1K3-1in2

談話構造による文書の特徴づけ

Characterization of Documents by Focusing on Discourse Structure

佐藤瞳 Hitomi SATO　金川絵利子 Eriko KANAGAWA　岡留剛 Takashi OKADOME

関西学院大学大学院理工学研究科
Graduate School of Science and Engineering, Kwansei Gakuin University

This study defines discourse trees for Japanese texts and characterizes texts by discourse structure to classify texts by the discourse structure. A tree kernel with the discourse trees enables us to capture the text similarities.

1. はじめに

文書の特徴づけを行なう手法は，文書の真贋判定や著者判定，クラスタリングなど様々な技術で用いられて，その多くが文や単語に着目している．しかし，文や単語が異なっていても類似していると思われる文書があることから，文書の構造も文書を特徴づける要因の一つであると考えられる．本研究では，文書の構造に着目し，それを木構造で表わした談話構造木を定義，談話構造木により文書を特徴づける．さらに文書に対する談話構造木を用いて，文書の類似性について検討する．

2. 関連研究

文書の構造をとらえる理論として修辞構造理論 [1] がある．これは，英語文書の構造を一般的に解析する目的で考え出された理論であり，文章を部分領域に分割し，領域間の二項関係で文書全体を表わしている．また，文書の自動要約の分野では，修辞構造理論に基づいて作成された構造木を，談話単位間の依存関係が明示された構造木に変換するという手法が考案されている [2]．しかし，現在のところ，日本語における談話構造木の明確な定義やコーパスは存在していない．黒橋らの研究 [3] では，日本語の科学技術文を対象に談話構造を自動で抽出する手法が示されている．本研究では，この文献における談話構造をもとに類似性の比較に適した構造木を設計する．

3. 談話構造木

文書は，文や節などの談話単位から成り立っている．談話単位をノードとし，談話単位間の関係をエッジとして文書を木構造で表わしたものを，談話構造木と呼ぶ．本研究では文献 [3] を参考に，類似性を比較するための談話構造木を設計する．以下でその構造と談話単位間の関係について示す．

3.1 構造

本研究では，句点で区切られた文を談話単位とし，1 段落につき 1 つの談話構造木を考える．このとき，箇条書きで示された文も 1 文とみなす．図 1 に構造木の例を示す．図の S_0 は初期節点を示し，S_i は段落内における i 番目の文を示す．S_0 に対応する文は存在しない．各談話単位は S_1 から順に木に加え，新たな文は，それまでに生成された木のいずれかのノードに接続する．エッジには談話単位間の文間関係を付与する．ただし，初期節点 S_0 に接続する場合は文間関係は「初期化」とする．本研究で考える談話構造木は並列関係を木の形で表わす．

連絡先: 氏名：佐藤　瞳
所属：関西学院大学大学院理工学研究科
住所:〒 669-1337 兵庫県三田市学園 2-1
メールアドレス：Hitomi.Sato@kwansei.ac.jp

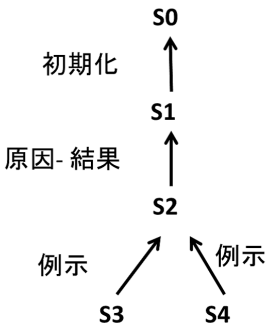

図 1: 談話構造木の一例

3.2 単位間関係

本研究では黒橋らの研究 [3] を参考に，実際に複数の論文について談話構造分析を行ない，論文と論説文を対象とした 10 種の文間関係を選定する．関係の種類を表 1 に示す．ただし，表中の S_i と S_j については，ある文間関係で接続される 2 文において，前の文を S_i 後の文を S_j とする．

単位間関係名	S_i , S_j の関係
対比	S_i と S_j が対比関係にある
主題連鎖	S_i の主題が S_j の主題となっている
焦点-主題連鎖	S_i の焦点要素が S_j の主題となっている
詳細	S_j が S_i の詳細を説明している
理由	S_j が S_i の理由となっている
原因-結果	S_i の結果，S_j となる
例示	S_j が S_i の例を示している
質問	S_j が S_i に対する質問となっている
順序	S_i の後，S_j が起こる
定義	S_i を S_j と定義している

表 1: 文間関係の種類

4. 木カーネル

4.1 木カーネル

木カーネル (subset tree kernel) は木構造に対して定義されたカーネルである [4]．2 つの木構造に共通して含まれる部分木

の個数をカーネル値とする．木カーネルは，2 つの木構造 T_1，T_2 に対し以下の式で定義される．

$$K_A(T_1, T_2) = \langle \phi(T_1), \phi(T_2) \rangle = \sum_{S \in \tau} \phi_S(T_1)\phi_S(T_2),$$

ここで，S は比較する部分木，τ は T_1，T_2 がもつ部分木の集合である．また，$\phi_S(T)$ は，木 T が S を部分木として含むときは 1，含まないときは 0 となる．

4.2 文書類似度

文書の類似度を木カーネルを用いて定義する．本研究では 1 つ目の文書の各談話構造木と 2 つ目の文書の各談話構造木に対して総当たりで木カーネル値を計算し，その平均を文書類似度とした．2 つの文書 D_1，D_2 は有限個の談話構造木の集合であると仮定する．このとき，文書 D_1，D_2 の文書類似度 $K_D(D_1, D_2)$ を以下のように定義する．

$$D_1 = \phi \text{ または } D_2 = \phi \text{ のとき } K_D(D_1, D_2) = 0,$$

$$K_D(s, D_2) = \frac{1}{|D_2|} \sum_{\bar{s} \in D_2} K_i(T_s, T_{\bar{s}}),$$

$$K_D(D_1 \cup s, D_2) = \frac{1}{|D_2| + 1}(K_D(D_1, D_2) + K_D(s, D_2)),$$

ここで $K_i(s, \bar{s})$ は段落 s と $\bar{s}$ の木カーネル値，$|D|$ は文書 D の談話構造木の数，すなわち段落数を示す．文書類似度は，2 つの談話構造木間で共通する部分木数の平均である．

5. 実験

提案した談話構造木の定義に基づいて文書から木を生成し，文書の類似性について検討する．本稿では Web で公開されている論文を対象とし，3 名の著者による論文それぞれ 3 文書，それぞれ異なる著者の論文 5 文書の計 14 文書について議論する．提案した談話構造木の定義に基づき文書から人手で木を生成する．さらに，生成した木から文間関係ラベルの種類を削除し，構造のみを表わす木を生成する．

それぞれのデータに対し文書類似度を算出する．カーネル値の算出には Mocshitti[4] の作成したプログラムを改変したもの利用する．全ての文間関係ラベルを用いた場合とラベルの種類を削除した場合の結果をそれぞれ表 2 と表 3 に示す．なお，3 名の著者をそれぞれ A・B・C とすると，表の文書 6・7・8 が著者 A，文書 9・10・11 が著者 B，文書 12・13・14 が著者 C の文書である．

	文書1	文書2	文書3	文書4	文書5	文書6	文書7	文書8	文書9	文書10	文書11	文書12	文書13	文書14
文書1		1.34	1.31	1.53	1.38	1.56	1.32	1.67	1.63	1.47	1.47	1.51	1.61	1.48
文書2	1.34		1.26	1.39	1.29	1.55	1.42	1.66	1.77	1.51	1.43	1.58	1.66	1.46
文書3	1.31	1.26		1.48	1.39	1.43	1.43	1.46	1.53	1.36	1.39	1.52	1.50	1.42
文書4	1.53	1.39	1.48		1.72	1.80	1.48	1.85	1.74	1.65	1.52	1.73	1.89	1.71
文書5	1.38	1.29	1.39	1.72		1.85	1.42	1.97	1.64	1.48	1.52	1.82	2.04	1.67
文書6	1.56	1.55	1.43	1.80	1.85		1.62	2.19	2.02	1.80	1.87	2.09	2.46	1.88
文書7	1.32	1.42	1.43	1.48	1.42	1.62		1.64	1.62	1.46	1.53	1.67	1.70	1.51
文書8	1.67	1.66	1.46	1.85	1.97	2.19	1.64		2.28	1.94	2.05	2.21	2.63	1.97
文書9	1.63	1.77	1.53	1.74	1.64	2.02	1.62	2.28		1.91	1.87	2.05	2.39	1.94
文書10	1.47	1.51	1.36	1.65	1.48	1.80	1.46	1.94	1.91		1.62	1.79	2.14	1.76
文書11	1.47	1.43	1.39	1.52	1.52	1.87	1.53	2.05	1.87	1.62		1.94	2.37	1.67
文書12	1.51	1.58	1.52	1.73	1.82	2.09	1.67	2.21	2.05	1.79	1.94		2.50	1.90
文書13	1.61	1.66	1.50	1.89	2.04	2.46	1.70	2.63	2.39	2.14	2.37	2.50		2.25
文書14	1.48	1.46	1.42	1.71	1.67	1.88	1.51	1.97	1.94	1.76	1.67	1.90	2.25	
平均	1.48	1.49	1.42	1.65	1.63	1.86	1.53	1.96	1.88	1.68	1.71	1.87	2.09	1.74

表 2: 全ての文間関係ラベルを用いた場合の文書類似度

6. 議論

表 1 より，文書類似度は文書 1 と文書 2・文書 3 が低く，文書 8 と文書 13 が高い傾向にあることがわかる．このことから，

	文書1	文書2	文書3	文書4	文書5	文書6	文書7	文書8	文書9	文書10	文書11	文書12	文書13	文書14
文書1		7.75	5.44	6.37	6.49	7.63	5.90	9.66	9.46	8.36	8.27	8.23	9.73	7.10
文書2	7.75		5.33	6.27	6.40	7.54	5.77	9.13	9.26	8.23	7.72	8.00	9.49	6.98
文書3	5.44	5.33		4.60	4.83	5.44	4.41	6.53	6.47	5.87	5.60	5.90	6.77	5.12
文書4	6.37	6.27	4.60		5.59	6.30	4.94	7.56	7.77	6.87	6.35	6.67	7.81	5.79
文書5	6.49	6.40	4.83	5.59		6.49	5.08	7.67	7.57	6.91	6.43	6.89	7.70	5.87
文書6	7.63	7.54	5.44	6.30	6.49		6.01	9.31	9.16	8.31	7.91	8.87	10.45	7.18
文書7	5.90	5.77	4.41	4.94	5.08	6.01		7.14	7.11	6.34	6.13	6.69	7.76	5.55
文書8	9.66	9.13	6.53	7.56	7.67	9.31	7.14		11.33	10.15	9.83	10.34	12.40	8.72
文書9	9.46	9.26	6.47	7.77	7.57	9.16	7.11	11.33		10.24	9.38	9.64	11.59	8.55
文書10	8.36	8.23	5.87	6.87	6.91	8.31	6.34	10.15	10.24		8.50	9.08	10.89	7.73
文書11	8.27	7.72	5.60	6.35	6.43	7.91	6.13	9.83	9.38	8.50		8.99	10.78	7.29
文書12	8.23	8.00	5.90	6.67	6.89	8.87	6.69	10.34	9.64	9.08	8.99		12.84	8.03
文書13	9.73	9.49	6.77	7.81	7.70	10.45	7.76	12.40	11.59	10.89	10.78	12.84		9.49
文書14	7.10	6.98	5.12	5.79	5.87	7.18	5.55	8.72	8.55	7.73	7.29	8.03	9.49	
平均	7.72	7.53	5.56	6.37	6.45	7.74	6.06	9.21	9.04	8.27	7.94	8.48	9.82	7.19

表 3: ラベルを削除した場合の文書類似度

文書 1 と文書 2・文書 3 は他の文書と比較して特徴的な構造をもつ文書であり，文書 8 と文書 13 は一般的な構造であると推測できる．また表 2 より，文間関係を無視して構造のみに着目すると，文書 3 と文書 4・文書 7 が低い傾向にあり，文書 8 と文書 13 が高い傾向にあることがわかる．これらのことから，特に文書 1 と文書 2 は文間関係に特徴があり，文書 4 と文書 7 には構造的な特徴があると推測される．

それぞれの文書について考察する．文書 1 の全ての文間関係ラベル中「詳細化」ラベルの割合は 17 %であり，平均の 30 %を大きく下回っている．また文書 2 は「例示」ラベルが 14 %であり，平均の 5 %を上回っている．文書 4 と文書 7 は枝分かれ構造が多く，他の文書と比較して直鎖状の部分木が少なくなっている．また，文書 3 は 1 文または 2 文で構成された段落が多いという特徴があった．これらの特徴が結果に表われたと考えられる．

また，著者ごとの特徴について，今回の実験では全文書で文書類似度の傾向がほとんど一致しており，著者ごとの特徴を捉えることができなかった．これは分析データ数が少なく，詳しい類似性を捉えることができないことが原因だと考えられる．

7. おわりに

本研究では，関連研究を参考に談話構造木を定義し，文書を談話構造に着目して特徴づけを行なった．また，. 論文を対象として人手で談話構造木を生成し，文書類似度を算出して比較した．その結果，談話構造木を用いて文書の特徴を捉えることができることを確認した．

参考文献

[1] Mann, W. C. and S. A. Thompson (1988). Rhetorical structure theory:toward a functional theory of text organization. *Text,* 8(3), 243-281 .

[2] 吉田康久・鈴木潤・平尾努・永田昌明 (2014)．係り受け木に基づく談話構造の提案． **言語処理学会第 20 回年次大会発表論文集**，468-471.

[3] 黒橋禎夫・長尾眞 (1994)．表層表現中の情報に基づく文章構造の自動抽出．**自然言語処理，1**，1, 3-20.

[4] Moschitti, M. (2006). Efficient convolution kernels for dependency and constituent syntactic trees. *Proceedings of the 17th European Conference on Machine Learning* (*ECML*2006), 318-329.