Vol.2020-ICS-198 No.4
2020/3/7

ユーザの要求に応じた 大規模な記事情報を表形式に要約するシステムの構築

目片亮太郎 [1]　渡部広一 [2]　土屋誠司 [2]

概要：近年，情報技術の発展により我々は多くの記事情報を入手することが可能となった．一方で，自身が必要とする情報を的確に入手することが困難となっている．そこで，本稿では記事情報を表形式に要約する手法を提案する．表が持つ項目に格納された単語から複数の情報を比較しやすくなり，的確に情報を選択できると考えられる．膨大な記事情報に対してユーザの入力をすることで，必要な記事を表形式に要約するシステムを構築する．

キーワード：記事情報，要約，表形式

According to User's Request Construction of a System to Summarize Large-Scale Article Information in a Table Format

RYOTARO MEKATA[†1]　HIROKAZU WATABE[†2]
SEIJI TSUCHIYA[†2]

Abstract: In recent years, the development of information technology has made it possible to obtain much article information. On the other hand, it is difficult to obtain the information that they need. In this paper, we propose a method of summarizing article information in a tabular format. It is considered that multiple information can be easily compared from the words stored in the items of the table, and the information can be selected accurately. We construct a system to summarize necessary articles in tabular form by inputting a large amount of article information by the user.

Keywords: Article information, Summarize, Tabular format

1. はじめに

近年，情報技術の発展によりユーザはニュースから大量の記事情報を入手することが可能となった．しかし記事情報の膨大化に伴い，必要とする情報を的確に選択することが困難となっている．そのため，膨大な量の記事情報から必要とする情報を的確に抽出する技術が求められている．その手法として，検索，分類，要約が用いられている．記事情報からキーワードを指定して検索することで，それに関連性のある情報を取得できる．また，記事情報を内容毎に分類することで，分野別ごとにグループ化してまとめることが可能である．しかし１つの記事の中にも多くの情報が存在するため，グループ化しても，そこから瞬時に求める内容を得ることは困難である．そのため，検索，分類によって抽出された記事情報をよりコンパクトにまとめるために，それらを自動で要約する技術が求められている．

情報のコンパクト化の一般的な手段として表に示す手法が挙げられる．表はデータの構造や意味を整理し，系統毎に表示することができることから，概要を素早く理解可能であることや論点が明らかになり，情報を素早く伝えられるといったメリットが挙げられる．このことから，大規模な情報の中からユーザが必要な情報を的確に取得するのに効果的であることが考えられる．そこで西口らによりテキスト情報を表形式要約するシステム[1]が考案された．このシステムはある一記事を入力し，それに対してユーザが知りたい情報をシステムに単語を入力することで，記事情報を表形式に要約して出力するシステムである．

表形式に要約することで様々な利点が考えられる．まず，表には格納された単語の大まかな概要を示す項目が存在する．これにより，複数の記事におけるに細かな内容の違いを項目に格納されている単語から比較することができるという点である．例えば項目として「スポーツ」があるとすれば，ある記事では「サッカー」という単語が入る場合や，別の記事では「ラグビー」と入る場合があることが考えられる．このように同一の項目の中に異なる単語が格納されることからユーザが必要とする内容を瞬時に見分けることが可能となる．また，項目の存在によって，助詞などを省いた名詞だけで内容を把握することができるため，文以上

1 同志社大学大学院理工学研究科情報工学専攻
Doshisha University Graduate School of Science and Engineering
2 同志社大学
Doshisha University

に圧縮して記事内容を要約することが期待できる．

西口らによって構築されたシステムが単一の記事にしか対応できなかった問題点に対して，複数件の記事を対応させることを可能した中西らが開発した表形式要約システム[2]が存在する．しかしこのシステムは，複数の入力できる記事数が最大でも 3 件であるため，大規模な情報の中からユーザが必要とする記事を提示するという目的に沿うことができず，また，生成された項目に対して格納された単語が一つしかない場合があることで，必要のない項目が多数生成される問題があった．そこで本研究では，入力できる記事件数の制限をなくし，ユーザの入力に沿って必要とする情報を適切に表示する表形式要約システムの構築を目指す．

2. 関連技術

2.1 MeCab

MeCab[3]とは，入力された文に対して形態素解析を行い，単語の切り出しや品詞を同定し，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．形態素解析の際には多数の Web 上の情報を用いて構築された，mecab-ipadic-NEologd[4]という辞書を用いる．mecab-ipadic-NEologd 辞書には専門用語，固有名詞など，約 308 万語の語彙を持つ． 「西川口駅で人身事故」という例文を MeCab にかけた結果を図 1 に示す．

```
西川口駅で人身事故
西川口駅 名詞，固有名詞，地域，一般，*，*，…
で　　助詞，格助詞，一般，*，*，*，で，デ…
人身事故 名詞，一般，*，*，*，*，ジンシンジコ…
```

図 1　MeCab の実行例

Figure 1　Execution example of MeCab

2.2 時語知識ベース

時語知識ベースとは，時間判断システム[5]内に存在する知識ベースのことである．時語知識ベースには，「大正」や「月曜日」などの明示的に時間を表現する単語が全 588 語登録されている．

2.3 場所語知識ベース

場所語知識ベースとは，場所判断システム[6]内に存在する知識ベースのことである．知識ベースはシソーラスの一部から構築されており，シソーラスの中から場所と考えられる部分として，親ノードの「場所」と「家屋」に繋がっているノードとリーフを取り出して構築している．

3. 表形式要約システム

表形式要約システムは，大規模な記事から取得した名詞を取得するシステム(以下，記事名詞取得システム)とその名詞とユーザの入力を基に表を作成するシステム(以下，表作成システム)で成り立っている．

3.1 記事名詞取得システム

記事名詞取得システムでは，まずニュース記事の本文の入力を行う．入力されたニュース記事に対して形態素解析を行い，名詞である単語を抽出する．抽出した名詞の中で必要なものに関しては名詞の複合化を行い，それらの名詞をテキストファイルに出力する形で記事名詞知識ベースを作成する．図 2 に名詞取得システムの全体の流れを示す．

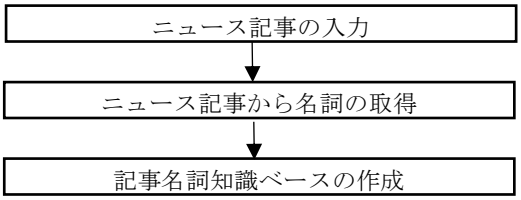


図 2　記事名詞取得システム

Figure 2　Article noun acquisition system

3.1.1 ニュース記事の入力

はじめにニュース記事を入力する．このとき用いる記事は「CD-毎日新聞[7]」の記事を使用する．「CD-毎日新聞」は記事ごとに見出しや本文だけでなく各記事のキーワードとなるような語が付与されている．図 3 に記事データの例，各行のタグの意味を表 1 に示す．

```
＼ＩＤ＼０００００００３０
＼Ｃ０＼１７０１０１００３
＼ＡＤ＼０１
＼ＡＥ＼Ｎ
＼ＡＦ＼１７０１０１Ｍ０１
＼Ｔ１＼皇室：特別立法、天皇の意思明記せず　…
＼Ｓ１＼　　’１７．１．１　朝刊　１頁　…
＼Ｓ２＼　政府は、天皇陛下に限り退位を認める…
＼Ｔ２＼　政府は、天皇陛下に限り退位を認める…
＼ＫＡ＼皇室
＼ＡＡ＼コウシツ
```

図 3　記事データの例(一部省略)

Figure 3　Example of article data (partially omitted)

表 1　「CD-毎日新聞」におけるタグ
Table 1　Tags in “CD-Mainichi Newspaper”

タグ表記	意味
ID	記事 ID
C0	索引記事番号
AD	掲載面種別コード
AE	写真・図の有無
AF	掲載日付とページ
T1	記事の見出し文
S1	掲載日付，朝夕刊，ページ番号，写真・図の有無，文字数
S2	記事本文の第 1 段落
T2	記事本文の全文
KA	記事見出しキーワード（漢字）
AA	記事見出しキーワード（カタカナ）

本研究では, 記事の内容をもとに表作成をしていくため, 記事本文及び，その記事の掲載日時を利用する．

3.1.2 ニュース記事から名詞の取得

ニュース記事に対して MeCab による形態素解析を行い，名詞である単語を取得する．このとき名詞が出現した直後に名詞が出現した場合，それらの名詞は 1 つの名詞であるとして複合化を行う．例えば「酒気帯び」という名詞の直後に「運転」という名詞が出現した場合，それらを複合化して「酒気帯び運転」という名詞を取得する．

3.1.3 記事名詞知識ベースの作成

前節で取得した名詞を記事ごとに，またそれぞれに掲載日時における情報を加えてテキストファイルに書き出すことで記事名詞知識ベースを作成する． 例えば，記事 A の名詞群を A，掲載日時を a，記事 B の名詞群を B，掲載日時を b としたとき，テキストファイルには「a，A」と「b，B」のように書き出し, それぞれの間に「**********」を入れて記事ごとの名詞を分ける．実際の記事から記事名詞知識ベースを作成する例を図 3 に示す．

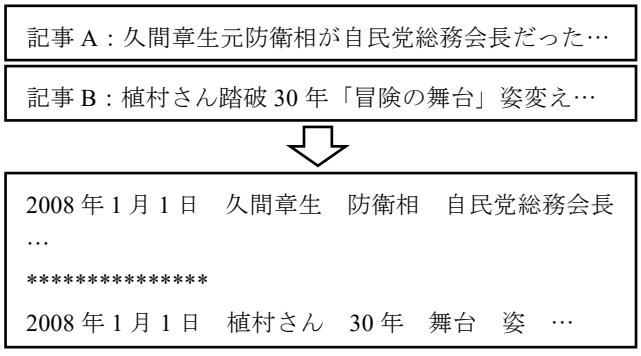


図 3　記事名詞知識ベースを作成する例
Figure 3　Example of creating an article noun knowledge base

3.2 表作成システム

表作成システムでは，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表に出力する．まずユーザ入力により記事の検索，抽出を行う．このとき抽出された記事がない，または 1 件のみであった場合表を作成することはできないため，表を作成しない．抽出された記事が 2 件以上である場合，抽出された記事の名詞を用いて，表に出力する項目やその中に格納する名詞(以下，重要語)を取得し項目となる名詞を生成したあと，それらを表に格納し，出力する．図 4 に表作成システムの全体の流れを示す．

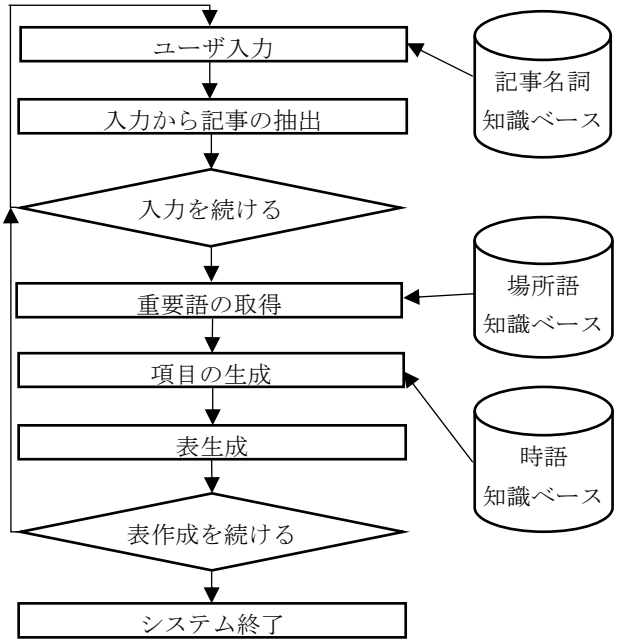


図 4　表作成システムの全体の流れ
Figure 4　Overall flow of the table creation system

3.2.1 ユーザ入力

はじめにユーザがシステムに必要とする情報のキーワード入力を行う．このとき入力するキーワードは複合化されていない名詞単体であり，「1」や「12」などの数詞単体や「交通事故」や「サッカースタジアム」などの複合名詞は入力を規制する．入力した単語は記事名詞知識ベースを参照して検索し，表記が一致したものがあれば，その単語が用いられた記事を抽出する．抽出を終えるとシステムからは，全体の記事の中から絞り込んだ件数と絞り込んだ記事から表を作成するかを問われる．作成する場合は絞り込んだ記事の単語の中から重要語を抽出動作に移行し，作成しない場合は前回の入力で抽出した記事から再度，検索する．図 5 にユーザの入力を行い，記事の検索を行っていく例を示す．ここで，画面出力の例においてシステム側の出力は「S:」ユーザ側の入力は「U:」の隣に出力されているものとする．

S：単語を入力してください
U：交通
S:256 件抽出できました. これで表を作成しますか？
U：いいえ
S：再度検索します
S：前回の入力：交通
U：事故
S：4 件抽出できました．これで表を作成しますか？
U：はい
S：重要語の抽出に移ります

図 5　ユーザ入力時の出力画面の例
Figure 5　Example of output screen at user input

3.2.2 重要語の取得

重要語はユーザ入力によって同じ記事から抽出する際にも違ったものになることが考えられる．例えば、ユーザ入力に「大阪」があるものとして，ある入力で「芸人」とあれば必要とされる重要語としては，大阪でのイベントや会場の場所などが必要と考えられるが,別の入力で「事故」とあれば必要とされる重要語としては，事故に関する被害状況や被害内容が重要語となることが考えられる．このように入力によって必要とされる重要語は変化することが考えられるため，抽出される単語から臨機応変に重要語を取得するアルゴリズムが必要である．

重要語を取得する際にユーザ入力で取得した名詞の中で出力するべきものを選びだすことに条件が必要である．そこで本研究では重要語となる単語であるものとして以下の条件を設定する．

- 記事間で同じ項目の名詞を持つ
- 同じ項目を持つ名詞が全体記事に一定数存在する

表の特徴として系統毎にデータを表示することが挙げられるが，この特徴を本研究では記事における単語それぞれにほとんど同一の概要を持つ，すなわち同じ項目にそれぞれの記事の単語が格納されていることを条件の一つとして設定する．これにより，項目内の単語を比較することで記事における細かな内容の違いを比較することを可能にする．また，上記のような項目が存在したとしても，全体の記事の中で少数しか満たしていない場合であると項目はあまり意味をなさない．例えば，ユーザ入力により 10 件の記事を抽出できたとして，1 の条件を満たす記事が 2 件しかない場合，満たさない 8 件においてはあまり意味のない項目される．これを防ぐための重要語取得における具体的なアルゴリズムを以下に示す．

1. ユーザ入力によって抽出したある記事の単語と他の全ての記事の単語を比較する
2. 単語同士で表記一致した単語があれば，比較元である単語を抽出する
3. 抽出した単語が全体の記事に出現する頻度を算出する
4. 3 で算出した値が 7 割以上のものを重要語として取得する

ユーザ入力で抽出された記事の中で共通された内容を持ち，なおかつユーザ入力に対応した重要と思われるキーワードは各記事でも同様に用いられていると考えられる．そこで本システムでは，記事間で共起した語をその記事の重要語として取得する．まず,「交通事故」と「事故」,「人身事故」や「事故」というように互いに部分一致した単語は重要語候補として一時的に抽出する．これを全記事ごとの単語同士で比較して行う.そして,抽出された単語の内,部分一致した回数が全体の記事と比較して 7 割以上のものは最終的に重要語として取得する．

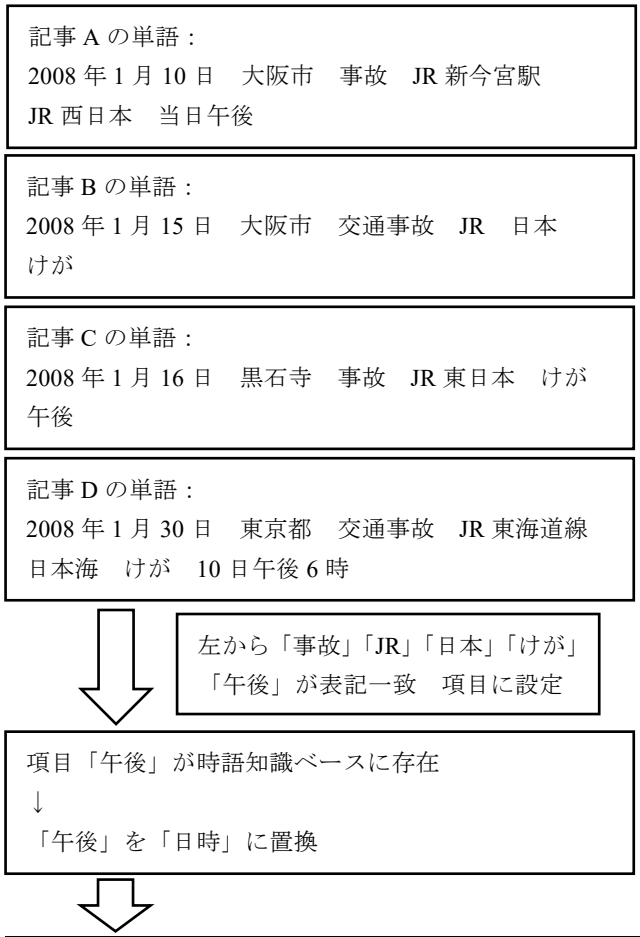
また，上記で取得した重要語のほかにユーザが必要と考える情報があると思われる．例えば，日時における情報は事件の日付やイベントの開催日時などを示す重要な情報である．しかし，新聞記事では日ごとに情報が更新されるため，年月日の詳しい情報が省略されて記載されることが多い．そのため，日が経った後に改めて記事を見返すといつの記事かがわからない場合が多い．そこで，本システムでは掲載日時の情報を重要語として扱うことで日時における情報の欠落を解消する．他にも「イベント行事」や「事件・事故」などは場所に関する情報は重要であると考えられる．しかし，本システムにおける重要語取得アルゴリズムでは表記一致することが著しく乏しいため，情報をとることができる場合が少ない．そこで，これらの情報は場所語知識ベースを参照することにより取得する．記事の単語において場所語知識ベースに存在する，または部分的に表記一致するものは場所に関する語として取得する．

3.2.3 項目の生成

項目を生成する条件としては各記事から取得された重要語の概要を表したものになっていることである．表の項目におけるメリットは「項目を見ただけでその中に格納された単語の意味を大まかに理解することができる」ことであるため，それを重視して項目を生成する必要がある．

本システムの項目は，前節で取得した各記事の重要語同士で表記一致した部分を項目として扱う．例えば,「交通事故」「事故」ではどちらも「事故」の部分で一致しているため，これらの項目は「事故」と設定し，他の重要語においても同様に設定する．また掲載日時や場所語に関する重要語にはそれぞれ「掲載日時」,「場所」の項目を生成する．項目の中で「午後」や「10 日」などの時間や日時に関係する語は出力される項目として必要な情報ではあるが，限定的な表現となるため，時語知識ベースを参照し，知識ベースに存在した場合はそれらの項目は「日時」に置き換える．

重要語から項目を生成し，表を作成する例を図 6 に示す．



掲載日時	場所	事故	JR	日本	けが	日時
2008 年 1 月 10 日	大阪市	事故	JR 新今宮駅	JR 西日本		当日午後
2008 年 1 月 5 日	大阪市	交通事故	JR	日本	けが	
2008 年 1 月 16 日	黒石寺	事故	JR 東日本	JR 東日本	けが	午後
2008 年 1 月 30 日	東京都	交通事故	JR 東海道線	日本海	けが	10 日午後 6 時

図 6　項目の生成・表出力の例

Figure 6　Example of item generation and table output

4. 評価

本システムの精度評価は，表作成システムにおいてどれほどユーザが必要とする情報を抽出し，なおかつ記事の内容を大まかに理解できるかについて評価実験を行っていく．記事は「CD-毎日新聞」の各年度から 1 万件，計 10 万件を使用し，計 10 人の被験者で実験を行う．

4.1 評価実験

まず被験者がシステムにユーザ入力を行い，表を作成する．表作成に成功した場合，被験者が作成した表に出力されてほしいと考える理想的な項目を記録する．失敗した場合，再度表を作成しなおす．これを 5 回繰り返し，作成されたそれぞれの表に対して評価を行っていく．評価基準は以下の 4 点を設けている．

1. 被験者が希望した理想的な項目が実際に出現した項目にどのぐらいの割合出現したか
2. 大まかに内容を推測できる記事が，表に出力された記事にどのぐらいの割合存在するか
3. 実際に出力された項目の中で被験者が必要と考える項目は全体でどのぐらいの割合存在するか
4. 被験者が 3 で必要とした項目に適切に格納された重要語は全体でどのぐらいの割合存在するか

以上の評価基準をそれぞれ評価基準 1，評価基準 2，評価基準 3，評価基準 4 とする．

まず，評価基準 1 では理想的な項目の出現する割合を評価基準として設定している．本研究ではユーザが必要とする記事を表形式に要約するシステムの構築を目的としているため，本来であれば被験者が記事全体の中から適切に表形式要約に用いた記事が用いられたかを評価するべきである．しかし，評価実験では 10 万件の記事を用いているため，それらの中から被験者が必要とする記事を見つけ出すことは現実的ではないと思われる．そのため，被験者が必要な記事を取得できたかを評価するために，あらかじめ「理想的な項目」を被験者が設定することで評価方法を代用する．被験者が設定する理想的な項目とはすなわち，被験者が記事から必要とする情報の中でも重要な位置づけがされていることが推測されるため，この理想的な項目の出現する割合を調べることで被験者が必要とする記事を取得できるかを評価する．

評価基準 2 では大まかに内容が理解できる記事が出現する割合を評価する．これにより記事の要約結果として正しく行えているかを評価する．記事のキーワードを示す重要な単語を適切に抽出することで，必要最低限の単語数で記事の内容を示せているかが評価できると考えられる．そこで評価実験では，実際に出力された表の各行に出力された重要語を確認し，本文の意味が大まかに理解できるかものが全体のうち何件あるかを評価する．

評価基準 3 では必要な項目が適切に出現する割合を評価する．例えば「駅での人身事故」の記事において「場所」に関する項目は必要と思われるが，「電車の種類」などに関する項目はあまり重要でないことが推測される．このような必要のない項目が多数生成されてしまうと表の冗長化の原因となり，またそれに対して先の例にある「場所」を示

す項目のような重要な記事内容を表す項目が存在しなければ，記事の要約を行うことが困難であるため，表から記事内容を一目で理解することが困難になる恐れがある．そのため，必要最低限の項目で表を構築されることが重要であることが考えられる．そこで，出力された項目の中で被験者が必要と考える項目の割合を測ることで評価する．

評価基準 4 では適切に格納された重要語の割合を評価する．例えば項目「被害」の中に被害状況を理解できる単語が格納されている場合は正しく項目が生成されていると思われるが，「被害」という単語が多数格納されている場合は正しい単語が格納されていないことがわかる．そこで，評価基準 3 で必要とされた項目に格納された重要語のうち，適切に格納された重要語の割合を求めることで表に格納された重要語の適切性を評価する．

4.2 評価結果

被験者 10 人が 5 回評価基準ごとに評価し，その評価結果を評価基準ごとに平均化した値と，平均化したものの平均値を表 2 に示す．

表 2　評価結果
Table 2　Evaluation results

	評価基準 1(%)	評価基準 2(%)	評価基準 3(%)	評価基準 4(%)
被験者 1	13.3	26.7	21.4	64.9
被験者 2	23.3	51.6	29.9	68.5
被験者 3	56.7	50.0	39.1	71.0
被験者 4	60.0	2.86	51.2	31.3
被験者 5	0.0	0.0	25.9	73.9
被験者 6	4.17	62.9	76.4	83.6
被験者 7	10.0	24.7	24.4	94.7
被験者 8	13.3	56.9	61.0	30.6
被験者 9	65.3	77.0	67.5	41.5
被験者 10	31.7	10.0	25.8	46.6
平均	27.8	36.3	42.2	60.6

5. 考察

評価基準 1，2，3 ではいずれも 50%を下回る結果が得られた．実際に出力された表には被験者が求める項目や必要と考える項目に意味が近いものが多く生成された．この理由として，項目は各記事で取得した重要語の表記一致から取得していることより，取得した重要語において項目は違うが，意味が近いものが存在すれば，必然的に意味が近い項目が生成されることが原因に挙げられる．このように重複して意味が近い項目が生成されると，表の項目が不適切に多く生成されるほか，冗長化された表が生成されるため，記事内容の理解にも悪影響を及ぼすことになる．そのため，意味が近い項目は一つの項目にまとめることや，ユーザが考えた理想的な項目に沿って項目を新たに生成する手法を検討する必要がある．

また評価基準 4 では唯一 50%を超える 60.6%となった．こちらの理由としては，先述したように重要語の表記から項目を生成しているため，項目と重要語の関係性が全く異なるといった表が生成されることはあまりなかった．しかし項目と重要語がまったく同じ名称で構成されたものが表に格納される場合もあったため，改めて項目の生成手法の検討が必要と考えられる．

6. まとめ

本稿では，ユーザの要求に応じた大規模なニュース記事の表形式要約システムを構築した．その結果，ユーザが求める理想的な項目が出力される割合は 27.8%，表から記事の内容が理解できた割合は 36.3%，ユーザが考える必要な項目が格納されていた割合は 42.2%，適切な項目に正しく格納された重要語の割合は 60.6%となった．このことから，本システムの課題として，意味が近い項目や重要語を複数生成していることが挙げられた．そのため，今後の展望として意味の近い項目や重要語に対して単語を臨機応変に生成していく手法を検討することでより適切な表形式要約を実現することが期待できる．

謝辞　本研究を進めるにあたり，ご指導頂きました本学の渡部広一教授，土屋誠司教授に心から感謝致します．また，アンケート調査や研究活動における諸問題の解決にご協力してくださった知識情報処理研究室の皆様に厚く御礼申し上げます．

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] 中西隆博，芋野美紗子，土屋誠司，渡部広一，“シソーラスノードへの自動割り付けを用いたニュース記事の表形式要約手法”，研究報告知能システム(ICS), 2017-ICS-186, pp1-7,2017-02-24.
[3] “MeCab“，<http://taku910.github.io/mecab/>，(参照 2020-01-08).
[4] GitHub - neologd/mecab-ipadic-neologd: Neologism dictionary based on the language resources on the Web for mecab-ipadic, https://github.com/neologd/mecab-ipadic-neologd（参照 2020-01-08）.
[5] 岩瀬元秀，渡部広一，河岡司：“文の意味理解に基づく常識的時間判断システムの構築”， 信学技報，AI2006-52, Vol.106，No.587，pp.1-8，2007.
[6] 杉本二郎， 渡部広一， 河岡司，“概念ベースを用いた常識場所判断システムの構築”，情報処理学会自然言語処理研究会資料，Vol.2003， No.4 ， pp.81-88，2003.
[7] 毎日新聞社，CD-毎日新聞記事データ集 2008，2009，2010，2011，2012，2013，2014，2015，2016，2017 年版，日外アソシエーツ