談話の削除不可能性に基づく 教師なし 談話核性分類

西田 典起　　中山 英樹
東京大学大学院情報理工学系研究科

{nishida, nakayama}@nlab.ci.i.u-tokyo.ac.jp

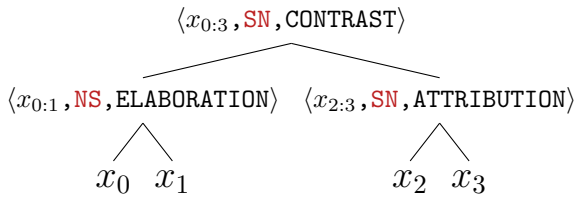
1 はじめに

一貫性のある文章において，その言語単位 (節や文，段落など) は統語的，意味的，また語用論的に相互結合しており，孤立したテキスト 領域は存在しない．そのような文章の一貫性は談話構造として表現され，様々な応用技術において談話構造の有用性が知られている．

談話構造解析とは，入力文章の談話構造を 計算機によって自動的に同定する技術であり，近年の技術的進歩 [7, 5, 8] に関わらず，談話構造解析は依然として困難なタスクのままである．その原因の一つとして，人手による 談話構造アノテーションが訓練データとして不足していることが挙げられる．これは，談話構造のアノテーション 作業が専門性が高く 複雑であり，金銭的また時間的なコスト が大きいことに因る．

この問題に対して，教師なし 談話構造解析というアプローチが考えられる．教師なし 談話構造解析では，入力文章の談話構造を 人手による 教師情報に依らずに同定することを目指す．本研究では，図 1 に示すような 修辞構造理論 (RST) [9] に基づく 構成素構造を 仮定する．葉ノード は節を 中心とした重複のないテキスト 領域であり，内部ノード は談話構成素，核性 (従属関係) ラベル，談話関係ラベルの三つの要素からなる．著者ら [12] は教師なし 談話構成素構造解析手法を 提案した．しかし，それは文章の階層性にのみ焦点を 当てており，核性ラベルや談話関係ラベルの予測までは行っていない．

そこで本研究では，[12] によって 与えられるようなラベルなし 木構造を 所与のものとし，その各内部ノード に対して 核性ラベルを 教師なし 手法によって 付与することを目指す．本研究では特に，結合される 二つのテキスト 領域間の従属関係として，単純化のため次の二つのクラスを 仮定する： *Nucleus-Satellite* (NS), *Satellite-Nucleus* (SN)．*Nucleus* (核) はより 中心的な情報を 表すテキスト 領域であり，*Satellite* (衛星) は核を 補足するテキスト 領域である．この仮定は二分木化を 行い，また 重要度が等しいテキスト 領域が結合する場合に左側を 核として 扱うことで常に 成り 立つ．



[Newport officials didn't respond Friday to requests]$_{x_0}$ [to discuss the changes at the company]$_{x_1}$ [but earlier, Mr Weekes had said]$_{x_2}$ [Mr. Hollander wanted to have his own team on the board.]$_{x_3}$

図 1: 本研究で仮定する，RST [9] に基づく 談話構造の例．本稿では特に，結合するテキスト 領域間の核性 (従属関係) の教師なし 分類に焦点を 当てる．

Carlson ら [2] は RST Discourse Treebank (RST-DT) を 構築する際に，核性を 決定するために 以下のテストを 提案している．

- 削除テスト : テキスト 領域が核ならば，それを 削除することで残ったテキスト 領域の一貫性は著しく 弱まる．衛星ならば，それを 削除しても，一貫性は弱くなるものの，残ったテキスト 領域の一貫性は保たれる．

本手法は，削除テスト を 自動的に 行うことによって，核性の教師なし 分類を 行う．

談話の削除可能性 (または削除不可能性) を 計算するために，本手法では抽出型文書要約における 文の重要度尺度と，シンプルな 再帰アルゴリズムを 統合する．核性ラベルは，各テキスト 領域の削除不可能性を 比較することによって 付与される．

RST-DT を 用いた 実験結果から，提案手法がベースラインの 教師なし 分類手法を 上回ることを 示した．

また，提案手法のもとで相補的な尺度を統合することで，より高い精度で談話の削除不可能性を推定できることを示した．

2 削除不可能性の計算

ここでは，どのようにして談話 (テキスト領域，または談話構成素) の削除可能性 (削除不可能性) を計算するのか説明する．

削除テストに類似したアイデアとして，Mareček ら [11] は教師なし依存構造解析のための句の削減可能性に着目している．彼らは文から n-gram を削除し，文の残された部分がコーパス中で文として出現するかどうかをチェックした．もしある n-gram を削除しても，残された部分が出現する傾向にあるなら，その n-gram は削減可能性が高いということになる．

しかし，このような削除テストを談話レベルで行うことは難しい．なぜなら，データスパースネスの問題があるため，たとえ削除可能性の高いテキスト領域を削除したとしても，残された文章がコーパスの他の部分で出現する可能性は非常に低いからである．

そこで，本稿では抽出型文書要約技術と再帰アルゴリズムを統合することで，テキスト領域の削除 “不” 可能性を計算する方法を提案する．

2.1 文の削除不可能性

抽出型文書要約では，入力文章から重要な文をいくつか抽出することで，要約を生成する．典型的には，(1) まず文章内の各文を特徴ベクトルに変換し，(2) そして文脈を考慮するアルゴリズムによって各文の重要度を計算し，(3) 最後にそれらのスコアに基づいて上位 n 文を選択する．

本稿では，抽出型文書要約のステップ (1), (2) によって計算される文章中の各文の重要度を，文の削除不可能性とする．特に，以下の 5 つの重要度尺度を比較する: SumBasic [14], Latent Semantic Analysis (LSA) [6], centroid-based 尺度 [13], LexRank [4], ヒューリスティック尺度 (Heuristic-Location, Heuristic-Length). Heuristic-Location は文章中における文の位置が早いほど重要度が高いものとし，Heuristic-Length は文が含む単語数が多いほど重要度が高いものとする．その他の尺度の詳細については，紙面の都合上割愛する．

2.2 テキスト領域の削除可能性

本手法では，文の削除不可能性から始めて，より大きな談話単位の削除不可能性をボトムアップかつ再帰的に求める．文の系列 $s_{i:j}$ に対応する部分木を

$$T(s_{i:j}) = (T(s_{i:k}), T(s_{k+1:j})) \tag{1}$$

のように記述する．部分木 $T(s_{i:j})$ の削除不可能性は，再帰関数 G を用いて

$$G(T(s_{i:j})) = \mathrm{Pool}(G(T(s_{i:k})), G(T(s_{k+1:j}))) \tag{2}$$

と定義できる．ただし Pool はプーリング関数とする．予備実験において Average Pooling と Max Pooling を比較した結果，Max Pooling

$$G(T(s_{i:j})) = \max(G(T(s_{i:k})), G(T(s_{k+1:j}))) \tag{3}$$

がより核性分類の正解率が高いことがわかったので，以下の実験では Max Pooling を採用した．

長さが 1 のスパン (i.e., $i = j$) に対しては, G を

$$G(T(s_{i:i})) = g(s_i) \tag{4}$$

のように定義する．ただし $g(s_i)$ は 2.1 節で挙げた文の重要度尺度を用いて求める文 s_i の重要度 (削除不可能性) とする．

3 核性分類

テキスト領域の削除不可能性が計算されれば，テキスト領域 $s_{i:k}$ と $s_{k+1:j}$ の間の核性を，それらの削除不可能性を比較することによって行う．テキスト領域 $s_{i:k}$ と $s_{k+1:j}$ が結合して構成される内部ノードの核性ラベル $l_{i,j}$ は，

$$l_{i,j} = \begin{cases} \texttt{NS} & (g_{i,k} - g_{k+1,j} \geq 0) \\ \texttt{SN} & (g_{i,k} - g_{k+1,j} < 0) \end{cases} \tag{5}$$

によって決定する．ここで $g_{i,k} = G(T(s_{i:k}))$, $g_{k+1,j} = G(T(s_{k+1:j}))$ である．文レベルの談話構成素構造の内部ノードについては，マジョリティラベルである NS を自動的に付与する．

4 実験

4.1 データと評価尺度

本研究では，図 1 に示すような RST に大きく基づく談話構造を仮定している．そのため，RST-DT [3]

のテストセットを評価に用いた．また，提案手法は訓練データを必要としないため，RST-DT の訓練セットを検証用とした．

本稿では従来の評価手法である RST-PARSEVAL [10] に基づき，核性分類精度を評価する．ただし，RST-DT の談話構造は二分木に変換し，二つの子ノードの核・衛星に基づいて親ノードの核性ラベル (i.e., NS または SN) を求めた．二つの子ノードともに核であった場合は (i.e., NN)，NS ラベルとした．Standard Accuracy (SA) は核性分類に関する従来の Micro F1 スコアに対応し，テストセット中で正しくその核性ラベルが予測されたスパン (内部ノード) の割合として定義される．各核性クラスごとの正解率をより詳細に調べるため，クラスごとの再現率 (i.e., NS-R, SN-R) も計算する．また，NS ラベルが SN ラベルよりも圧倒的に多いというテストセット中でのクラスアンバランス問題を緩和するため，Balanced Accuracy (BA) [1] も計算する．Balanced Accuracy は，クラスごとの再現率の平均として定義される．

前段のタスクにあたる EDU 分割やラベルなし談話構成素構造解析のエラーは，核性分類のためのアルゴリズムをより純粋に評価する際のノイズになりうる．そこで，実験ではゴールドスタンダードの EDU 分割とラベルなし談話構成素構造を用いた．

4.2 ベースライン

クラスタリングは最も単純な教師なし分類手法のひとつである．本稿ではクラスタリングによる教師なし分類器をベースラインとする．具体的には，クラスタ数 $K = 2$ とした K-Means 分類器を RST-DT の訓練セット中の談話構成素を用いて構築する．談話構成素の特徴ベクトルは，それを構成する二つのテキスト領域をそれぞれ事前学習済み単語ベクトルの総和として表し，それを結合することで表した．未知の談話構成素の核性ラベルは，各クラスターの中心ベクトルとの距離に基づいて予測する．K-Means によって得たクラスターそれぞれを NS と SN のどちらに対応させるかは恣意性があるため，K-Means 分類器の結果に関しては二通りのスコアを示す．

手法	NS-R	SN-R	SA	BA
K-Means	91.25	2.35	78.65	46.80
	8.75	97.65	21.35	**53.20**
SumBasic	42.95	52.35	44.29	47.65
LSA	38.68	66.47	42.62	52.57
Centroid-based ··· (a)	49.08	61.76	51.88	56.00
LexRank	49.08	61.76	50.88	55.42
Heuristics-Location ··· (b)	100.00	0.00	85.82	50.00
Heuristics-Length	49.56	64.71	51.71	**57.13**
0.8× (a) + 0.2× (b)	57.63	59.42	57.88	**58.52**

表 1: RST-DT のテストセットを用いた教師なし核性分類の実験結果.

4.3 結果と考察

表 1 に核性分類における各手法のスコアを載せる．表 1 の上段は K-Means 分類器のスコアを示す．下段は，異なる重要度尺度を用いたときの提案手法のスコアを示す．

上段から，K-Means 分類器が非常にバランスの偏ったクラスターを形成し，マジョリティクラスタと同じラベルを予測しすぎる傾向にあることがわかる．マイノリティクラスタの再現率 (e.g., 2.35%, 8.55%) はマジョリティクラスタの再現率 (e.g., 91.45%, 97.65%) よりも著しく低く，これは両クラスが等しく重要な状況では望ましくない．

下段から，centroid-based 尺度と LexRank が，SumBasic または LSA よりも高い Balanced Accuracies であることがわかる．前者二つの尺度とは異なり，SumBasic と LSA は事前学習済みのベクトル表現を活用しておらず，入力文章におけるカウントベースの統計情報にのみ依存している．これらの結果は，たとえ核性分類のために学習されたものでなくても，大規模なコーパスを用いて獲得した単語知識，文知識が，談話の削除不可能性を推定し核性を同定するのに有用であることを示唆する．Heiristic-Length は単一の尺度としては最もスコアが高く，これは単語数が多いテキスト領域ほど削除することが難しく，核である傾向が高いという直観に一致している．

また，グリッドサーチによって，二つの異なる重要度尺度を異なる重みによって組み合わせたときの結果についても調べた．本稿では，組み合わせ後の文の削

除不可能性を

$$\lambda_{m_1} g_1(s_i) + \lambda_{m_2} g_{m_2}(s_i) \quad (6)$$

のように定義した．ここで $g_{m_1}(s_i)$ と $g_{m_2}(s_i)$ はそれぞれ尺度 m_1 または m_2 に基づいて計算された文の重要度であり，λ_{m_1} と λ_{m_2} は各尺度の重みとする．

実験の結果，centroid-based 尺度 ($\lambda = 0.8$) と Heuristic-Location ($\lambda = 0.2$) の組み合わせが，他のすべての組み合わせを上回り，Balanced Accuracy で 58.52% に到達することがわかった．これは，各尺度単体でのスコア (i.e., 56.00%, 50.00%) を上回っている．この結果は，これらの尺度が異なる観点から文の重要度を表しており，相補的な尺度を組み合わせることによってより高い精度で文の削除不可能性が推定できることを示している．実際，centroid-based 尺度は位置情報を考慮できないのに対し，Heuristic-Location は位置に基づいて文の重要性を計算することができる．

興味深いことに，Heiristic-Length は単体では最良の尺度であるにも関わらず，centroid-based 尺度と Heiristic-Length の組み合わせは，centroid-based 尺度と Heiristic-Location の組み合わせを上回ることができなかった．これは，centroid-based 尺度と Heiristic-Length が相補的ではないからであると予想される．なぜなら，本稿において，単語ベクトルの和を用いる centroid-based 尺度は既に文長 (単語数) に関する情報を保持しているからである．

5 おわりに

本稿では，Carlson ら [2] の削除テストに基づき，談話の削除不可能性によって談話核性分類を行う教師なし手法を提案した．談話の削除不可能性は，抽出型文書要約における文の重要度尺度と再帰アルゴリズムを統合することによって計算した．提案手法がベースラインを上回ることを示し，また相補的な尺度を統合することでより高い精度で談話の削除不可能性を推定できることを確認した．

謝辞

本研究は，独立行政法人情報通信研究機構 (NICT) の委託研究「多言語音声翻訳高度化のためのディープラーニング技術の研究開発」および JSPS 科研費 JP19K22861，JP18J12366 の成果として得られたものです．

参考文献

[1] Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. The balanced accuracy and its posterior distribution. In *ICPR*, 2010.

[2] Lynn Carlson and Daniel Marcu. Discourse tagging reference manual. In *Technical Report ISI-TR-545*. University California Information Sciences Institute, 2001.

[3] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, 2001.

[4] Güneş Erkan and Dragomir R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, Vol. 22, No. 1, pp. 457–479, 2004.

[5] Vanessa Wei Feng and Graema Hirst. A linear-time bottom-up discourse parser with constraints and post-editing. In *ACL*, 2014.

[6] Yihong Gong and Xin Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR*, 2001.

[7] Yangfeng Ji and Jacob Eisenstein. Representation learning for text-level discourse parsing. In *ACL*, 2014.

[8] Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. Codra a novel discriminative framework for rhetorical analysis. *Computational Linguistics*, Vol. 41, No. 3, pp. 385–435, 2015.

[9] William C. Mann and Sandra A. Thompson. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, Vol. 8, No. 3, pp. 243–281, 1988.

[10] Daniel Marcu. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, Vol. 26, No. 3, pp. 395–448, 2000.

[11] David Mareček and Zdeněk Žabokrtský. Exploiting reducibility in unsupervised dependency parsing. In *EMNLP*, 2012.

[12] Noriki Nishida and Hideki Nakayama. Unsupervised discourse constituency parsing using Viterbi EM. *Transactions of the Association for Computational Linguistics*, to appear.

[13] Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. Centroid-based text summarization through compotitionality of word embeddings. In *MultiLing*, 2017.

[14] Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing and Manegement*, Vol. 43, No. 6, pp. 1606–1618, 2007.