E-022

修飾関係を利用してニュース記事を表形式に要約する手法
The Method for Summarizing News Articles in Table Format Using Modified Relationships

中西 隆博† 吉村 枝里子‡ 土屋 誠司‡ 渡部 広一‡
Takahiro Nakanishi Eriko Yoshimura Seiji Tsuchiya Hirokazu Watabe

1 はじめに

近年, 情報技術の発展によりユーザは大量の情報を入手可能となった一方で, 求める情報を的確に選択することが困難となっている. 情報を的確に選択する手段の一つとして表形式による要約が挙げられる. 表形式で要約することで, 文章形式での要約と比べて複数の情報を比較しやすくなり, 適切な情報を選択しやすくなると考えられる. よって, 大量の情報を扱う際には表形式での要約の方が適していると考えられる. そのため, 本稿では表形式での要約を扱う. 表形式での要約手法は西口らによって「ニュース記事の表形式要約」手法[1]が提案されている. この手法において, 項目に格納される語は 3.2.3 節で述べる初期名詞と呼ばれる名詞と修飾関係にある語のみであり, 要約に必要な情報を項目に格納しきれていない. そこで本稿では, 修飾関係を利用して適切な情報を項目に格納する手法を提案し, 西口らの既存手法と比較し, 評価を行う.

2 関連技術

NTT シソーラス

NTT シソーラス[2]とは, 単語の意味や概念を分類, 整理して階層的に表したものである. NTT シソーラスには一般名詞の意味的用法を表す約 2700 個のノードの上位下位関係・全体部分関係が木構造で表され, 約 13 万語のリーフが登録されている.

3 既存手法

3.1 使用する知識ベース

3.1.1 項目知識ベース

NTT シソーラスから表の項目名として適切であると目視で判断したノードの語を格納した項目知識ベースを使用する. 項目知識ベースには 2048 語が格納されている.

3.1.2 初期項目と付随項目群

ニュース記事を表現するために特に重要だと考えられる項目（初期項目）の候補と付随項目群があらかじめ分野別に目視で設定されている. 付随項目群とは初期項目以外に表に追加する項目群である.

3.2 既存手法の流れ

1 つの記事の内容を行, 複数記事の内容に共通する項目を列としてまとめて出力する. 既存手法の全体の流れを図 1 に示す.

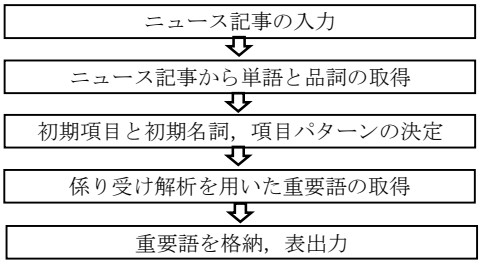

図 1 既存手法の全体の流れ

†同志社大学大学院理工学研究科
Graduate School of Science and Engineering, Doshisha University
‡同志社大学理工学部
Faculty of Science and Engineering, Doshisha University

3.2.1 ニュース記事の入力

初めに複数記事の入力を行う. 入力するのは同分野の記事とし, 各記事の見出しと本文, それらの記事の分野を入力する. 既存手法ではスポーツ, 音楽, 災害などの 16 の分野を設定している.

3.2.2 ニュース記事から単語と品詞の取得

ニュース記事に対して茶筌[3]による形態素解析を行い, 単語とそれぞれの品詞を取得する. 形態素解析の結果において, 名詞の後に名詞が続く場合は複合語と判断し, 1 つの名詞として扱う.

3.2.3 初期項目と初期名詞，付随項目群の決定

記事から取得した単語より, 初期項目に格納する初期名詞を選択する. NTT シソーラスを用い, 見出し文から取得した名詞の上位ノードを取得する. 初期項目の候補が, 取得した上位ノードに存在する場合はその名詞を初期名詞とする. 初期項目の候補が複数存在する分野の場合, 出力する表の初期項目は, 入力された記事の中に最も多く存在するものとする.

3.2.4 係り受け解析を用いた重要語の取得

南瓜[4]を用いて, 各記事の本文において 3.3 節で決定した初期名詞が修飾している語と修飾されている語を重要語として取得する.

3.2.5 重要語の格納と表出力

3.4 節で取得した重要語より, 出力する表の付随項目群に対応する語を選択し, 表を出力する. 各重要語の上位ノードをシソーラスより取得し, 項目パターン内の項目と一致するノードが存在する場合, その語を項目に対応する内容であると判断し, 表に格納する. 重要語に対応する項目が付随項目群に存在しない場合, その語の上位ノードに項目知識ベース内に存在する語があればそれを新たな項目とする. その後, 新たな項目は表の右端に生成し, その項目に対応する重要語を格納する. 図 2 に既存システムによって生成される表の例を示す.

記事 1：JR 京浜東北線で運転見合わせ…
記事 2：大阪府北部で震度 3 の地震…
記事 3：ジャカルタで洪水、3 万人避難…

⇩

災害	場所	日時	効果	害
人身事故	大宮駅		影響	
地震	大阪府北部	15 日		
洪水	ジャカルタ	16 日	影響	浸水被害

図 2 生成される表の例

3.3 既存手法の問題点

付随項目群の項目および, 項目知識ベース内のノードの中には, 既存の手法では適切に項目の中身を格納できないと考えられる語がある. 例えば, 「原因」という項目には原因の内容ではなく, 「原因」という言葉がそのまま格納されてしまう. これは, 重要語の上位ノードをそのまま項目名として使用しているために起きる.

4　提案手法

4.1　修飾関係を利用した格納手法

3.6 節の問題点に対し，係り受け解析により項目に対応する重要語と修飾-被修飾の関係にある語を項目内に格納することで正確な項目と中身を格納できると考えられる．この際，重要語の含まれる文節が修飾する文節内の名詞を重要語と置換して表の中身として格納する．

例えば，「彼の総資産は 1 億円だ」という例文において，「資産」という項目を生成する場合，「資産」というノードに対し，前述した手法を用いることで，「資産」項目に対応する重要語の「総資産」が含まれる文節が修飾する文節に含まれる名詞の「1 億円」を「資産」という項目の中に格納することができる．図 3 に提案手法の流れを示す．図 4 に例文から生成される表を示す．

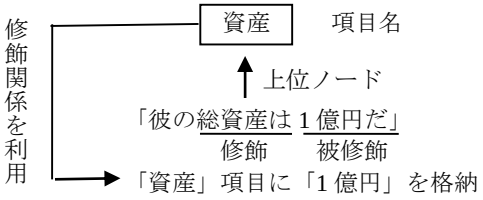

図 3　提案手法の流れ

既存手法

人物	資産
彼	総資産

⇨

提案手法

人物	資産
彼	1 億円

図 4 生成される表の例

4.2　提案手法の適用条件

4.1 節で提案した格納手法を適用する条件を述べる．茶筌による形態素解析で重要語が未知語となっておらず，かつ重要語の含まれる文節が修飾する文節内に名詞が存在する場合に提案手法を適用する．ただし，初期項目と「日時」項目は既存手法で適切に項目を格納できているため適用しない．茶筌による解析で未知語とされる語の多くは記事の内容に深く関わる固有名詞だったため，置換しない．

5　評価方法

朝日新聞社の Web ニュースサイト[5]からニュース記事の見出し文，およびその記事本文を分野ごとに 15 件ずつ，合計 240 件取得し，テストセットとして扱う．このニュース記事 240 件から既存手法と提案手法によって表を出力した後，被験者 3 人で生成した表は記事を要約していると言えるか評価を行った．生成した表とニュース記事の見出し文及び本文を比較し，表から本文の内容が理解でき，項目名と項目の中身が全て適切である場合は○，表から本文の内容が理解でき，項目名か項目の中身に不適切なものが 1 つ以上ある場合は△，表から本文の内容が理解できない場合は×とした．そして，○は 3 点，△は 2 点，×は 1 点として扱い，3 人の合計 9 点満点の内，7 点以上を○，5 点，6 点を△，4 点以下を×として評価を行った．

6　評価結果

評価結果を表 1 に示す．

表 1　評価結果

	○（%）	△（%）	×（%）
既存手法	10.8	41.3	47.9
提案手法	15.8	42.9	41.3

表 1 より，既存手法と比べて提案手法では○が 5.0%増加，×が 6.6%減少している．これより，提案手法により精度が向上したと言える．

7　考察

図 4 に産業分野の記事および既存手法と提案手法による出力結果の例を示す．

イトーヨーカ堂も大量値下げ　食品など 1 千品目
イトーヨーカ堂は 12 月 1 日から全国の約 165 店で、食品など 1 千品目を 10～40%値下げする。…

⇩

既存手法の出力結果

会社	日時	領土	食品	値下げ
イトーヨーカ堂	12 月 1 日	全国	食品	値下げ

提案手法の出力結果

会社	日時	領土	食品	値下げ
イトーヨーカ堂	12 月 1 日	165 店	1 千品目	値下げ

図 4　産業分野の記事の例

図 4 より，「領土」や「食品」項目の中身が，提案手法では詳細に表示されている．これにより本文の内容がより正確に把握できるため，精度が向上した．一方で，提案手法では，複合語が格納されている項目の項目名の約 6 割が不適切であった．提案手法では複合語が格納されている項目の項目名は複合語の前半の名詞から順に上位ノードを取得し，項目名としている．しかし，多くの複合語では後ろの語から上位ノードを取得した方が適切に項目名を決定できると考えられる．そのため，複合語の項目名を決める際にどちらの名詞からノードを取得すべきかの順番を考慮することで精度の向上が期待できる．

また，提案手法では多義語の判別が出来ない．例えば，金融分野のニュース記事において「赤字」や「黒字」という語は「利益・損害」項目に格納されるべきであるが，出力した表では「文字」項目に格納されており，不適切であると評価した．そこで，多義を持つ語である場合は，入力された記事の分野と関連性の高い語を選択し項目とすることで，精度向上が期待できる．

8　おわりに

本稿では，修飾関係を利用してニュース記事を表形式に要約する手法を提案した．修飾関係を利用して項目の中身を置換することで，既存の手法では項目に格納されなかった情報を格納できるようになり，項目数を変えることなく，詳細な内容の表を生成することが可能となった．その結果，既存手法と比べて○が 5.0%増加，×が 6.6%減少した．

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] 西口駿祐，芋野美紗子，土屋誠司，渡部広一：“ユーザの要求に応じたニュース記事の表形式要約”，情報科学技術フォーラム FIT2011，E-006，pp.207-208，2011.

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

[3] Chasen-形態素解析器，http://chasen-legacy.sourceforge.jp/，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座（松本研究室），2011.

[4] 松本裕治：“形態素解析システム「南瓜」”，http://chasen.naist.jp/chaki/t/2005-08-29/doc/

[5] “asahi.com：朝日新聞社の速報ニュースサイト”，http://www.asahi.com/