Vol.2017-ICS-186 No.2
2017/3/3

シソーラスノードへの自動割付けを用いたニュース記事の表形式要約手法

中西隆博[†1]　土屋誠司[†2]　渡部広一[†2]

本稿では，ニュース記事から，その記事の内容を理解するうえで必要だと考えられるキーワード（時間，場所，人名，出来事など）を選出し，表形式でまとめる手法の提案を行っている．既存の表形式要約手法では，シソーラスに存在しない言葉（未知語）を適切に扱うことが出来なかった．そこで，未知語をシソーラス上のノードに割付けることで，その意味を推定して項目名を獲得した．その結果，提案手法では，既存手法より精度が向上した．

Tabular summarization method of news articles using automatic assignment to thesaurus node

TAKAHIRO NAKANISHI[†1]　SEIJI TSUCHIYA[†2]　HIROKAZU WATABE[†2]

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．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]（以下既存手法とする）が提案されている．この手法において，表の項目名は 2.2 節で述べる項目知識ベース内の語から選ばれるが、新語や固有名詞などのシソーラスに定義されていない語（未知語）は項目名を適切に生成できていなかった．そこで本研究では, 未知語をシソーラス上のノードに割付けることで, 適切な項目名を獲得可能な表形式要約システムの構築を目指す.

2. 関連技術

以下に本稿で使用する技術を示す.

2.1 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，整理して用語を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示され，約 13 万語のリーフが登録されている．

2.2 項目知識ベース

項目知識ベースとは，ＮＴＴシソーラスに存在するノードから目視で判断し，適切と考えられるノードを格納した知識ベースである．「遊び」「人」「宝石」など，2048 語が格納されている．

2.3 茶筌

茶筌[3]とは，入力された文に対して形態素解析を行い，品詞を決定しシソーラスに基づく意味情報を出力するシステムである．形態素解析とは，自然言語処理技術の 1 つであり，自然言語で書かれた文を，意味を持つ最小の言語単位（形態素）の列に分割し，それぞれの品詞を判別することである.

2.4 南瓜

南瓜[4]とは，係り受け解析器の 1 つである．係り受け解析とは，文法規則により文の構造を句・文節を単位として解析することである．ここで，句とは文を構成する 2 つ以上の語の塊のことである．また，文節とは日本語を意味のわかる単位で区切ったものであり，文を読む際に自然な発音によって区切られる最小の単位である. 日本語において, 文における任意の 1 つの文節は，少なくともその文節の後の 1 つの文節と係り受け関係を持つ特徴がある．係り受けとは語の間にある修飾―被修飾の関係であり，文中のこれらの関係を解析するのが係り受け解析である.

2.5 概念ベース

概念ベース[5]とは複数の国語辞書や新聞などから機械的に構築した語（概念）とその意味特徴を表す単語（属性）の集合からなる知識ベースである．概念と属性のセットにはその重要性を表す重みが付与されている．概念ベースには，約 9 万語の概念が収録されており，1 つの概念に約 30 個の属性が存在する. ある概念 A は属性 a_i とその重み w_i の対の集合として式（1）で表される.

$$A = \{(a_1, w_1), (a_2, w_2), \cdots, (a_i, w_i), \cdots, (a_m, w_m)\} \quad (1)$$

†1 同志社大学大学院　理工学研究科
Graduate School of Science and Engineering, Doshisha University

†2 同志社大学　理工学部
Faculty of Science and Engineering, Doshisha University

任意の 1 次属性 a_i は，その概念ベース中の概念表記の集合に含まれている語で構成されている．したがって，1 次属性は必ずある概念表記に一致するため，さらにその 1 次属性を抽出することができる．これを 2 次属性と呼ぶ．概念ベースにおいて，「概念」は n 次までの属性の連鎖集合により定義されている．

2.6 関連度計算方式

関連度計算方式[6]とは，概念ベースに定義されている 2 つの概念間の関連の強さを定量的に表現する手法である．関連度は 0.0 から 1.0 の間の実数値で表され，概念間の関連が強いほど大きな数値となる．例えば概念「自動車」に対して，「車」,「学校」の関連の強さを表 1 に示すように数値化できれば，コンピュータは「自動車」と関連が強いのは「車」であるということを判断できる．

表 1　関連度計算方式の例

基準概念	対象概念	関連度
自動車	車	0.912
	学校	0.001

本節では，本稿が提案する手法で利用した重み付け手法である tf・idf[7]と SWeb-idf[7]について述べる．

2.6.1 tf・idf

tf・idf による重み付けとは，語の頻度と網羅性に基づいた重み付け手法である．文書 d における索引語 t の重み $w(t,d)$ は以下の式（2）によって得られる．

$$w(t,d) = tf(t,d) \cdot idf(t) \qquad (2)$$

$tf(t,d)$ は文書 d における索引語 t の出現頻度である．また，$idf(t)$ は検索対象文書数 N と索引語 t が出現する文書の数 $df(t)$ によって決まり，式（3）によって定義される．

$$idf(t) = \log\frac{N}{df(t)} + 1 \qquad (3)$$

2.6.2 SWeb-idf

SWeb-idf（Statics Web Inverse Document Frequency）とは，Web 上の語の idf を統計的に調べた idf 値である．まず，無差別に選んだ固有名詞 1000 語を作成する．この作成した 1000 語に対して個々に Google で検索を行い，1 語につき 10 キャッシュを取得する．よって，得られたキャッシュ数は 10000 キャッシュとなる．この 10000 キャッシュを Web の全文書空間と見なし，その中での語の出現割合を求める SWeb-idf は，式（4）で求められる．これらにより得られた語とその idf 値をデータベースに登録した．なお $df(t)$ 項は，全文書空間（10000 キャッシュ）に出現する概念 t の頻度である．

$$SWeb\text{-}idf(t) = \log\frac{N}{df(t)} \quad (N = 10000) \qquad (4)$$

3. 既存手法

既存手法では，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．まず複数のニュース記事の見出しと本文を入力する．入力されたニュース記事に対して形態素解析を行い，単語を取得する．次に分野ごとにあらかじめ複数設定した，ニュース記事から生成される可能性の高い項目の候補（初期項目候補）から，出力する表において最も重要な項目（初期項目）を選択し，初期項目に格納する名詞（初期名詞）を決定する．そして，ニュース記事から取得した単語のうち，記事と関連の強い単語を重要語とする．その後，表を生成する際の項目のパターンを決定する．そして各項目とそれに対応する重要語を表へ格納し，出力する．図 1 に既存手法の全体の流れを示す．

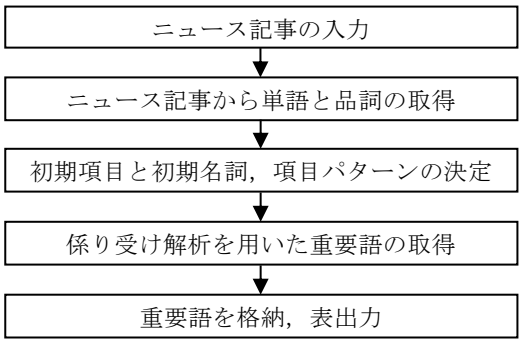

図 1　既存手法の全体の流れ

3.1 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．既存手法で扱う分野は，国際分野，政治分野，災害分野，裁判分野，産業分野，市況分野，事件・事故分野，文化分野，訃報分野，芸能分野，金融分野，音楽分野，映画分野，スポーツ分野，漫画・アニメ分野，科学分野の 16 分野である．

3.2 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による形態素解析を行い，単語を取得する．形態素解析により，各形態素の品詞も得ることができる．その結果，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．例えば，「携帯」と「電話」はそれぞれ名詞であるので，この 2 つの単語が連続した場合は「携帯電話」という 1 つの名詞として取得する．図 2 に「ＪＲ京浜東北線で運転見合わせ　西川口駅で人身事故」という文から単語を取得する例を示す．

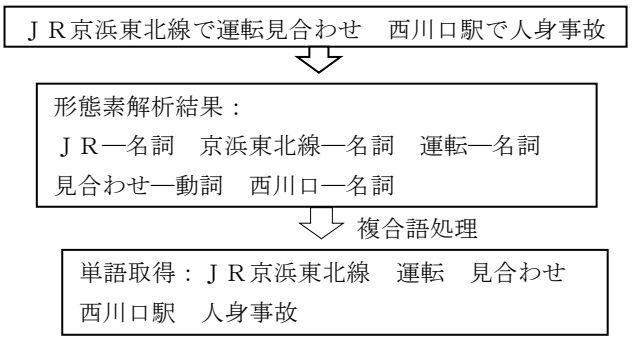

図 2　ニュース記事からの単語取得の例

3.3 初期項目と初期名詞，項目パターンの決定

既存手法ではニュース記事を 16 分野に分類しており，各分野で最も重要である項目が存在すると考えられる．よって，あらかじめ分野別に生成されると考えられる項目の候補（初期項目候補）と初期項目候補ごとの項目パターンを目視で設定する．ここで設定する項目パターンとは初期項目以外に表に追加する項目のことである．記事から生成される初期項目候補は，複数存在する分野もある．そのような分野においては，それぞれの初期項目候補に優先順位を設定する．3.2 節で取得した単語より，初期項目に格納する初期名詞を選択する．ニュース記事において，見出し文は本文の内容を文章に要約して表したものと考えられるので，初期名詞は入力された見出し文の名詞から選択する．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．初期項目候補が，取得した上位ノードに存在する場合はその語を初期名詞とする．初期項目候補は分野によっては複数存在する.そのような分野の場合，入力された複数記事の中に最も多く存在する項目を，出力する表における初期項目として決定する．そして，決定した初期項目に対応する名詞を各記事における初期名詞として決定する．図 3 に各分野とそれに対応する初期項目候補を示す．初期項目候補が複数設定されている分野においては，図 3 の左から順に優先順位が設けられている．記事ごとの初期項目候補が同数である場合，出力する表の初期項目は初期項目候補の優先順位を用いて決定する．

分野　初期項目候補
国際→「国」「地域」
政治→「人物」「選挙」
災害→「災害」「交通」「事故」
裁判→「罪」「訴訟」
産業→「会社」
市況→「株」「相場」
事件・事故→「容疑」「事故」
文化→「人物」「国」
訃報→「人物」「職業」
芸能→「人物」
金融→「会社」「人物」
音楽→「歌手」「楽器」
映画→「タイトル」
スポーツ　→「スポーツ」
漫画・アニメ→「タイトル」
科学→「組織」「人物」

図 3　各分野の初期項目候補

例えば，災害分野の場合は優先順位の高いものから順に「災害」項目，「交通」項目，「事故」項目となる．各初期項目候補に設定された項目パターンが左から順に表に格納される．例えば，災害分野の記事の見出し文である「ＪＲ京浜東北線で運転見合わせ　西川口駅で人身事故」では，「事故」の上位ノードに「災害」が存在するので，「事故」が初期名詞となる．この際，複合語処理によって「人身事故」を取得するので，初期名詞は「人身事故」となる．この時，災害分野の入力において初期項目として「災害」が選択された場合，出力される表の項目には左から「災害」「事故」「場所」「日時」という項目パターンが入る．次に，各初期項目候補の項目パターンを図 4 に示す．

国際分野　国 ―人物 ―組織 ―事件 ―日時
地域 ―人物 ―組織 ―事件 ―日時
政治分野　人物 ―選挙 ―結果 ―日時
選挙 ―人物 ―組織 ―地域 ―結果 ―日時
災害分野　災害 ―事故 ―場所 ―日時
交通 ―災害 ―事故 ―場所 ―原因 ―日時
事故 ―災害 ―場所 ―原因 ―人物 ―日時
裁判分野　罪 ―訴訟 ―人物 ―結果 ―事件 ―日時 ―地域
訴訟 ―罪 ―被告 ―結果 ―日時 ―地域
産業分野　会社 ―結果 ―日時
市況分野　株 ―状況 ―場所 ―時間 ―日時
相場 ―状況 ―場所 ―時間 ―日時
事件・事故分野　容疑 ―事件 ―処分 ―人物 ―日時 ―事故
事故 ―容疑 ―処分 ―人物 ―場所 ―日時
文化分野　人物 ―出来事 ―日時
国 ―出来事 ―日時
訃報分野　人物 ―原因 ―享年 ―職業 ―日時
職業 ―人物 ―原因 ―享年 ―日時
芸能分野　人物 ―年齢 ―場所 ―出来事 ―日時
金融分野　会社 ―人物 ―日時 ―予算 ―決算 ―資産 ―金利
人物 ―会社 ―日時 ―予算 ―決算 ―資産 ―金利
音楽分野　歌手 ―場所 ―出来事 ―日時
楽器 ―場所 ―出来事 ―日時
映画分野　タイトル ―人物 ―出来事 ―日時

図 4　項目候補ごとの項目パターン

3.4 係り受け解析を用いた重要語の取得

南瓜を用いて初期名詞と他の文節間の関係を調べることで本文より重要語を取得する．まず，各記事において，3.3 節で決定した初期名詞を本文から検索し，本文中でその名詞が係っている語と係られている語を重要語として取得する．取得した重要語が名詞以外の場合は，表に格納する単

語としては不適切であると考え,既存手法では使用しない.

3.5 重要語の格納と表出力

3.4 節で取得した重要語より,出力する表の項目それぞれに対応する語を選択し，表を出力する．ＮＴＴシソーラスを用いて各重要語の上位ノードを取得し，項目と一致する語が存在する場合，その項目に対応する語と判断し，表に格納する．取得した重要語に対応する項目が存在しない場合は,ＮＴＴシソーラスの上位ノードから項目を生成する．この際, 2.2 節で述べた項目知識ベースを使用する．重要語の上位ノードに，項目知識ベース内に存在する語があればその語を新たな項目として生成する．その後，新たに生成した項目は表の右端に追加し，その項目に対応する重要語も格納する．例として，災害分野の記事を既存システムにかけた場合の出力結果を図 5 に示す.

記事 1 : 見出し文「ＪＲ京浜東北線で運転見合わせ 西川口駅で人身事故」記事本文：京浜東北線は、西川口駅での人身事故の影響で、…

記事 2：見出し文「大阪府北部、京都府南部で震度３の地震 震源は奈良県」記事本文：１５日、奈良県を震源とする…

記事 3 : 見出し文「ジャカルタで洪水、3 万人避難　死者は 7 人に」記事本文：雨期に入っているインドネシアの首都、ジャカルタでは…

⇩ 初期項目・初期名詞・項目パターンを決定

初期項目	初期名詞	出力する表の初期項目
記事 1災害	人身事故	
記事 2災害	地震	⇨ 災害
記事 3災害	洪水	

⇩ 重要語の取得

記事 1　人身事故　影響　大宮駅
記事 2　地震　大阪府北部　15 日
記事 3　洪水　ジャカルタ　16 日　影響　浸水被害

⇩ 表出力

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 5　出力結果の例

3.6 既存手法の問題点

既存手法では，表の項目内に格納される語はシソーラスのノードによって決定される．そのため，新語や固有名詞などのシソーラスに定義されていない語（未知語）は表にする際に適切に扱えていなかった.

4. 提案手法

以下に本稿で提案する手法を示す.

4.1 Web 検索を利用したシソーラスノードへの割付け

3.6 節の問題点に対し，Web 検索を利用して未知語をシソーラス上の適切なノードに割付ける[8]ことで適切な項目に格納できると考えられる．Web 検索によって未知語の属性と重みを取得し，それらを利用することで未知語をシソーラス上で所属すべきノードへと割付ける．割付けられたノード名を未知語の項目名として使用する.

4.2 シソーラスノードへの割付けの流れ

まず，ニュース記事より取得した未知語とシソーラスのノードの属性を概念ベースに存在する語から取得する．ここで，シソーラスのノード属性にはノードが属するリーフを使用する．次に，取得した属性群を用いて関連度計算を行い，未知語と最も関連度の高いノードを所属すべきノードとして未知語を割付ける．図 6 に未知語をシソーラスのノードへ割付ける流れを示す.

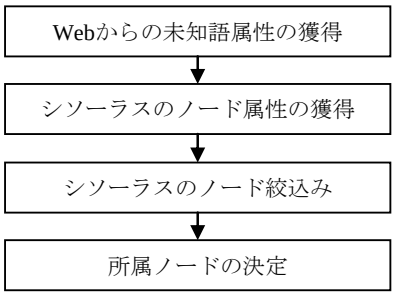

図 6　ノード決定の流れ

まず，未知語を入力した後に，未知語とノードの性質を比較する（関連度計算を行う）ために未知語属性とノード属性を獲得する．次に，獲得した属性群を用いて関連度計算を行い，所属ノード候補を絞り込む．さらに，シソーラスが持つ情報を利用（所属ノード決定手法）して，未知語が所属すべきノードを決定する.

4.3 Web からの未知語属性の獲得

4.3.1 未知語の概念化

未知語を入力して Google で検索を行い, 検索結果ページを取得する．不要な情報を取り除いた文書群の形態素解析を行い，自立語を抽出する．最後に概念ベースに存在する語を未知語の属性とする．そして，得られた属性の頻度に SWeb-idf の値を掛け合わせたものを属性の重みとし，重み順に並び替える．なお，SWeb-idf のデータベースに存在しない属性は SWeb-idf 値の最大値を掛け合わせている．表 2 に未知語の属性と重みの例を示す.

表 2　未知語「クイニーアマン」の属性と重み

属性	重み
パン	382.88
バター	249.17
菓子	116.73

4.3.2 未知語属性の拡張

未知語属性の拡張として，未知語の属性を 2 次属性まで展開して獲得する． 4.3.1 節で説明した手法により獲得した未知語の属性(1 次属性)をキーワードとして再び Google で検索を行い，属性 (2 次属性) の獲得を行うものである．表 3 に未知語を 2 次属性まで展開した様子を示す．

表 3　未知語「クイニーアマン」の属性展開

	属性 重み	属性 重み	属性 重み
1 次属性	パン 382.88	バター 249.17	菓子 249.17
2 次属性	パン 1457.08	バター 2062.55	菓子 1805.81
	酵母 107.04	発酵 149.49	和菓子 235.57
	レシピ 86.68	サンド 128.25	ケーキ 204.19

得られた 2 次属性をそのまま加えると，2 次属性の影響力が大きくなるため，1 次属性の重みの大きさを考慮している．具体的には，2 次属性の重みに 1 次属性の重みの比率を掛け合わせている．例えば，表 3 における 1 次属性「パン」の重みの比率は以下のように計算される．

382.88/748.78≒0.51

(「パン」の重み/1 次属性の重みの合計)

1 次属性「パン」の属性である「パン」，「酵母」，「レシピ」の重みは以下のように計算される．

「パン」: 1457.08*0.51=743.11

「酵母」: 107.04*0.51=54.59

「レシピ」: 86.68*0.51=44.21

最終的に得られる未知語「クイニーアマン」の属性を表 4 に示す．

表 4　2 次属性まで展開して得られた未知語

属性	重み	属性	重み
パン	1125.99	レシピ	44.21
バター	929.81	サンド	42.32
菓子	405.66	和菓子	37.69
酵母	54.59	ケーキ	32.67
発酵	49.48		

4.4 シソーラスのノード属性の獲得

各ノードに属する全てのリーフに対して概念ベース参照を行い，リーフを概念とする語の 1 次属性とその重みを取得する．そして，これらを足し合わせたものを属性集合として取得する．この作業を全てのノードに対して行い，シソーラスのノード属性を取得する． 次に，取得したシソーラスの全ノード属性内で，tf・idf を利用して各属性の重みを求める．これをシソーラスのノード属性とする．表 5 に例として「時計」のノード属性を示す．

表 5　ノード「時計」の属性（一部）

ノード属性	重み
懐中時計	4733.49
掛時計	3476.39
置時計	2791.44

4.5 シソーラスノードの絞り込み

4.3 節で説明した手法を用いて属性を獲得した未知語と 4.4 節で説明した手法を用いて属性を獲得したノード属性に対して関連度計算を行う．なお，関連度の閾値を 0.0 から 0.05 まで 0.001 刻みで実験を行った結果，最も高い精度を得られた 0.02 以上の関連度を持つノードを所属ノード候補とする．

4.6 所属ノード属性の決定

4.6.1 ノード動詞

シソーラスは単語を体系的に配置しており,「同一のノードに属するリーフは助詞を伴う動詞の係り受けに同様の語を取る」という関係が存在する．ノード動詞とはこの関係を利用して，ノードに設定したキーワードのことであり，ノード決定の補助に利用する．例えば，未知語が「マイルドセブン」，所属ノード候補が「たばこ」の場合,「マイルドセブンを吸う」というキーワードの検索を Google で行ったときの HIT 数を求める．表 6 にノード動詞の例を示す．

表 6　ノード動詞（一部）

ノード	ノード動詞
飲物	を飲む
菓子	を食べる
カメラ	で撮影
たばこ	を吸う
歌手	を歌う

4.6.2 共起ヒット

関係のある 2 語はある文書に共に出現すると考えられる．そこで, 未知語とノード名の And 検索による HIT 数を調べてノード決定の補助を行う．例えば，未知語が「マイルドセブン」，所属ノード候補が「たばこ」である場合,「マイルドセブン」と「たばこ」で And 検索を Google で行ったときの HIT 数を求める．

4.6.3 所属ノードの決定手法

ノード動詞と共起ヒットを利用して，未知語の所属ノードの決定を行った．ノード決定の計算式は以下の式 5 に示したものであり，所属ノード候補 $node_i$ の中でノード得点 $NodeValue_i$ が最も高いノードを所属ノードとする． MR が未知語と $node_i$ の関連度， $VerbHit(node_i)$ は未知語にノード動詞を連結したキーワードの検索を Google で行ったときの HIT 数, $CoincidenceHit(node_i)$ は未知語とノード名の And 検索を Google で行ったときの HIT 数を表す．

$$NodeValue(node_i) = MR(node_i) \cdot \log(VerbHit(node_i)) \cdot \log(CoincidenceHit(node_i)) \quad (5)$$

4.7 文書データベースによる属性獲得

文書データベース[9]とは，文書から抽出された情報を登録し，全文検索機能を提供する．初期に存在する文書として，2012 年 6 月時点の Wikipedia 日本語版の全文を記事ごとに分割し，それぞれが文書データベース内の概念として登録されている．文書データベースによる属性獲得は Web 検索を用いた属性獲得よりも処理時間が高速である．そこで，4.1 節から 4.6 節までのノード割付け処理において，Web 検索の代わりに文書データベースを用いて属性獲得処理を行う．

5. 評価

ノード割付け処理と表形式要約の評価を行う．

5.1 ノード割付け処理

5.1.1 ノード割付け処理の評価方法

朝日新聞社の Web ニュースサイト[10]から 2013 年度のニュース記事を取得し，その中から茶筌によって未知語又は複合名詞と判断された 275 語をテストセットとして扱う．テストセットの各未知語の入力に対して，適切なシソーラスのノードに割付けられた語は正解とし，不適切なノードに割付けられた，又は結果が出力されなかった語を不正解として精度を算出する．使用したテストセットと評価の例を表 7 に示す．

表 7　テストセットと評価（例）

未知語	割付け結果	評価
サッポロ一番	麺類	○
吉永小百合	役者	○
iPhone	通信機器	○
N-ONE	乗り物(本体(移動(陸圏)))	○
亀田興毅	君主	×

5.1.2 ノード割付け処理の評価結果

Web 検索と文書データベースのそれぞれを用いた場合で，ノードの割付け処理を行った評価結果を図 7 に示す．

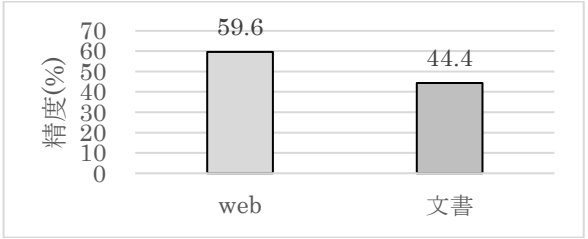

図 7　ノード割付け処理の評価

図 7 より，Web 検索を用いた場合では精度が 59.6%であったのに対して，文書データベースを用いた場合では 44.4%となった． Web 検索を用いた場合には全てのテストセットで出力があったのに対して，文書データベースを用いた場合にはテストセットの語から属性を獲得することが出来ずに出力が得られないものが 49 件(17.8%)あったためである．そのため，文書データベースから属性を獲得できた語のみでの精度は Web 検索と同等であると考えられる．

5.2 表形式要約

5.2.1 表形式要約の評価方法

朝日新聞社の Web ニュースサイト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 10 件ずつ，合計 160 件取得し，テストセットとして扱う．このニュース記事から既存手法と文書データベースによるノード割付け処理を適用した手法（以下，提案手法と呼ぶ）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かを評価した．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 3 点，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 2 点，表から本文の内容が理解できない場合は 1 点とした．そして， 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×として評価した．

5.2.2 表形式要約の評価結果

評価実験を行った結果の全分野を合わせた精度を表 8 に示す．

表 8　表形式要約の精度

	○(%)	△(%)	×(%)
既存手法	11.3	41.9	46.9
提案手法	29.3	43.1	27.5

表 8 より，既存手法と比べて提案手法では○が 18.0%増加，×が 19.4%減少している．これより，提案手法によって精度が向上したと考えられる．また，両手法において○の割合が低かった原因として，評価基準が考えられる．今回の評価実験では，表から本文の内容が理解できた場合でも，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△と評価している．そのため出力する項目数が増えるほど○と評価し辛くなったと考えられる．また，×と評価した要因として，重要語として名詞以外を使用しなかったことで，動詞などで記事の内容を表すものが項目に反映されなかったことが考えられる．既存手法の分野ごとの精度を表 9 に，提案手法の分野ごとの精度を表 10 に示す．

表 9　分野ごとの既存手法の精度

分野	○ (%)	△ (%)	× (%)	分野	○ (%)	△ (%)	× (%)
国際	20	40	40	訃報	40	60	0
政治	0	30	70	芸能	0	30	70
災害	20	30	50	金融	10	50	40
裁判	0	50	50	音楽	10	40	50
産業	0	30	70	映画	10	30	60
市況	20	60	20	スポーツ	10	60	30
事件 事故	30	50	20	漫画 アニメ	10	40	50
文化	0	30	70	科学	0	40	60

表 10　分野ごとの提案手法の精度

分野	○ (%)	△ (%)	× (%)	分野	○ (%)	△ (%)	× (%)
国際	30	30	40	訃報	60	30	10
政治	10	40	50	芸能	30	40	30
災害	30	50	20	金融	20	50	30
裁判	20	40	40	音楽	30	50	20
産業	30	40	30	映画	30	40	30
市況	40	50	10	スポーツ	30	50	20
事件 事故	20	60	20	漫画 アニメ	30	30	40
文化	40	30	30	科学	20	50	30

6. 考察

表 9 より，既存手法において，分野によって大きく精度に偏りがあることが分かった．芸能，文化，産業分野などは人物名や製品名などの未知語を適切に処理可能となったことで，既存手法では中身を格納できていなかった項目が埋まり，精度が向上したと考えられる．また，表 10 より，提案手法では，15 分野で精度の向上が見られる．表 11，表 12 に既存手法と提案手法による出力結果の例を示す．

表 11　既存手法の出力結果の例

会社	日時	契約	増加減少	増加減少	入れ
ドコモ	12 月	契約数	増加	27 万件	導入

表 12　提案手法の出力結果の例

会社	日時	契約	増加減少	増加減少	通信機器	入れ
ドコモ	12 月	契約数	増加	27 万件	iPhone	導入

表 11，表 12 より，提案手法では「通信機器」項目が新しく生成され，「iPhone」が格納されている．ノード割付け処理によって未知語「iPhone」を「通信機器」として扱うことが可能となったことで，既存手法では扱うことの出来なかった未知語を適切に表に反映している．そのため，本文の内容がより正確に把握可能となった．評価結果より，提案手法にはいくつかの改善案が考えられる．一つ目は，ノード割付けに使用する属性の選別である．オートフィードバックや文書データベースを用いた属性獲得では，属性を獲得する未知語の名称内に含まれる名詞が極端に重みの高い属性として表れやすい傾向にある．例えば，「ドラゴンクエスト」という未知語の場合，「ゲーム」という属性よりも，「ドラゴン」という属性の方が高い重みとなる．そのため，未知語の名称内に含まれる名詞の属性については，その重みの扱い方を考慮すべきである．次に，多義語の判別である．例えば，金融分野のニュース記事において「赤字」や「黒字」という語が存在した．金融分野における「赤字」や「黒字」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．最後に，記事の解析方法である．提案手法では，重要語のみに対して修飾語および被修飾語を取得している．しかし，重要語が含まれる文以外にも記事を説明するのに必要な語が存在すると考えられる．よって，先に取得した重要語の類義語や，重要語が修飾している語がさらに修飾している語などに範囲を広げて取得することで精度向上が期待できる．

7. おわりに

本稿では，未知語をシソーラス上のノードに割付けることで，適切な項目名を獲得可能な表形式要約手法を提案した．その結果，既存手法と比べて○が 18.0％増加，×が 19.4％減少した．未知語を適切に扱うことが可能となったことで，詳細な内容の表を生成することが可能となった．

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 小島一秀，渡部広一，河岡司，“連想システムのための概念ベース構成法－語間の論理関係を用いた属性拡張”，自然言語処理，Vol.11，no.3，pp.21-38，2004.

6 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.

7 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，Vol.2D1-01，Jun.2004.

8 後藤和人，渡部広一，河岡 司：“Web を用いた未知語検索キーワードのシソーラスノードへの割付け手法”，情報処理学会第 68 回全国大会講演論文集，4N-3，2006.

9 斎木淳：“文書データベースを用いた未定義語の自動属性獲得”，同志社大学理工学部インテリジェント情報工学科卒業論文，2012.

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