談話構造を利用した学術論文の自動要約生成

中須賀 謙吾　鶴岡 慶雅
東京大学大学院 工学系研究科
{sky58, tsuruoka}@logos.t.u-tokyo.ac.jp

1　はじめに

文章自動要約とは、与えられた文章に記述された情報を簡潔にまとめた短い文章を自動的に生成することであり、自然言語処理の重要な課題の一つである。文章自動要約における様々な課題・手法が研究されてきたが、その中の一つに、文章の文や句の間の役割の関係を表す談話構造が要約作成において有用であるという研究が存在する [1, 2]。

一方で、文章要約の課題の一つとして学術論文の要約が存在する [3, 4]。学術論文の方が一般の文章と比べて論理的に文章が展開していくため、より談話構造の情報が要約精度の向上において有益になると考えられる。また、多くの学術論文にはそれぞれ Abstract が付いており、これはその論文の要約と捉えることができるため [5]、教師データを十分な量用意することができる。そのため、要約に用いられやすい特徴量を学習することができると考えられる。

そこで本論文では、学術論文の要約において談話構造を利用し、要約精度の向上を試みる。そうして学術論文要約の研究に寄与するとともに、この課題を通じて談話構造の有用性を検証する。それによって、談話構造の他の応用先に関しても効果を上げられることが期待される。

本論文では、論文中の各文から特徴量を抽出し、その文がどれだけ要約に使われ易いかをその論文の Abstract を参考に学習することで要約を生成するという手法を提案する。その際に本文の談話構造を利用した特徴量を用い、学術論文要約において談話構造が精度の向上に有用であるかを検証する。

結果として、学術論文の自動要約生成において、談話構造が精度向上に有効に働く可能性があることが明らかとなった。

2　関連研究

文章自動要約の課題の一つに、学術論文を要約するという課題が存在する [3]。学術論文の要約を生成することは、特に研究者にとってある課題・手法・研究分野といったものを理解するのに助けとなる。学術論文特有の特徴として、引用の情報を利用した手法が存在する [4]。具体的には、要約対象となる学術論文を引用している他の論文における、その論文を引用している文の情報も要約に利用する。他の論文を引用している文には、引用した論文の目的や手法、貢献などが簡潔に書かれていることが多く、要約の生成において有用であると考えられるためである。また、学術論文にはほとんどの場合 Abstract が書かれているが、この Abstract は論文の要約と捉えることができる。したがって、Abstract を再現するような文章を作ることで、学術論文の要約は作ることができると考えられる [5]。

一方、文章自動要約において談話構造を利用する研究も存在する。談話構造とは、文章中の文や句の間の役割の関係や話題の推移といった構造を表すもので、質問応答や会話応答において文脈の把握に役立つことが知られている [6]。Louis らは、抽出的単一文章要約手法において、談話構造をもとにした特徴量が有意な役割を果たすかどうかに関して研究を行った [1]。そして実験の結果、談話構造を元に生成された特徴量が、文が要約に用いられたかそうでないかを分ける有意な特徴であることが分かった。

3　提案手法

学術論文における Abstract は、その論文の要約と捉えることができる。そのため、Abstract を要約の教師データとして用いることができると考えられる [5]。一般的に Abstract は分量がかなり少ないが、要約の分量を変化させることで要求に沿った要約を生成するといった応用が可能であると考えられる。

そこで本論文では、本文中のそれぞれの文がどれだけ Abstract に用いるのに適しているかの値（本論文では Abstract 適応度と呼ぶことにする）を推定する関数を学習し、その適応度が高い文を並べることで Abstract を自動生成する手法に関して実験を行う。

本手法では、論文中の各文とその論文の Abstract の文との類似度を測ることで、それぞれの文の Abstract

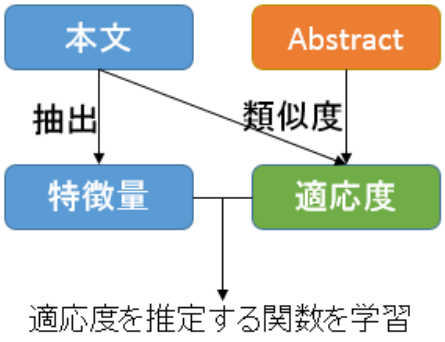

図 1: 学習データの論文からの関数の学習

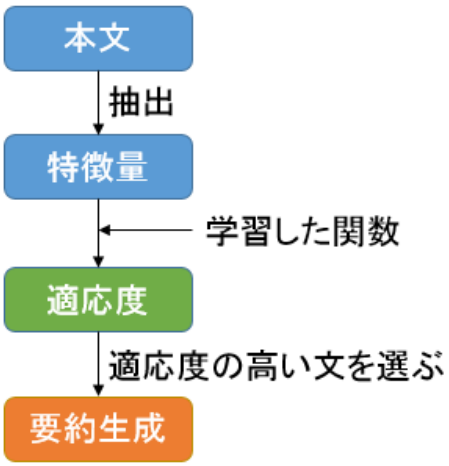

図 2: 学習した関数を用いての Abstract 生成

適応度を定める。そして、論文中の各文から特徴量を抽出し、線形回帰を用いて各文の特徴量からその文の Abstract 適応度を求める関数を学習する（図 1）。

論文から Abstract を生成する際には、論文中の各文から抽出した特徴量と学習した関数を用いてそれぞれの文の適応度を推定する。そして、その適応度の合計が最大となるように文を選択することで Abstract を生成する（図 2）。

そして、この手法においてそれぞれの文から特徴量を抽出する際に本文の談話構造を利用した特徴量を用いることを提案する。そして、その特徴量を用いた場合と用いなかった場合の精度を比較することで、学術論文の自動要約における談話構造の有用性を検証する。

3.1 Abstract 適応度

本文中の各文の Abstract 適応度は式 (1) のように定める。

$$\max_{s^* \in abs} Sim(s, s^*) \tag{1}$$

ここで、abs は Abstract に含まれる文の集合である。つまり、Abstract に含まれる文の中で最も類似度の高い文との類似度をその文の Abstract 適応度とする。また、類似度関数 $Sim(s, s^*)$ は

$$Sim(s, s^*) = \frac{1}{|s^*|} \sum_{w \in s, w^* \in s^*} same(w, w^*) \tag{2}$$

ただし、

$$same(w, w^*) = \begin{cases} 1 & (w, w^* \notin stopword, w = w^*) \\ 0 & (otherwise) \end{cases}$$

とする。つまり、stopword を除いて二つの文で一致している単語の数を Abstract に含まれる文の長さで割ったものを類似度関数に用いている。

3.2 学習モデル

それぞれの文の特徴量から Abstract 適応度を推定するモデルには、線形回帰を用いる。線形回帰では、特徴量ベクトル $\boldsymbol{x}$ から値 y を $y = \boldsymbol{w}^T \boldsymbol{x}$ といった形で表す。

学習の際には、学習用の論文データからそれぞれの文の特徴量 $\boldsymbol{x}_i$ と Abstract 適応度 y_i を取り出して学習データ $\{\boldsymbol{x}_i, y_i\}(i = 1, 2, \cdots, n)$ を用意する。そして、その学習データにおいて誤差関数

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \boldsymbol{w}^T \boldsymbol{x_i})^2 + \alpha |\boldsymbol{w}|^2 \tag{3}$$

を最小化するように重みベクトル $\boldsymbol{w}$ を学習する。ここで第 2 項は過学習を防ぐ正則化項で、α を変化させることで学習データへの適合度合いを調整できる。

各文の特徴量ベクトルの要素は、以下に述べるものを用いる。

3.3 談話構造に基づいた特徴量

本手法では、論文中のそれぞれの文から談話構造に基づいた特徴量を抽出するために、論文の各セクションごとに談話構造を解析し、RST 形式の談話構造木にパーズする。RST 形式の談話構造木は、図 3 のような形式となっている。

まず、文章は Elementary Discourse Unit (EDU) という単位に分割される。この単位は、文または句といった意味のまとまりである。そして、Root ノードはその文章の全ての EDU の集合に当たり、それが次々に分解されていく。つまり、各ノードは文章中の EDU の連続した集合に相当している。そして、リーフノードにはそれぞれの EDU が相当している。

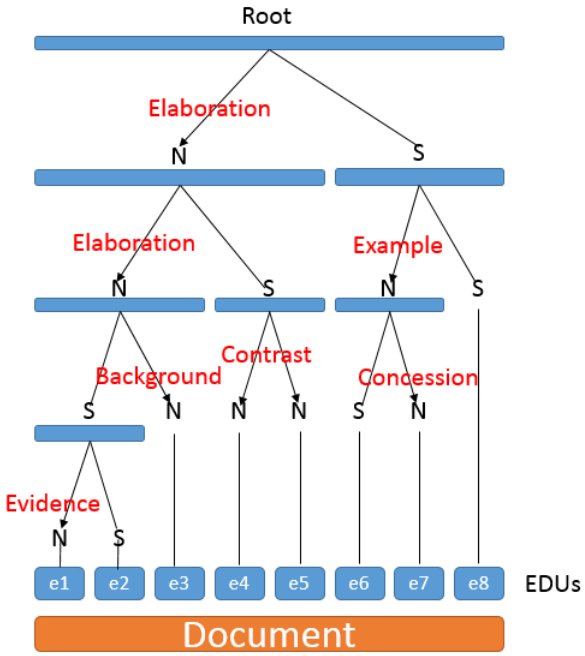

図 3: RST コーパスの談話構造木の一例

そして、各々のノードは Nucleus 又は Satellite という 2 つの状態のどちらかに属している。Nucleus は Satellite より構造上重要な役割を持っているとされる。そして、ノード間には、Cause, Attribution などといった 16 のクラスに分けられた関係が記されている。

この談話構造木から特徴量を抽出する方法は、Louis らの研究 [1] でも用いられていた Ono Penalty と呼ばれる指標を採用した。Ono Penalty はこの研究以前に小野ら [7] によって提案された特徴量であり、その EDU に相当するリーフノードから根元のノードに至るまでに Satellite ノードを通った回数である。

図 3 の Discourse Tree においては、e1 の Ono Pelanty は 1 となり、e2 の Ono Penalty は 2 となる。

この Ono Penalty を元に、それぞれの文の Discourse Score を

$$DiscourseScore(s) = \frac{1}{\max_{e \in s} OnoPenalty(e) + 1}$$

と定める。ここで e は文 s に含まれる EDU である。つまり、その文に含まれる EDU の最大の Ono Penalty の逆数をその文の Discourse Score としている。

3.4 その他の特徴量

前節で述べた特徴量以外には、それぞれの文から以下のような特徴量を抽出する。

1. タームの頻度：本文中に多く登場する単語はその論文において重要な単語であると考えられる。よって、そのような単語を多く含む文も重要な文であると考えられる。そこで、各文の Term Score を式 (4) のように定める。

$$TermScore(s) = \sum_{w \in s} tf(w) \tag{4}$$

ここで、$tf(w)$ とは単語 w の頻度で、単語 w がストップワード（a, the, for などの一般的な単語）の場合は 0 になり、それ以外の単語の場合は単語 w が文章中に現れる頻度となる。

特徴量ベクトルの要素として用いる際には、その文が含まれる論文中の文の中での最大の Term Score で割ることで $[0, 1]$ の範囲に正規化する。

2. 本文中の文の位置：単純ながら強力な単一文章要約手法として、文章の先頭から指定された長さだけ取り出してそのまま要約とする LEAD 法が存在することからも分かるように、文章の先頭に近い場所に存在する文は要約として使われ易い。そこで、文章の先頭からその文までの距離の逆数を特徴量として用いる。

3. セクション中の文の位置：論文はセクションごとに分けられていることが多いが、各セクションの最初の方に登場する文はより重要度が高いと考えられる。ただ、本文全体での位置と比べて、セクション中では冒頭の文だけ飛び抜けて重要度が高いとはづらい。そこで、文がセクションの先頭として現れたら 1、末尾として現れたら 0 として、その間は線形に分配して特徴量として用いる。

3.5 Abstract の生成

論文中のそれぞれの文の Abstract 適応度を求めた後は

1. 文の長さを「重さ」、Abstract 適応度を「価値」としたナップザック問題として定式化する
2. 動的計画法を用いてナップザック問題を解き、指定された単語長に収まる範囲で適応度の合計が最大となるように文の集合を選ぶ
3. 選んだ文を元の論文に登場する順番で並べる

という方法で Abstract を生成する。

4 実験

4.1 実験設定

データは、Teufel らが公開している Argumentative Zoning Corpus [8] を用い、談話構造の解析には Feng

表 1: 実験結果

Method	ROUGE-1 F-score
LEAD 法	0.230
談話構造なし	0.231
提案手法	**0.271**

らの談話構造解析器 [9] を用いた。このコーパスに含まれる 80 の学術論文のうち談話構造の解析を行えた 50 論文を実験に使用し、40 論文を学習用、10 論文をテスト用に用いた。その際、3 単語以下からなる文はすべて取り除いた。学習の際は、式 (3) における係数 $\alpha = 0.02$ とした。また、Abstract の指定単語長は全て 150 単語とした。

評価には、ROUGE と呼ばれる評価手法 [10] を用いる。ROUGE は、文章要約の際によく用いられる評価法であり、単語の n-gram の一致度などの観点から 2 つの文章の類似度を評価する手法である。人間が作成するなどした要約の「正解」とシステムが自動生成した要約を ROUGE を用いて比較することで、その自動要約システムの評価を行っている。

そして、本手法の他に、ベースラインとして LEAD と呼ばれる本文を頭から指定された単語長だけ選ぶ手法、および談話構造の有用性を検証するために談話構造を元にした特徴量を使用しなかった手法に関しても実験を行い本手法と性能を比較した。

4.2 実験結果

実験結果を表 1 に示す。談話構造を元にした特徴量を使用していない場合では LEAD 法と殆ど同等の結果となったが、談話構造を元にした特徴量を使用することで精度が改善することが分かった。

以下に例として、Brennan らの " A Centering Approach to Pronouns " という代名詞の処理に関する論文 [11] から本手法が生成した要約として選んだ文のうち、要約として有用であると思われたものを抜粋して示す。

- The (Grosz et al. 1986) centering model is based on the following assumptions.
- Therefore, we propose the following extension which handles some additional cases containing multiple ambiguous pronouns: we have extended rule 2 so that there are two kinds of shifts.

5 おわりに

文章自動要約の一つの課題として学術論文の要約が存在し、また文章要約において文章の談話構造が利用できるという研究が近年存在する。そこで本論文では、学術論文の Abstract を自動生成する課題に文章の談話構造を利用し、その有用性を検証した。結果として、談話構造を元にした特徴量を使用することで精度が改善する可能性があることが明らかとなった。

今後の課題は、本論文では談話構造における Nucleus と Satellite の関係のみ考慮していたが、Cause, Attribution などの隣り合ったノード同士の関係といった談話構造の他の要素も考慮すること。また、引用の情報などの学術論文特有の特徴をうまく組み合わせること。そして、単純に文ごとに適応度を定めるだけではなく、Abstract として適した文同士の組み合わせを考慮するといったことが挙げられる。

参考文献

[1] Annie Louis, Aravind Joshi, and Ani Nenkova. Discourse indicators for content selection in summarization. SIGDIAL '10, pp. 147–156, 2010.

[2] Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. EMNLP '13, pp. 1515–1520, 2013.

[3] Nitin Agarwal, Kiran Gvr, Ravi Shankar Reddy, and Carolyn Penstein Rosé. Towards multi-document summarization of scientific articles: Making interesting comparisons with scisumm. WASDGML '11, pp. 8–15, 2011.

[4] Vahed Qazvinian, Dragomir R. Radev, Saif M. Mohammad, Bonnie Dorr, David Zajic, Michael Whidby, and Taesun Moon. Generating extractive summaries of scientific paradigms. *J. Artif. Int. Res.*, 2013.

[5] Danish Contractor, Yufan Guo, and Anna Korhonen. Using argumentative zones for extractive summarization of scientific articles. COLING '12, pp. 663–678, 2012.

[6] Bonnie Webber, Markus Egg, and Valia Kordoni. Discourse structure and language technology. *Natural Language Engineering*, pp. 437–490, 2012.

[7] Kenji Ono, Kazuo Sumita, and Seiji Miike. Abstract generation based on rhetorical structure extraction. COLING '94, pp. 344–348, 1994.

[8] Argumentative zoning corpus. `http://www.cl.cam.ac.uk/~sht25/AZ_corpus.html`.

[9] Vanessa Wei Feng and Graeme Hirst. Text-level discourse parsing with rich linguistic features. ACL '12, pp. 60–68, 2012.

[10] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. ACL '04, pp. 74–81, 2004.

[11] Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. A centering approach to pronouns. ACL '87, pp. 155–162, 1987.