引用文脈の分散表現を利用した学術論文の被引用文章要約の一手法

田邊　俊介† 　太田　学† 　高須　淳宏††

† 岡山大学大学院自然科学研究科　〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
E-mail: †tanabe@de.cs.okayama-u.ac.jp, ohta@cs.okayama-u.ac.jp, ††takasu@nii.ac.jp

あらまし　学術論文では，一般に手法やデータなど研究の根拠が書かれた文献を参考文献として引用する．そのためこれらの参考文献を読むことは論文理解の助けとなるが，多くの参考文献の確認は読み手にとって負担となる．そこで本稿では，引用文に基づき被引用論文中の重要な文を抽出し，その要約を提示する方法を提案する．まず，論文中で参考文献を引用している文章を“引用箇所”として抽出する．そして，doc2vec により文章を分散表現で表し，文章間の類似度に基づいて各文のスコアを算出して重要文を選択する．その重要文を元に生成した被引用文章の要約を評価するために，被引用論文のアブストラクトとの一致度を測り，アンケートを用いた被験者実験を行う．

キーワード　文書要約，分散表現，閲覧支援，引用文脈

1. はじめに

学術論文には多くの場合参考文献があり，これらは先行研究等と関係がある．また著者が行った実験の結果を強調する，主張の根拠を示す，など様々な理由から参考文献は引用される．これらはいずれも閲覧する論文の内容を補完し，読み手が内容をより理解する助けとなる．しかし，1 つの論文に対して百件以上の参考文献が記載される場合もあり，それら全てを閲覧するには膨大な時間がかかる．

そのため論文中の引用に対して，読者の役に立つ情報を含む適切な文章を被引用論文内で特定し提示する手法が提案されている．石井らは [1]，論文本文に含まれる引用を表す文字列を含む文を“引用箇所”，それに対する読み手に提示するのにふさわしい文章を“被引用箇所”とし，まず人手で収集した手がかり語を用いて引用箇所の引用意図を推定した．そして，その引用意図をもとに被引用箇所として適切と考えられる被引用論文中の節を推定し，手がかり語を用いてその節の中の適切な被引用箇所を特定する方法を提案した．しかしこの方法では主に 2 つの問題があった．1 つは被引用箇所の特定のために人手で集めた手がかり語を用いるため，様々な分野の論文への応用が困難だった．もう 1 つは被引用箇所が 1 つの節だけではなく複数の節にまたがって出現する場合に対応できなかった．我々は，引用箇所と被引用論文内の文を doc2vec によりベクトル化し，引用箇所の特徴ベクトルと最も類似した特徴ベクトルを持つ文を被引用箇所と推定する手法を提案した [2]．これは手がかり語を用いないため，1 つ目の問題は改善できたといえる．しかし，2 つ目の問題に対応できていない．そこで本稿では引用文脈の分散表現を利用して学術論文の全文から重要文を選び，被引用文章を要約する手法を提案する．

本稿の構成をは以下の通りである．2 節で本研究に関連する研究を紹介し，3 節で引用文脈を用いた被引用文章の要約方法について説明する．4 節で生成した要約の適切性等に関する評価実験を示し，5 節で本研究についてまとめる．

2. 関連研究

2.1 論文閲覧支援システム

学術論文は現在，電子化し PDF 形式で公開することが一般的である．そのためこれらの閲覧の利便性を向上させることを目的としたシステムとして様々なものが提案されている．

阿辺川ら [5] は Sidenoter と呼ばれる論文閲覧支援システムを提案した．このシステムは，PDF 形式の論文を画像に変換し，Web ブラウザ上での閲覧を可能とするものである．このシステムでは，文書の拡大縮小表示，背景色変更などが可能な閲覧機能，全文検索に加えて，論文のタイトル，著者，会議名などの属性を指定した検索機能，表示している論文の補足情報をページの左右に表示する脚注表示機能を使用することができる．特に脚注表示機能では，Wikipedia をリソースとした見出し語の説明の表示や，TF-IDF（term frequency-inverse document frequency）によって重み付けされたベクトル空間モデルでの類似性を用いた関連論文の列挙など，本文と結び付けられた外部の関連情報を表示できる．

2.2 文書要約

近年，Web 上では SNS やニュースなど，入手可能なテキストデータが爆発的に増えている．そこで，膨大な情報の重要な部分だけを収集することで情報探索コストを下げるという観点から，文書要約の技術が注目されている．これらは学術論文にもあてはまる．様々な文書要約手法が提案されているが，それらは大きく 2 種類に分類される．1 つは要約する文書から重要と思われる文のみを抽出して結合し要約を作成する抽出型という方式であり，もう 1 つは入力文書をもとに，元々含まれていない単語や構文などの表現を用いた新たな文章で要約を作成する生成型という方式である．両方式ともに利点，欠点が存在するが，本研究では抽出型の要約手法を採用する．

Mihalcea ら [6] は，文章をグラフ構造で表しそれを元に文章を要約する TextRank を提案した．この手法はウェブページの重要度を決定するための PageRank アルゴリズムを文章に応用

したものである．PageRank はリンクされているページは良いページであり，多くのページからリンクされているページからのリンクは高い評価を与えるという考えを元に Web ページを評価するというものである．この手法に対し，TextRank ではノードを単語や文，エッジを文の類似度に置き換えてグラフを作成する。そして，新たに定義したアルゴリズムによりエッジの重みを考慮したノードの重要度の算出を行い、文をランク付けする．

Carbonell と Goldstein ら [7] は Maximum Marginal Relevance (MMR) を用いた抽出型要約手法を提案した．重要度のみでランク付けをする手法では同じような文章を選択し，冗長な要約を作成する可能性が高くなる．これは好ましくない．そのため MMR は，すでにスコア付けされた文のリストから類似度の高い文のスコアを下げて再び文の重要度のランク付けをする．この手法により，冗長性が除去でき，多様な文を選択できる．

中須賀ら [8] は，学術論文における談話構造に基づいた特徴量を利用した要約手法を提案した．談話構造とは文章中の文や句の間の関係などを表すものであり，学術論文においても重要な情報であると考えられている．この手法では論文中のアブストラクトを利用している．論文中の各文とその論文の類似度を求めることでそれぞれの文のアブストラクト適応度を定め，線形回帰を用いて各文の特徴量からその文のアブストラクト適応度を推定した．

2.3 doc2vec

doc2vec [3] は，任意の長さの文章をベクトル化する Paragraph Vector の 1 つである．この手法は，文章中の単語を前後関係を元に学習し分散表現を得ることができる word2vec [4] を拡張したものである．3 層のニューラルネットワークを使用し，文中に出現する単語を予測するよう学習することで，その文章の特徴ベクトルを獲得する．学習モデルには，word2vec の CBOW を拡張した PV-DM と Skip-gram を拡張した PV-DBOW の 2 種類がある．doc2vec は，文書の感情極性判定や文書検索の評価実験において，長文，短文に関わらず，単語の語順を考慮しない Bag-of-Words などのそれまでの手法を大幅に上回ることが報告されている．

3. 被引用文章要約手法

3.1 概　　要

被引用文章要約の方法について手順を説明する．初めに論文中の引用箇所を特定し，それをもとに被引用論文中の各文のスコア付けをする．そして，そのスコアにより重要文を収集し，並び替えて被引用文章の要約を作成する．まず，3.2 節で引用箇所の特定方法について述べる．そして，3.3 節で引用箇所をもとにした doc2vec による類似度を用いた文のスコア付けについて述べ，3.4 節で類似度のスコアを用いた被引用文章の要約の作成方法を述べる．

3.2 引用箇所の特定

論文中の引用箇所を特定する方法について説明する．

Adding to the locational aspect, the temporal aspects are also considered in this work. Three recent workshops, NTCIR GeoTime [7], GeoCLEF [15], and GIR [21] addressed the combination of geographic and temporal search. In the most recent NTCIR-8 GeoTime workshop, INESC group from Lisbon, Portugal [14] shows the best performance among the participants. They used geographic resources and TIMEXTAG system [23] to extract geographic expressions from topics and documents.

図 1　引用箇所特定の例（[12] の中の 1 段落を引用）

（1） pdftotext [9] を用い，論文 PDF ファイルをテキスト化する．

（2） ParsCit [10] を用いてテキスト化した論文から本文のみを取り出す．

（3） 引用を示す“ [1] ”のような表現（以下，参考文献マーカ）を手がかりとして引用箇所を特定する．

ParsCit [10] は CRF を用いて論文のテキストから書誌情報を抽出する．このプログラムのオプションである，テキストに論文構成要素タグを付与する extract_section コマンドを使用し，bodyText タグが付与された部分を論文本文とする．この論文本文には論文題目や図表，参考文献などは含まれない．

また，本研究では O’Conner の手法 [11] を参考に，引用箇所を特定する．まず，参考文献マーカを R，R を含む 1 文を S，S_{-1} を S の直前の 1 文，S_1 を S の直後の 1 文とし，次の手順で引用箇所を選択する．

（1） S を選択する．

（2） S の中に複数の R が存在する場合，文中の句点や接続詞で区切り，それぞれを新たに S として選択し手順を適用する．そして，各 S に 1 つの R のみが存在するようにする．

（3） S の先頭 12 語に接続を表す単語（this, these, those, such, same, similar, similarly, they, their, former, latter）（以下，コネクタ）があれば S_{-1} も選択する．

（4） S_{-1} の先頭 7 語にコネクタがあれば S_{-2} も選択する．

（5） S_{-2} の先頭 4 語にコネクタがあれば S_{-3} も選択する．

（6） S_1 の先頭 7 語にコネクタがあれば S_1 も選択する．

（7） S_1 が選択され，S_2 の先頭 4 語にコネクタがあれば S_2 も選択する．

（8） S_{-1} と S が同じ段落にあり S_{-1} に 14 語未満ならば S_{-1} も選択する．

（9） S と S_1 が同じ段落にあり S_1 に 14 語未満ならば S_1 も選択する．

（2）の手順は O’Conner の手法の手順に本稿で新たに加えた手順である．本稿では，各引用箇所に R は 1 つのみ存在させる．よって，上記の引用箇所特定の手順で選択する条件を満たしている S_{-1} や S_1 があったとしても，他の R を含む文ならばその文は選択しない．また，“ [1, 2, 3] ”の様にまったく同じ場所に複数の R が出現する場合は，それ以降にそれぞれの参考文献について詳しく述べている箇所が存在すると考えられ

るため，本実験では利用しない．引用箇所特定の例を図 1 に示す．異なる色の下線は異なる引用箇所を示す．新たに加えた手順（２）により，同一文中にある“ [7] ”，“ [15] ”，“ [21] ”がそれぞれ別々の引用箇所となっている．

3.3 doc2vec による文のスコアリング

doc2vec [3] により生成された特徴ベクトルの類似度を算出することで，文の類似度を測ることができる．そこで，この方法を用いて引用箇所の文章と被引用論文中の各文の類似度を測定し，それを文の重要度のスコアとする．本研究では，このスコアが高いほど引用箇所の説明として重要であるとみなし要約に利用する．

3.4 被引用文章の要約

文章は長ければ重要な文章が含まれる可能性も高くなる．しかし，本稿では読者の負担を考え，より簡潔な文章の作成を目的としている．実験では，被験者に提示する要約として，100 単語の要約と 250 単語の要約を提示する．前者は 3 文程度の要約，後者は英語論文のアブストラクトとほぼ同じ単語数を目指したものである．まず，重要文の選択方法を以下に述べる．

（1） スコアによってランク付けされた被引用論文中の各文から最も重要な文 s を選択し，重要文集合 S に加え，s を被引用論文から削除する．

（2） S の各文の単語数を数え，その和が閾値（100 または 250）を超えていれば重要文選択を終了する．

（3） s と，被引用論文中の各文との類似度を算出し，類似度が閾値 (0.8) を超えた文を削除し，（1）の手順に戻る．

重要文集合 S が確定したら，集合中の各文を被引用論文中の出現順に並べ替える．それを被引用文章の要約とする．

4. 評 価 実 験

実験では，提案手法で生成した要約の適切性を評価する．まず，評価指標 ROUGE（Recall-Oriented Understudy for Gisting Evaluation）[14] を用いて被引用論文のアブストラクトと生成した要約の差異について述べる．次にアンケートによる被験者実験を行い，生成された文章が被引用文章の要約として適切かを評価する．

4.1 実験データ

NTCIR13（NII Testbeds and Community for Information access Research）[13] に投稿された 75 件の論文を学習データとし．特徴ベクトルを 100 次元，窓サイズを 4，学習モデルを PV-DM，学習回数を 50 とし，モデルを作成する．まず，CRF を用いて論文のテキストから書誌情報を抽出する ParsCit を用いて XML データを抽出する．その後，このプログラムのオプションの，テキストに論文構成要素タグを付与する extract_section コマンドを使用し，bodyText タグが付与された部分を論文本文とする．そして，文章を 1 文ごとに区切ったものを入力として doc2vec を学習する．

表 1　ROUGE-N による結果

	ROUGE-1			ROUGE-2		
	Recall	Precision	F 値	Recall	Precision	F 値
100 単語	0.24	0.27	0.25	0.07	0.09	0.08
250 単語	0.33	0.50	0.37	0.12	0.17	0.14

4.2 ROUGE による評価

ROUGE [14] は基準となる文章と，自動生成された文章との一致度を測る指標である．ROUGE には様々な種類があるが，本稿では ROUGE-N によって被引用論文のアブストラクトから人手で抜き出した要約と，提案手法によって生成された被引用文章の要約を比較し、これらにどの程度の差異があるかを測った．文章は 100 単語と 250 単語の長さのものを作成した．ROUGE-N は N-gram 単位での出現単語の一致度を用いた評価である．スコアは，基準となる文章のなかに一致する単語がどのくらいあるか (Recall)，自動生成された文章のなかに一致する単語がどのくらいあるか (Precision)，そして Recall と Precision の調和平均（F 値）を求める．本稿では uni-gram 単位の評価である ROUGE-1 と，bi-gram 単位の評価である ROUGE-2 を用いて評価する．

表 1 は，NTCIR13 [13] に投稿された論文の引用箇所で特定できた 20 件の被引用論文を用いて作成された被引用文章の要約を，論文中の概要と比較した結果となる．この表から，提案手法による要約は，ある程度概要と同じ語彙を含むが，文章としては似ていないことがわかる．

4.3 被験者実験による評価

岡山大学大学院自然科学研究科の大学院生 5 名と岡山大学工学部情報系学科の学部生 2 名に対してアンケートを行い，提案手法で生成した文章が引用箇所に対する被引用文章の要約として適切であるか評価した．まず，NTCIR13 [13] に投稿された論文の中から，5 つの引用箇所に対して 100 単語程度と 250 単語程度の文章を作成する．作成する文章は，各引用文に対する被引用論文のアブストラクトから得られる文章（A），[2] で我々が提案した被引用箇所（B），本稿の提案手法により生成した被引用文章の要約の文章を要約（C）の 3 種類である．

- A : 各引用箇所に対する被引用論文のアブストラクトから，引用箇所の説明として適切であると思われる文章を人手により抜き出し，100 単語程度と 250 単語程度の文章を作成した。
- B : [2] では，引用箇所と被引用論文内の文を doc2vec によりベクトル化し，引用箇所の特徴ベクトルと最も類似した特徴ベクトルを持つ文とその前後を被引用箇所と推定する手法を提案した．そこで，[2] の提案手法で最も評価の高かった，「引用箇所を O'Conner の手法に従って得た文章とし，被引用論文の本文全文から特定する手法」を用いて文章を作成した．本手法で得られる 3 文の文章を 100 単語程度の文章．1 段落の文章を 250 単語程度の文章として被験者に提示する．
- C : 本稿で提案した手法により 100 単語程度の文章と 250 単語程度の文章を作成した．

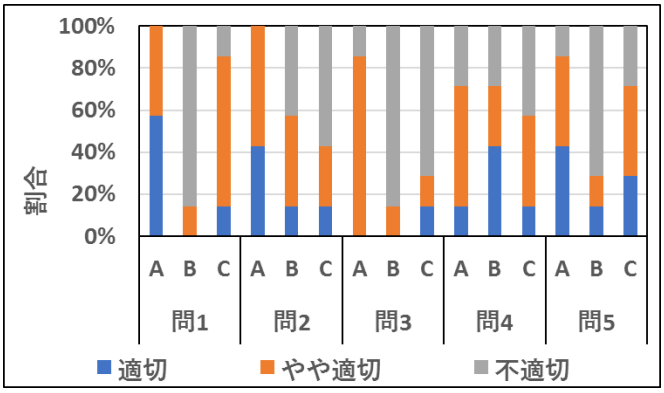

図 2　各問における要約（100 単語）の適切性評価の被験者による回答

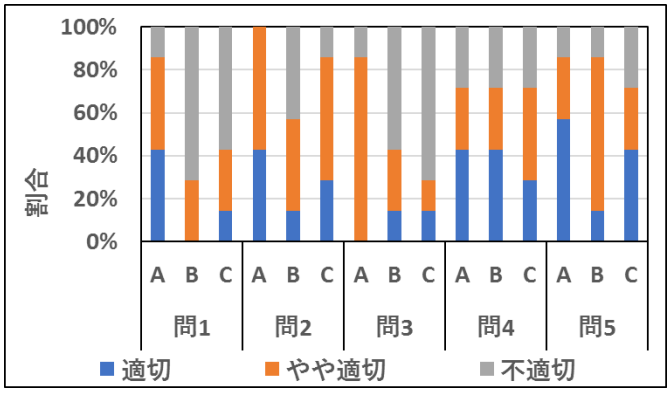

図 3　各問における要約（250 単語）の適切性評価の被験者による回答

これらの文章を被験者に提示し，それぞれについて被引用文章として適切かどうか「適切である」,「やや適切である」,「不適切」の中から選択させた．また，各文章を適切と思わる順にランク付けさせた．またこの実験に用いた 5 つの引用箇所は，石井ら [1] が提案した引用意図の“Data”，“Group”，“Equation”，“Result”，“Method”として特定された引用箇所であり，それぞれ問 1 から問 5 として使用した．各引用意図の意味を以下にまとめる．

- Data

実験等のデータが特に引用したい内容であるもの．

- Group

タスクやフォーラム，ワークショップの詳細が特に引用したい内容であるもの．

- Equation

計算式の詳細が特に引用したい内容であるもの．

- Result

既存研究の実験結果が特に引用したい内容であるもの．

- Method

研究に使用する，または著者の手法と比較する既存の手法や考えが特に引用したい内容であるもの．

要約の候補として文章を提示する順番は無作為に変えた．これらの評価を行うことによって，引用箇所に対して生成した被引用文章の要約に対する被験者の印象を調査する．

4.4　考　　察

4.3 節で行った被験者実験の結果をグラフにまとめる．図 2，図 3 は各文章が要約として適切かどうかの問に対して，被験者が「適切である」,「やや適切である」,「不適切」の中からどれを選択したかの割合をまとめている．100 単語の文章（図 2），250 単語の文章（図 3）のどちらの問に対しても，A の被引用論文の概要から得られる文章は，ほとんど「適切である」,「やや適切である」と評価されていた．しかし，A でも「適切である」がほとんど選択されていない問もある．特に引用意図が Result である問 3 では全く選択されていない．これは，概要に直接実験結果が記述されていなかったためである．また [2] の

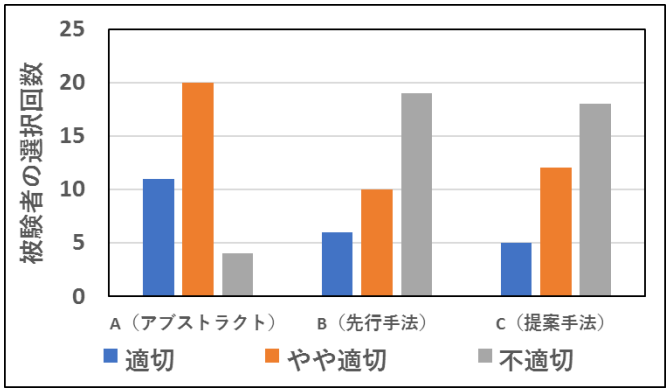

図 4　要約（100 単語）の適切性評価の回答の合計

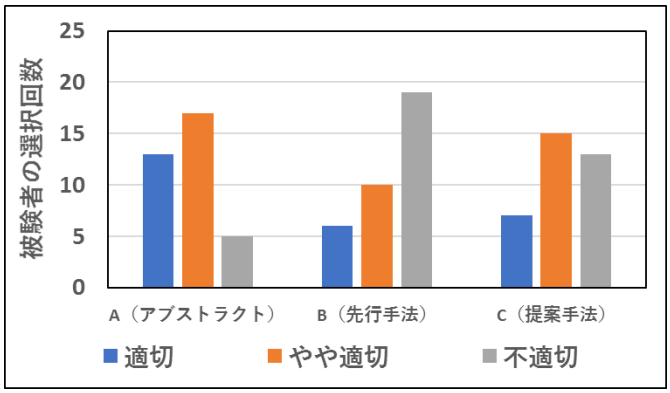

図 5　要約（250 単語）の適切性評価の回答の合計

手法である B や，提案手法 C では，ほとんど「不適切」が選択されている問もあれば，3 分の 2 以上の被験者が「適切である」または「やや適切である」を選択している問もある．

図 4，図 5 は各文章が被引用文章の要約として適切かどうかとの問に対して，被験者が「適切である」,「やや適切である」,「不適切」の中からどれを選択したかの回数の合計である．問 1 から問 5 の合計でも，A の文章が高評価であることがわかる．また，図 4 と図 5 を比較したとき，A や B がほとんど変化していないのに対し，C の文章において「不適切」として選択された回数に大きく差が出ており，100 単語より 250 単語の方が評価が高くなっている．C は冗長な文章を削除するためより多くの情報を含む．3 文程度の短い文より，少し長めの 250 単語

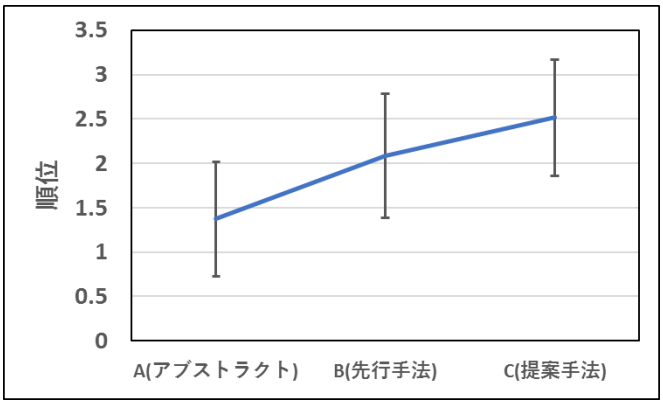

図 6　各文章の適切性の順位の平均と誤差範囲（100 単語）

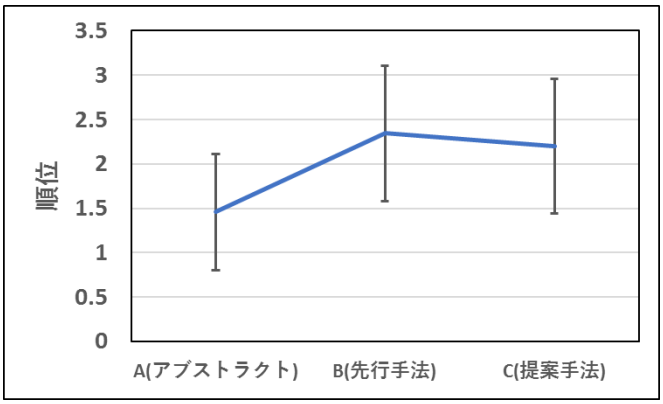

図 7　各文章の適切性の順位の平均と誤差範囲（250 単語）

程度で文章を生成するならば，B のように重要文の前後のみを提示するより，論文の様々な場所から文を収集して提示するほうがよりよい印象を与える可能性がある．

図 6，図 7 は，提示された各文章を被引用文章の要約として適切だと思わる順に 1 位から 3 位までランク付けをさせた結果の平均順位と誤差範囲を示している．平均誤差は $\pm\sigma$（標準偏差）としている．この 2 つの図から，図 4，図 5 と同様に，重要度の順位に関しても A の文章の評価が高いことがわかる．また，100 単語程度の文章であれば，C より B の方が評価が高いが，250 単語程度の文章になるとそれが逆転している．これも 250 単語程度の要約ならば，論文の 1 箇所から抽出した文章より，論文の様々な場所から重要度順に文を収集して提示したほうが，読み手にある程度評価の高い印象を与えることを示している．

5. ま と め

本研究では，引用文を利用して被引用論文中の重要文を抽出し，その要約を被引用文章として提示する方法を提案した．具体的にはまず，論文中で参考文献を引用している文章を“引用箇所”として抽出する．そして，doc2vec により引用箇所の文章と被引用論文中の各文の類似度を計算し，それを文の重要度のスコアとした．スコアの高い重要文を冗長性を考慮しながら選択し，被引用論文での出現順に選択した重要文を並び替えて，被引用文章とした．提案手法によって得られた要約を評価するためにアンケートを用いた被験者実験を行った．結果として，被引用論文のアブストラクトから人手により抜き出した文章の評価が最も高かった．一方，引用したいものが数式や論文の結果などの場合，アブストラクトから抜き出した文章だけは不十分であり，論文全体から文を選択した方が有効なことがあることを確認した．

今後の課題としては，被引用文章要約の可読性の向上などがあげられる．

謝　　辞

本研究の一部は，科学研究費補助金基盤研究 C（課題番号 18K11989) および科学研究費補助金基盤研究 B（課題番号 15H02789）の援助による．ここに記して深謝する．

文　　献

[1] 石井仁子，太田学，高須淳宏，“引用意図を利用した学術論文閲覧支援のための適切な被引用箇所の特定情報処理学会研究報告”，第 7 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2015），F3-5，2015.
[2] 田邊俊介，太田学，高須淳宏，安達淳，“doc2vec による学術論文の被引用箇所推定の一手法”，第 10 回データ工学と情報マネジメントに関するフォーラム（DEIM Forum 2018），G4-5，2018.
[3] Le，Q. and Mikolov，T.，“Distributed Representations of Sentences and Documents”，CoRR，abs/1405. 4053，pp. 1-9，2014.
[4] Mikolov，T.，Sutskever，I.，Chen，K.，Corrado G. and Dean，J.，“Distributed representations of words and phrases and their compositionality”，Advances in Neural Information Processing Systems，pp. 3111-3119，2013.
[5] 阿辺川武，相澤彰子，“内部構造解析機能と脚注表示機能を備えた論文閲覧システム”，人工知能学会，インタラクティブ情報アクセスと可視化マイニング第 7 回研究会，pp. 13-18，2014.
[6] Mihalcea，R.，Tarau，P.，“Textrank: Bringing order into texts”，Emprical Methods in Natural Language Processing (EMNLP)，pp. 404-411，2004.
[7] Carbonell，J.，Goldstein，J.，“The use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries”，the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval，pp.335-336，1998.
[8] 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 回年次大会発表論文集，pp. 569 - 572，2015.
[9] pdftotext，http://www.foolabs.com/xpdf
[10] ParsCit，http://aye.comp.nus.edu.sg/parsCit
[11] O’Conner，J.，“Citing statements: Computer recognition and use to improve retrieval”，Information Processing & Management，Vol. 18，No. 3，pp. 125-131，1982.
[12] Yoonjae，J.，Gwan，Jan.，Kyung-min，K.，Sung-Hyon，M.，“Geo-temporal Information Retrieval Based on Semantic Role Labeling and Rank Aggregation”，NTCIR-9 Workshop Meeting，pp. 22-28，2011.
[13] NTCIR，http://research.nii.ac.jp/ntcir

[14] Lin, C., "Rouge: A package for automatic evaluation of summaries", ACL-04 Workshop, pp. 74-81, 2004.