2004－DBS－134　(II)　(88)
2004／7／15

格文法を用いた複数文書融合手法

渡邊　拓也[†]　太田　学[‡]　片山　薫[‡]　石川　博[‡]

†‡東京都立大学大学院工学研究科 〒192-0397 東京都八王子市南大沢 1-1

E-mail: †jam@love.eei.metro-u.ac.jp, ‡{ohta, katayama, ishikawa}@eei.metro-u.ac.jp

あらまし　膨大なデータから効率的に情報を得るために、複数文書要約手法が研究されている。我々は文献[1]の論文で、形態素解析・係り受け解析を用いて文書から意味情報を抽出し、意味情報を疑似自然言語に融合する手法を提案した。本手法ではさらに格文法の概念を応用し、融合精度と読みやすさの向上を図った。深層格を求めるのに単語の概念まで理解する必要がある時は、検索エンジンのヒット数から単語の概念を推測し、深層格を求めた。この手法を用いる事により、表層格・動詞が全く同じでも深層格を判別できた。深層格を融合に用いる事により、強調構文・自動詞と他動詞の違いなど、言い回しの違いを吸収した融合を行えた。

キーワード　複数文書融合，複数文書要約，格文法

A Page Fusion Method using Case Grammar

Takuya WATANABE[†]　Manabu OHTA[‡]　Kaoru KATAYAMA[‡]　and　Hiroshi ISHIKAWA[‡]

†‡Graduate School of Engineering, Tokyo Metropolitan University, 1-1 Minami-Osawa, Hachioji-shi Tokyo, 192-0397

E-mail: †jam@love.eei.metro-u.ac.jp, ‡{ohta, katayama, ishikawa}@eei.metro-u.ac.jp

Abstract　Multi-document summarization techniques are researched to get information efficiently from a huge amount of text data. We have already proposed techniques which use morphological analysis and dependency structure analysis, extract semantic information from documents and integrate it into pseudo natural language. In this article, combined this Multi-Document Fusion method and case grammar, and aimed at an improvement of fusion precision and readability. If understanding a word concept is required to analyze deep case, we guess it using search engine's hit counts.

Keyword　Multi-Document Fusion，Multi-Document Summary，Case Grammar

1. はじめに

情報が氾濫している現在、それらを効率的に理解するための手法は有用だと考えられる。我々は文献[1]で複数文書融合手法に着目した。

本手法により、多くの情報を持った新聞記事の自動生成が可能である。また、融合された文書内容と同時に「元の文書それぞれがどのような内容を含んでいたか」を提示する事が可能なので、新しい検索エンジンインターフェースの構築も可能である。

本手法はそれが目指す所が他の複数文書要約手法と大きく異なる。他の複数文書要約手法は文書の重要な部分を提示する事を目標にしているが、本手法は「重要な部分を強調し、重要ではない部分も同時に提示する」という事を目標にしている。

この事を表したのが図１である。１つの文書は１つの楕円で表してある。文書内容には重なっている部分もあればそうでない部分もある。他の手法は重なっている部分の提示を目標にしているが、本手法は重なっている部分をまとめながら全ての情報を提示する事を目標としている。

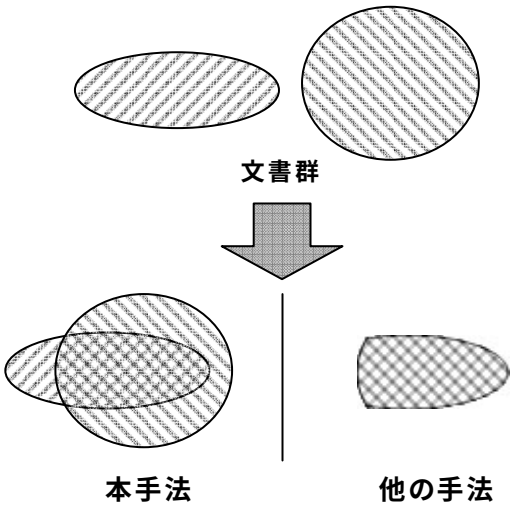


図１．他の手法との比較

現在は新聞記事など、文法的に洗練されている文書の融合に着目している。将来的にはブログなど、文法が曖昧な文や口語にまで適用文書を拡張する予定である。また、融合する文書集合としては、よく似た文書で構成される集合を想定している。

本手法は、格文法の深層格の概念を導入し、言い回しの違いを吸収できるようにした。深層格とは、文節の意味的な役割の事である。また、述語節に関しては、助動詞の特性も表示するようにした。

深層格は、言葉の意味的な情報がわからないと決定できない時がある。このような時は、検索エンジンのヒット数を用いて深層格を決定した。

以降、２章では関連研究、第３章では提案手法、第４章では実験について述べ、第５章でまとめを行う。

2. 関連研究

2.1. Abstraction

Abstraction 手法の代表的なものに newsblaster[2]があり、自然言語処理により意味解析を行っている。これは文書群を（ｉ）よく似た文書、（ii）特定の人物に対する伝記、（iii）その他、に分類し、それぞれ異なる手法で要約を生成する。我々の提案手法が対象としている「（ｉ）よく似た文書」に絞り、処理の流れを以下に示す。

（１）文のトピックを特定
（２）同一トピックの文をグルーピング
（３）文から、「単語を頂点、助詞や時制を頂点の属性とする有向グラフ」を生成
（４）有向グラフを既存の自然言語自動生成システムに入力し、自然言語を生成

英語の自然言語処理により意味解析を行うので日本語には適用できない。また、語の機能と意味両方の一致のみを扱う。

本研究は日本語の自然言語処理により、意味解析を行う。グルーピングは文単位ではなく、意味情報単位で行う。語の機能のみの一致も扱う。

2.2. Extraction

Jade Goldstein ら[3]は、段落・文など、何らかの単位の節に重要度を付加し、重要な節から順に抜粋している。重要度は（ｉ）クエリー又はユーザープロファイルとの類似度、（ii）トピック網羅性、（iii）文書内の位置、（iv）節を含む文書の生成された時間、（ｖ）既に抜粋された節との相違度、により計算する。但し、言い回しの違い等を吸収し、表現する事はできない。

本研究は節の重要度算出は行わない。節の意味解析を行うので、節間のわずかな違いを吸収し、同時提示可能である。

大野ら[4]は、あらかじめ指定したキーワードを用いてネットオークション上の評価コメントを要約している。しかしこれは、形式的なあいさつを多く含む分野にのみ適用可能な手法である。

本研究は自然言語処理により融合・要約を実現するので適用分野が限定されない。

3. 提案手法

3.1. 概要

本手法は文書を形態素まで分解し、意味的なまとまりをグルーピング、冗長度を排除しながら融合する。

3.1.1. 提案手法の流れ

提案手法の処理の流れを以下に示す。

STEP1:分解

最初に文書を文に分解する。chasen[5]を

用いて文を形態素に分解し、cabocha[6]を用いて形態素を文節毎にまとめる。

STEP2:解析

各文節の意味的役割（「主節」など）を求める。意味的にまとまった文節群（以下意味情報と呼ぶ）を求める。

STEP3:グルーピング

等しい内容の意味情報をグルーピングする。

STEP4:融合

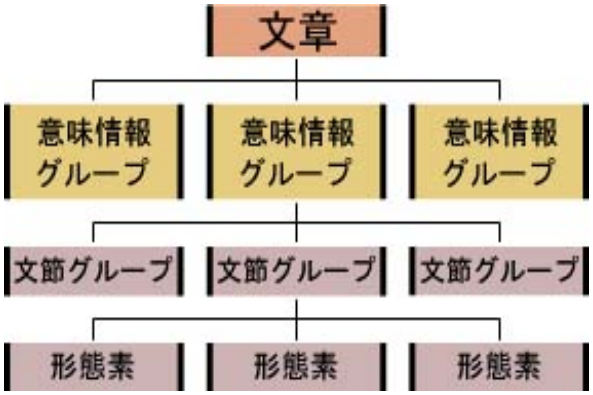
意味情報グループを一つにまとめて表示する。

STEP5:提示

融合された意味情報グループを並べて提示する。

3.1.2. 内部データ

内部データのデータ構造を図２に示す。



文節の例：「特捜戦隊は」
形態素の例：「特捜」「戦隊」「は」

図２． 内部データ構造

3.2. 分解

3.2.1. 解析ソフトによる分解

chasen を用いて各文を形態素に分解し、cabocha を用いて形態素を文節毎にまとめる。cabocha は同時に、（ｉ）文節の係り受け関係、（ii）文節毎の主たる形態素、を解析する。

3.2.2. 文節をまとめなおす

連続した文節で、お互いに依存した関係にある文節は、一つの文節としてまとめなおす。例えば、以下の例がある。

フセイン元大統領＋（６６）を
→　フセイン元大統領（６６）を

ここで、「２文節をまとめる事で生成された新しい文節」の主たる形態素を求め直す必要がある。「まとめあげる連続文節」と、「その時新しく生成された文節の主たる形態素」を以下に示す。

（１）「（」で始まると文節と、その直前の文節。
→主たる形態素は前の文節の主たる形態素
（２）「できる」「可能」という文節、その直前の「事が」「事の」という文節、さらにそのまた直前の文節。この場合計３文節をまとめあげる。
→主たる形態素は最も前の文節の主たる形態素
（３）「名詞＋の」という文節と、その直後の文節
→主たる形態素は後の文節の主たる形態素

前述の例は（１）の場合に相当し、主たる形態素は文節「フセイン元大統領」の主たる形態素（このばあい「大統領」）と、なる。

3.3. 解析

3.3.1. 文節の役割を解析

提案手法で定義している文節の役割を以下に示す。

（１）主格	（２）対象格
（３）道具格	（４）場所格
（５）源泉格	（６）目標格
（７）時間格	（８）原因
（９）述語節	（１０）接続節
（１１）修飾節	
（１２）be 動詞の補語	

cabocha により求めた主たる形態素を用い、文節がそれぞれの役割を持つ確率を求める。候補となる役割が一つしかない文節は、その役割の確率が１である。主たる形態素と文節の役割の対応を表１に示す。

表１． 主たる形態素による文節の役割

主たる形態素	文節の役割
動詞	述語節
名詞（時間を表す）	時間格
名詞（その他）	助詞等により算出
接続詞	接続節
その他	修飾節

ただし、文節が文の終端に位置している時は、「主たる形態素が動詞でない文節」の役割を「be 動詞の補語」とした。

主たる形態素が名詞の時は、助詞などを用いて文節の役割を求める。助詞と役割の対応を表２に示す。

表２． 助詞による文節の役割

助詞	文節の役割
が、は	主格
を	対象格
に	対象格、場所格、 目標格、原因
から	源泉格
と	対象格、修飾節
で	主格、道具格、場所格、 原因、修飾節
へ	目標格
まで	目標格
より	源泉格、修飾節
により	原因

対応する「文節の役割」が複数ある助詞は、検索エンジンのヒット数を利用して文節の役割を決定する。助詞「で」の例を示す。

例１：次の文が含む文節の役割を求める。（文節の区切りはスペースで示す。）

> 基地で 手話で ゴミ問題で
> みんなで 徹夜で 語り合った。

「語り合った。」は主たる形態素が「語り合う」という動詞なので、述語節になる。

その他の５文節は全て助詞「で」を持つ。それぞれの主たる形態素に「にて」「を用いて」「が原因で」「の集まり」「のまま」という語を後ろにつけ、検索する。また、主たる形態素の後ろにつける５フレーズそれぞれも、単独で検索する。

「『後ろにつけるフレーズ単独のヒット数』に対する『主たる形態素の後ろにフレーズをつけた時のヒット数』の割合」を求める。この割合を対応する役割の確率とする。

それぞれのフレーズと、対応する役割を以下に示す。

「にて」　→　場所格
「を用いて」　→　道具格
「が原因で」　→　原因
「の集まり」　→　主格
「のまま」　→　修飾節

後ろにつけるフレーズ単独のヒット数を表３に、主たる形態素の後ろにフレーズをつけた時のヒット数を表４に、後ろにつけるフレーズと主たる形態素の共起する「割合」を表５に、表５から求まる役割の確率を表６に示す。

なお、文節の判定毎に何回も検索するのでは非常に時間がかかる。よって、検索結果をＤＢに保存し、２回目からは検索を行わない事で対処している。

また、強調構文・受動態の文・自動詞を伴う文は固有の処理を行っている。

表３． フレーズ単独のヒット数

フレーズ	ヒット数[単位：万]
にて	814.0
を用いて	51.9
が原因で	35.0
の集まり	16.7
のまま	133.0

表４． 主たる形態素＋フレーズの検索ヒット数

	基地	手話	問題	みんな	徹夜
にて	3070	58	729	77	100
を用いて	2	201	555	0	0
が原因で	2	3	2040	1	13
の集まり	10	7	61	257	3
のまま	42	11	252	44	1050

表５． フレーズと主たる形態素の共起する割合

	基地	手話	問題	みんな	徹夜
にて	3.77	1.12	20.83	4.61	0.75
を用いて	0.00	3.87	15.86	0.00	0.00
が原因で	0.00	0.06	58.29	0.06	0.10
の集まり	0.01	0.13	1.74	15.39	0.02
のまま	0.05	0.21	7.20	2.63	7.89

（単位：１／１００００）

表６． 役割の確率

	基地	手話	問題	みんな	徹夜
場所格	0.98	0.21	0.20	0.20	0.09
道具格	0.00	0.72	0.15	0.00	0.00
原因	0.00	0.01	0.56	0.00	0.01
主格	0.00	0.03	0.02	0.68	0.00
修飾節	0.01	0.04	0.07	0.12	0.90

3.3.2. 助動詞の役割を解析

述語節が助動詞を含む時は、その助動詞の役割を求めておく。助動詞の種類、助動詞の直前の形態素により、以下の役割を割り当てた。一つの文節が複数の役割を割り当てられる事もありうる。

使役　受身　可能　尊敬　自発　丁寧
希望　打消　断定　過去　完了　推定
存続　確認　推量　意志　例示　伝聞
様態

3.3.3. 意味情報の生成

意味情報は種となる文節に、関連する文節を肉付けする形で生成する。文献[1]では以下の２種類の文節を種として用いた。

・動詞
・主格と be 動詞の補語のペア

提案手法では以下の２種類の文節を種として用いている。

・動詞
・be 動詞の補語

意味情報生成と同時に、前節で求めた「役割毎の確率」から文節の「役割」を求める。具体的には、最終的に求まった役割に対応する確率をスコアと考え、その総和が最も高くなるように求める。この時、１文に同じ格が複数現れないようにする。

例２：例１の文節の役割を求める。（例１の６文節が１つの意味情報を生成したとする。）

最終的に役割は以下のように求まる。

基地で　→　場所格
手話で　→　道具格
ゴミ問題で　→　原因
みんなで　→　主格
徹夜で　→　修飾節

この時のスコアは

0.98 + 0.72 + 0.56 + 0.68 + 0.90 = 3.84

と、なる。これは最終的な役割の全組み合わせにおいて、最も高い数値である。

3.4. グルーピング

3.4.1. 意味情報のグルーピング

２つの意味情報が以下に示す条件のうちどれかを満たしていれば、その２つの意味情報をグルーピングする。

（１）主格と be 動詞の補語が共に等しい
（２）主格と述語節が共に等しい
（３）一方の意味情報が含む役割の２／３以上が等しい

例３：以下の２つの意味情報がグルーピングされるかどうかを判定する。（それぞれの意味情報は線で囲う。その中で文節の区切りはスペースで示す。）

基地で 手話で ゴミ問題で
みんなで 徹夜で 語り合った。

基地で テレパシーで
徹夜で 語り合った。

上の意味情報が含む文節の役割は例２のように求められる。下の意味情報も同様に文節の役割を求められ、例２で求めていない「テレパシーで」の役割は道具格となる。

上の意味情報は主格「みんなで」を含むが、下の意味情報は含まない。よって、条件（１）、（２）は満たされない。

２つの意味情報で等しい役割は、場所格・修飾節・述語節の３つである。含む役割数の２／３は上の意味情報では４、下の意味情報では約２．７である。下の意味情報が含む役割の２／３以上が等しいので、（３）の条件を満たす。

よって、この２つの意味情報はグルーピングされる。

3.5. 融合･提示

文献[1]では役割毎に融合を行った後に各役割を並べ替えていた。

融合処理の流れを以下に示す。

（１）意味情報グループに含まれる意味情報から、役割毎の配置順を決定

（２）役割毎の配置順に従い、意味情報内の文節を意味情報毎に並べる

（３）並べられた文節を形態素に分割し、意味情報グループ全体を融合

融合結果は図３のような表となり、列が１つの形態素に、行が１つの意味情報に対応している。図３は例３の融合結果を示している。

このように、１つの意味情報を１つの行と対応させた事により、文献[1]の手法で大きな問題であった「間違った内容の提示」を防いでいる。例えば、以下の例文の融合結果は図４のようになる。

１．デカレンジャーは宇宙から来た。
２．コスモレンジャーは宇宙から来た。
３．タイムレンジャーは未来から来た。

さらに、多くの意味情報に含まれる形態素のセルは大きく表示されるので、全体像の把握が可能である。複数文書要約の代表的な手法である重要語抽出は重要な語を残し、重要ではない語は完全に隠してしまう手法である。本手法は重要な語は強調し、重要ではない語は完全には隠さないので、閲覧者は好みに応じ、概要だけを把握したりより詳しい情報を得たりする事が可能である。

4. 実験と考察

4.1. 実験手法

「言い換えのあれこれ」[7]で紹介されている例文に対して本手法を適用した。ただし、単語のみの言い換え、英文に関しては、本手法の適用範囲外であるので除いた。

これらの例文に対して本手法を適用し、読みやすさが向上されているかどうかをチェックした。ただし、動詞や助動詞の活用を扱っていない事による影響は除いて考えた。

4.2. 実験結果

同じ意味を表すと思われる例文群は 95 あり、本手法が少しでも読みやすさを向上できたと考えられる例文群は 59 あり、中でも 19 の例文群は本手法が非常に効果的に働いていたと考えられた。ただし、全く効果がない、または本手法がかえって読みやすさを低下させてしまっている例文も 36 存在した。このうち cabocha や chasen の出力が明らかに正しくないのは 2 存在した。

文節の一致判定にシソーラスを用いていないため、「公算が大きい」と「可能性が大きい」等を同じ意味だと見なす事ができなかった。

比較文にも対応していないため、「兄は弟より太っている」と「弟は兄より痩せている」を同じ意味だと見なす事ができなか

基地	で	手話 テレパシー	で	ゴミ	問題	で	みんな	で	徹夜	で	語り合う	た	。

図３． 例３の融合結果

デカ
コスモレンジャーは 宇宙 から来るた。
タイム 未来

文献[1]の手法

デカ コスモ タイム	レン	ジャー	は	宇宙 未来	から	来る	た	。

提案手法

図４． 文献[1]の手法と提案手法の比較

１．[時]昨夜、 [原]ダンスのために [修]一番 [修]良い
[対]スーツを [動]着ていたのは [修]ジョンだ
２．[時]昨夜、 [主]ジョンは[原]ダンスのために
[修]一番 [修]良い [対]スーツを [動]着ていた
３．[時]昨夜、 [主]ジョンが[原]ダンスのために
[動]着ていたのは [修]一番 [修]良い [修]スーツだ
４．[主]ジョンが[原]ダンスのために [修]一番
[修]良い [対]スーツを [動]着ていたのは [時]昨夜だ
５．[時]昨夜、 [主]ジョンが[修]一番 [修]良い
[対]スーツを [動]着ていたのは [修]ダンスのためだ
６．[時]昨夜、 [主]ジョンが[修]一番 [修]良い
[対]スーツを [動]着ていたのは [原]ダンスのために [修]だ

文節の前には役割を括弧で囲って示している。
時：時間格　原：原因　修：修飾節　対：対象格　動：述語節　主　主格

図５．強調構文の例文と文節の役割

<table>
<tr><td rowspan="4">昨夜</td><td rowspan="2">、</td><td rowspan="2">ジョン</td><td>は</td><td rowspan="2">ダンス</td><td rowspan="2">の</td><td rowspan="2">ため</td><td rowspan="2">に</td><td rowspan="2">良い</td><td rowspan="2">スーツ</td><td rowspan="2">を</td><td rowspan="4">着る</td><td rowspan="4">て</td><td rowspan="4">いる</td><td rowspan="4">た</td></tr>
<tr><td>が</td></tr>
<tr><td rowspan="2">、</td><td rowspan="2">ジョン</td><td rowspan="2">が</td><td rowspan="2"></td></tr>
<tr><td>良い</td><td>スーツ</td><td>を</td></tr>
</table>

図６．強調構文の融合結果

１.[主]ジョンが[対]私のコンピュータを [動]壊した
２．[主]私のコンピュータが[動]壊れた

図７．自動詞と他動詞の融合例（入力）
表記法は図５と同様

<table>
<tr><td rowspan="2">ジョン</td><td rowspan="2">が</td><td rowspan="2">私</td><td rowspan="2">の</td><td rowspan="2">コンピュータ</td><td>を</td><td>壊す</td><td rowspan="2">た</td></tr>
<tr><td></td><td>が</td><td>壊れる</td></tr>
</table>

図８．自動詞と他動詞の融合例（出力）

った。

強調構文を用いた例文は cabocha の出力が明らかに正しくなかったため、手動で文節を分解し、融合し直してみた。その例文と、各文節の役割を図５に、融合結果を図６に示す。

図６より、強調構文のように助詞の形態がバラバラであっても、格文法の概念を用いる事である程度の融合ができている事がわかる。ただし、強調構文の終端に見られる「～だ」という文節の格を正しく判別できていないため、少し見にくくなっている。

自動詞と他動詞をうまく融合した例の、例文を図７に、融合結果を図８に示す。

4.3. 考察

動詞の活用などを処理できていないので、本来は「着ていた」と表記するべき所と「着るていた」と表記してしまうなど、読みにくい部分もある。しかし、

「情報量過多のきらい傾向が」のように、形態素の違いをうまく吸収できた例もある。

全体的に、１文から１つの意味情報が生成される場合は、本手法を適用する事により非常にすっきりした表現を行えた。しかし、現在は融合された意味情報群をただ列挙する形で提示しているので、１文から複数の意味情報が生成される場合は必ずしもすっきりした表現とはいえない。「言い換えのあれこれ」に出てくる例文はほとんどが単文であるが、新聞記事など実際の文章は長文を多く含む。そのため、この問題を解決する事が読みやすさの向上に及ぼす影響は大きいと考えられる。

5. おわりに

本稿では文献[1]で提案した複数文書融合手法に格文法の概念を適用し、融合精度と読みやすさの向上を図った。深層格の決定に単語の概念理解が必要な時は、検索エンジンのヒット数から単語の概念を推測した。

今後の課題を以下に示す。

- 動詞と助動詞の活用の表現
- 意味情報提示方法の洗練
- 意味情報生成手法の洗練
- シソーラスを利用した文節一致判定
- 自然言語処理技術の洗練
- 文書のグルーピングの自動化

謝辞

本研究の一部は，文部科学省科学研究費補助金特定領域研究(2)[情報学:A02](課題番号:16016273)による。

文　献

[1] 渡邊拓也,太田学,片山薫,石川博,”分野に依存しない複数文書要約手法の提案,”DEWS2004, March 2004

[2] Newsblaster: http://www1.cs.columbia.edu/nlp/newsblaster/

[3] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz, Multi-Document Summarization by Sentence Extraction, "Proc. ANLP/NAACL Workshop on Automatic Summarization, pp.40-48, Seattle, USA, Apr.2000.

[4] 大野華子,楠村幸貴,土方嘉徳,西田正吾,”社会的関係を用いた評価コメントの自動要約方法とその有効性の検証,” 情報処理学会研究報告，Vol.2004, No.45, no.(8), pp.67-74, May 2004.

[5] Chasen: http://chasen.aist-nara.ac.jp/.

[6] Cabocha: http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

[7] 言い換えのあれこれ: http://cl.naist.jp/lab/kura/ParaCorpus/