ボトムアップ)(該当する全ての段落を選択)
- 主題に関して話が進んでいる段落 (上位 3 つを選択)
- 主題に関して話が進んでいない段落 (上位 3 つを選択)

表 1: 実験で用いた文章の段落数と文の数

	段落数	一段落あたりの文数
A-1	11	4.181
A-2	12	4.167
A-3	13	4.923
A-4	15	4.067
A-5	9	3.556
B-1	13	2.769
B-2	8	5.250
B-3	12	2.083
B-4	12	2.667
B-5	9	4.000

4.2 実験結果と考察

表 2 に話の組み立てに関するシステムの出力と被験者の回答の一致不一致とシステムの不具合の割合を示す．これより，システムの出力と被験者の回答の一致で 6 割前後の値が出ていることが分かった．また，表 2 のシステムの不具合とは，システムの出力と被験者の回答の不一致の割合の中で，類似度が高くないトップダウン，ボトムアップの段落を出力した場合，または，類似度が高い段落間を出力していない場合のことを指す．今回は，類似度の閾値を 0.3 として考え，トップダウン，ボトムアップのシステムの不具合の値は順に 0.126，0.038 と不一致の中でも低い割合となっているのでシステムに問題はなく，文章を読んだ被験者の読み取りミスの可能性が高いといえる．

表 2: 話の組み立てに関するシステムの出力と被験者の回答の一致不一致とシステムの不具合の割合

	一致	不一致	システムの不具合
トップダウン	0.593	0.407	0.126
ボトムアップ	0.635	0.365	0.038

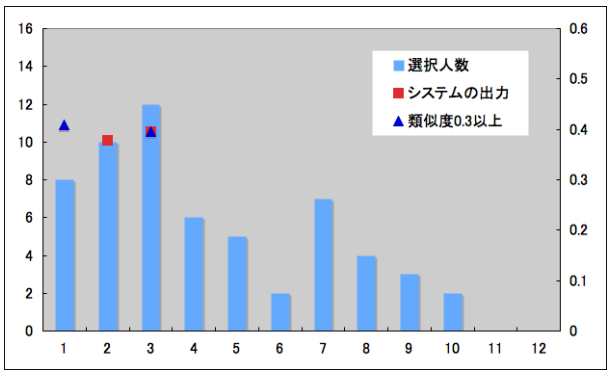

図 3: A-2 のシステムが出力したトップダウンと被験者の選択

図 3 に A-2 のシステムが出力したトップダウンと被験者の選択を示す．左縦軸は被験者の選択人数，右縦軸は類似度，横軸は段落番号を表している．なお，ここでの類似度は各段落でトップダウン判定した段落間の平均類似度を指し，システムが出力していない場合は，類似度 0.3 を超える全ての段落の平均類似度を取ったものとする．図 3 より，システムがトップダウン判定した段落は類似度が高く，また，被験者も多く選択していることが分かった．第 1 段落は，システムがトップダウン判定していないが，類似度は 0.3 を越えるので，場合によってはリンクを追加する処理を加えると考える．

次に，図 4 に A-1 のシステムが出力したボトムアップと被験者の選択を示す．グラフの見方は図 3 と同様で，左縦軸は被験者の選択人数，右縦軸は類似度，横軸は段落番号を表している．システムがボトムアップと判定した段落は第 9 段落と第 11 段落で，第 11 段落に関しては被験者は全員選択したが，第 9 段落に関しては選択した人が少なかった．システムは出力したが，類似度を見ると 0.2 を超えていなく，低い値となっているため被験者が選択できなかったと考えられる．この結果より，トップダウンの時とは逆にリンクを削除する処理を加える必要もあると考えられる．

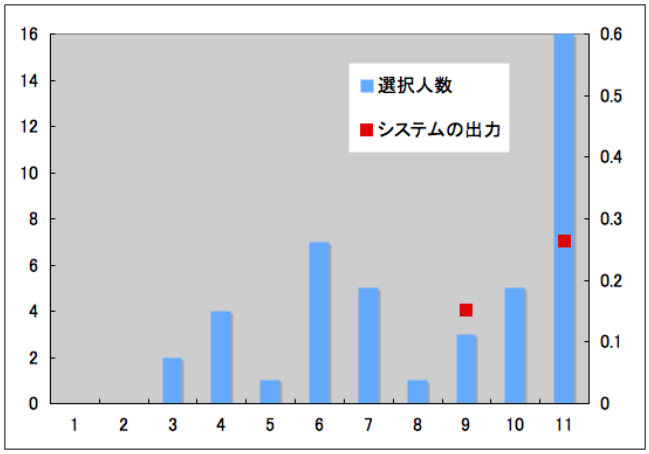

図 4: A-1 のシステムが出力したボトムアップと被験者の選択

表 3: 話の展開速度に関するシステムの出力と被験者の回答の一致不一致の割合

	一致	不一致	システムの不具合
主題に関して話が進んでいる段落	0.544	0.456	0.081
主題に関して話が進んでいない段落	0.597	0.403	0.050

表 3 に話の展開速度に関するシステムの出力と被験者の回答の一致不一致の割合を示す．システムの不具合は表 2 の話の組み立ての場合と同様に閾値を設け，話が進んでいる段落に関しては 20 単語，話が進んでいない段落に関しては 5 単語とした．値は，話が進んでいる段落，進んでいない段落の順に 0.081，0.050 と低い値となっているのでシステムに問題はなく，被験者の読み取りミスと考えられる．今回，主題に関して話が進んでいる段落，主題に関して進んでいない段落ともに上位 3 段落を選択してもらったので，話の組み立てに比べれば一致の割合が低かったと考えられる．

以下，例として A-1 の話の展開速度に関する各単語数と被験者の選択数を示す．

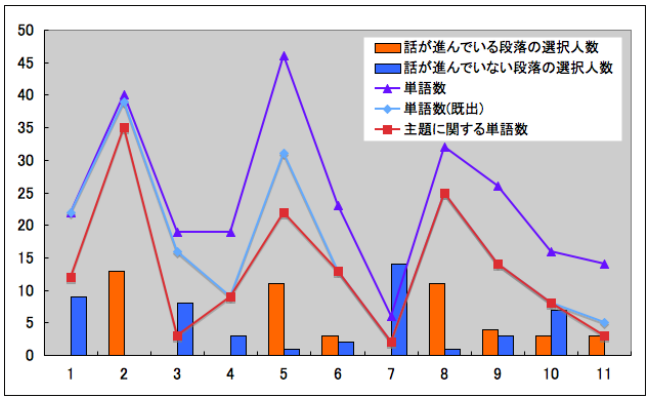

図 5: A-1 の話の展開速度に関する各単語数と被験者の選択数

図 5 の縦軸は単語数と被験者の選択人数，横軸は段落番号を示している．折線グラフより，単語数は各段落で使われた単語数，単語数（既出）はある段落に着目し，その段落より前の段落で使われていない単語数，主題に関する単語数は，ある段落に着目し，主題の単語と一緒に使われた単語，かつ，その段落より前の段落で使われていない単語数のことを表す．棒グラフは左側から順に主題に関して話が進んでいる段落の選択人数，話が進んでいない段落の選択人数を表している．A-1 では，被験者はシステムが出力した主題に関して話が進んでいる段落，主題に関して話が進んでいない段落ともに選ぶことができたと分かった．この結果より，各段落の主題に関する単語と一緒に使われた単語数が他の段落の単語数と近い数字でない，かつ，主題に関して話が進んでいる単語数が多い，主題に関して話が進んでいない単語数が少ない場合は上手くいくと考えられる．また，主題の単語と一緒に使われている単語数を見ることによって，単語数を数えるよりも話の展開速度を捉えることが出来ると確認できた．

5. 結論

文章の話の組み立てと展開速度による段落間関係を評価するシステムの評価を行った．SAT システムの妥当性を評価する実験により，話の組み立てと展開速度それぞれに関して文章の理解や作成の支援に用いられる可能性を確認した．今後は SAT システムの精度を上げるために段落間のリンクを追加，または，削除する処理を加えてシステムを改良していく．

また，今回の実験では，システムの出力結果を被験者に見せていないので，システムを改良した後に実験を行う際に，システムの出力結果を被験者に見せ，出力結果を支持するか，または，被験者自身の回答を支持するか，答えてもらうようにする．

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: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)

[2] 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)

[3] 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)

[4] 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)

[5] 佐藤圭太，西原陽子，砂山渡:光と影を用いたテキストのテーマ関連度の可視化，人工知能学会論文誌，Vol.24，No6，pp.479-487，(2009)

[6] 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)

[7] 砂山渡，谷内田正彦:文章の特徴を表すキーワードを発見して重要文を抽出する展望台システム，電子情報通信学会論文誌，D-I，情報・システム，I-情報処理，pp.146-154，(2001)

[8] 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)