分野に依存しない複数文書要約手法の提案

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。複数文書要約に、単一文書要約で広く用いられていた抜粋処理を拡張して適用するには限界がある。このため、文書を意味レベルまで解析し、それらを表す文脈を再構築する手法が研究され始めている。本稿では、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して複数文書の要約文とする手法を提案する。意味情報の融合には、意味情報を構成する文節の内容と役割、それぞれの類似度を用いる。話題の分野ごとに特化したパターンマッチング技術を用いないので、様々な分野に適用可能である。

キーワード 情報統合，テキストマイニング，知識発見，情報検索，ユーザインタフェース

A Domain Independent Multi-document Summarization Technique

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 realize that extending extraction techniques used by many single-document summarization tasks has limitations. Therefore, techniques analyzing documents semantically, and reconstructing them into contexts are researched recently. The technique s proposed in this paper use morphological analysis and dependency structure analysis, extract semantic information from documents and integrate it into pseudo natural language. To fuse semantic information, we use both semantic and functional similarity of clauses. This technique is domain independent, because it uses no domain sensitive pattern matching techniques.

Keyword Information Integration, Text Mining, Knowledge Discovery, Information Retrieval, User Interface

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。本研究では複数文書要約手法に着目した。

本手法により、多くの情報を持った新聞記事の自動生成や、検索結果をグルーピングする検索システムにおけるクラスタ内容簡易表示などが可能である。

現在は文法的に洗練されている新聞記事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。

現在までに複数文書要約手法は多く研究されており、それらは Abstraction と Extraction に大別される[1]。

Abstraction は文書の意味解析を行い、新たに文書を生成する手法である。意味解析としては、一般に（１）話題分野を特定したパターンマッチング、（２）自然言語処理、が行われている。

Extraction は文書の構成要素に重みを付け、重要なものから順に抜粋していく手法である。単文書要約で広く用いられたが、複数文書要約で用いるには限界があると考える。抜粋処理だけでは、用いられる文節の「違い」を表現できないからである。

本研究は Abstraction に分類され、その中でも自然言語処理を用いて意味解析を行う。本研究の特徴は、以下のようになる。

- 未知語を名詞として扱い（ほとんどの専門用語は名詞だと考えられる）、助詞を中心にした自然言語処理を行う。これにより、特定の専門用語を知る必要はなく、分野に依存しない。
- 語の意味の一致のみでなく、語の機能のみの一致も扱う。これにより、異なる文節の存在を吸収した表現が可能
- 出力は自然言語ではなく、疑似自然言語である。これにより、視覚的にわかりやすい表現が可能

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。

言い回しの違い等を吸収し、表現する事ができない。

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。

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[4]を用いて文を形態素に分解し、cabocha[5]を用いて形態素を文節毎にまとめる。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。

STEP2:解析

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し、意味情報グループ毎に等しい内容の文節をグルーピングする。

STEP4:融合

意味情報グループ毎に文節グループを融合し、融合された文節グループを用いて意味情報グループを融合する。

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以下に示す。

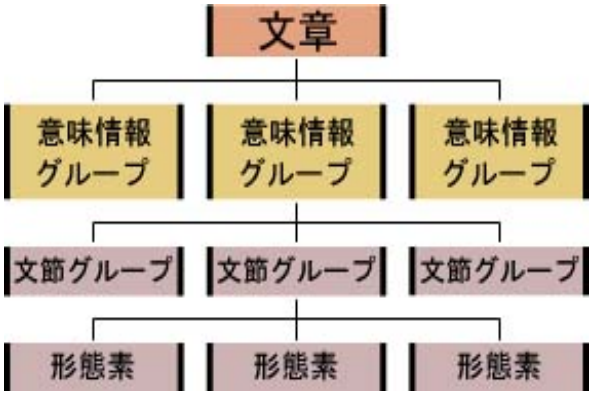


図１．内部データ構造

意味情報の例：図３参照、文とは異なる

文節の例：「岡田さんは」「出版した」

形態素の例：「岡田」「さん」「は」

3.2. 解析

3.2.1. 文節の機能決定

文節を構成する形態素により、文節を機能毎に以下のグループに分ける。

・　主節

格助詞の「が」、係助詞の「は」を含む文節。

例：もんじゅが　岡田さんは

・　述語節

主たる形態素が動詞である文節。または、主たる形態素がサ変接続の名詞で、動詞の「する」が後に続いている文節。

ただし、名詞節（主たる形態素が名詞である文節）に係っている文節は除く。

例：投げる　　出版する

・　目的節

格助詞の「を」を含む文節。

例：本を

・　接続節

主たる形態素が接続詞である文節。

例：そして　　しかし

・　修飾節

上記以外の文節

例：大幅に　筋トレで

3.2.2. 意味情報抽出

意味情報抽出は次の２段階の処理で行う。

まず主節と、それが係る文節の組み合わせを求め、意味情報の種とする。次に、意味情報の種に述語節、目的節、修飾節で肉付けを行う。

（１）意味情報の種の生成

先頭から文節を走査する。意味情報の種が生成されるパターンは２つある。

（パターン１）主節が見つかれば、その主節が係っている文節と組み合わせ、意味情報の種とする。ここで、主節が修飾節に係っていれば、この修飾節を目的節として扱う。（英語で言えば、この修飾節が be 動詞の補語に相当する。）

（パターン２）他の意味情報に含まれていない述語節が見つかれば、直前の意味情報生成で使用した主節と組み合わせ、意味情報の種とする。これは、主節が省略されていると考えるためである。

（２）意味情報の種の肉付け

意味情報に、以下の条件を満たす目的節、修飾節、名詞節に係っている述語節を付加する。

・　他の意味情報に含まれていない。
・　該当する意味情報を構成するいずれかの文節に係っている。

この操作を、付加する文節がなくなるまで繰り返す。

例１：「岡田さんは大幅に減量し、本を出版した。」という文から意味情報を抽出する。文節の係り受け関係、算出された機能を図２に示す。

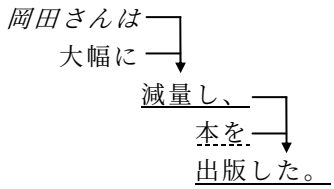


図２．例１の例文の係り受け関係、文節機能

※　文節機能：斜体→主節　実下線→述語節
　　点下線→目的節　飾り無し→修飾節

※　各行は文節を表し、矢印は係り受け関係を表す。

抽出例を以下に示す。

意味情報の種生成１　先頭から走査し、主節「岡田さんは」を見つける。「岡田さんは」は述語節「減量し、」に係っているので、「岡田さんは」と「減量し、」が意味情報の種となる。

意味情報の肉付け１　修飾節「大幅に」は「減量し、」に係っているので、この意味情報に含める。ほかに含める文節がないので、「岡田さんは」、「大幅に」、「減量し、」の３文節で１つの意味情報を構成する事になる。

意味情報の種生成２　「大幅に」から走査し、まだ意味情報に含まれていない述語節「出版した。」を見つける。直前の意味情報抽出で使用された主節「岡田さんは」と「出版した。」が意味情報の種となる。

意味情報の肉付け２　目的節「本を」は「出版した。」に係っているので、この意味情報に含める。ほかに含める文節がないので、「岡田さんは」、「本を」、「出版した。」の３文節で１つの意味情報を構成する事になる。

3.3. グルーピング

3.3.1. 文節の一致度判定

グルーピング操作について説明する前に、２文節の一致度を判定する操作について説明する。

文節Ａと文節Ｂの一致度判定により、①ＡとＢは等しい（Ａ＝Ｂ）、②ＡがＢを包含する（Ａ⊃Ｂ）、③ＡはＢに包含される（Ａ⊂Ｂ）、④ＡとＢは無関係（Ａ≠Ｂ）、のいずれかの関係を算出する。

文節の文法的機能（「文節の機能決定」よりも細かく分類した機能）と内容（文節から補助的な形態素を排除したもの）両方の一致度を求め、それらの一致度から文節の一致度を判定する。

以下、名詞・接尾動詞以外の動詞・形容詞・副詞・未知語を意味語と呼ぶ事にする。

（１）文法的機能の一致度判定

文節内で「最後に現れる意味語よりも後ろに存在する助詞の並び」を「終端助詞シーケンス」と呼ぶ事にする。

「文節の機能決定」で求めた機能を、終端助詞シーケンスを用いてさらに細かく分類したものが文法的機能である。

以下に、文法的機能の一致度とその判定条件を示す。

①Ａ＝Ｂ

・　Ａ、Ｂ共に主節
・　Ａ、Ｂ共に述語節
・　Ａ、Ｂに目的節
・　Ａ、Ｂ共に修飾節で、終端助詞シーケンスが等しい（共に終端助詞シーケンスを持たない場合も含む）
・　Ａ、Ｂ共に修飾節で、一方の文節の終端助詞シーケンスが「の」で、もう一方の文節は終端助詞シーケンスを持たない

②Ａ⊃Ｂ（③Ａ⊂Ｂ）

・　①の条件を満たさず、Ｂ（Ａ）が終端助詞シーケンスを持たない
・　①の条件を満たさず、Ｂ（Ａ）の終端助詞シーケンスが「の」である

④Ａ≠Ｂ

・　①～③以外

（２）内容の一致度判定

（２－１）形態素のフィルタリング

文節内で、意味語以外の形態素を破棄する。

（２－２）比較

２文節それぞれにおいて、「フィルタリングされた全形態素数に対する、２文節で語幹が重複している形態素数の割合α」を求める。

以下に、内容の一致度とその判定条件を示す。

①Ａ＝Ｂ

・　Ａ、Ｂ共に、αが閾値を越えた

②Ａ⊃Ｂ（③Ａ⊂Ｂ）

・　Ｂ（Ａ）のみ閾値を越えた

④Ａ≠Ｂ

・　Ａ、Ｂ共に、αが閾値を越えない

閾値は経験的に０．５９に設定している。

また、内容の一致判定には以下の例外法則も適用する。

・　Ａ、Ｂ共に相手が持っていない固有名詞を持って
・　いる時、無条件にＡ≠Ｂと判定する。
・　Ａ、Ｂ共に相手が持っていない数字を持っている時、無条件にＡ≠Ｂと判定する。

（３）文節の一致度判定

表１に、文法的機能と内容の一致度の各組み合わせに対する文節の一致度を示す。

		文法的機能			
		①Ａ＝Ｂ	②Ａ⊃Ｂ	③Ａ⊂Ｂ	④Ａ≠Ｂ
内容	①Ａ＝Ｂ	①Ａ＝Ｂ	①Ａ＝Ｂ	①Ａ＝Ｂ	④Ａ≠Ｂ
	②Ａ⊃Ｂ	①Ａ＝Ｂ	②Ａ⊃Ｂ	②Ａ⊃Ｂ	④Ａ≠Ｂ
	③Ａ⊂Ｂ	①Ａ＝Ｂ	③Ａ⊂Ｂ	③Ａ⊂Ｂ	④Ａ≠Ｂ
	④Ａ≠Ｂ	④Ａ≠Ｂ	④Ａ≠Ｂ	④Ａ≠Ｂ	④Ａ≠Ｂ

表１．文節の一致度判定条件

ただし、同じ意味情報から生成された文節は無条件にＡ≠Ｂと判定する。

例２：主節Ａ「岡田-さん-は」と主節Ｂ「岡田-は」を比較する時（形態素はハイフンで区切ってある。）

共に主節であるので、文法的機能はＡ＝Ｂと判定される。

意味語以外の形態素が排除され、文節Ａは「岡田-さん」、文節Ｂは「岡田」と、なる。

両方の文節で重複している形態素は「岡田」の１つである。文節Ａ、Ｂそれぞれについて、αは１／２、１／１と、なる。文節Ｂのαのみが閾値０．５９を越えるので、内容の一致度はＡ⊃Ｂと判定される。

文法的機能の一致度がＡ＝Ｂで、内容の一致度がＡ⊃Ｂと判定されたので、文節の一致度はＡ＝Ｂと判定される。

例３：内容の一致度の例外法則適用例を以下に示す。

・　「佐藤弁護士事務所」≠「山本弁護士事務所」
・　「第１５７特別国会」≠「第１５８特別国会」

3.3.2. 意味情報のグルーピング

以下の条件のいずれかを満たす意味情報群をグルーピングする

（ｉ）主節と述語節の一致度がＡ≠Ｂでない

（ii）共に述語節が存在せず、お互いの主節と目的節の一致度がＡ≠Ｂでない

英語で言えば、(ｉ)は主語と do 動詞が等しい場合、（ii）は主語と be 動詞の補語が等しい場合、と、いえる。

例４：以下に示す意味情報はグルーピングされる。

（ｉ）の例

「*岡田さんは*　大幅に　減量した」と

「*岡田さんは*　筋トレで　減量した」

（ii）の例

「新総裁の　*任期は*　０６年９月までの　３年」と

「*任期は*　３年」

※文節機能の表記法は図２と同様。文節はスペースで区切り、出現順に羅列。係り受け関係の表記は省略。

3.3.3. 文節のグルーピング

「文節の一致度判定」の結果を用い、意味情報グループ内で、各意味情報を構成する文節をグルーピングする。処理の流れを以下に示す。

（１）完全一致によるグルーピング

文節の一致度がＡ＝Ｂと判定された文節をグルーピング

ただし、この操作でＡ＝Ｂでない文節をグルーピングしないようにする。つまり、Ａ＝B_1、Ａ＝B_2、B_1≠B_2の時、ＡとB_1、B_2のいずれか一方のみをグルーピングする。

（２）包含関係によるグルーピング

前ステップで生成された文節グループを包含関係によりグルーピングする。

ここで、文節グループ同士の関係を定義する。文節グループＸ内の任意の文節ｘと文節グループＹ内の任意の文節ｙの間に②ｘ⊃ｙの関係があった時、Ｘ⊃Ｙとする。Ｘ⊂Ｙの関係も同様。包含関係が生成されなければＸ≠Ｙとする。

Ｘ⊃Ｙ、Ｘ⊂Ｙとなる文節グループをグルーピングする。ただし、この操作でＸ≠Ｙである文節グループをグルーピングしないようにする。

また、同じ意味情報に所属する文節を持つ文節グル

ープ同士もグルーピングしないようにする。例外的に、同じ意味情報に所属する文節同士が隣り合っていればグルーピングする。

例５：意味情報グループ内に文節Ａ、Ｂ、Ｃ、Ｄ、Ｅが存在し、Ａ＝Ｂ、Ｃ＝Ｄ、Ｂ⊃Ｃの時。
まず、ＡとＢ、ＣとＤがグルーピングされる。
「ＡとＢのグループ」⊃「ＣとＤのグループ」であるので、２つのグループをまとめる。

結果、Ａ～Ｄのグループと、Ｅのみのグループが生成される。

3.4. 融合

3.4.1. 文節グループの融合

不要な形態素を破棄し、文節グループ内の文節を構成する形態素を並び替える。
連続関係が推測される形態素は連続して並べる。形態素を無理にシーケンシャルに並べると新しい語を作ってしまう可能性があるので、前後関係がはっきりしない形態素は併記する。
処理の流れを以下に示す。

（１）助詞、記号、句読点を破棄
形態素同士の接着剤として働くような形態素はいったん破棄する。

（２）形態素の連続関係を集計
「ある形態素の次にどの形態素が現れるか」という情報を集計する。この時、同じ形態素が複数の文節に現れていれば一つの形態素として扱う。よって、「一つの形態素の次に複数の形態素が現れる」という事もありうれば、「複数の形態素の次に１つの形態素が現れる」と、いう事もありうる。
形態素Ａ、Ｂ、Ｃに「Ａ→Ｂ、Ｂ→Ｃ、Ａ→Ｃ」というような連続関係があった時、Ａ→Ｃという連続関係は破棄する。これは、Ａ→Ｃという連続関係はＢが省略された事により発生したと考えるためである。

（３）形態素を並べる
形態素の連続関係から「形態素をノード、連続関係をリンクとする有向グラフ」を生成し、このグラフを基に形態素を並べる。

（４）助詞の復元
表記された文節グループの最後に助詞を付与する。付与する助詞は文節グループの役割（3.4.2(1)参照）により、以下のように決定される。
（ｉ）述語節グループ
助詞を付与しない。
（ii）それ以外
文節グループを構成する文節で最も多く用いられている終端助詞シーケンスを付与。

例６：主節「岡田-さん-は」（形態素はハイフンで区切ってある。）と主節「岡田-社長-は」と主節「岡田-俊夫-社長-は」の３文節を融合する。
不要な形態素を破棄し、３文節はそれぞれ「岡田-さん」、「岡田-社長」、「岡田-俊夫-社長」と、なる。
以下の連続関係が集計される。
「岡田→さん」、「岡田→社長」、「岡田→俊夫」、「俊夫→社長」
「岡田→社長」という連続関係は、「岡田」と「社長」の間の「俊夫」が省略された事により発生したと考え、破棄する。
「岡田 ⇒ 俊夫→社長 / ⇒ さん」という有向グラフが生成され、

形態素は岡田 俊夫社長 / さん と、並べられる。

述語節グループではないので（3.4.2.(1)参照）、最も多く用いられている終端助詞シーケンス「は」を付与し、この文節グループの最終的な融合結果は

岡田 俊夫社長 / さん は と、なる。

3.4.2. 意味情報グループの融合

意味情報グループ内の文節グループを並び替える事により行う。処理の流れを以下に示す。

（１）文節グループの役割を求める
文節グループを以下の７つのグループに分類する。
①主節グループ
②述語節グループ
③目的節グループ
④主節の修飾節グループ
⑤述語節の修飾節グループ
⑥目的節の修飾節グループ
⑦意味情報全体の修飾節グループ
処理の流れを以下に示す。
（１－１）文節グループを①～③のグループと修飾節グループ（④～⑦のどれに分類されるかは未決定）の４つに分類
（１－１－１）文節の役割を再計算
「文節の機能決定」で、助詞などから文節の役割を求めた。この役割を基に、以下のように文節の役割を再計算する。（文節グループの役割とは違うので、注意されたい。）

- 主節・述語節：「意味情報のグルーピング」で、意味情報のグルーピングに用いられた主節・述語節

・　目的節：再計算された述語節に係り、「文節の機能算出」で目的節と判定された文節
・　修飾節：上記以外の文節

（１－１－２）重み付け多数決

文節グループを構成する文節の役割から重み付け多数決を取り、文節グループの役割とする。

重みは「主節・述語節・目的節：２票　修飾節：１票」と、する。複数の役割が共に１位になってしまった時は、主節、述語節、目的節、修飾節の順番で優先的に決定する。

（１－２）修飾節グループを④～⑦に分類

まず、文節グループ同士の係り受け関係について定義する。「文節グループＡを構成する文節のいずれかが、文節グループＢを構成する文節のいずれかに係っている時、文節グループＡは文節グループＢに係っている」と、する。

文節グループ同士の係り受け関係を用い、修飾節グループの分類条件を以下に示す。

④主節の修飾節グループ
・　主節グループにのみ係っている

⑤述語節の修飾節グループ
・　述語節グループにのみ係っている

⑥目的節の修飾節グループ
・　目的節グループにのみ、又は目的節グループと述語節グループにのみ係っている

⑦意味情報全体の修飾節グループ
・　上記以外

ただしここでは、他の修飾節グループを介して間接的に係っている係り受け関係も、直接存在する係り受け関係も同様に扱う。

（２）同じ役割同士で並べる

前ステップで同じ役割に分類された文節グループ同士を並べ替える。「文節グループの融合」の（２）、（３）と同様の処理を行う。ただし、文節グループの並び替えでは文節グループの連続関係だけでなく、係り受け関係も集計し、係り受け関係が存在する文節グループはなるべく連続するように並び替える。

（３）役割に応じて並べる

前ステップで文節グループの並び替えを行った「文節グループのグループ」を並べる。この操作は①～⑦の役割を単純に以下の順番に並べる事により行う。

⑦意味情報全体の修飾節グループ
④主節の修飾節グループ
①主節グループ
⑥目的節の修飾節グループ
③目的節グループ
⑤述語節の修飾節グループ
②述語節グループ

例７：図３に示す２つの意味情報を融合する。

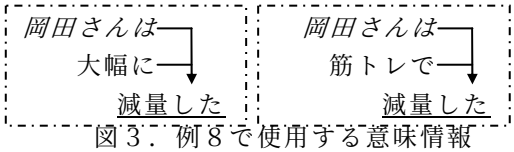


図３．例８で使用する意味情報

※表記法は図２と同じ

「文節のグルーピング」により、以下の４つの文節グループが生成されている。(括弧内は文節グループの名前)

・　２つの「岡田さんは」によるグループ　（Ａ）
・　２つの「減量した」によるグループ　（Ｂ）
・　「大幅に」によるグループ　（Ｃ）
・　「筋トレで」によるグループ　（Ｄ）

まず、２つの「岡田さんは」、２つの「減量した」はそれぞれ主節、述語節なので、文節機能再計算後も変わらず主節、述語節である。述語節に係る文節の中で、目的節と判定された文節がないので、残る「大幅に」、「筋トレで」は修飾節となる。

文節機能の多数決により各グループが取った票数を表２に示す。

	主節	述語節	目的節	修飾節
A	4	0	0	0
B	0	4	0	0
C	0	0	0	1
D	0	0	0	1

表２．文節機能の多数決結果

表２より、Ａは①主節グループ、Ｂは②述語節グループ、ＣとＤは修飾節グループ、と、判定される。

Ｃ、Ｄは述語節グループにのみ係っているので、⑤述語節の修飾節グループに判定される。

①～⑦のグループ毎に文節グループを並び替える。

①主節グループと②述語節グループは要素が１つなので並べる必要はない。⑤述語節の修飾節グループ内に係り受け関係はないので、以下のように単純に併記される。

C
D

①～⑦各グループの並び替えを行う。①主節グループ、⑤述語節の修飾節グループ、②述語節グループ、の順に並べられ、最終的な意味情報グループ融合結果は、文節グループを融合し、以下のように表現される。

岡田さんは 大幅に 筋トレで 減量した

3.5. 提示

意味情報グループの前後関係を推測し、構成する意味情報が多いグループから順に１つずつ、融合された意味情報グループを並べる。

意味情報グループを構成する意味情報数に閾値を与え、提示判定を行えば、要約結果の圧縮量を調整可能である。

以下、意味情報ｘ１とｘ２で構成されている意味情報グループＸを「Ｘ｛ｘ１、ｘ２｝」と、表す。また、ソースＳでｎ番目（先頭から走査して、意味情報が生成された順番）に生成された意味情報ｘを「ｘ【Ｓ，ｎ】」と、表す。

すでにｎ個の意味情報グループが並べられている時、（ｎ＋１）個の挿入位置が考えられる。挿入位置を決める処理の流れを以下に示す。

（１）「ソースにおいて前か後ろか」と、いう情報のみを用いて挿入位置を決める。これは、ソース毎に意味情報量が異なるので、意味情報量の多いソースの影響が支配的にならないようにするためである。

（２）前ステップで決まらなければ、「ソースにおいてどれだけ前か後ろか」と、いう情報を用いて挿入位置を決める。

（３）前ステップで決まらなければ、情報量（文節量とする）の多い意味情報グループが前にくるように挿入する。

例８：意味情報グループＸ｛ｘ１、ｘ２、ｘ３｝と、意味情報グループＹ｛ｙ１、ｙ２｝の前後関係を比較する。

ｘ１【Ｓ、ｓ】、ｘ２【Ｔ、ｔ】、ｘ３【Ｕ、ｕ】、ｙ１【Ｓ、ｓ＋３】、ｙ２【Ｔ、ｔ－１】と、表現できるとする。

（１）比較可能な意味情報はｘ１とｙ１、ｘ２とｙ２である。ｘ１とｙ１を比較すればＸの方が前、ｘ２とｙ２を比較すればＹの方が前、と、なり、多数決で決まらない。

（２）ｘ１とｙ１を比較すればＸの方が３意味情報分前、ｘ２とｙ２を比較すればＹの方が１意味情報分だけ前、と、なり、Ｘの方が前だという結論に至る。

4. 実験と考察

4.1. 予備実験

５つの新聞社のサイトから、同じ日に書かれたイラクのフセイン元大統領拘束についての記事を集め、実験を行った。

生成された意味情報グループの例を図４に、その融合結果を図５に示す。

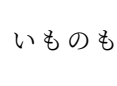
4.2. NTCIR のデータを用いた実験

NTCIR-3 SUMM[6]の文書データ集合（ＣＤ－毎日新聞[7,8]より抜粋）に対して本手法を適用し、その出力結果と NTCIR-3 SUMM の要約データとを比較した。

この文書データ集合には、同じ事柄について様々な日に書かれた記事の集合や、同じトピック（例：地震）について書かれた様々な内容の記事の集合や、同じ内容について近い日に書かれた記事の集合などがある。

4.3. 考察

4.3.1. 予備実験の考察

文節の融合については、や  など、読みにくいものも存在する。だが、  のように、形態素の違いをうまく吸収できたものも存在する。

文節のグルーピングについては、「イラクの」と「イラク北部の」がグルーピングされてしまい、混乱を招くが、６０の文節が３０の文節グループに圧縮されている事がわかる。

意味情報グループの融合については、文節グループが併記され、文節グループの違いをうまく表現できていると考える。

4.3.2. NTCIR のデータを用いた実験の考察

融合対称の記事集合内において、各記事の書かれた日が離れていたり、又は内容が異なっていた場合、異なる事象を示す意味情報をグルーピングしてしまっていた。例えば、異なるマグニチュードを並列表記した場合があった。ただ、異なる事象を一つにまとめる事で、文書集合全体の概略が理解できるとも考えられる。

近い日に同じ内容について書かれた記事の集合は、「イスタンブール」と「最大都市イスタンブール」を融合させ「最大都市イスタンブール」とだけ表示させたり、同じ「７６０万スウェーデン・クローナ」の円換算を２種類並列表示するなど、有効な例も見られた。

全体として、本手法の出力が自然な要約結果になっているとは言い難いが、出力を眺める事で記事集合の概略はつかめる。

内部で行われる各グルーピング処理の整合性については、組み合わせの数が多いため、人手による検証は困難である。よって、現在検証法を検討中である。

5. おわりに

本稿では、自然言語処理技術を用いた、分野に依存しない複数文書要約手法について提案した。

具体的には以下の事を提案した。

- 意味情報の抽出
- 文節のグルーピング

・　意味情報のグルーピング
・　文節の融合
・　意味情報の融合
・　意味情報グループの並び替え

今後の課題を以下に示す。

・　文節のグルーピング

（１）助詞などを基にしたより詳細な役割分類、（２）シソーラスを用いた表現の揺れの吸収、（３）日本語コーパスから生成された形態素のＤＦを用いた内容の一致度判定、（４）「同日」「昨年」など、時間を表す文節を絶対時間に変換したグルーピング、を行いたい。

また、文節一致度を決める形態素一致度の閾値の妥当性を確かめたい。

・　文節の融合

過去形、受け身など、助動詞を融合結果で表現したい

・　意味情報の融合

併記している文節の持つ「包含」「背反」「枚挙」「不定」等の関係を特定し、融合結果に生かしたい。

・　意味情報の並び替え

意味情報の前後関係を算出するだけでは不十分であり、意味情報の複文関係や接続節の情報などを用いた並び替えを行いたい。

・　手法全体

（１）文法が曖昧な文や口語にまで適用文書を拡張するための自然言語処理技術の洗練、（２）文書のグルーピングの自動化、を行いたい。

文　献

[1] U. Hahn, and I. Mani, The callenges of automatic summarization, IEEE Computer, vol.33, no.11, pp.29-36, Nov.2000

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

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

[6] NTCIR: http://research.nii.ac.jp/ntcir/index-ja.html

[7] CD-毎日新聞'98:毎日新聞社

[8] CD-毎日新聞'99:毎日新聞社

イラク統治評議会に　よると、　イラクを
占領下に　置く　*米軍当局は*　１４日、　米軍の
バグダッド制圧以降、　８ヶ月以上も
行方不明と　なっていた　イラクの
サダム・フセイン元大統領　（６６）を
イラク北部の　ティクリートで　生きたまま
拘束した

米軍部隊は　元大統領を　拘束したが

イラク統治評議会に　よると、　イラクを　占領下に
置く　*米軍当局は*　１４日、
サダム・フセイン元大統領　（６６）を
イラク中部ティクリートで　生存したまま
拘束した

複数の　情報を　統合すると、　*米軍は*
ＰＵＫなどの　民兵組織とともに、　元大統領の
故郷　北部ティクリートの　隠れ家を　急襲、
銃撃戦の　末、　拘束したという

イラク統治評議会に　よると、　イラクを
占領する　*米軍当局は*　１４日、　イラクの
サダム・フセイン元大統領　（６６）を
イラク北部の　ティクリートで　拘束した

図４．意味情報グループ例
（表記法は例４と同様）

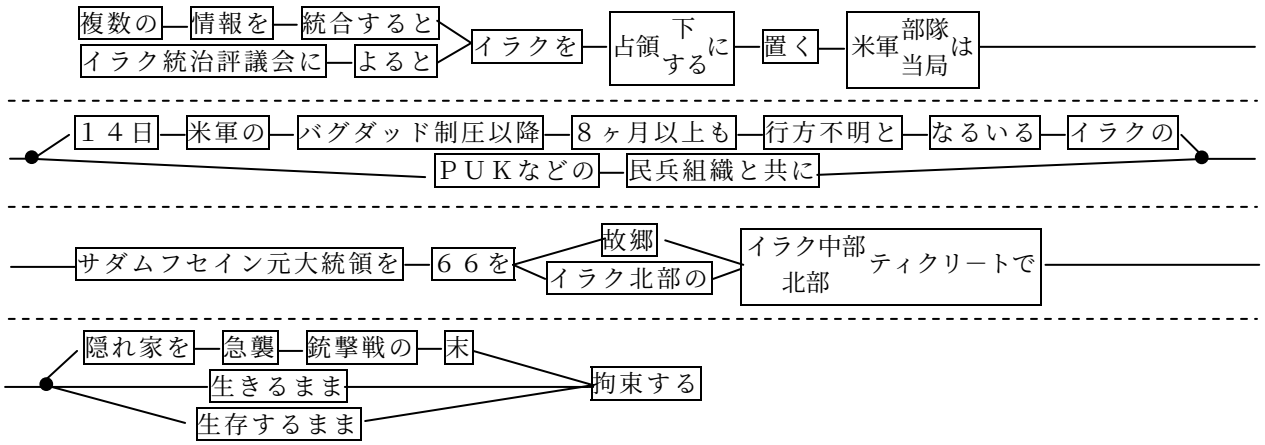


図５．意味情報グループ融合例
（文節は線で囲い、「行間」は点線で示す。）