修　　士　　論　　文

要約抽出文の解析結果を用いた教師付き学習による学術論文要約

Supervised scientific paper summarization using sentence extraction analysis

指導教員　　鶴岡 慶雅 准教授

東京大学工学系研究科
電気系工学専攻

氏 名　　**37-146467** 中須賀謙吾

提 出 日　　平成 **28** 年 **2** 月 **4** 日

概 要

文章自動要約は自然言語処理の重要な課題である。現在利用できる情報は日に日に増えており、より短い文章を読むことでそういった情報を得ることができるようになるのは多くの人にとって助けとなる。また、要約という課題を研究することで、文章・言語の理解といった他の課題にもつながる。

文章自動要約においても他の自然言語処理の課題と同様に教師データを用いた教師付き学習において精度の向上が試みられているが、要約の教師データは十分に用意するのが難しいという問題点が存在する。

また、学術論文の自動要約生成という課題が存在する。学術論文には多くの場合 Abstract が付与されており、Abstract は要約の正解と捉えることができる [1]。そのため、学術論文を要約の対象とすることで、教師付きのデータを大量に用意することができる。そういった条件のもとで実験を行うことで、一般の文章自動要約に関しての有用な知見が得られることが期待できる。また、学術論文の自動要約生成という課題自体も、主に研究者などにとってその分野の理解に役立つ。

本研究では、Abstract を教師データとして用いる手法として、その論文の本文中から Abstract との類似度ができるだけ大きくなるような文集合を取り出し、それらの文を正解データとして扱う手法を提案する。また、そのようにして取り出した文集合を解析することで、要約手法の精度向上の余地や有用な特徴量に関して考察を行う。

実験の結果、提案手法を用いて取り出した文集合はその論文の要約として十分に高い質であることが判明した。また、その文集合を正解データとして学習を行うことで、学習を行わない手法より精度の高い要約の生成が可能であるという結果が得られた。また、教師データの量を増やすことで要約の精度が向上するということが明らかとなった。

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　背景

現在、インターネット上には利用することのできる大量の情報が溢れており、その中から欲しい情報を素早く集める等、それらの情報を自動で効率よく扱うために自然言語処理の需要は増している。

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。例えば、以下のような文章が与えられたとする。

> Early on the morning of May 27, a strong earthquake with a magnitude of 6.3 struck the Special region of Yogyakarta in the central part of Java, Indonesia. According to a report on June 5, 5,782 people have died and over 36,000 people have been injured. More than 87,000 houses have collapsed and approximately 340,000 people are how homeless. …

これは 2006 年 5 月にインドネシアで発生した地震に関するニュースである。この文章から、インドネシアで地震が発生した、それによって死者が数多く発生した（既に 5 千人以上の死者が確認されている）といった情報を簡潔にまとめ、

Strong earthquake kills thousands in Indonesia.

といった短い文章を自動的に生成するという課題が文章自動要約である。

文章自動要約が重要な問題である理由として、まず世の中の人々にとって必要とされていることが挙げられる。膨大な量の文章を読むことは現実的に不可能であり、要約によって必要な情報をより短い文章を読むことで理解できるようになれば、人々にとって大きな助けになると考えられる。

また、要約を生成するには言語の深い理解を必要としている、要約手法にはまだ解くべき問題や課題が数多く残っているといった学術的な動機も挙げられる。

文章自動要約には、文章中に書かれている文をそのまま抽出することで要約を作成する抽出型と、新しい文を生成して要約を作成する非抽出型、単一文章の要約と同じような事柄について書かれた複数文章の要約、ある特定の観点に関して要約を行うクエリ指向型、それまでに発行された文章の情報は既知として、最新の文章において新たに判明した情報を要約するアップデート型など様々な課題・手法が存在する。

文章自動要約も他の多くの自然言語処理の課題と同様に教師付き学習を用いた手法が試みられている [6]。しかし、文章要約にはあまり大きな教師データのコーパスが存在しない。そのため、様々な特徴量を試そうとしても上手く学習が行えないといった問題点が生じている。

また、近年学術論文の要約を行う研究が出てきた [7, 8]。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。

学術論文には多くの場合に Abstract が付いており、これは論文の著者が書いたその論文の要約と捉えることができる。そのため、Abstract を要約の正解としてみなすことができると考えられる [1]。

1.2 本研究の目的

本研究では、Abstract を教師データとして用いた教師付き学習を用いることで、学術論文の自動要約生成を行い、その精度の向上させることを目的とする。

1.3 本研究の貢献

まず、抽出的手法においての達成可能な精度の最高値を推定し、精度向上の余地が大いにあることを示した。また、具体的にどういった特徴を持つ文が要約として選ばれたかを確認した。また、教師データを増やすことで要約精度が向上する余地があることを示した。

1.4 本論文の構成

まず 1 章にて本研究の背景や目的を説明する。続いて 2 章では関連研究として教師付き学習の手法や文章自動要約に関する研究について説明する。3 章では提案手法の説明を行い、また実際にどういった文が最適な要約として選ばれるかといった点に関して考察を行う。4 章では実験の手法とその結果からの考察を行い、最後に 5 章で本研究の結論と今後の課題について述べる。

第2章　関連研究

2.1　教師付き学習

教師付き学習とは、機械学習の手法の一つであり、その中でも最も基礎的な手法である。ある問題を解くために、その問題と正解の組が列挙された「教師データ」を用いて、そのデータから正解を導く方法を学習するといった手法である。

教師付き学習の典型的な課題として分類問題がある。分類問題は、例えば教師データが

$$(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \cdots$$

のように与えられるので（ここで各 $\mathbf{x_i}$ は入力となるベクトル、y は分類結果となるラベルである）、新たな未知の入力 $\mathbf{x}$ に対する分類結果 y を求められるような分類器を学習する、といったものである。

そういった分類問題を解く教師付き学習の主な手法としては、パーセプトロン学習 [9] や条件付き確率場 [10] などが存在する。

2.1.1　パーセプトロン学習

パーセプトロン学習は、正解のラベル付きのデータから学習する教師有り学習においてよく用いられる手法の一つである。パーセプトロン学習は、学習データを利用して、正例・負例を分離する平面を見つけるものである。分離平面の例を図 2.1 に示す。

入力を x、特徴ベクトルを $\mathbf{\Phi}(x)$、重みベクトルを $\mathbf{w}$ とすると、出力は以下の式 (2.1) で表され、$+1$ なら正例、-1 なら負例と判断される。

$$y(\mathbf{x}) = f(\mathbf{w}^T \mathbf{\Phi}(x)) \tag{2.1}$$

ここで、$f(x)$ は以下の式 (2.2) で表されるステップ関数である。

$$f(a) = \begin{cases} +1 & (a \geq 0) \\ -1 & (a < 0) \end{cases} \tag{2.2}$$

この重みベクトル $\mathbf{w}$ を学習するパーセプトロン学習アルゴリズムは以下のようになる。

1. 重みベクトルを $\mathbf{w} = \mathbf{0}$ と初期化する
2. 学習データからランダムにサンプルを選択し、現在の重みベクトルを用いて分類を行う。

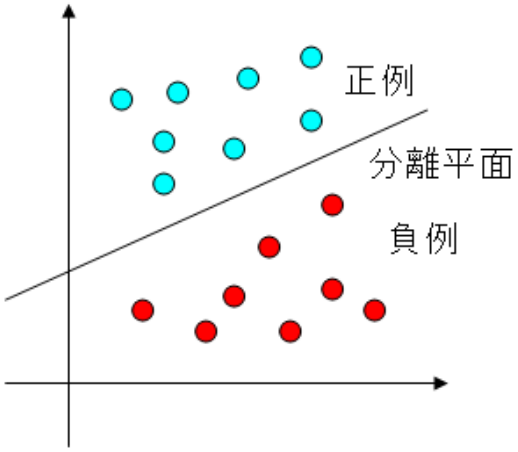

図 2.1. 二次元座標における分離平面の例

3. 分類が間違っていたら以下の式で重みベクトルを更新する
   - 正しくは正例のものを負例と分類した場合　$\mathbf{w} \leftarrow \mathbf{w} + \mathbf{\Phi}(x)$
   - 正しくは負例のものを正例と分類した場合　$\mathbf{w} \leftarrow \mathbf{w} - \mathbf{\Phi}(x)$
4. すべてのサンプルを正しく分類できるまで (2) に戻り繰り返す

2.1.1.1 平均化パーセプトロン

学習データが線形分離可能であれば前述のアルゴリズムにより、分離平面を有限のステップで見つけられることが保証されている。しかし、データが線形分離不可能（図 2.2 のように分離平面が存在しない場合）である時は、パーセプトロン学習は収束しない。また線形分離可能である場合でも、ステップ数が非常に大きくなることがある。加えて、学習データに対する分離平面がテストデータに対しても分離平面となっているとは限らない。学習データに対して収束するまで学習を行うと、過学習になる可能性もある。

そこで、実際の学習においてはパーセプトロンを改良した平均化パーセプトロンがよく用いられる [11]。上に述べたパーセプトロン学習において、過去の重みベクトルの和を記憶しておき、学習データの全てのサンプルを読み込んだ後に重みの平均化を行う。このようにすることで学習データによる過学習を防ぎ、また収束を待たないので線形分離不可能な場合でも学習を終了させることができる。サンプル数を N、i 番目のサンプルを読み込んだ後の重みベクトルを $\mathbf{w_i}$、過去の重みベクトルの和 $\mathbf{v} = \sum_{i=1}^{N} \mathbf{w_i}$ とすると、最終的な重み $\mathbf{w}$ は

$$\mathbf{w} = \frac{\mathbf{v}}{N} \tag{2.3}$$

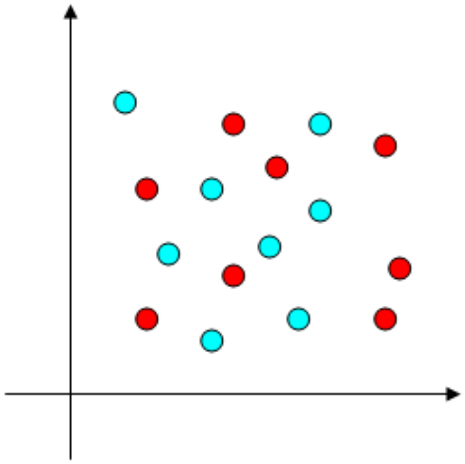

図 2.2. 線形分離不可能な例

となる。

2.2 文章自動要約

要約とはある文章の内容を簡潔に短い文章にまとめることであり、文章自動要約は要約を自動で行うという課題である。

文章自動要約が重要な問題である理由として、まず世の中の人々にとって必要とされていることが挙げられる。膨大な量の文章を読むことは現実的に不可能であり、要約によって必要な情報をより短い文章を読むことで理解できるようになれば、人々にとって大きな助けとなると考えられる。

また、要約を生成するには言語の深い理解を必要としている、要約手法にはまだ解くべき問題や課題が数多く残っているといった学術的な動機も挙げられる。

2.2.1 抽出的手法と非抽出的手法

文章自動要約においても様々な手法が研究されているが、まず手法は大きく「抽出的手法」と「非抽出的手法」の 2 つに分けられる。

抽出型手法とは、文章中の一部の文をそのまま抽出して並べることで要約を生成するという手法である。例えば、図 2.3 のように、文章から赤字の文のみを抽出して要約を生成するという手法が抽出的手法である。

そのため、抽出的手法では、文章中の重要な文を何からの方法で推定し、抽出を行う。推定において、グラフベースの手法を用いた研究 [12] や、文章の談話構造を用いた研究 [13] が存在する。

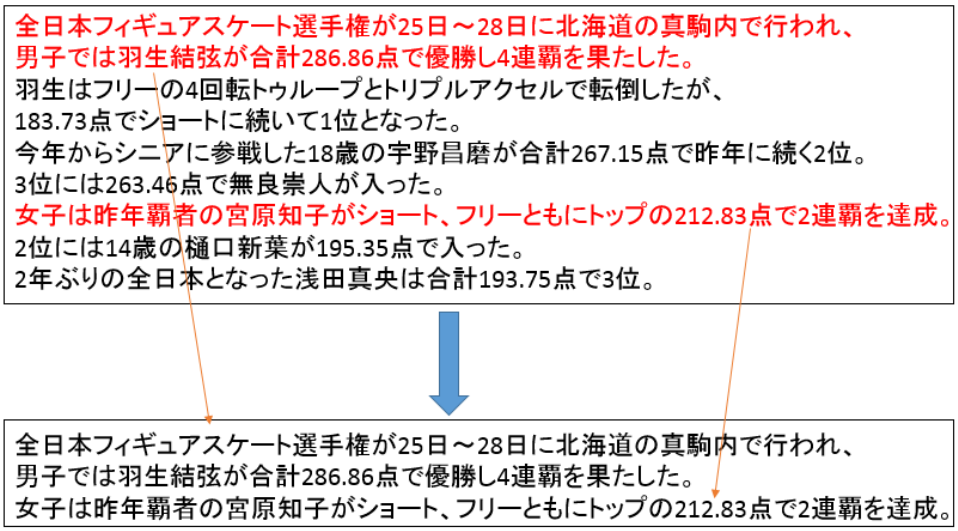

図 2.3. 抽出的手法による要約の例

また、要約に含まれる情報を保ったままできるだけ要約の長さを短くするために、抽出した文をその文に書かれている内容を保ったまま短くする文圧縮と組み合わせる手法も存在する [14]。

一方で、文章中の文を抽出するのではなく、新たに文を生成することで要約を行うのが非抽出的手法である。非抽出的手法による要約では、アテンションベースの手法 [15] などが挙げられる。図 2.3 と同じ文章は、非抽出的手法を用いると例えば図 2.4 のように要約される。

単に文を選ぶだけであった抽出的手法と比べ、新たに文を生成する非抽出的手法はより困難な手法である。しかし、図 2.3 と図 2.4 を見比べることから分かるように、非抽出的手法の方がより短い要約に文章の内容を収めることができる可能性がある。そのため、抽出的手法だけでは高品質な要約は作れないのではないかといった問題提起も存在する [16]。

2.2.2 文章自動要約と教師付き学習

抽出的手法も用いる場合も非抽出的手法を用いる場合も、どういった文や節・単語が重要であるかといったことを推測するために、文章の各特徴量をパラメータ化して推測を行う。その際どういった特徴に注目するか、どのように重みづけを行うかといった点に関しては人間がルールベースで行うことも出来るが、他の多くの自然言語処理の課題と同様に教師データから学習する（要約の場合は、正解の要約に近い要約を生成するように）ことで、より正確に重みの推定を行うことができると考えられる。

しかし、要約にはそこまで大きな教師データは存在しない。例えば、同じく自然言語処理の主要な課題である機械翻訳の場合と比べてみると、科学論文から取り出した ASPEC コーパス [17] が約

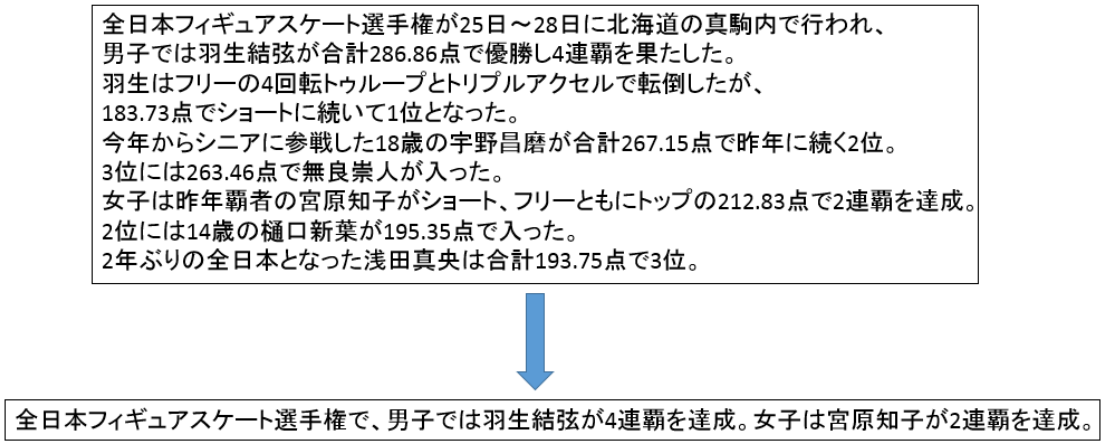

図 2.4. 非抽出的手法による要約の例

300 万対訳文、特許文章から取り出した NTCIR コーパス [18] も同様に約 300 万対訳文を有している。一方文章要約の場合は、文章要約に関する主要な国際会議である DUC (Document Understanding Conference) [19] では毎年要約の課題・テストデータを公開しているが、例えば DUC2004 においては約 4000 要約文と少ない量になっている。

そのため、様々な素性を試そうとしても上手く学習が行えないといった問題点が生じている。

2.2.3 学術論文要約

学術論文には Abstract が付いており、Abstract がその論文の要約であるともいえる。しかし、Abstract よりも長い要約が必要である、ある分野の研究をまとめた要約が必要である、またはその論文の筆者以外の視点でのその論文のまとめが必要であるといった場合などに、学術論文要約の必要性が生じる。

2.2.3.1 Citation 情報を利用した学術論文要約

学術論文特有の特徴として、Citation の情報を利用した手法が存在する [20, 21, 22, 23]。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を要約に利用する (図 2.5)。

例えば、以下の文は Qazivinian らの論文 [22] の中の一文であるが、

> (Romanello et al., 2009) use Conditional Random Fields (CRF) to from extract references from unstructured text in digital libraries of classic texts.

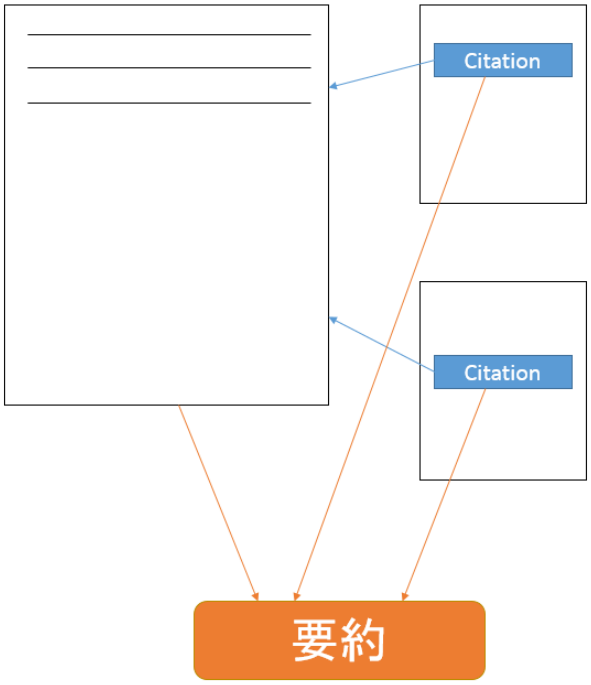

図 2.5. 他論文における引用部分を利用した要約生成

この一文から、Romanello らの論文では非構造化テキストから参照の情報を抽出する目的で Conditional Random Fields を用いた手法を提案ししてことが分かる。このように、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非常に有用であるためである。

2.2.3.2 学術論文中の文の役割解析

学術論文におけるそれぞれの文は、背景について述べている文、先行研究について述べている文、提案手法について述べいている文、といったように、一般的にどの論文でも同じような役割に分類することができる。

Teufel らは、ArgmentatizeZoning という、学術論文の中のそれぞれの文を予め定めておいた役割に分類するという課題を提案した [24, 25]。ArgmentativeZoning においては、学術論文中の文は以下の 7 つのカテゴリーに分類される。

- Own (OWN): 自分の行った研究に関する中立的な記述
- Other (OTH): 他の研究に関する中立的な記述

- Backgorund (BKG): 学術的な背景に関する記述
- Contrast (CTR): 自分の研究と他の研究の比較に関する記述（他の研究の弱点を明確に記した記述）
- Basis (BAS): 他人の研究に基づいた自分の研究に関する記述
- Aim (AIM): その論文・研究での目的に関する記述
- Textual (TXT): その論文の文章的な構成に関する記述（「2 章で関連研究、3 章で提案手法に関して述べる」などといった記述）

例えば、以下の文章は Poznański らの機械翻訳に関する論文 [26] の Abstract であるが、

> The lexicalist approach to Machine Translation offers significant advantages in the development of linguistic descriptions. However, the Shake-and-Bake generation algorithm of (Whitelock, 1992) is NP-complete. We present a polynomial time algorithm for lexicalist MT generation provided that sufficient information can be transferred to ensure more determinism.

この Abstract の第 1 文は、機械翻訳に対して語彙的なアプローチの手法が現れたという他人の研究について述べている。第 2 文は、その手法は計算量が NP 完全であるという弱点を述べている。第 3 文は、この論文では計算量が多項式時間である手法を提案する、という研究の目的について書かれている。よって、それぞれの文を ArgumentativeZoning の 7 つのカテゴリーに分類すると以下のようになる。

- **OTH:** The lexicalist approach to Machine Translation offers significant advantages in the development of linguistic descriptions.
- **CTR:** However, the Shake-and-Bake generation algorithm of (Whitelock, 1992) is NP-complete.
- **AIM:** We present a polynomial time algorithm for lexicalist MT generation provided that sufficient information can be transferred to ensure more determinism.

また、こういった論文中の各文の役割の推定を機械学習を用いて行う研究も行われている [27, 28, 29, 30]。

論文中の各文の役割を自動的に判別するだけでも論文を読む人にとって有益であるが、この ArgmentativeZoning の情報を学術論文の要約に利用する研究も行われている [1]。

2.2.3.3 Abstract を教師データとして用いた学術論文要約

2.2.2 節で述べた通り、文章要約における教師データの数は多くは存在せず、またそういった教師データを人手で用意することも労力を要する。しかし、前述の通り学術論文には多くの場合に Abstract が付いており、これは論文の著者が書いたその論文の要約と捉えることができる。そのため、Abstract

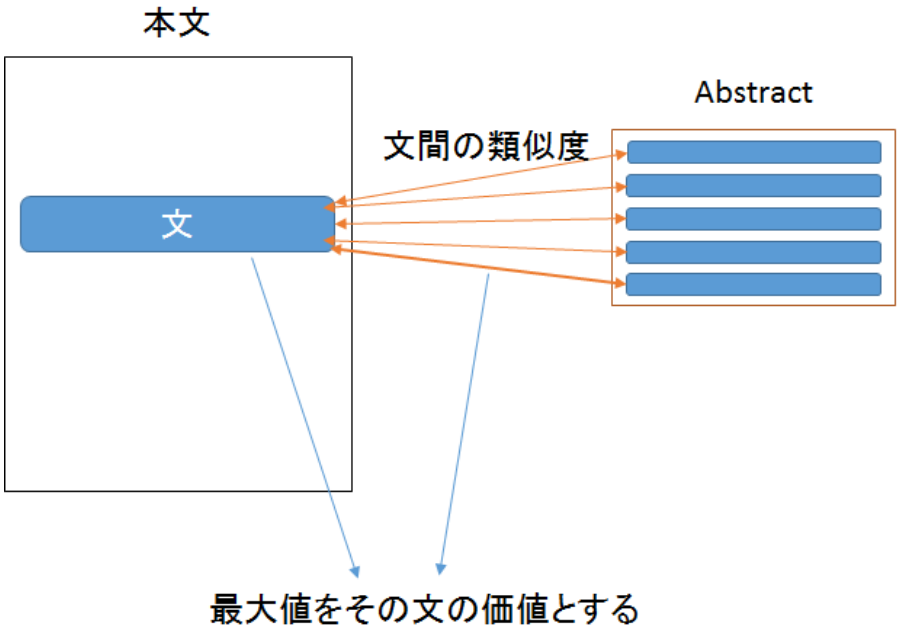

図 2.6. 先行研究 [1] における Abstract の利用方法

を要約の正解としてみなすことができると考えられる [1]。従って、学術論文はそれ自体が要約における正解付きの教師データとみなすことができる。

従って要約の対象に学術論文を取ることで、大量の学習データを要する文章要約の手法を試すことができる、といった点も存在する。

Contractor らの研究 [1] では、論文中に含まれる各文の「価値」を Abstract に含まれる文との類似度の最大値とする (図 2.6) といった方法で Abstract を教師データとして用いている。

そして 2.2.3.2 節で述べた文の役割などの特徴量を各文から抽出し、その特徴量から各文の価値を予測する回帰問題として学習を行う。そして、要約の際には学習をもとに各文の価値を予測し、その価値の大きい文を抽出して要約を生成している。

2.2.3.4 発表スライドの自動生成

また、学会発表などでは論文に関する発表スライドが用いられるが、論文からその発表スライドを自動生成するという研究も行われている [31]。

この課題も、出力を文章ではなくスライドとした学術論文の要約であると捉えることもできる。

2.2.4 文章自動要約の評価指標

また、文章自動要約の研究を行う際には、システムが出力した要約の質を評価する評価指標が必要となる。

もちろん、人間が見て質を評価するというのが分かりやすい指標であり、人間の実感に沿った評価を下すことができるが、大量の要約に対してそのように一つ一つ人間が評価を行うのは現実的ではない。

そこで、自動的に要約の評価を行うシステムが求められるが、要約の研究においては ROUGE と呼ばれる評価手法 [32] が主に用いられている。ROUGE は、2 つの文章の類似度をいくつかの基準によって評価する。そして、正解の要約（大抵の場合は人間が作成して用意する）とシステムが出力した要約の類似度を ROUGE で評価することによって要約の質を評価する。

ROUGE の評価指標としてよく用いられるのが以下に述べる ROUGE-N スコアである。

ROUGE-N スコア

ROUGE-N スコアは、2 つの文章の N-gram の一致度を測定する。正解の要約に含まれる N-gram の集合を $GramN_{ref}$、システムが出力した要約に含まれる N-gram の集合を $GramN_{sys}$ とおくと、ROUGE-N スコアは式 2.4,2.5,2.6 のようになる。

$$ROUGE_{N_{precision}} = \frac{|GramN_{ref} \cap GramN_{sys}|}{|GramN_{sys}|} \tag{2.4}$$

$$ROUGE_{N_{recall}} = \frac{|GramN_{ref} \cap GramN_{sys}|}{|GramN_{ref}|} \tag{2.5}$$

$$ROUGE_{N_{F1}} = \frac{2ROUGE_{N_{recall}}ROUGE_{N_{precision}}}{ROUGE_{N_{recall}} + ROUGE_{N_{precision}}} \tag{2.6}$$

要約の研究では、ROUGE-1 スコアおよび ROUGE-2 スコア、つまり unigram と bigram の一致度が主に評価指標として用いられる。

また、こういった ROUGE の評価指標を元に 2 つの文章の類似度を自動で測定できるツールが公開されている [33]。このツールでは、ストップワード [1] を除くかどうかといった細かいオプションを指定した上での測定を行うことができる。

[1] あまりに一般的で多く使われる単語のことで、自然言語処理においては検索精度・比較精度などの向上のためこういった単語を除外して処理を行うことも多い。助詞などの機能語 (日本語では「は」「の」など、英語では「a」「the」など) は殆どの場合ストップワードに該当する。

第3章　学習用教師データの生成

本章では、教師付き学習の際にどのように論文の Abstract を利用するのかを説明する。またそれに伴い、文章自動要約における抽出的手法の可能性や有用性に関しても調査・考察を行う。

3.1　使用データコーパス

本研究では、bioMed コーパス [34] に含まれる論文を使用データとして用いた。

BioMed コーパス [34] とは、生命・医学系の論文が含まれたコーパスである。生命・医学系の各ジャーナルに投稿された論文が、xml 形式で採録されており利用できるようになっている。

3.1.1　コーパスの前処理

コーパスに含まれる各論文に、以下の手順で前処理を行った。

1. 各論文から Abstract を取り出す。
2. 各論文の本文をセクションごとに取り出す。その際、各セクションの名前も同時に取り出す。
3. 取り出した Abstract とセクションごとの本文を Enju-parser [35] を用いてパースし、各単語の語幹や品詞タグといった情報を取り出す。

このようにすることで、各文のセクション中の位置や、その文の属するセクションの名前が分かる。また、例えば “And” と “and” といったように同じ単語だが文章中での表現が異なってしまったものも同じ単語と正しく認識できるようになり、各単語の品詞も利用できるようになる。

3.1.2　データコーパスの中身

3.1.1 節の処理を行い、bioMed コーパスに含まれる論文のうち 9763 本を今回使用するデータとして用いた。うち 7763 本を学習用データ、1000 本を開発用データ、1000 本をテスト用データとして用いた。それぞれの含まれる文、単語の数は表 3.1 のようになった。

そして、学習データ中における本文と Abstract の長さの比は表 3.2 のようになる。

文の数で考えても単語の数で考えても Abstract の長さは本文の長さの約 6%となっている。しかし、本文の長さに関わらずジャーナルの指定などで Abstract の長さがある程度決まっているということも考えられる。

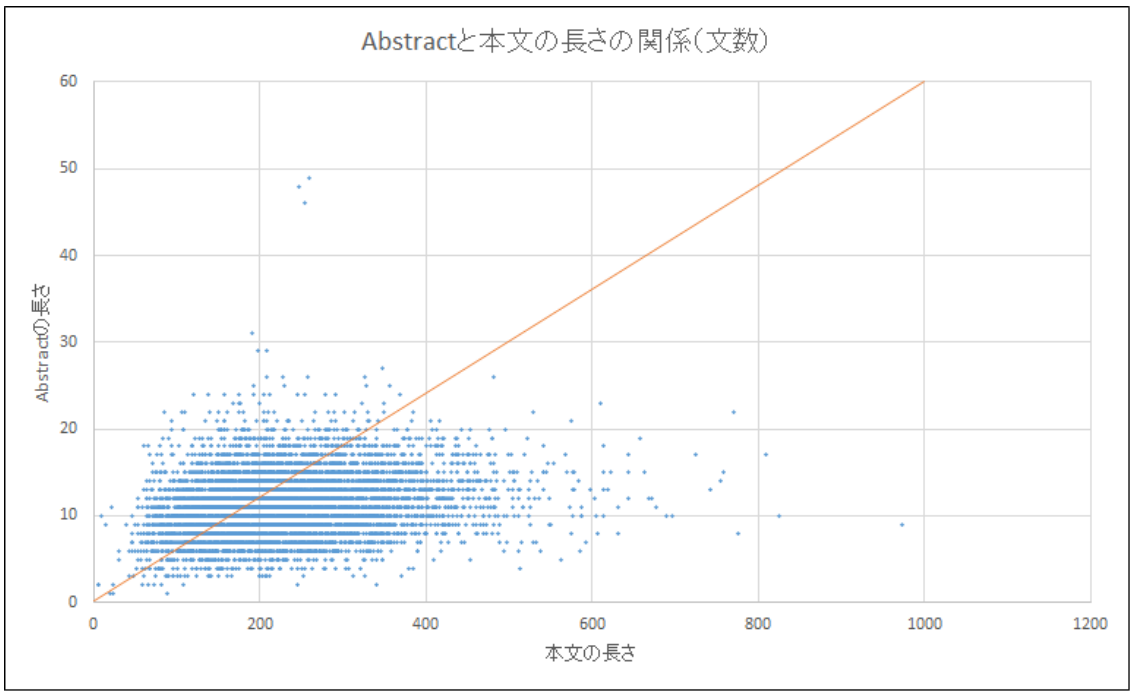

図 3.1. Abstract と本文の長さの比（文数）

そこで、Abstract と本文の長さの分布に関しての調査を行った。文の数に関しての散布図を図 3.1、単語の数に関しての散布図を図 3.2 に記す (どちらの図も、グラフ上に Abstract の長さが本文の長さの 6%となる直線が併記されている)。

このグラフからも、全体として Abstract の長さは本文の長さの約 6%とみていいことが分かる。

3.2 教師データの生成手法

本研究では、Abstract を要約の教師データとして扱い、教師付き学習により学術論文の要約を行う。学術論文という、教師データが自然に付与しているタスクで自動要約の手法を試すことで、他の

表 3.1. データ中に含まれる文・単語の数

	論文の数	含まれる文の数	含まれる単語の数
学習データ	7763	2303445	39218694
開発データ	1000	242855	5371863
テストデータ	1000	269276	6106989

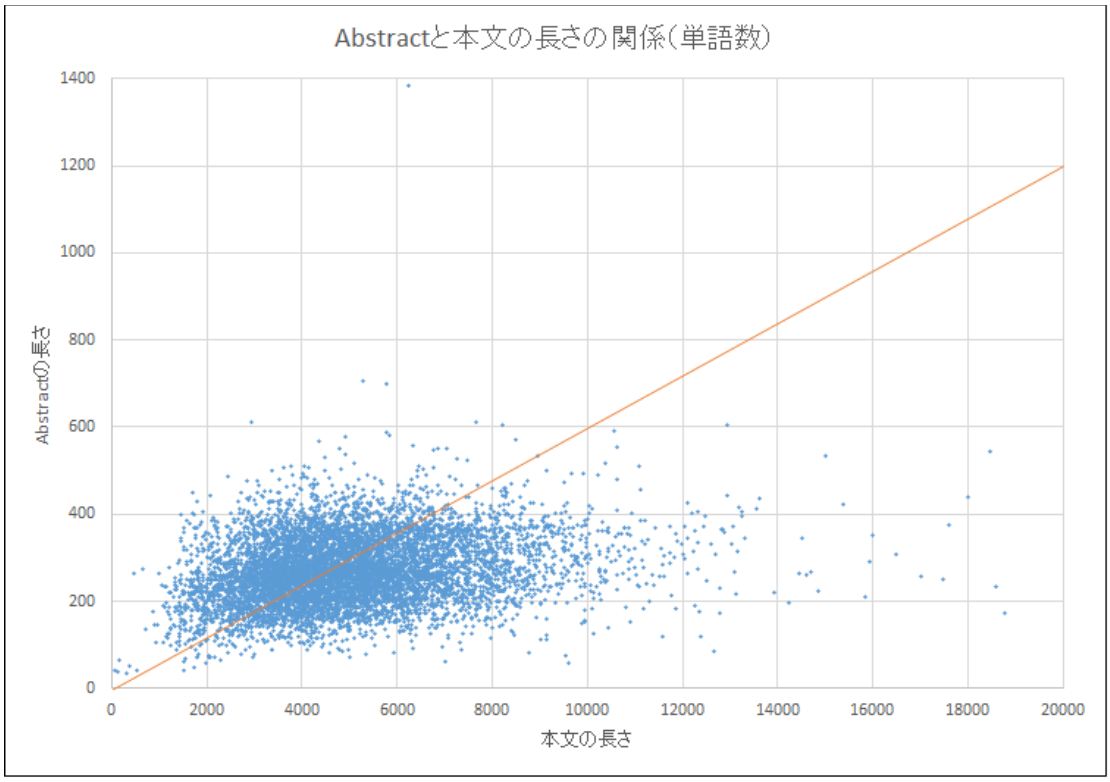

図 3.2. Abstract と本文の長さの比（単語数）

自動要約のタスクへの貢献にも寄与できると考えられる。

本研究では、抽出的手法により学術論文の自動要約生成を行う。その際、Abstract を要約の教師データとして用いる。そのため、抽出的手法で達成できる最も Abstract に近い要約を生成することが究極的な目標である。そこで、本文から取り出した文集合の中で Abstract をできるだけ再現するようなものを生成し、その文集合を学習を行うことで再現するという手法を提案する。

このようにすることで目標とすべき出力に直接近づけるように学習することができるため、より Abstract に近い要約を生成することができると考えられる。また、Abstract をできるだけ再現するような文集合を生成し、その精度や中身を解析することで、抽出的手法の可能性・限界といった点や学

表 3.2. 学習用データに含まれる本文と Abstract の長さの比

	Abstract	本文	本文に対する Abstract の割合の平均
文の数	87552	1791314	5.64%
単語の数	2109497	39218694	6.18%

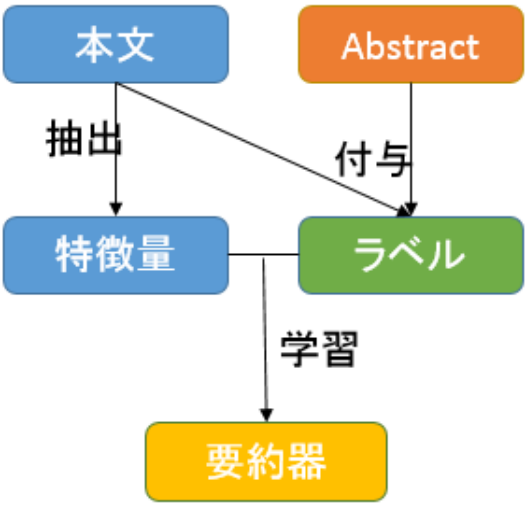

図 3.3. 提案手法における学習方法

習の際に用いるべき特徴量の考察を行うことができると考えられる。以下に今回行った手法を説明する。

学習の際には、まず論文の本文中の各文から特徴量を抽出する。そして Abstract を参照してその文が要約に選ばれるべき文であるかどうかのラベルを付与し、そのラベルと特徴量を教師データにし要約器の学習を行う (図 3.3)。

そして要約を生成する際には、学習の際と同様に論文の各文から特徴量を抽出する。そしてその特徴量を元に要約器が要約に選ばれる文を選び、選ばれた文を元論文に現れる順番に並べて要約を生成する (図 3.4)。

ラベルの付与に関しては、以下のような手法を提案する。

1. 論文の本文から、Abstract をできるだけ再現するような文の集合を取り出す
2. 手順 1 で取り出された文には正例、それ以外の文には負例のラベルを付ける

この手法を図で表すと図 3.5 のようになる。

また、「Abstract をできるだけ再現する」という点に対する定義は、Abstract と文集合の ROUGE-1 スコアで評価した類似度をできるだけ大きくする、ということとする。

3.3　正解データとなる文集合の生成アルゴリズム

本節では 3.2 節で述べた、論文の本文中から Abstract との類似度 (ROUGE-1 F1 スコア) が最大となるような文集合を取り出すアルゴリズムについて説明する。

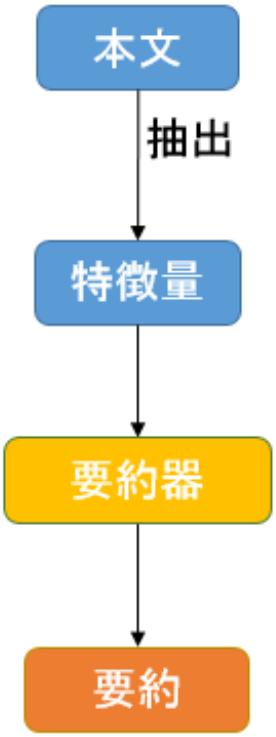

図 3.4. 提案手法における要約方法

全ての文集合について Abstract との類似度を求めれば類似度最大の文集合を求めることができるが、全ての文集合の数は $2^{(\text{本文中の文の数})}$ 存在するため、それら全てに関して類似度を調べるのは現実的ではない。そこで本論文では、以下に記すアルゴリズムを用いて正解データとなる文集合を生成する手法を提案する。

式 2.4,2.5,2.6 より、Abstract に含まれる単語の集合を Abs、選んだ文集合に含まれる単語の集合を Sum と置くと、ROUGE-1 F1 スコアは式 3.1 のようになる。

$$ROUGE_{1_{F1}} = \frac{2 * |Abs \cap Sum|}{|Sum| + |Abs|} \tag{3.1}$$

式 3.1 より、選んだ文集合に含まれる単語の数が同じであれば、その文集合と Abstract の両方に含まれる単語の数が多いほどスコアは高くなる。よって、単語の数ごとに、その中で「文集合と Abstract の両方に含まれる単語」が最大となる文集合を見つければよい。

そこでまず、本文に含まれる文をそれぞれ $s_1, s_2, \cdots, s_N$ とした時、$s_1, s_2, \cdots, s_i$ の中から合計の長さ（単語数の合計）が j になるように文集合を選んだ時に最も Abstract との単語の積集合の大き

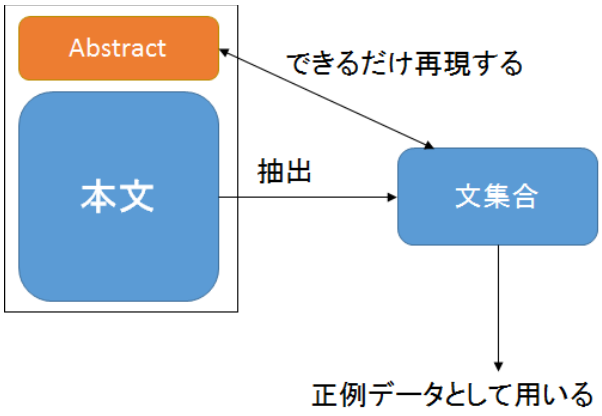

図 3.5. 教師データ生成のためのラベル付けの手法

さが（疑似的に）最大となるものを $Best_{i,j}$ と定義する。

そして各 $Best_{i,j}$ をどのように求めるかに関しては、ナップザック問題を解く際に用いる動的計画法と同様なアルゴリズムを用いる。まず、各 $Best_{i,j}$ を表すテーブルを図 3.6 のように表す（ここで、$Best_{i,j}$ は図 3.6 の中では $Best(i,j)$ と記している）。

そして、この各 $Best_{i,j}$ を、i の小さい方から順に求めていく。i の小さい方、つまりより小さな集合の中での疑似最適な文集合を求め、その結果を元により大きな集合の中での疑似最適な文集合を求める、といったことを繰り返す（図 3.7）。

そして全ての $0 \leq i < M, 0 \leq j \leq W$ に対して $Best_{i,j}$ が求まっている時、それを利用して $Best_{M,k}$ はを求める。まず $s_1, s_2, \cdots, s_M$ の中から合計の単語数が k となる疑似最適な文集合を選ぶ場合、文 s_M を含む場合と含まない場合の 2 通りが考えられるので、それぞれについて考える。

文 s_M を含まない場合

$s_1, s_2, \cdots, s_{M-1}$ の中から合計の単語数が k となる疑似最適な文集合が求めるものとなる。

文 s_M を含まない場合

$s_1, s_2, \cdots, s_{M-1}$ の中から合計の単語数が文 s_M を足して k となる、つまり $k - |s_M|$ となる最適な文集合に s_M を加えたものが求めるものとなる。

よって、この 2 つのうち Abs との積集合の大きさが大きい方が $Best_{M,k}$ となる。

これを図で表すと図 3.8 のようになる。

具体的には、Algorithm3.1 に記したアルゴリズムで疑似最適な文集合を求める。

このようにして各 $Best_{i,j}$ を求めると、それぞれの文集合を要約に選んだ時の ROUGE-1 F1 スコアは式 3.1 のようになるので、その中で最も ROUGE-1 F1 スコアが高くなる文集合 $BestSum$ は式

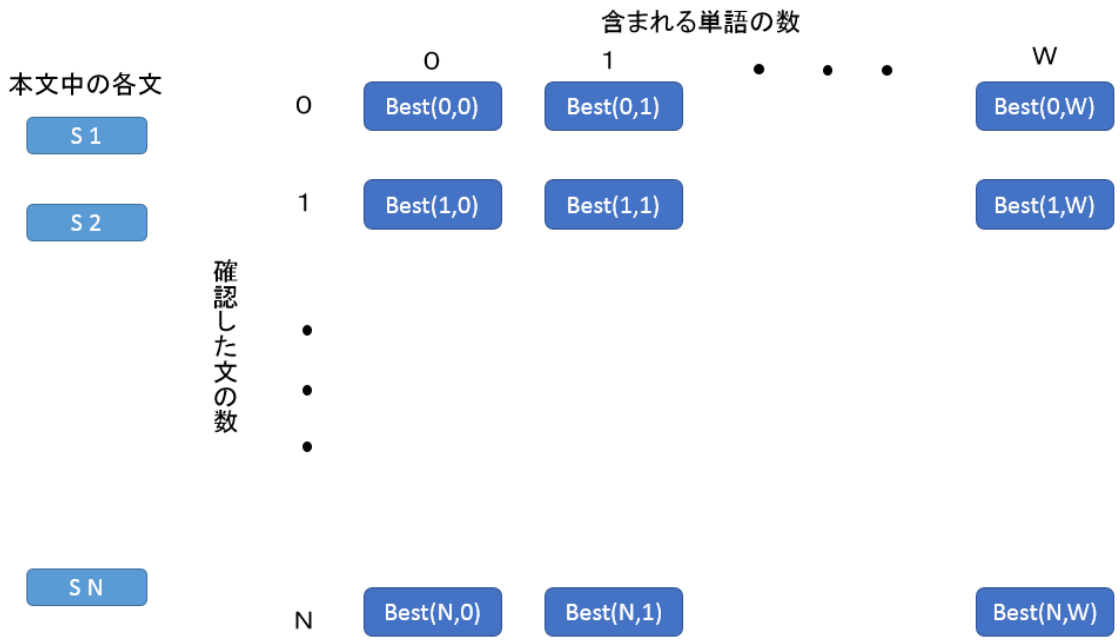

図 3.6. 各 $Best_{i,j}$ を求めるためのテーブル

3.2 のようになる。

$$BestSum = \underset{Best_{i,j}(1\leq i\leq n, 1\leq j\leq m)}{\arg\max} \frac{2 * |Abs \cap Best_{i,j}|}{|Best_{i,j}| + |Abs|} \tag{3.2}$$

この $BestSum$ を、本手法における「正解データとなる文集合」として用いた。

3.3.1　提案手法において取り出した文集合の最適性に関して

この問題は、集合被覆問題の一種である。集合被覆問題とは、ある集合 U およびその部分集合の族 $S_1, S_2, \cdots, S_n$ が与えられた時、U の要素を全て含むように $S_1, S_2, \cdots, S_n$ から最小個数の部分問題を選ぶという問題である。

ここで、今回の問題においては、集合 U は Abstract に含まれる単語の集合 Abs のことで、各部分集合 $S_1, S_2, \cdots, S_n$ は本文中の各文に含まれる、Abs にも含まれる単語の集合である。

集合被覆問題は NP 困難であることが知られており [36]、多項式時間で最適解を求めるアルゴリズムは現在のところ見つかっていない。

本研究で提案した手法におけるアルゴリズムでも、もしある地点までの最適解が全体でも最適となっていれば動的計画法を用いることができて全体での最適解も求められる。ただし今回の場合は、必ずしもある地点までの最適解が全体での最適につながるとは限らない。

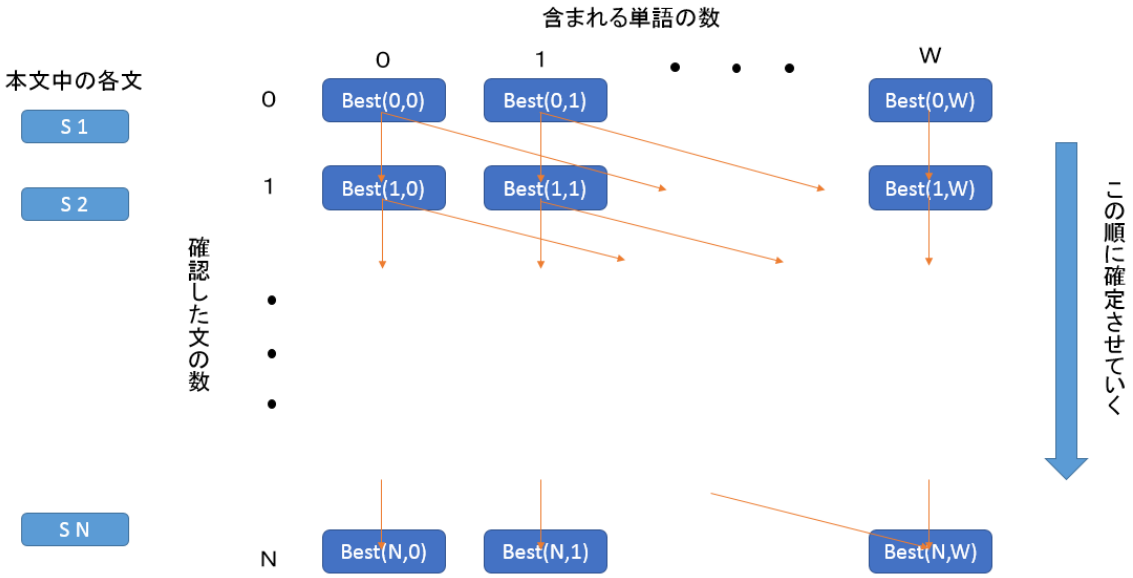

図 3.7. 全ての $Best_{i,j}$ を求めるため処理の流れ

例えば、図 3.9 に記した 1 文からなる Abstract と 3 文からなる body の極めて簡単な例に関して考えてみる。(ここで、各青丸は一つの単語を表し、同じアルファベットのものは同じ単語を、違うアルファベットのものは違う単語を指すものとする)

この時の $Best_{2,4}$、つまり S_1, S_2の中から選んだ文集合のうちで単語数が 4 となるものの中でAbsとの積集合の大きさが最大となるものは、S_2ではなく S_1を選んだ場合である。

一方、$Best_{3,7}$、つまり全ての文から選んだ文集合のうちで単語数が 7 となるものの中で Abs との積集合の大きさが最大となるものは、S_2と S_3を選んだ場合である (図 3.11)。しかし、説明したアルゴリズムだと$Best_{3,7}$ は $Best_{2,4}$ に S_3を加えたものとなるため、$Best_{3,7}$ には最適な文集合が入らなくなってしまう。そのため、各 $Best_{M,k}$ を「疑似最適な文集合」と称した。

そのため、これよりも ROUGE スコアが高くなる文集合が生成される可能性がある。そこで、各 $Best_{M,k}$ を文集合一つではなく、その地点での Abs との積集合の大きさ上位 B 個を保持するビームサーチにした場合に関しても実験を行った。

そして、用意したデータの中の一部の約 400 論文に対してビームサーチを用いた場合と精度を比較した。その結果を表 3.3 に記す。ここで、ビーム幅 1 はつまりビームサーチを使っていないということである。

この表より確かにビーム幅を増やすことにより精度が向上していることが分かる。しかし、その変化の幅がわずかであること、また 3.4 に述べるがビーム幅 1 の精度でも十分に優れていることから、本研究ではビームサーチを使わずに（つまりビーム幅 1 で）生成した文集合を正解データとして用いた。

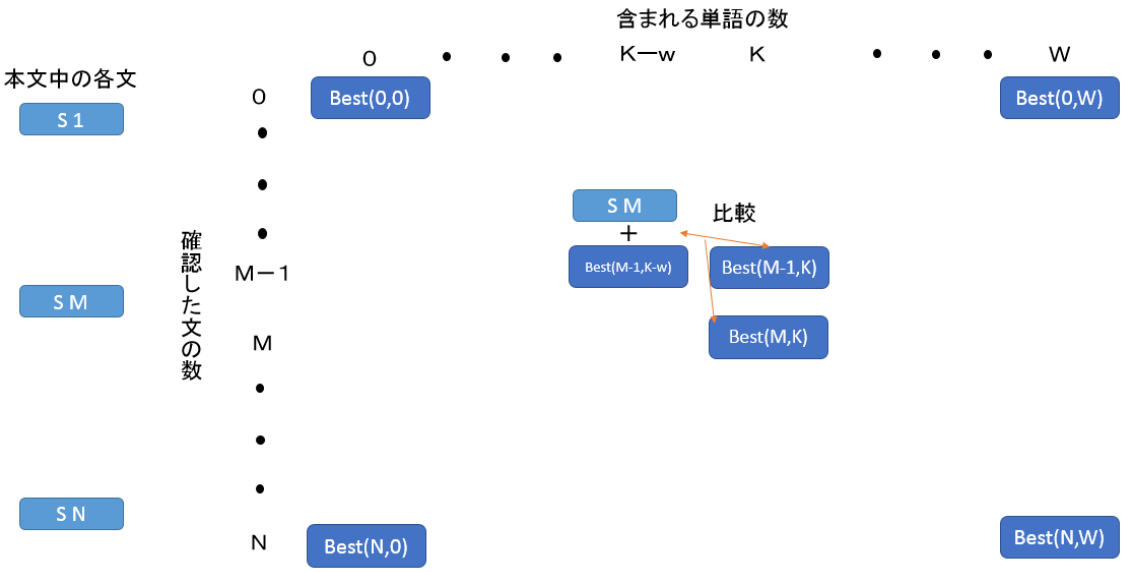

図 3.8. 各 $Best_{i,j}$ を求めるため処理

3.4　正解データとなる文集合の解析

3.3 節で述べた手法を用いて、3.1 節で紹介したコーパスに含まれる論文における正解データとなる文集合を求めた。その際の要約の精度は表 3.4 のようになった。

この値は、今後この文集合を教師データとして学習した時に目標となる値であり、また抽出的手法による要約であってもこの精度までは到達できるという値でもある。2.2.1 節でも述べた通り、抽出的手法による要約ではあまり質の高い要約を生成できないのではないかという問題提起もある。しかし、設定や条件は各々異なるものの、抽出的手法による要約の ROUGE-1 スコアは多くの場合 0.3 や 0.4 といった辺りである [13, 1] ことを考えると、抽出的手法による要約にもまだ改善・精度向上の余地が多いに残されているということが分かる。

表 3.3. ビーム幅の変化による文集合の精度の変化

ビーム幅 B	ROUGE-1 Precision スコア	ROUGE-1 Recall スコア	ROUGE-1 F1 スコア
1	0.660	0.670	0.663
3	0.660	0.673	0.664
5	0.660	0.674	0.665
7	0.662	0.673	0.665
30	0.664	0.679	0.669

Algorithm 3.1 正解データとなる文集合を生成するアルゴリズム

Ensure: $Best_{i,j}(1 \leq i \leq N, 1 \leq j \leq W)$ に $s_1, s_2, \cdots, s_M$ の中から合計の単語数が k となる疑似最適な文集合を入れる
  $S = s_1, s_2, \cdots, s_N$: 本文中の各文
  $Best_{i,j}(1 \leq i \leq N, 1 \leq j \leq W) \Leftarrow \emptyset$
  for $i = 1\ to\ N$ **do**
    for $j = 0\ to\ W$ **do**
      $Best_{i,j} \Leftarrow Best_{i-1,j}$
      if $|s_i| \leq j \& |Abs \cup Best_{i,j}| < |Abs \cup Best_{i-1,j-|s_i|} + s_i|$ **then**
        $Best_{i,j} \Leftarrow Best_{i-1,j-|s_i|} + s_i$
      end if
    end for
  end for

次に、この手法で具体的にどのような要約が生成されたかを見て行く。図 3.12 に記したのはうつ病の男女による性差に関する研究について書かれたある論文 [2] の Abstract および、3.3 節で説明した手法を用いてこの論文の本文から抽出した疑似最適な文集合である。

Abstract では、背景としてうつ病には性差があると言われているが、それは自分の状態・不満を報告する人に性差があるためにそう観測されているに過ぎないのではないかといった問題提起がまずなされている。次に使用データ、用いた手法に関して説明し、最後に測定結果から実際にうつ病には性差があったという結論を述べている。

一方、提案手法で取り出した文集合では、まずこの論文ではうつ病の性差に関する比較を行うという旨の文があり、次に別要因により統計の結果が変わってしまっているのではないかという問題定義の文がある。そして用いたデータや手法に関する文があり、測定結果と、そこからうつ病に性差は存在したと結論付ける文がある。Abstract には含まれていなかった今後の研究課題に関する文も存在するが、そういった点も含めてこの論文の要約と称して差し支えない内容となっている。

また、別の論文 [3] に関しても Abstract と取り出した文集合に関しても見てみる（図 3.13）。

この論文においても、Abstract で説明されているタンパク質の性質に関する考察と、それに基づいた新たな分類・名前付けといったこの論文の内容が、取り出した文集合においても表れていることが分かる。

これらの例より、実際に提案手法で取り出した文集合を「正解」として教師データに用いることは十分に妥当であると言える。

表 3.4. 正解データとなる文集合の精度

ROUGE-1 Precision スコア	ROUGE-1 Recall スコア	ROUGE-1 F1 スコア
0.647	0.658	0.650

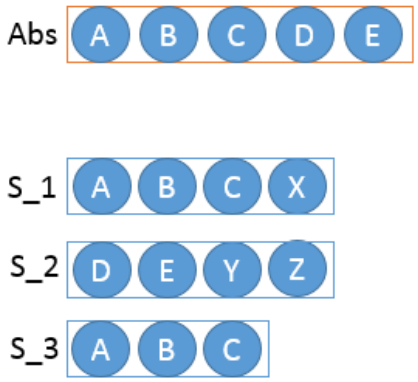

図 3.9. ある地点までの最適解が全体でも最適となっていない例 (1)

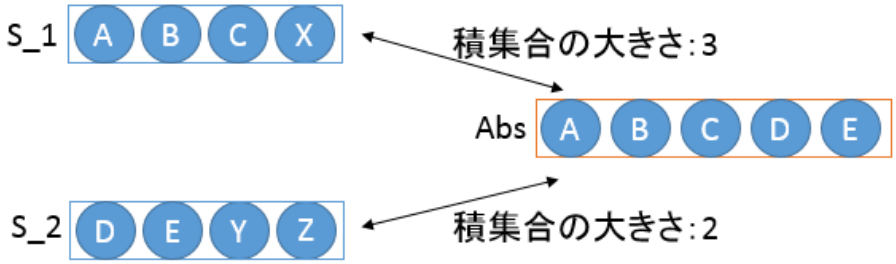

図 3.10. ある地点までの最適解が全体でも最適となっていない例 (2)

また、BioMed コーパス中の論文における主な 10 のセクション名に関して、各セクションからどれだけ最適な文集合に文が選ばれたかといった点について調べた。その結果は表 3.5 のようになった。

ここから、Conclusion のセクションが最も選ばれる割合が高いことが判明した。その結論を述べていることからその論文全体をまとめた内容が多く含まれているためと思われる。一方 Methods のセクションが最も選ばれる割合が低くなっていた。用いた手法に関する詳細な、具体的な説明はあまり要約として適していないためと思われる。

そして、各セクションの先頭に位置する文、およびセクションを文の数で 4 分割した時にそれぞれの場所に属する文の中で、どれだけの文が最適な文集合に文が選ばれたかといった点に関しては表 3.6 のようになった。

ここから、セクションの先頭の文は他の文と比べて正解データの文集合に選ばれる割合が高いことが判明した。

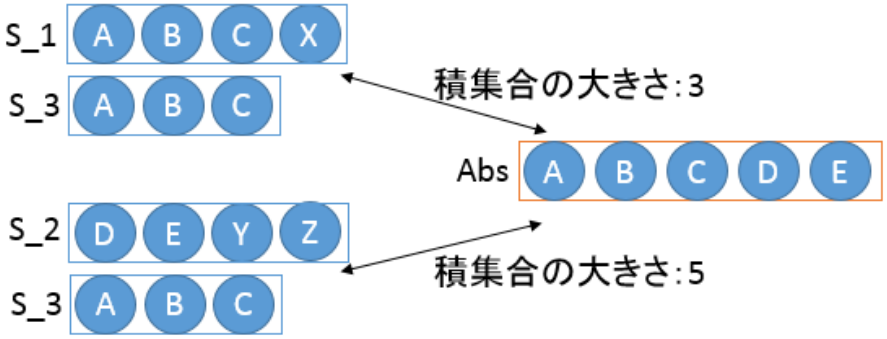

図 3.11. ある地点までの最適解が全体でも最適となっていない例 (3)

表 3.5. 各セクション名ごとの含まれる文が正解データの文集合に選ばれた割合

セクション名	全ての文の数	正解データの文集合に選ばれた文の数	選ばれた割合
Results	547353	30601	5.59%
Methods	509289	13212	2.59%
Background	280097	35271	12.59%
Discussion	273094	19986	7.32%
Results and discussion	268802	13051	4.86%
Materials and methods	86099	2626	3.05%
Conclusion	69269	14716	**21.24%**
Authors' contributions	41864	1059	2.53%
Introduction	28030	2758	9.84%
Stastistical analysis	7020	320	4.56%

表 3.6. セクション中の文の位置ごとの含まれる文が正解データの文集合に選ばれた割合

セクション名	全ての文の数	正解データの文集合に選ばれた文の数	選ばれた割合
セクションの先頭	146025	15676	**10.74%**
セクションの最初の 4 分の 1	632573	46851	7.41%
セクションの 2 つ目の 4 分の 1	557086	29815	5.35%
セクションの 3 つ目の 4 分の 1	592243	31423	5.31%
セクションの最後の 4 分の 1	521536	35910	6.89%

International research consistently finds gender differences in depression, but do women genuinely experience more complaints or are the findings contaminated by group-specific elements unrelated to depression but affecting its measurement? The study of gender differences in depression depends on the measurement quality of the instrument used to evaluate depression.
In the present study we test the measurement equivalence of a shorter version of a commonly used instrument in mental health research, the Center for Epidemiologic Studies - Depression Scale (CES-D), using data from the Belgian sample of the third round of the European Social Survey (N = 1794).
Evidence for measurement invariance can be established within the multigroup confirmatory factor analysis framework.
This method allows us to evaluate a nested hierarchy of hypotheses to test different levels of cross-group measurement invariance: configural, metric, scalar and residual invariance, and clarifies under what conditions meaningful comparisons between the male and female respondents can be made.
The best fitting factor model is then used to estimate the 'true' prevalence of depressive symptoms for both groups.
In our study measurement equivalence is established at all levels, indicating that the current depression scale allows defensible quantitative gender comparisons.
Our data also confirm the epidemiological finding that women report more complaints of depression than men.

In the present study, we aim to evaluate the measurement invariance of a depression scale as a tool for making cross-gender comparisons.
Our estimates of gender differences in depression in the general Belgian population will therefore reflect true differences between men and women, rather than being contaminated by possible group-specific attributes unrelated to depression.
In the present study, we made use of the Belgian sample of the third round of the European Social Survey (ESS 3) 16, organised in 2006 and 2007.
Depression is assessed by the Center for Epidemiologic Studies - Depression Scale or CES-D 17.
The latter is a measurement equivalence approximation that can be tested with confirmatory factor analysis (CFA), a special case of structural equation modelling.
However, configural invariance is not sufficient to defend quantitative group comparisons.
This model is fitted to male and female data via multigroup analysis using Maximum Likelihood estimations.
In the present study we established factorial invariance at all levels: dimensional, configural, metric, scalar and residual invariance.
Our study based on the ESS 3 data of the general population in Belgium confirms the consistent epidemiological finding that women report more complaints of depression than men.
Further research is needed to determine the extent to which these factors influence responses to self-report instruments.
Measurement equivalence tests show that the scale allows defensible cross-gender comparisons leading to prevalence estimates that are not contaminated by group-specific elements unrelated to depression.

図 3.12. ある論文 [2] の Abstract（上）および提案手法で取り出した文集合（下）

Shroom is a recently-described regulator of cell shape changes in the developing nervous system.
This protein is a member of a small family of related proteins that are defined by sequence similarity and in most cases by some link to the actin cytoskeleton.
At present these proteins are named Shroom, APX, APXL, and KIAA1202.
In light of the growing interest in this family of proteins, we propose here a new standard nomenclature.

We now know that the Apx protein of Xenopus is not in fact the orthologue of human APXL.
Instead, we propose that the new nomenclature be based upon the name 'Shroom' as this is now the most thoroughly studied member of the family.
Several papers suggest that these related proteins play diverse and important roles in the development of the nervous system and other tissues.
Future studies will be required to show if sequence similarity among Shroom protein family members is mirrored by conservation of their cellular and molecular function.

図 3.13. ある論文 [3] の Abstract（上）および提案手法で取り出した文集合（下）

第4章　教師付き学習による要約生成の実験

本章では、3 章で生成した学習データをもとに要約を生成する手法を説明し、その手法を用いた自動要約の実験の結果を述べる。

4.1　要約の実験設定

本手法で要約を行う手順に関して、以下に改めて記す。

まず、学習データ中の各論文から、3.3 章で述べた手法を用いて最適な文集合を取り出す。次に、各論文における本文中の文の中で、その最適な文集合に選ばれたものを正例、それ以外の文全てを負例として扱い、2.1.1.1 節で説明した平均化パーセプトロンを用いて重みベクトルの学習を行う。

ただし、今回対象とする課題においては、負例が正例に対して極めて多くなっている（最適な文集合として選ばれた文よりも、選ばれなかった文の方が極めて多いため)。よって、そのままパーセプトロン学習を行った場合、多くの特徴量は（たとえその特徴量を持っている場合は正例になりやすかったとしても）その特徴量を持つ例のほとんどが負例となるために、重みが負に学習されてしまう。

そこでそういった問題点を解決するために、今回は正例を増やして学習を行った。学習データに存在する正例それぞれに関して全く同じ学習例を 5 例ずつ追加した (つまり、正例が元の 6 倍に増えていることとなる、図 4.1)。

要約を生成する際は、テストデータ中の各論文における本文中の文それぞれから特徴量を抽出し、重みベクトルとの内積の大きい方から上位 6%(3.2 節で調べた Abstract と本文の長さの比より) の文を取り出し、それらの文を本文に現れる順番に並び替えて要約とする。

各文の特徴量に関しては、以下のものを用いた

1. 文が属するセクションの名前

bioMed コーパスに含まれる論文の中でよく表れる、表 4.1 に記した 10 個の名前のセクションに含まれる文に関しては、そのセクション名を特徴量として用いた。

2. 文の属するセクションの中での位置

セクション中の文の位置を特徴量として用いる。具体的には、まずセクション中の先頭であるかどうか、末尾であるかどうかを特徴量として用いた。そして、セクションの中を文の数で四分割し、そのどの位置に含まれるかも特徴量として用いた。

3. 文に含まれる単語

文の中に含まれる単語 1 つ 1 つをそれぞれ特徴量として用いた。

4. 文に含まれる動詞

文の中でも含まれる動詞は、特にその文の役割・意図をよく表していると考えられる。例えば、show という動詞が含まれていたら、その文は結果を示す重要な文ではないかと考えられる。そのため、Parse した結果動詞と判定された単語に関しては、上で述べた特徴量にさらに加えて用いた。

4.2 比較手法

また、本研究で提案手法との比較のために用いた手法に関しても説明を行う。単純なルールベースの手法と、教師データをもとにした学習を行わない先行研究の手法の 2 つを

4.2.1 LEAD 法を元にしたルールベース手法

文章要約におけるルールベースの手法として、LEAD 法と呼ばれるものが存在する。LEAD 法とは、文章の先頭から指定された長さだけ抜き出して要約にする（図 4.2）という手法である。これは、多くの文章で冒頭に結論やその文章の大枠が記されているために用いられた手法である。

LEAD 法は極めて単純な手法であるが、単一文章要約ではそれなりに強力なベースライン手法であり、最近の研究でも比較手法として用いられている [13]。

しかし、学術論文で文章の先頭から取り出した場合には、多くの場合 Introduction や Background の章から取られるだけになってしまい、あまり質の高い要約にはならない。そこで本研究では、各セクションの先頭から文を 1 文ずつ取り出して、それを並べることで要約を作成するという手法（図 4.3）を比較のためのベースライン手法として用いる。

各セクションはそれぞれで一つの文章として完結しており、一般の文章と同様に先頭の文にそのセクションの大枠が書かれていることが多い。実際に正解データの文集合を見ても、セクションの先頭の文は他の文と比べて正解データの文集合に選ばれやすいことが分かる (表 3.5)。そのため、こういったルールベースで、提案手法と比較するには十分なベースライン手法になり得ると考えられる。

4.2.2 学習を用いない先行研究の手法

もう一つの比較手法として、教師無しで要約を生成する Liu らの手法 [37] を用いた。

Liu らの手法では、式 4.1 を最大化するような文集合を取り出し要約を生成する。

$$f(V) = \sum_{i \in V \setminus S} \sum_{j \in S} \omega_{i,j} - \lambda \sum_{i,j \in S, i \neq j} \omega_{i,j} \tag{4.1}$$

ここで、V はその文章に含まれる文集合、S は要約として取り出した文集合、ω_{ij} は文 i と文 j の類似度を表す 0 以上の値、λ は 0 以上の値で重み調整用のパラメータである。

つまり、この手法では文章の中身をよく表すために要約として選ばれなかった文との類似度の合計をできるだけ大きく、そして要約における冗長性をなくすために要約に選ばれた分同士の類似度の合計をできるだけ小さくするような要約を生成している。

式 4.1 が最大となる文集合を効率的に見つけるのは困難であるが、Liu らはスコア関数 4.1 の列モジュラ性を手法を用いて要約を生成した。

本実験においては、文同士の類似度 ω_{ij} に関しては Liu らの論文中に記された TF-IDF スコアによるコサイン類似度を、パラメータに関しても同様に Liu らの論文中に記された $r = 0.3, \lambda = 4$ を用いた。

4.3 実験結果

4.3.1 学習データ量と要約精度の変化

まずは、各特徴量学習データを増やした場合の精度の変化に関しての実験を行う。学習のイテレーション回数を 5 回に固定して、教師データとして用いる論文の数を変化させたときの、開発データにおける要約精度の変化を図 4.4 に記す。

図 4.4 より、学習を行ったことで比較手法よりも精度が向上したこと、および教師データの量を増やすことで精度の向上が見込めることが分かる。

4.3.2 各特徴量の有用性

次に、各特徴量の有用性を検証するための実験を行う。特徴量ごとに

- その特徴量を使用せずに学習した場合の開発データにおける要約精度
- 全ての特徴量を使用して学習した場合の開発データにおける要約精度

の二つを比較し、その特徴量の有用性を検証する。

各特徴量を除いた場合の要約精度の変化を表 4.2 に記す。

結果、文中の単語の特徴量を使用しなかった場合は用いた場合より精度がわずかながら上がったが、それ以外の特徴量を使用しなかった場合は全ての特徴量を使用した場合よりも精度が下がるという結果となった。セクションの名前の特徴量を使用しなかった場合が最も精度の下がり幅が大きく、セクションの名前の特徴量が要約生成の学習において今回用いた特徴量の中で最も重要であることが明らかとなった。

4.3.3 重要な単語

本節では、学習データで学習させた結果、文中の動詞の特徴量で重みが大きかったものはどういったものであったか、つまりどういう動詞を含む文が要約に選ばれやすいかについて検討を行う。

重みベクトルの重み上位 10 個の単語を表 4.3 に記す。

この中で、guess や progress といった単語は、仮説に関する考察や貢献に関する説明といった、論文の主張の部分によく表れているために重みが大きくなったと考えられる。delay,aid,inhaled,dredge や classify といった単語は、実験に関する説明で重要であるため重みが大きくなったと考えられる。特に、aid や inhaled(吸い込む)、dredge(浚渫する) といった単語は、使用したコーパスが生命・医学系であるための特有の単語であると思われる。

4.3.4 テストデータにおける結果

そして、テストデータにおける実験結果は表 4.4 のようになった。

これより、テストデータにおいても提案手法が比較手法を上回る精度を達成していることが明らかとなった。

4.3.5 実際に生成された要約

本節では、実際に生成された要約を見ていく。まずはショウジョウバエの X 染色体に関する論文 [4] の Abstract と提案手法で生成した要約を並べたものを図 4.5 に記す。

Abstract を見ると、ショウジョウバエの X 染色体が男性性徴の消失を示しているという背景が書かれている。次に、当論文では Retrogene[1] を解析することでそれを検証するということが書かれている。その次に検証の結果今までの研究と同様のことが示せたと結論が書かれており、その後の文で実験・検証の中身が記されている。

一方生成した要約では、最初の文でショウジョウバエの Retrogene が興味深い挙動を示すという点が書かれている。次にショウジョウバエと X 染色体、男性性徴に関する説明・考察の文が続いている。そして最後から 2 番目の文では今までの研究と同様のことが示せたということが書かれており、最後の文ではその結果を元にした考察が書かれている。

また、ほかの例として、遺伝子と文化に関する論文 [5] の Abstract と提案手法で生成した要約を並べたものを図 4.6 に記す。

Abstract では、まず遺伝子と文化にの間には関係があると思われてきたもののそれを調査する手段がなかった、といった背景が書かれている。次にその中で当論文で調査する仮説の具体例と、それを検証するために調べた遺伝子の名前を述べている。そして遺伝子の調査の結果が書かれて、最後に仮説が示されたことが書かれている。

一方生成した要約では、背景に関しては触れられていないが、最初の 2 文で検証する仮説の具体例を述べている。3 文目は実験結果とそれによってもたらされる結論、4 文目は結論からの文かに関

[1] 生物学の用語で、RNA で逆転写された DNA のこと

する考察、5 文目には論文で用いた手法が書かれ、最後の文には結論が書かれている。重複している部分もあるが、この要約も十分に読むことで論文の内容を理解できる要約であると言える。

4.4 考察

学習の結果、各セクションの先頭を集めるという比較手法を上回ることはできたので、最低限の学習は行えていたことが分かる。しかし、正解データとして用いた文集合と比べると大きく劣っており、まだまだ制度改善の余地が大いにあることが分かる。

4.3.2 節の実験結果を見ると、主に有効に働いた特徴量は文の属するセクションの名前やその文のセクション中の位置といった特徴量で、文中の単語に関する特徴量はあまり有効に働いていないことが分かる。そのため、そういった文の中身に関してより要約として選ばれるかが決まる際に重要な特徴量を利用することが必要と考えられる。

例えば、細胞周期の制御に関するある論文 [38] における要約生成において、教師データでは正例だったにも関わらず要約に選ばなかった文として以下のようなものが存在した。

> These results indicate that Dap expression contributes the G1 cell cycle arrest of Su(H) mutant cells.
>
> These observations suggest that Dap accumulation is unlikely to contribute to the G1 arrest of notch mutant cells and that the G1 cell cycle arrest of notch and Su(H) mutant cells in the SMW probably involve distinct mechanisms.

こういった “These results indicate that” や “These observations suggest that” は明らかに要約として選ばれるべき文の特徴を示していると思われるが、実際にそれぞれの単語の重みはあまり大きくは学習されていない。よって、単純に単語一つ一つを特徴として用いるのではなく、節や主語・動詞の組み合わせといったように文の構造を元に特徴を取り出すことが必要であると考えられる。

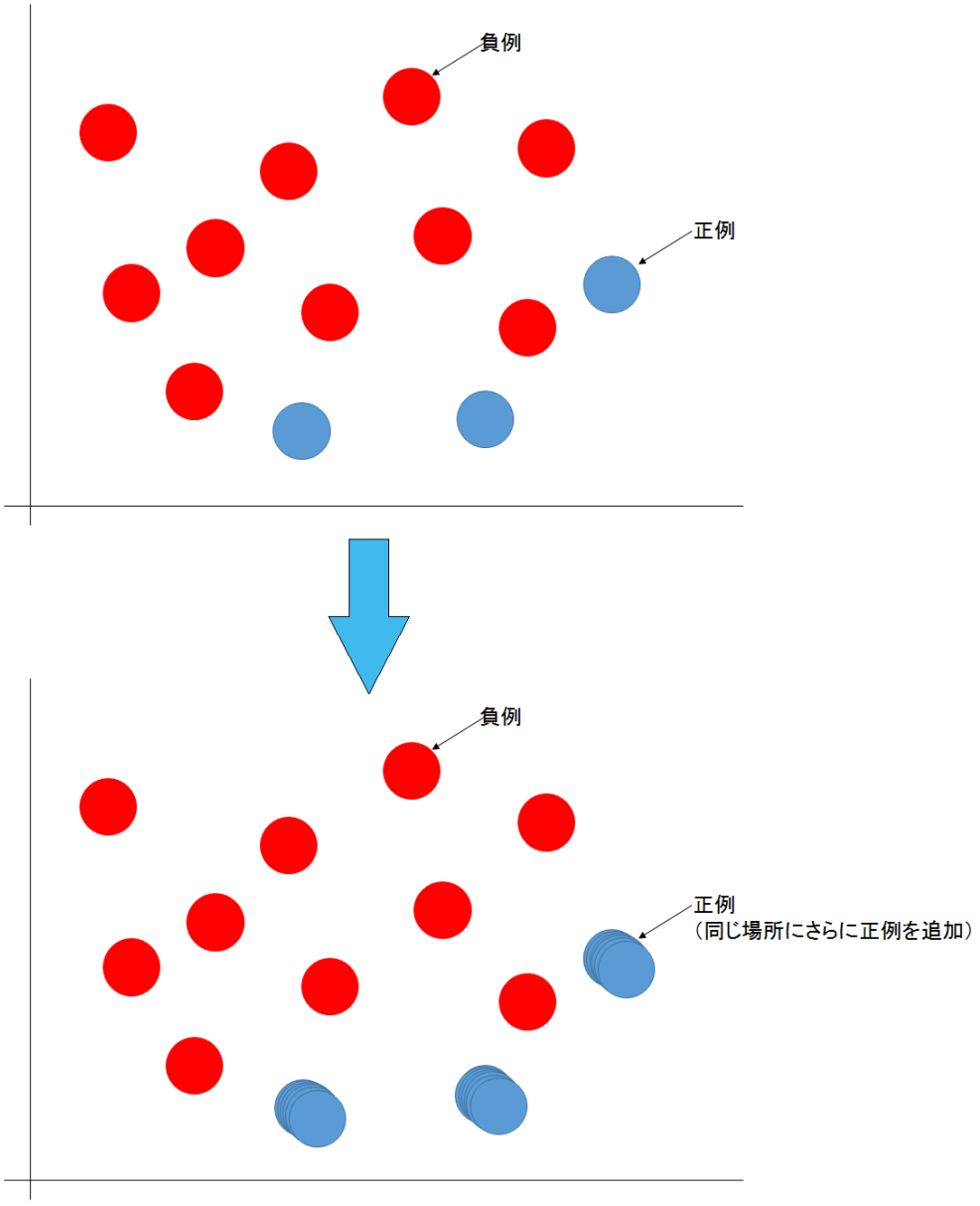

図 4.1. 正しく学習を行うための教師データの改善

表 4.1. 特徴量として用いたセクション名

Results
Discussion
Methods
Materials and methods
Introduction
Background
Results and discussion
Authors’ contributions
Stastistical analysis
Material and methods

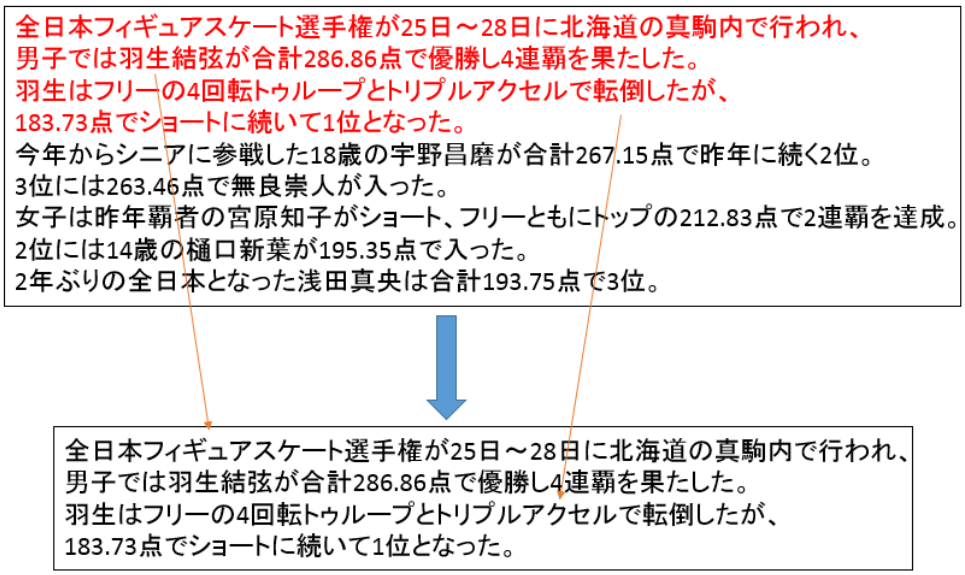

図 4.2. LEAD 法による要約の例

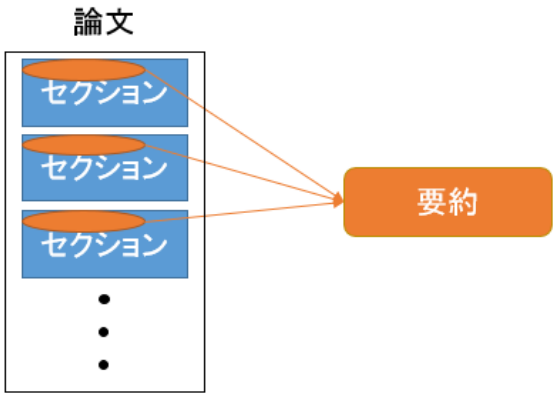

図 4.3. 本研究におけるベースライン手法の説明

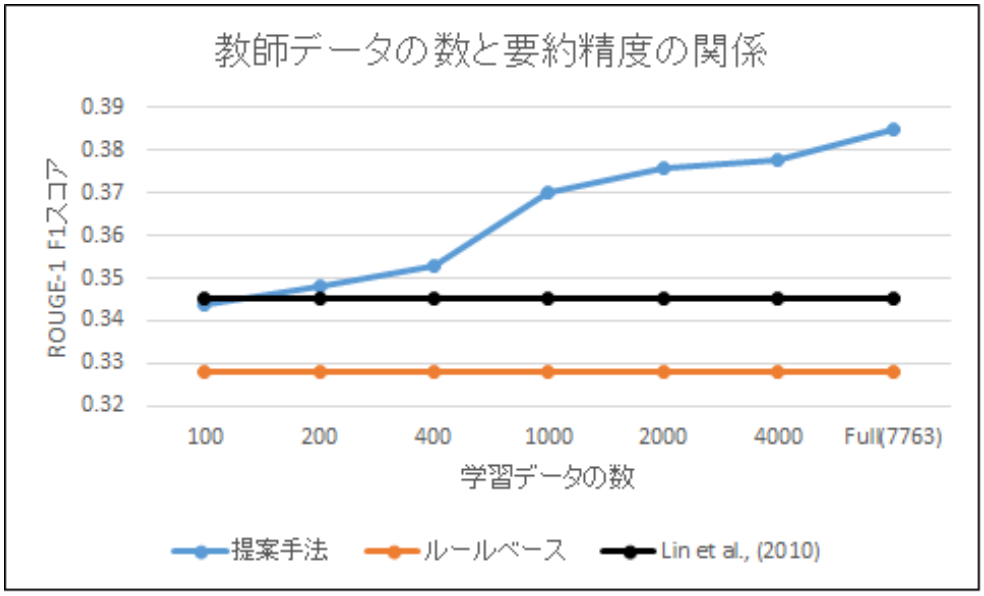

図 4.4. 学習データの量による要約精度の変化

表 4.2. 各特徴量を除いた場合の要約精度の変化

使用しなかった特徴量	ROUGE-1 F1 スコア
なし（全ての特徴量を使用）	0.3849
セクションの名前	**0.366**
セクション中の位置	0.378
文中の単語	0.3852
文中の動詞	0.382

表 4.3. 重みが大きく学習された動詞

guess
delay
aid
progress
inhaled
dredge
motorize
classify
orientate
walk

表 4.4. テストデータにおける各手法の要約精度

セクション名	ROUGE-1 Precision	ROUGE-1 Recall	ROUGE-1 F1score
提案手法	0.358	0.429	**0.378**
ルールベース	0.299	0.402	0.329
Lin et al., (2010)	0.320	0.376	0.337

The Drosophila X-chromosome shows a significant underrepresentation of genes with male-biased gene expression (demasculinization).
This trend is matched by retrogenes, which typically have a male biased gene expression pattern and show a significant movement bias from X-chromosomes to autosomes.
It is currently assumed that these patterns are best explained by selection, either mediated by male meiotic sex chromosome inactivation (MSCI) or sexually antagonistic forces.
We scrutinized the evolutionary dynamics of retroposition by focusing on retrogenes for which the parental copy has degenerated.
Consistent with a functional substitution of the degenerated gene by the retrogene, patterns of sequence evolution and gene expression were similar between retroposed and parental genes.
Like previous studies, our set of retrogenes showed a significant movement off the X-chromosome.
In contrast to data sets where retroposition caused gene duplication, the genes in our study showed primarily female-biased or unbiased gene expression.
Based on our results, the biased transposition pattern cannot be explained by MSCI and probably not by sexual antagonism.
Rather, we propose that the movement away from the X-chromosome represents a general property of retroposition in Drosophila.

In Drosophila retrogenes show interesting dynamics.
In combination with the preferential male biased gene expression of retroposed genes, their non-random integration pattern has been used to explain the "demasculinization of the X-chromosomes".
It has been suggested that male meiotic sex chromosome inactivation (MSCI) or sexual antagonism may be the evolutionary force driving the movement of retroposed genes, which ultimately leads to contrasting patterns in gene expression between X-chromosome and autosomes.
Taken together, the results of gene conservation, analysis of molecular evolution and gene expression suggest that transposition did not result in a functional alteration.
The preferential male-biased gene expression of retrogenes in combination with the under-representation of male-biased genes on the X-chromosome has been widely viewed as strong support for selection driving the pronounced movement bias of retrogenes from the X-chromosome to the autosomes.
While there is little doubt about the existence of the X chromosome inactivation during spermatogenesis, it is also clear that many genes with an expression in testis persist on the X-chromosome.
Consistent with previous results, our data suggests an excess of retrotransposition events out of X-chromosome in Drosophila, but do not show a male-biased or testis specific expression.
These results indicate that the biased transposition pattern cannot be due to MSCI or sexual antagonism, rather the pattern is a general property of retrotransposition in Drosophila.

図 4.5. ある論文 [4] の Abstract(上) と提案手法で生成した要約 (下)

Genes and culture are believed to interact, but it has been difficult to find direct evidence for the process.
One candidate example that has been put forward is lactase persistence in adulthood, i.e. the ability to continue digesting the milk sugar lactose after childhood, facilitating the consumption of raw milk.
This genetic trait is believed to have evolved within a short time period and to be related with the emergence of sedentary agriculture.
Here we investigate the frequency of an allele (-13910*T) associated with lactase persistence in a Neolithic Scandinavian population.
From the 14 individuals originally examined, 10 yielded reliable results.
We find that the T allele frequency was very low (5%) in this Middle Neolithic hunter-gatherer population, and that the frequency is dramatically different from the extant Swedish population (74%).
We conclude that this difference in frequency could not have arisen by genetic drift and is either due to selection or, more likely, replacement of hunter-gatherer populations by sedentary agriculturalists.

The Pitted Ware Culture (PWC) was a major Neolithic hunter-gatherer population in Northern Europe and was partly contemporaneous with the farming TRB population (TRB after the German word Trichterbecherkultur, i.
Funnel Beaker Culture).
The PWC are thought to have been present in Scandinavia between 5,400-4,300 years before present (BP), which is later than the suggested initiation of selection for the T allele.
In this study, we find that the frequency of the derived allele is low in the PWC (5%) compared to the frequency in the extant Swedish population, and that the change in frequency is incompatible with genetic drift as the sole explanation under a model of population continuity.
Thus, a genetic component interacting with culture, such as the ability to digest milk as an adult, could have been the result of the replacement of the hunter-gatherer population by an agricultural population.
If we decrease the population sizes in the simulation (and thereby increase the effect of genetic drift) we can determine how robust the result is to small population sizes.
Thus, it is unlikely that genetic drift caused the observed frequency in the PWC given the frequency in the extant Swedish population.

図 4.6. ある論文 [5] の Abstract(上) と提案手法で生成した要約 (下)

第5章　おわりに

5.1　本研究のまとめ

本研究では、Abstract を教師データとして用いた学術論文の要約生成に取り組んだ。それによって、学術論文の自動要約生成における精度改善にきよするとともに、教師データが自然に存在する課題で実験を行うことで、文章要約という課題に対する方策や改善の余地といった点に関する考察を行うことを目的とした。

まずは、Abstract との類似度をできるだけ大きくするように本から文の集合を取り出す手法を提案し、実験を行った。提案手法を用いて取り出した文集合は十分にその論文の要約とみなせることが明らかとなった。また数値的な評価においても、そのような文集合を要約とみなした時の精度は、現在の抽出的手法における要約精度を大きく上回っており、抽出的手法においても精度向上の余地が大いにあることが明らかとなった。また、その文の所属するセクションの名前やセクション中の位置から、どういった文が要約として選ばれやすいか、といった解析が可能となった。

次に、そうして取り出した文集合を正解データとした教師付き学習を用いて要約生成を行った。その結果ベースライン手法を上回る精度を達成した。また、教師データとする論文の数を増やすほどに要約精度が向上しており、一般的な文章自動要約においても教師データを多く用意することで精度の向上が見込めることが明らかとなった。

5.2　今後の展望

今回の要約生成の実験においては、要約に選ばれるかどうかに関して各文を独立に判定を行っていた。しかし、そういった場合は似たような内容に関して触れている文ばかりが抽出されてしまうという可能性も考えられる。そこで、そういった選ばれる分同士の関係性も含めて学習を行う手法を用いるといったことが今後の課題として挙げられる。

また、2.2.3.2 節で述べたように、学術論文の各文の役割を推定するという研究がある。そういった情報を要約に使われる文を推定する特徴量に用いることが今後の課題として挙げられる。またその際、要約の結果を元に、より正解と近い要約が生成されるように各文の役割を推定器の調整も行うという手法も考えられる。そうすることで、より要約を行うにあたって有用な役割の推定ができると期待できる。

また、本研究では使用できるコーパスの都合もあり生命・医学系の論文を実験用データとして用いたが、コンピュータサイエンスの論文など他のドメインの論文を用いることも挙げられる。またその際、他のドメインにおいては図 4.3 のように学習を行った結果上位にくる単語がどのようになるの

か、どういったセクション名の文が選ばれるのかといった点に関して、ドメイン間での違いを比較検討することに関しても価値があると考えられる。

参考文献

[1] Danish Contractor, Yufan Guo, and Anna Korhonen. Using argumentative zones for extractive summarization of scientific articles. In *Proceedings of the 24th International Conference on Computational Linguistics*, Vol. 12 of *COLING '12*, pp. 663–678, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[2] Sarah Van de Velde, Katia Levecque, and Piet Bracke. Measurement equivalence of the ces-d 8 in the general population in belgium: a gender perspective. *Archives of Public Health*, Vol. 67, No. 1, p. 15, 2009.

[3] Olivier Hagens, Andrea Ballabio, Vera Kalscheuer, Jean-Pierre Kraehenbuhl, M Vittoria Schiaffino, Peter Smith, Olivier Staub, Jeff Hildebrand, and John B Wallingford. A new standard nomenclature for proteins related to apx and shroom. *BMC cell biology*, Vol. 7, No. 1, p. 18, 2006.

[4] Muralidhar Metta and Christian Schlötterer. Non-random genomic integration-an intrinsic property of retrogenes in drosophila? *BMC evolutionary biology*, Vol. 10, No. 1, p. 114, 2010.

[5] Helena Malmström, Anna Linderholm, Kerstin Lidén, Jan Storå, Petra Molnar, Gunilla Holmlund, Mattias Jakobsson, and Anders Götherström. High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern europe. *BMC evolutionary biology*, Vol. 10, No. 1, p. 89, 2010.

[6] Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pp. 224–233, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[7] Jingqiang Chen and Hai Zhuge. Summarization of scientific documents by detecting common facts in citations. *Future Generation Computer Systems*, Vol. 32, pp. 246–252, 2014.

[8] Nitin Agarwal, Kiran Gvr, Ravi Shankar Reddy, and Carolyn Penstein Rosé. Towards multi-document summarization of scientific articles: Making interesting comparisons with scisumm. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, WASDGML '11, pp. 8–15, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[9] Michael Collins and Brian Roark. Incremental parsing with the perceptron algorithm. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004.

[10] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA, 2001.

[11] Michael Collins. Discriminative training methods for hidden markov models: theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol. 10 of *EMNLP '02*, pp. 1–8, Stroudsburg, PA, USA, 2002.

[12] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, Stroudsburg, PA, USA, July 2004. Association for Computational Linguistics.

[13] Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*, Vol. 13 of *EMNLP '13*, pp. 1515–1520, Stroudsburg, PA, USA, 2013. Association for Computational Linguistics.

[14] 富田紘平, 高村大也, 奥村学. 重要文抽出と文圧縮を組み合わせた新たな抽出的要約手法 (翻訳・要約・抽出). 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2009, No. 2, pp. 13–20, 2009.

[15] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pp. 379–389, Stroudsburg, PA, USA, September 2015. Association for Computational Linguistics.

[16] Jackie Chi Kit Cheung and Gerald Penn. Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1233–1242, Stroudsburg, PA, USA, 2013. Association for Computational Linguistics.

[17] Aspec, asian scientific paper excerpt corpus. `http://lotus.kuee.kyoto-u.ac.jp/ASPEC/`.

[18] Ntcir-9:patientmt ntcir-project. `http://ntcir.nii.ac.jp/PatentMT/`.

[19] Document understanding conference. `http://duc.nist.gov/`.

[20] Vahed Qazvinian and Dragomir R. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pp. 689–696, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[21] Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, pp. 584–592, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[22] Vahed Qazvinian and Dragomir R. Radev. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pp. 555–564, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[23] 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.*, Vol. 46, No. 1, pp. 165–201, January 2013.

[24] Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pp. 110–117, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[25] Simone Teufel and Marc Moens. Discourse-level argumentation in scientific articles: human and automatic annotation. In *In Proceedings of the ACL-1999 Workshop Towards Standards and Tools for Discourse Tagging*. Citeseer, 1999.

[26] Victor Poznański, John L Beaven, and Pete Whitelock. An efficient generation algorithm for lexicalist mt. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '15, pp. 261–267, Stroudsburg, PA, USA, 1995. Association for Computational Linguistics.

[27] Yufan Guo, Roi Reichart, and Anna Korhonen. Unsupervised declarative knowledge induction for constraint-based learning of information structure in scientific documents. *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 131–143, 2015.

[28] Diarmuid O Séaghdha and Simone Teufel. Unsupervised learning of rhetorical structure with un-topic models. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, COLING '14, pp. 2–13, Stroudsburg, PA, USA, 2014. Association for Computational Linguistics.

[29] Andrea Varga, Daniel Preotiuc-Pietro, and Fabio Ciravegna. Unsupervised document zone identification using probabilistic graphical models. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, LREC '12, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA).

[30] Douwe Kiela, Yufan Guo, Ulla Stenius, and Anna Korhonen. Unsupervised discovery of information structure in biomedical documents. *Bioinformatics (Oxford, England)*, Vol. 31, No. 7, pp. 1084–1092, April 2015.

[31] Yue Hu and Xiaojun Wan. Ppsgen: Learning to generate presentation slides for academic papers. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pp. 2099–2105. AAAI Press, 2013.

[32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pp. 74–81. Association for Computational Linguistics, 2004.

[33] Rouge: Recall-oriented understudy of gisting evaluation. `http://www.berouge.com/Pages/default.aspx`.

[34] Biomed central. `http://www.biomedcentral.com/`.

[35] Enju - a fast, accurate, and deep parser for english. `http://www.nactem.ac.uk/enju/index.ja.html`.

[36] Uriel Feige. A threshold of ln n for approximating set cover. *J. ACM*, Vol. 45, No. 4, pp. 634–652, July 1998.

[37] Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pp. 912–920, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[38] Madina J Sukhanova and Wei Du. Control of cell cycle entry and exiting from the second mitotic wave in the drosophila developing eye. *BMC developmental biology*, Vol. 8, No. 1, p. 7, 2008.

本研究に関する発表文献

- 中須賀謙吾, 鶴岡慶雅. 談話構造を利用した学術論文の自動要約生成. 言語処理学会第 21 回年次大会, 2015 年 3 月.
- 中須賀謙吾, 鶴岡慶雅. Abstract を教師データに用いた学術論文の自動要約生成. NLP 若手の会第 10 回シンポジウム, 2015 年 9 月.

謝辞

本研究を行うにあたって、多くの方々にお世話になりました。

指導教員である鶴岡慶雅准教授には、研究の内容や方向性についてのご指摘、ミーティングでの議論、また発表や論文執筆におけるアドバイスなど、研究に関わる全ての事項にわたって指導して頂きました。

また、博士課程の学生である橋本和真先輩には、技術的な内容にとどまらず、研究に関して様々な助言を頂きました。

その他研究室を同じくした皆さまにも、毎回のミーティングで貴重な意見を下さったり、また普段は雑談したり一緒にゲームをしたりと、研究生活のあらゆる面でお世話になりました。

この場を借りて皆様に厚く御礼申し上げます。