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研究会(第9回)
SIG-AM-09-03

照応解析と動詞シソーラスに基づく ニュース概要把握のための図解生成システム

Generating Illustrated Diagram to Support Understanding of News Summaries Based on Anaphora Resolution and Verb Thesaurus

廣田 暖貴[1*]　白松 俊[1]　岩田 彰[1]

Haruki Hirota[1], Shun Shiramatsu[1], Akira Iwata[1]

[1]名古屋工業大学 大学院工学研究科
[1]Graduate School of Engineering, Nagoya Institute of Technology

Abstract: The present study is aiming to automatically generate news summaries using illustrated diagram that enables users to understand an overview of the news. To generate illustrated diagram, zero anaphora resolution is needed because the elements of the statements are often omitted in Japanese sentences. In this study, we propose a method for zero anaphora resolution focusing on human nouns that appear in diagram based on the centering theory, and a method for hierarchical management of illustrated diagram using the verb thesaurus. We conducted an experiment to compare summaries generated by the system and summaries of existing conventional services. The experimental result indicated that the illustrated diagram is useful to understand the overview.

1　はじめに

本研究ではニュース記事を入力とし，テキストと図を複合的に用いた図解を要約として自動生成する．これにより，ニュースの概要把握を支援するシステムの開発を目指す．しかし，日本語の文章では，主語や目的語など文の要素が省略されることが多く，テキストを図解に変換するためには省略された要素を特定する必要がある．また，直感的な図解を生成するためには，動詞に適した表現をする必要があるが，動詞ひとつひとつに図を登録すると管理コストが膨大になってしまう．

そこで本研究では，図解に出現する格要素および人や組織の名詞に着目したゼロ代名詞補完手法と，動詞シソーラスを用いた図解の階層管理手法を提案する．また，提案手法を用いて生成した要約について評価実験を行い，要約での図解の有用性を確認する．

図解を用いない一般的な自動要約では，情報のソースを受け取り，そこから内容を抽出し，最も重要な内容をユーザに，簡約した形で，かつ，ユーザやアプリケーションの要求に応じた形で掲示する[1]．

*連絡先：　名古屋工業大学大学院工学研究科
創成シミュレーション工学専攻
〒466-8555 愛知県名古屋市昭和区御器所町
E-mail: 25413561@stn.nitech.ac.jp

表 1：テキストと図の性質比較

性質	テキスト	図
概要の一覧性	△	◎
詳細な記述力	◎	△
理解形態	ボトムアップ処理	トップダウン処理

原文書に含まれる情報を短時間に理解できることが求められるため，読み手にとってわかりやすく表現することが重要である．既存の自動要約サービスで用いられているメディアであるテキストは，詳細な意味や種々の抽象概念を表現できるという利点をもつ．しかし，構成される要素から全体へと理解していくため，直感性・概覧性に欠け，理解に時間を要するといった欠点を持っている．表 1 に，テキストと図の性質を示す．この問題は，表現している情報の内容概略を直感的に把握することができる図を複合的に用いることによって克服できると考えられる．

2　関連研究

すでに，テキストから図的メディアを生成する研究[2]，[3]が行われている．関連研究[2]では，物語テキストに含まれる各発話文についての話し手と聞き手を同定し，その会話の中身から登場キャラクタ同士の関係を推定し，関係図の自動構築を行っている．本研究では物語中に含まれる会話に限定することな

く，Web ニュースのテキストを対象とし，構文・格解析により人物の関係を図にしている点で異なる．関連研究[3]では，会議などの会話内容のテキストを入力として，ノードとエッジからなる DT-MAP と呼ばれるグラフを作成し，会話内容を表現している．本研究では，テキストを並べたグラフの作成ではなく，図とテキストを複合的に用いることで，要約を生成している点で異なる．

本研究では，これらの関連研究とは異なる着眼点として，図解のタイプを定義して動詞シソーラス上で継承させて管理できるようにした点と，ゼロ代名詞の補完の際に図解生成に必要な格を考慮した点に注力して研究した．

3 図解生成手法

3.1 図解の定義

テキストと図を複合的に用いることによって文書内の情報を効率的に提示することができるが，図を用いた表現には大きな自由度があり，内容が十分に伝わる形式を選ぶことが必要となる．そのために，図解としてどのようなものを生成するべきかを考慮する．ニュースでは「首相が記者団に発表する」や「王子が日本を訪問」など,「人物」が「対象」に「何かをした」といった記事が多く見られる．そこで，動作の内容や動作の主体と客体を図化すると，テキストのみによる表現よりも読みやすい要約となり，概要把握の補助になると考えた．

そのためには,「誰が」,「何を」,「誰に」,「どうした」といった図解を生成する必要がある．ここで,「誰に」という動作の客体が存在しない場合もあるため，客体をもたない図解と，客体をもつ図解に分けて生成する.図 1 のように,それぞれの図解の type を single，pair と定義する．

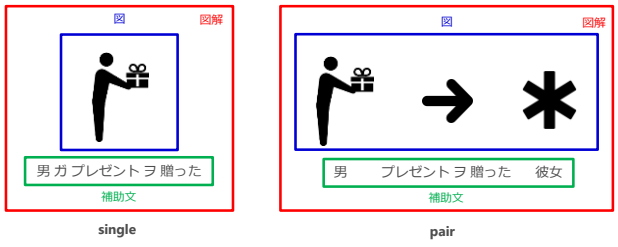

図 1：図解の type と各部の名称

図解に用いる図は，補助文に含まれる動詞に適したピクトグラムを人手で選択している．

ニュースの内容を表現する図解を生成するためには，次のような流れの処理が必要となる．

(1) ニュースを端的に表す補助文の生成
(2) 動詞に適した図の生成

3.2 補助文の生成

補助文を生成するには，ニュース記事から動詞を抜き出し，その動詞を基に図を説明する補助文を生成する．動詞を中心とした文の構造を把握するために，述語項構造解析を行う．述語項構造は，述語と項（述語と格関係にある単語）を同定するものである．動詞・形容詞などの「述語」は，文の中心で動作・状態を表す要素である．そして「項」(名詞＋格助詞）は述語が表す事態に関係する人，ものを表現する要素である．述語項構造を用いることで文中の各述語が表す意味を補う働きをする項を同定し，文の意味の骨格を表すことが可能となる[4]．本研究では既存の解析ツールである KNP[5]を使用することで，文の構造を取得している．表 2 は「太郎は学校へ行ってサッカーをした。」という文の解析結果の例である．この例文には「行く」と「する」の二つの動詞がある.「行く」という動詞は，ガ格に「太郎」，ヘ格に「学校」が相当する．一方,「する」という動詞は，ヲ格に「サッカー」が相当する．しかし，ガ格が取れていないため，これを図解にすると，主語のない意味のわからない図解を生成してしまう．そのため，ガ格に「太郎」という名詞を補完するために，述語項構造解析と同時に照応解析を行う必要がある．

表 2：述語項構造の例

	ガ格	ヲ格	ヘ格
行く	太郎		学校
する	φ	サッカー	

照応とはある表現が同一文章内の他の表現を指す機能をいい，指す側の表現を照応詞，指される側の表現を先行詞という．日本語の場合は述語の格要素の位置に出現している照応詞が頻繁に省略される．この省略された格要素をゼロ代名詞（記号 φ で表す）といい，ゼロ代名詞と照応関係となる場合をゼロ照応と呼ぶ[6]．このため，ニュース記事のような自然言語文をそのまま用いると，主体や客体が抜けたわかりづらい図解を生成してしまうという問題がある．

3.3 図の生成

文の内容を直感的に理解できる図解を生成するには，動詞の意味に合った図を生成する必要がある．しかし，図の種類が多くなると，膨大な数の動詞ひとつひとつに，意味に適した図を人手で登録することになり,管理コストが膨大になってしまう.逆に,図の種類が少なく，同じような図が繰り返し用いられていると，動詞の意味に適した図解にならず，理解の妨げになってしまうことや，直感的な理解に役

立たない問題がある．

3.4 研究目的

3.2 節と 3.3 節から，ニュース記事の内容を端的に表す図解を生成するためには二つの課題を解決する必要がある．

(1) ニュース記事をそのまま入力として与え格解析をするだけでは，ゼロ代名詞が頻繁に出現するため，主体や客体の抜け落ちた図解を生成してしまう

(2) 動詞の意味に合った図が少なすぎると直感的に理解できる図を生成できず，逆に多すぎると図と動詞を対応づける管理コストが膨大になる

本研究は，この二つの課題を解決し，生成物の要約としての有用性を確認することを目的とする．

そこで，本研究ではセンタリング理論を応用し図解出現要素に着目したゼロ代名詞補完手法と，動詞シソーラスを用いた動詞に対応する図の階層管理手法を提案する．

4 提案手法

4.1 ゼロ代名詞補完手法

センタリング理論は英語の代名詞の照応関係を決定する手法として Grosz[7]らによって提案され，大規模な知識を必要とせず，計算機上で実現容易であるなどの利点を持つ．文の中心になっているものをセンターと呼び，談話中でセンターが連続している場合，つまり話題が連続している場合には代名詞が使われているはずである，という基本規則を利用して照応解析を行っている．本研究では特に，図解に現れる格要素や人物に特化した．

4.1.1 センターの定義

談話単位中の各発話 U には，前向き中心(forward-looking-center) $C_f(U)$ と後向き中心(backward-looking-center)$C_b(U)$が結びついている[8]．C_fは発話U_iで実現される名詞リストを次発話U_{i+1}での参照されやすさで並べたもので，C_fのうち現在の話の中心になっている特別な要素がC_bである．C_fの要素は次のランキングで順序付けられる．

主題 ＞ ガ格 ＞ 二格 ＞ ヲ格 ＞ その他

4.1.2 センターの制約

発話列$U_1, \ldots, U_m$からなる談話単位中の各発話U_iについて，以下の制約が成り立つ[8]．

(a) ただ一つの$C_b(U_i)$が存在する．

(b) $C_f(U_i)$のあらゆる要素はU_iで実現されている

(c) $C_b(U_i)$は，$C_f(U_{i-1})$の要素のうちU_iで実現されているものの中で，$C_f(U_{i-1})$での序列が最も高かったものである．

4.1.3 ゼロ代名詞補完規則

センタリング理論に基づく，図解出現要素に着目したゼロ代名詞補完手法について説明する．KNP の出力する格をすべて必須格と考えてゼロ代名詞の補完を行うと，補完対象の誤りが頻繁に発生してしまう．そのため，ゼロ代名詞の補完の際に，図解生成に必要な格(図解出現格)を考慮し，ガ格，二格，ヲ格のみを補完対象とした．補完の規則として以下のとおり定めた．

(a) 図解出現格のみを補完対象とする

(b) 補完対象格の優先順序は以下の通りである

ガ格 ＞ 二格 ＞ ヲ格

(c) 述語項構造解析結果に，補完する語が含まれている場合はC_fの序列が次に高いものを補完する

4.1.4 提案手法の適用例

提案手法の適用例について説明する．

A) 彼は彼女に花を贈りました。

B) φ ガ とてもドキドキしました。

表 3：センターの遷移

	C_b	C_f
(A)	彼	彼<主題，人>，彼女<ヲ格，人>，~~花<二格，植物>~~
(B)	彼	彼<ガ格，人>

表 3(A)のように人の名詞にのみ着目すると，「花」が除外される．「彼」は「彼女」より序列が高いため，(B)のゼロ代名詞には「彼」が補完される．

4.2 階層管理手法

シソーラスを用いることで，階層構造で管理できるため，上位の動詞に登録されている図を下位の動

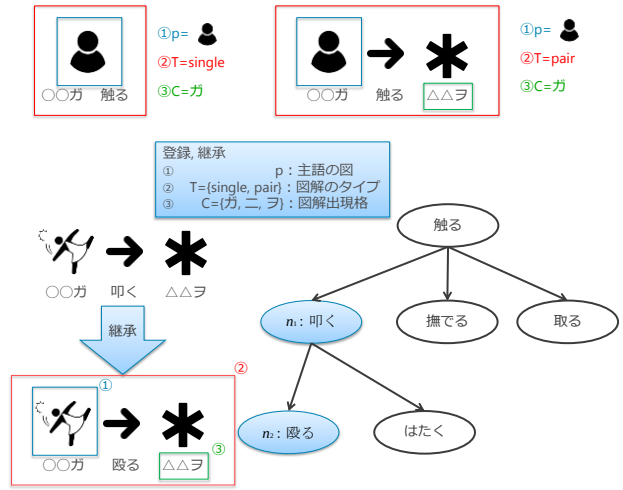

図 2: 階層管理手法

詞に継承することができる．階層管理手法の例を図 2 に示す．ある動詞に対して図が登録されている場合はその図を出力に用い，登録されていない場合は上位の動詞を参照し，上位の動詞に図が登録されている場合はその図を継承し，出力に用いる．

提案手法は以下の式で表すことができる．あるノードnの図をd(n)，タイプをtype(n)とし，図および図解のタイプの登録はそれぞれ以下の式で表現する．

$$\mathrm{reg\ d(n)} = \mathrm{p},\ \mathrm{reg\ type(n)} = (T, C)$$

上位ノードを$\mathrm{sup}(n_2) = n_1$と表すとき，type の継承は以下の式で表す．

$$\mathrm{type(n)} = \begin{cases} \mathrm{reg\ type}(n) & if\ \mathrm{type}(n) \neq \varphi \\ \mathrm{type(sup}(n)) & if\ \mathrm{type}(n) = \varphi \end{cases}$$

図の継承も同様に，以下の式で表現される．

$$\mathrm{d(n)} = \begin{cases} \mathrm{reg\ d}(n) & if\ \mathrm{d}(n) \neq \varphi \\ \mathrm{type(d}(n)) & if\ \mathrm{d}(n) = \varphi \end{cases}$$

次章より，提案手法を用いた要約生成処理について述べる．

5 要約生成処理

提案手法を用いた要約生成処理について説明する．簡単な処理の流れを図 **3** に示す．

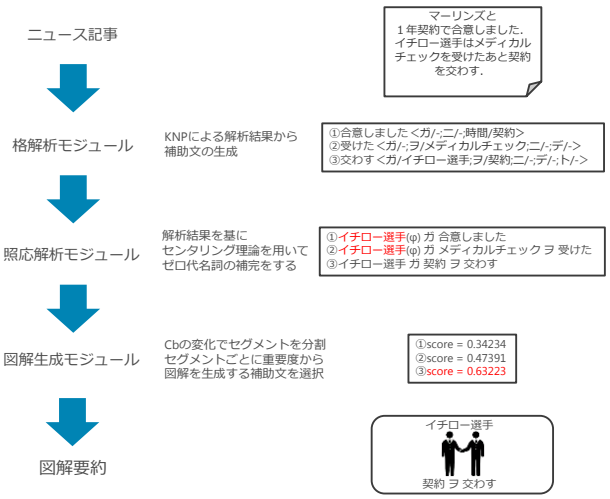

図 3: 処理の流れ

最初に入力テキストであるニュース記事を文単位に分割して，一文ずつ KNP で述語項構造解析を行う．次に，解析結果を用いて補助文を生成する．そして，生成した補助文のゼロ代名詞に人，組織・団体の名詞を補完する．続いて，補助文の重要度を計算し，セグメントごとの生成する図解を決定する．図解に用いる図は 4.2 節で述べた手法で決定する．生成する図解は図 4 のようなものになる．現在，補助文は簡易的なものを生成しているため，着目している動詞を含む一文を図解に並べて表示している．これをニュース記事全文に対して繰り返し行うことにより複数の図解を生成し，ニュース記事全体の要約を生成する．

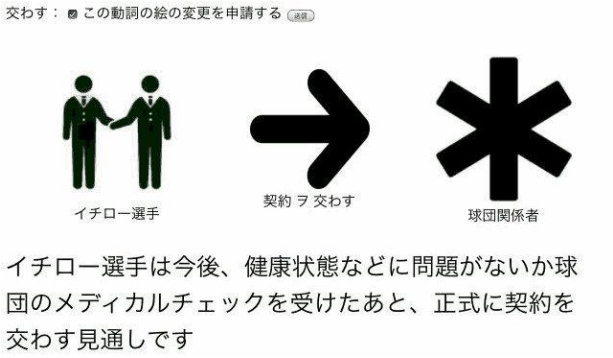

図 4：要約に用いる図解とニュース文

5.1 重要文選択

短時間で概要を把握するには，ニュースの中から要約に相応しい文を選び出す必要がある．重要文選択は以下の二つの仮定に基いて行っている．

1. C_bが変化するとセグメントを分割可能
2. TF-IDF の総和が大きい文が重要文

重要文選択の例を表 4 に示す．

5.1.1 C_bの変化によるセグメンテーション

本研究ではC_bは常に人や組織の名詞である．つまり，話の中心人物が変化すると，セグメントを分割する．セグメンテーションの例を表 4 に示す．各発話中，C_bとなっている要素を下線で示している． この例では三つのセグメントに分割される．

5.1.2 TF-IDF 重み付け

TF-IDF は，文書中の単語の重みとして広く使用される尺度である．単語の文書内での頻度を表す TF と，単語が現れる文書数に基づき語の珍しさを表す IDF を掛けあわせた尺度であり，その値が大きいほど各文書を特徴付ける語だと言える．ここでは TF-IDF の総和で文の重要さを表すという単純な手法をとる．すなわち，語wの TF-IDF 値をtfidf(w)で表す時，語$w_1, w_2, \cdots w_m$から成る文sの重要度は以下の式で求める．

$$score(s) = \sum_{i=1,\cdots,m} \mathrm{tfidf}(w_i)$$

表 4 の文(3)の場合，重要度は以下のように求める．

$$score(s_{(3)}) = \mathrm{tfidf("ジェリー")} + \mathrm{tfidf("頭")} + \mathrm{tfidf("ネズミ")}$$

表 4：重要文選択例

発話	score
(1)トムはいつもジェリーを追いかけています	0.123456
(2)ジェリーはトムと同じ家に住んでいます	0.345678
(3)ジェリーは頭のいいネズミです	0.678912
(4)飼い犬のスパイクとジェリーは仲良しです	0.456789
(5)スパイクはトムに仕返しをします	0.567891

このように(2), (3), (5)が各セグメント内で最も重要な文と考え，図解を生成し要約に用いる．

5.2 図と図解出現格の登録

シソーラスには日本語 WordNet[9]を利用した．日本語 WordNet は大規模な語彙データベースであり，語を類義関係のセット（synset）でグループ化している点に特徴があり，一つの synset が一つの概念に対応している．また，各 synset は上位下位関係などの多様な関係によって結ばれている．この synset に対して図と図解出現格を登録する．

ユーザが要約の図解を見て，わかりづらいと感じた場合，図 5 の中にあるチェックボックスにチェクを入れることで, 図の変更を申請することができる．図 5 は管理者の図と図解出現格登録画面の一部である．この画面では，ユーザがわかりづらいと感じた動詞の現在の図と，その一文が表示されている．管理者はその文を見て動詞に適した絵と図解出現格を選択して登録する．

図 5 : 図と図解出現格の登録

6 評価実験

本実験の目的は，提案手法を用いて生成する要約について，図解の有用性および既存サービスの生成する要約との比較による優位性の確認である．被験者として学生 6 名を対象にアンケート実験を行った．

比較する要約は，既存サービス，提案手法（自動解析），提案手法（手動解析）の三つである． KNP やセンタリング理論による解析は，生成する要約の品質に大きな影響を与える．つまり，解析の誤りにより図解生成に誤りが生じることがあり，ニュース記事の自動要約における図解の有用性が確認できない可能性がある．そのため，提案手法（自動解析）が出力した結果において誤りを含む解析結果を，人手で修正した解析結果で図解を生成したものが提案手法（手動解析）である．また，既存サービスには，自動で記事を 3 文に要約するニュースサービスである SLICE NEWS[10]を用いた．

6.1 実験の手順

実験は以下の手順で行った．

(1) テキストのみのニュース記事を提示
(2) 要約を被験者に提示
(3) 評価を用紙に記入
(4) 手順(2)と手順(3)を残りの要約に対して同様に行う
(5) 手順(1)〜(4)を提示する要約の順序を変え，3 つのニュース記事に対して行う

要約を読むたびにニュースの内容を把握していくため，順序が後になった要約の評価のスコアが高くなってしまうことを避けるため，手順(5)のように，ニュース記事ごとに提示する要約の順序を変えて実験を行った．また，本システムの想定する利用形態である，スマートフォン（iPhone 5S）を用いて要約の提示を行った．

評価項目は，二つの基本的な評価軸ごとに四つの項目を設定した合計八つであり,「とてもそう思う（5 点)」「そう思う（4 点)」「どちらともいえない 3 点」「あまりそう思わない（2 点)」「全く思わない（1 点)」の 5 段階で評価を行った．(▼は逆転項目の意．)

- 内容的品質: 現文書の内容を適切に反映した要約になっているか
  - ① 文章表現が適切である
  - ② 必要な情報が省略されている▼
  - ③ 同じ情報が繰り返されている▼
  - ④ 無関係な情報が含まれている▼
- 読解的品質: 読みやすい要約になっているか
  - ⑤ 読みやすい
  - ⑥ 登場人物の関係がイメージしやすい
  - ⑦ すぐに概要を把握できる
  - ⑧ ほしい情報がすぐに見つかる

これらの評価項目の合計点の平均スコアを求め評価を行う．

6.2 実験結果・考察

評価実験の結果を図 6 に示す．

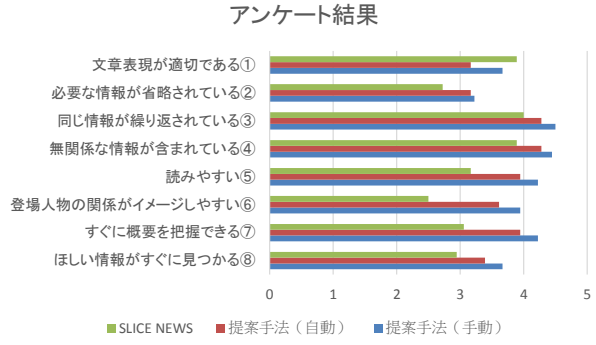

図 6: 評価実験結果

まず，特に手法間の差が顕著に見られた①，⑥，

⑦の項目について考察を述べたのち，全体の考察を述べる．

① 文章表現が適切である

八項目の中で唯一既存サービスが最も高いスコアを示した項目である．理由として，要約を生成する手法の違いが挙げられる．既存サービスはニュース記事を文に分解し，要約として相応しい文を選び，それらを繋げることで作る抽出的要約であり，作文は行っていない．それに対して，本システムでは述語項構造解析と照応解析の結果から，基本句単位に分解した後に作文や図解生成を行う生成的要約である．そのため，解析の誤りや作文の誤りによって誤りのある図解，すなわち要約が生成されたためであると考えられる．

⑥ 登場人物の関係がイメージしやすい

既存サービスは，元のニュース記事から重要な文を抽出し，それを繋ぎ合わせて要約を生成している．そのため，登場人物の関係を把握することに関しては工夫がなされておらず，元のニュース記事を読む場合となんら変化はない．それに対して，本システムでは二者間の関係を表す図解を複数生成しているため，既存サービスよりも高いスコアが得られたと考えられる．また，自動解析よりも手動解析のほうが高いスコアを得られた理由として，ゼロ代名詞補完の誤りによって自動解析では主体や客体の誤った人物関係を出力しており，イメージのしやすさの妨げとなっていたことが考えられる．

⑦ すぐに概要を把握できる

項目②の結果と合わせて本システムは，必要な情報を省略することなく，概要を把握しやすいという結果が得られたことから，図解が概要把握の手助けの一因になっているということが考えられ，自動要約における図解の有用性を確認することができた．

全体の結果では，項目①以外で，既存サービスよりも提案手法(自動)が，提案手法(自動)よりも提案手法(手動)が高いスコアを得られた．したがって，既存のテキストのみによる要約よりも本システムの図解を用いた要約は読み手にとってわかりやすい要約であったといえる．

また，提案手法(自動)よりも提案手法(手動)が高いスコアを得られたため，解析結果が要約品質に大きく影響することがわかった．述語項構造解析は KNP の結果に依存しているため，他手法によるゼロ代名詞補完は今後の検討課題である．

7 おわりに

本論文では，図解に出現する格要素および人や組織の名詞に着目したゼロ代名詞補完手法と，動詞シソーラスを用いた図解の階層管理手法を提案した．提案手法を用いてニュースの要約を自動生成し，評価実験を行った．

提案手法の生成する要約は，既存のニュース自動要約サービスの生成する要約と比較して，アンケート評価において内容的品質および読解的品質ともに高いスコアを得ることができた．特に，「登場人物の関係がイメージしやすい」，「すぐに概要を把握できる」の項目において顕著に見られた．また，既存のサービスと提案手法の生成する要約は同程度の文章量でありながら，提案手法のほうが概要を把握しやすいという結果が得られたということから，図解が概要把握に有用であることが確認できた．

今後の課題としては，ゼロ代名詞の補完精度を向上させるために，センタリング理論以外の機械学習による手法の検討を行う予定である．

謝辞

本研究の一部は，JSPS 科研費若手研究(B) (No.25870321) の助成を受けた．

参考文献

[1] Mani, I.: *Automatic Summarization*, John Benjamins Publishing (2001)

[2] 神代大輔，高村大也，奥村学: 物語テキストにおけるキャラクタ関係図自動構築，言語処理学会第 14 回年次大会発表論文集, Vol. 14, pp. 380-383 (2008)

[3] 二宮和弘，岡田信一郎，後藤寛幸，藤原祥隆: コミュニケーション支援のための会話内容の図式化ツールの開発，電子情報通信学会技術研究報告．TM，Vol. 104, No. 567, pp. 37-41 (2005)

[4] 岸邊賢太，横山晶一，井上雅史: 形容詞、複合動詞を扱う述語項構造解析システム，平成 22 年度第 6 回情報処理学会東北支部研究会，資料番号 10-6-B3-4, (2010)

[5] KNP, http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[6] 飯田龍，小町守，井之上直也，乾健太郎，松本裕治: 述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から，自然言語処理，Vol. 17, No. 2, pp. 25-50, (2010)

[7] Grosz., B. J., Joshi., A. K., and Weinstein, S.: Providing a Unified Account of Definite Noun Phrases in Discourse, In *Proc. of the 21st Annual Meeting of the Association for Computational Linguistics*, pp. 44-50 (1983)

[8] 佐竹正臣: 新聞記事の固有表現を対象とした参照関係の解析，JAIST 学術研究リポジトリ，http://hdl.handle.net/10119/1558 (2002)

[9] 日本語 WordNet, http://nlpwww.nict.go.jp/wn-ja/

[10] SLICE NEWS, http://slicenews.net