Webニュース記事本文を利用した見出し文の意味具体化手法

芋野 美紗子　　吉村 枝里子　　土屋 誠司　　渡部 広一
同志社大学　理工学部，大学院　工学研究科
{mimono, eyoshimura, tsuchiya, watabe}@indy.doshisha.ac.jp

1　はじめに

近年，人間と円滑なコミュニケーションが行える知的ロボットの実現に向けて様々な研究が行われている．人間同士のコミュニケーションはその多くが会話によってなされており，将来的にはロボットに対しても人間のような会話能力が求められると考える．人間が行う会話の種類は様々であり，ロボットも多種多様の会話を行えてこそ，より人間にとって親しみのある存在になれると考える．

ロボットにより人間らしい会話を行わせるための研究の一端として，ロボット側からの能動的な話題提供について考える．人間側の発話を受けて決まった返答を行うのではなく，ロボット側から新しい話題を持った発話をすることが出来れば，それはより人間らしい会話ではないかと考える．そのような会話をロボットに行わせるためには，話題の元となるリソースが必要となる．このリソースとして，Webニュースサイトで公開される新聞記事の見出し文に着目した．Web上で公開される新聞記事からは最新の時事情報を得ることができ，そこから新たな話題を提供することが出来る．見出し文は記事の内容を端的に表しており，記事本文より会話に適した形式であると考える．しかし見出し文はその端的さ故に具体的な情報に欠けるという問題がある．例えば“オリンパス元社長，社長職復帰断念”という見出し文からは，オリンパス元社長とはいったい誰なのか，いつ断念することを決めたのかといった情報は得られない．また，新聞の見出し文には体言止めや助詞の欠落が多く存在しており，そのままでは会話のリソースとして適さない場合が多い．

そこで本稿では新聞記事の見出し文をロボットの会話リソースとして利用するため，記事本文を用いて見出し文の意味の具体化を行う手法を提案する．端的に表された新聞記事の見出し文を会話リソースに足る文へと変換することで，人間らしいロボット会話のための研究の一端を示す．

2　概要

提案手法ではWebニュースサイトから得られる見出し文に対して語句の追加や置換を行い，意味の具体化を図る．まず各新聞社のWebサイトから得たニュース見出し文に対して構造解析を行い，格および動詞を分類する．分類した格の情報と大規模格フレームを用いて助詞の補完を行うことで，見出し文特有の端的な書式からより自然な文の形へ変換を行う．さらに，記事本文中の語句による見出し文の格の置換・追加を行うことで，より意味が具体化した文を出力する．図1に具体化処理の概要を示す．

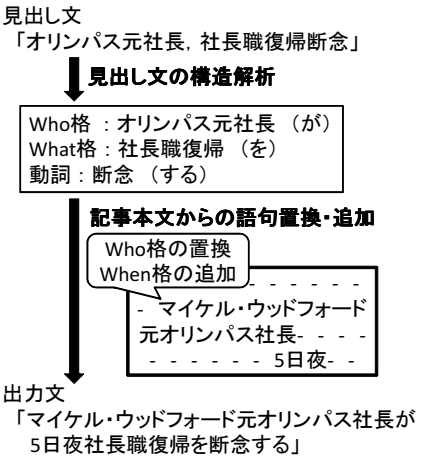

図 1: 具体化処理の概要

図1の様に，「オリンパス元社長，社長職復帰断念」という見出し文に対して構造解析による助詞の追加および本文を用いたWhen格の追加・Who格の置換を行うことで，「マイケル・ウッドフォード元オリンパス社長が5日夜社長職復帰を断念する.」の様に具体的かつ自然な文を出力する．

3 使用技術

3.1 概念ベースと未定義語の概念化手法

概念ベース [1] は複数の電子化国語辞書などの見出し語を概念として定義し，人間が持つ概念への常識的な知識をモデル化した知識ベースである．ある概念の定義は，属性と呼ばれる他の概念群と属性それぞれの重要さを表す重みによってなされる．概念ベースの具体例を表 1 に示す.

表 1: 概念ベースの具体例

概念	属性
夏	(夏場,0.34)(夏休み,0.11)(海,0.08)…
夏場	(夏季,0.25)(暑さ,0.18)(太陽,0.04)…

現在の概念ベースには 87242 の概念が定義されているが，定義されていない概念（未定義語）に関しては Web を用いた概念化手法（*AutoFeedback*：*AF*）が提案されている [2]．未定義語を用いて Web 検索を行い，その検索結果中に含まれる語を属性として付与することで未定義語の概念化が可能になる．

3.2 関連度計算方式

関連度計算方式 [3] は概念と概念の関連性を関連度とよばれる数値で定量的に表現する手法であり，その有効性が示されている．関連度は 0.0 から 1.0 の値を取り，概念間の関連が強いほど大きな値を示す．例えば概念「自動車」と概念「車」の関連度は 0.912 と非常に大きな値を示す．一方，概念「自動車」と概念「学校」の関連度は 0.012 となり，関連が薄い概念同士では小さな値となる．

関連度は概念動詞の属性の対応により算出される．互いが持つ属性の内，最も意味が近いもの同士の組を作った上でそれぞれの重みを用いて関連度を算出する．

4 提案手法

本稿で提案する手法では，まず Web ニュースサイトから取得したニュース見出し文の構造解析を行い，見出し文の分割および欠けている情報の調査を行う．次にその見出し文が示すニュース記事本文中の語句から見出し文で欠けている情報を取得・補完を行うことで，見出し文の意味の具体化を行う．

4.1 見出し文の分割とテーマの解析

「オリンパス前社長、関与認める　東京地検が任意聴取」のような見出し文では文の前半と後半で話の内容が変化している．このような見出し文に関しては前後で分割を行い，それぞれについて具体化処理を行う．ここでは見出し文中に全角空白もしくは三点リーダがある場合を分割の条件とした．また，新聞見出し独自の書式として「：」を用いたテーマの提示がある．例えば「民主党：９人離党届」という見出しは「：」の前の「民主党」が記事全体のテーマであることを示している．このような書式の場合,「：」を「に関して」という表記に置換をした上で，以降の文について具体化処理を行うこととした．

4.2 動詞の解析

見出し文において動詞がどの語句に当たるかを解析する．品詞解析には茶筅を用いるが，見出し文ではサ変名詞による体言止めの表現が多く使われているため，これらの語句は語尾に「する」を付与して動詞とみなす．また,「養成へ」「停止か」のようにサ変名詞の後ろに助詞「へ，か」が続く表現は「養成するかもしれない」のように変換を行った上で動詞と判断する．例えば「パレスチナ民兵１人死亡」という見出し文があった場合にはサ変名詞である「死亡」に「する」を接続して「死亡する」をこの見出し文の動詞と判断する．

4.3 格の解析

見出し文の格の情報を解析する．まず，前節で述べた処理で判明した動詞とその直前の語の間に助詞が存在しない場合はその補完を行う．具体的には大規模格フレームを用いて動詞と直前の語を繋ぐ助詞を検索し，最も頻度の高い助詞を選択する．例えば「パレスチナ民兵１人死亡」という見出し文は前節の処理から動詞が「死亡する」であると分かる．直前の語である「１人」と「死亡する」を繋ぐ頻度が最も高い助詞は「が」であり，よって「パレスチナ民兵１人（が）死亡（する）」という補完が行われる．この処理を行った上で，表 2 に示す分類規則に基づいて格の情報の解析を行う．なお，表中で用いている上位ノードの検索は日本語語彙体系 [4] により行った．

ここで表 2 に示した規則のうち，読点が含まれた場合の Who 格への分類が行われた際には，自然な文へ変換するために読点を助詞「が」へ置換する事とした．

表 2: 分類規則

条件	格の分類
助詞「が, は」，読点「、」が含まれる	Who
助詞「を」が含まれる	What
助詞「に」が含まれる	Whom
文節末の語の上位ノードに”場所”が存在	Where
文節末の語の上位ノードに”時間”が存在	When

例えば「パレスチナ民兵１人死亡」という見出し文は,「パレスチナ民兵/１人/死亡」という文節に分けられる．補完処理により「１人（が)」の文節は助詞「が」が含まれるため，これが Who であると分類される.「パレスチナ民兵」のようにどの分類にも含まれない文節が存在した場合には，係り受け関係にある「１人（が)」の文節への接続を行う．この場合,「パレスチナ民兵１人（が)」をまとめて Who と判断する．

4.4 動詞の追加

解析において，動詞に分類される語句が存在しなかった場合には動詞の追加を行う．見出し文が示す記事本文中から動詞およびサ変名詞を全て取得し，その中から見出し文に適した語句を選択する．見出し文末尾の語と記事本文中の動詞・サ変名詞全てとの組み合わせで格フレームを検索し，接続可能な動詞の中で最も頻度の高い語を適切な動詞みなす．例えば「浦和東・菊池が３発」という見出し文の場合，末尾の語「発」と本文中の動詞「決める」が格「を」によって接続される組み合わせの頻度も最も高い．よって「浦和東・菊池が３発を決める」のように動詞を追加する．

4.5 When・Where 格の追加

解析において，When もしくは Where 格の語句が存在しなかった場合は When・Where 格の追加を行う．記事本文中の語句の上位ノードを調べ，その中に「時間」が存在する場合は When 格,「地名」または「場所」が存在する場合は Where 格に追加する．複数の語が該当する場合は，本文中に最も早く出現した語を追加する．また，When 格に関しては”…月…日”の様な具体的な日時が存在する場合には，そちらを優先して追加する．この場合は数字の後ろに月もしくは日がある部位を表記一致により検索し判断する．

4.6 Who 格の置換

解析において Who 格の語句が得られた場合に，記事本文中の語句を用いて置換を行うことで見出し文における主体の意味具体化を図る．本文中の全語句を置換の候補とし，そこから最も Who の置換に適切な語句の選択を行う．

まず，動詞解析もしくは動詞の追加により得られた見出し文の動詞と，Who 格および置換候補語句それぞれとの共起ヒット件数を Web 検索により取得する．このとき「置換候補語と動詞」の共起ヒット件数が「Who 格と動詞」の共起ヒット件数と比べてあまりにも小さい場合にはその候補語が置換に適さないと判断できる．本稿では「置換候補語と動詞」の共起ヒット件数が「Who 格と動詞」の共起ヒット件数の一割に満たない場合は候補語句から除外した．

次に，置換候補語句と Who 格との関連度を算出する．3.2 節で述べたとおり，関連度の算出はそれぞれの概念が持つ属性と重みが必要となる．しかし新聞記事中の語句の多くは固有名詞や複数語句の集合などであり，概念ベースに定義されていない未定義語である場合が多い．そこで，未定義語に対しては AF を利用してこれらの語句の概念化を行った上で関連度の算出を行う．置換候補語句と Who 格との関連度が高いほど，それらの語句の間の関連性が高く置換に適していると判断できる．また，この関連度に下限の閾値を定めることで関連が無い語句による置換が行われないようにする．本稿では下限値を 0.1 と設定した．図 2 に Who の置換処理の具体例を示す．

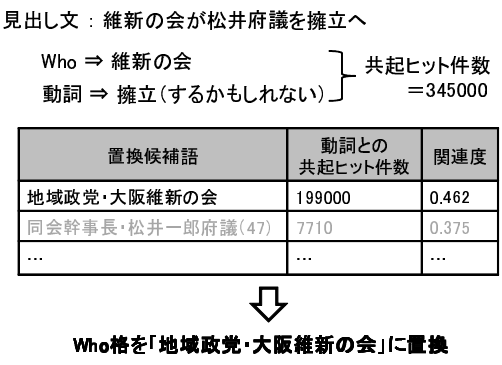

図 2: Who の置換処理の具体例

「維新の会が松井府議を擁立へ」という見出し文から，Who 格として「維新の会」，動詞として「擁立(するかもしれない)」という語句が得られる．これらの語句の共起ヒット件数は 345000 件となった．次に

表 3: 意味具体化結果

見出し文	意味具体化語
オリンパス前社長、関与認める　東京地検が任意聴取	菊川剛・前社長（７０）が２１日の 捜索前関与を認める． 東京地検特捜部が任意聴取する．
那覇西が初戦敗退	那覇西が 第９０回全国高校サッカー選手権大会第２日 埼玉県の埼玉スタジアムで初戦敗退する．

記事本文中から得られた候補語である「地域政党・大阪維新の会」と「同幹事長・松井一郎府議」それぞれの語句と動詞との共起ヒット件数を調査すると，後者の候補語句のヒット件数が 345000 件の一割に満たないため，候補から外れる．残った候補語句との関連度は 0.462 となっており，これは下限値 0.1 より大きい．よってこの候補語句と Who 格の間に十分な関連があると見なして置換を行う．結果として Who 格は「地域政党・大阪維新の会」に置換される．

5　評価

評価は Web ニュースサイトから各 40 文，計 120 文の見出し文を取得して行う．この見出し文 120 文と提案手法により意味の具体化を行った出力文のセットを被験者 3 名に提示し，出力文が元の見出し文の意味を具体化した上で日本語として不自然でないかの判断を行った．2 名以上が出力文を妥当とした場合に正解とする．また，不正解となった出力に関しては「構造解析の失敗」「Who 格の置換」「When 格の追加」「Where 格の追加」「動詞の追加」の内，どの部分に不備があるかを調査した．図 3 に評価結果を示す．

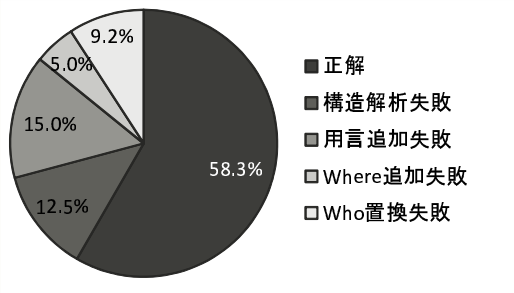

図 3: 評価結果

結果として評価文全体の 58.3%の見出し文について，日本語として自然な形で意味の具体化を行うことが出来た．表 3 に意味具体化の例を示す．

6　おわりに

本稿では人間らしい会話を行えるロボットのための研究の一端として，Web ニュースサイトの記事見出し文の意味具体化手法について提案した．記事本文から語句を補完することで端的な表現の見出し文に具体的な意味を付与した．*AF* による未定義語の概念化および関連度計算方式を用いることで，見出し文中の語句と記事本文中の語句の関連性を配慮した語の置換が可能となった．結果として評価に用いた見出し文 120 文の内，58.3%の見出し文について意味の具体化を行うことができた．これにより，ロボットと人間の自然な会話のための研究の一端を示せたと考える．

謝辞

本研究の一部は，科学研究費補助金（若手研究（B）24700215）の補助を受けて行った．

参考文献

[1] 奥村紀之, 土屋誠司, 渡部広一, 河岡司. 概念間の関連度計算のための大規模概念ベースの構築. 自然言語処理, Vol.14, No.5, pp.41-64, 2007.

[2] 辻 泰希 渡部 広一 河岡 司. www を用いた概念ベースにない新概念およびその属性獲得手法. 第 18 回人工知能学会全国大会論文集, 2D1-01, 2003.

[3] 渡部 広一 奥村 紀之 河岡 司. 概念の意味属性と共起情報を用いた関連度計算方式. 自然言語処理, Vol.13, No.1, pp.53-74, 2006.

[4] NTT コミュニケーション科学研究所. 日本語語彙体系. 岩波書店, 1997.