WWW 上の定義文における表現特徴を利用した用語説明文抽出のためのテンプレート自動生成について

土橋 惇一　　荒木 健治
北海道大学大学院 情報科学研究科
{jun1，araki}@media.eng.hokudai.ac.jp

1 はじめに

インターネットの普及によって，我々はウェブ上から膨大な量の情報を獲得できるようになった．近年では，ウェブ上の情報を大規模なコーパスや知識情報として質問応答システムに利用する研究が盛んに行われている[1]．質問応答システムの典型的な例として，「～とは何ですか？」のように，ある用語（被定義語）の定義や説明文を求めるような質問応答が挙げられる．

ウェブ上からある用語の定義を獲得するためには，ウェブ文章中から，その用語の定義であると考えられる箇所を特定し，抽出する必要がある．定義文を抽出する手法として，「とは」や「である」など，用語の説明文に特有の表現，またはそれらの表現の組み合わせに整合した部分を含む文を文章中から抜き出す手法がある．この，整合させる表現を複数組み合わせたものを「テンプレート」と呼び，これまでの我々の研究[2]や，類似研究[3]においてテンプレートを用いた用語説明文抽出が行われている．

文献[3]において，テンプレートは既存の用語辞典に記述されている説明文をサンプル文として半自動的に生成されている．既存の用語辞典は表現が統一されているため，そこから生成されるテンプレートもまた限定された表現のものとなる．これに対し，ウェブ上に存在する用語説明文を利用することも考えられるが，これらの用語説明文はその作成者によって表現に多様な揺れがあり，それらを用語辞典から生成されるテンプレートによって抽出することは困難である．そこで我々は既存の用語辞典に記述されている文ではなく，ウェブ上の文をサンプル文としてテンプレートの生成を行うことを考えた．

本稿では，ウェブ上に実在する定義文・説明文における表現の特徴を利用し，用語説明文を抽出するためのテンプレートを自動的に生成する手法を提案する．また，本手法によって生成されたそれぞれのテンプレートの妥当性を検討するための性能評価実験を行い，その考察について報告する．

2 提案手法

2．1 用語説明文の構造的特徴

日本語の一般的な用語説明文に見られる構造的な特徴として，「とは」や「は」のように被定義語の直後に出現する表現（これを「文中表現」と呼ぶ）と，「です」や「である」のように文末に出現する表現（これを「文末表現」と呼ぶ）の 2 種類の表現の存在が挙げられる．多くの用語説明文は，この 2 種類の表現の組み合わせを含んでおり，またその特徴はウェブ上の文章にも当てはまることが分かっている[4]．本研究では，その構造的特徴に着目し，ウェブ上の用語説明文において頻繁に出現する文中表現と文末表現をそれぞれ獲得し，それらを組み合わせてテンプレートを生成する．

文中表現および文末表現の獲得に利用するサンプル文の収集方法は次のとおりである．まず，用語事典「イミダス[5]」より無作為に用語を選択する．次に人手により，その用語を含み，かつ「イミダス」に記述されている定義文および説明文と意味内容の一致する文をウェブ中より特定し，1文単位で抽出する．このようにして，300 のサンプル文を収集した．

２．２　文中表現の獲得

サンプル文を「茶筅[6]」によって形態素解析する．次に，サンプル文中の被定義語の直後に続いて出現する 5 つの単語に関して，連続する単語の組の共起頻度を計算し，共起頻度の高い N 種類の単語の組を文中表現とする．

２．３　文末表現の獲得

２．２と同様に，サンプル文を形態素解析する．次に，サンプル文の文末に出現する 5 つの単語に関して，連続する単語の組の共起頻度を計算し，共起頻度の高い N 種類の単語の組を文末表現とする．

２．４　テンプレートの生成

２．２および**２．３**によって，6 種類の文中表現と 16 種類の文末表現を獲得した．これら 2 種類の表現を互いに組み合わせ，96 種類のテンプレートを生成した．

なお，獲得した文中表現および文末表現の例を表 1 に示す．

表１　獲得した表現の例

文中表現	とは，というものは，については，の意味は，の定義は，の定義とは
文末表現	を言う，を言います，と言う，と言います，である，を指す，を指します，・・・

３　性能評価実験

実験１：テンプレートの精度の評価

生成したテンプレートの抽出精度を測定する．具体的には「あるテンプレート A が入力用語（被定義語）500 語中，何語において適切な説明文を抽出できたか」を調べる．具体的な実験方法は次のとおりである．

まず，既存の専門用語辞典 3 編[8][9][10]より無作為に 500 語を選択し，検索エンジンのクエリとする．このとき，それぞれの専門用語辞典に記述されている説明文を本実験での正解文とする．用語説明が複数の文によって記述されている場合は，その各文全てをクエリに対する正解文とする．

次に，検索エンジンによってクエリを検索し，検索結果上位 200 件のページ中から，クエリと，各テンプレートに整合する文字列を含んだ文を 1 文単位で抽出する．なお，今回はテンプレートとの整合のみを抽出の条件とするため，抽出された文が重文もしくは複文であるかどうかについては考慮しない．また，その文の長さについても考慮しない．

抽出された文をそれぞれの正解文と比較し，意味内容の一致しているものを正解抽出文，一致していないものを不正解抽出文とするとする．比較の際の，意味内容の一致・不一致の判別は第一著者の主観による．このとき，あるテンプレート A の抽出精度は次の式によって表される．

$$P_A = \frac{\text{テンプレート}A\text{による正解抽出文の数}}{\text{テンプレート}A\text{による抽出文の総数}}$$

実験結果および，文中表現・文末表現の頻度の一部を表 2 に示す．

実験２：説明文抽出精度の評価

生成した全てのテンプレートを我々が作成した説明文抽出システムに実装し，その抽出精度を測定する．また，本システムの類似のシステムとし

て，Cyclone[7]との比較を行う．

Cyclone は文献[3]で提案された手法を用いて作成された検索サイトであり，前述のとおり既存の用語辞書から獲得したテンプレートを利用している．また，Cyclone では 1 文単位の抽出を行わず，テンプレートに整合した文から N 文，または，テンプレートに整合した文を含む段落，のようにある条件を満たす領域を特定し抽出する．

実験 1 と同様に，各種専門用語辞典[8][9][10]より無作為に 300 語を選択し，システムの入力用語とする．用語辞典に記述されている説明文を正解文とし，抽出された文のうち，正解文と意味内容が一致しているものを正解抽出文，一致していないものを不正解抽出文とする．意味内容の一致･不一致の判別は筆者の主観による．このとき，システムの抽出精度は次のように表される．

$$P_A = \frac{\text{正解抽出文の数}}{\text{抽出文の総数}}$$

なお Cyclone では抽出単位が 1 文ではなく複数文であるため，抽出された複数の文それぞれを 1 文単位で抽出された文として正解文と比較を行う．実験結果を表 3 に示す．

また，判断基準により不正解抽出文となったが，説明の補足文として適切である可能性を持つ文を補足候補文として，不正解抽出文における補足候補文の割合を算出した．算出結果を表 4 に示す．

4　考察

実験 1 において，8 割以上の精度を持つテンプレートは生成されなかった．しかし，不正解抽出文となった文をそれぞれ調査した結果，補足候補文が不正解抽出文の約 50%を占めていた．よって，これらの補足候補文を正解と考えた場合，テンプレートの精度は全体的に大きく増加すると考えられる．ただしこの場合，補足説明文として適切であるかどうかの判断基準が問題となる．

また，文中表現・文末表現のそれぞれの頻度とテンプレートの精度との間において明確な関係は見つけられなかった．例えば，文中表現と文末表現において最も高い頻度を持つ「とは」と「である」の組み合わせからなる「～とは…である」というテンプレートは，入力用語 500 語中 480 語の検索結果において適用されたが，多くの不正解抽出文を抽出してしまったため，結果としてテンプレートの精度は低下している．このように，出現頻度の高い文中表現および文末表現から生成された抽出テンプレートが，必ずしも高い精度を持っているわけではない．

表 2　テンプレートの抽出精度と表現の頻度

テンプレート	文中表現の頻度 文末表現の頻度	精度
～とは …を意味する	69% 16%	0.767
～とは …である	69% 38%	0.752
～の定義は …である	10% 38%	0.718
～とは …を指す	69% 9%	0.711
～とは …のことです	69% 12%	0.692

表 3　システムの抽出精度

抽出システム	精度
提案手法によるシステム	0.659
Cyclone	0.722

表 4　補足候補文の割合

抽出システム	正解候補文の割合
提案手法によるシステム	45%
Cyclone	16%

なお, 本実験で得られたテンプレートの精度を, ユーザに抽出した説明文を提示する際における, その文の信頼性を表す数値として利用することを検討している.

実験２において，提案手法によるシステムは Cyclone とほぼ同程度の精度であることが示された．また，本システムでは，補足候補文が多く抽出されており，その割合は不正解抽出文の 45%であった．これに対し，Cyclone での補足候補文の割合は不正解抽出文の 16%であった．今回は 1 つの辞典に記述されている説明文のみを正解文としたが，その辞典の説明は必ずしも過不足無く記述されているわけではない．複数の辞典の説明文を正解文としたとき，これまでの補足候補文が正解抽出文として判別される可能性も考えられ，その点で補足候補文を多く抽出することができる本システムは有効である．

システム全体の精度が低下する主な原因として, 実装しているテンプレートの種類の数が過剰であることが考えられる．本システムでは生成されたテンプレート全てを実装しているが，全テンプレート 96 種類中, 11 種類は精度が 3 割未満である．精度の低いテンプレートによる誤った文抽出がシステム全体の精度を低下させているものと考えられる．よって，全てのテンプレートを実装するのではなく，テンプレートの精度に閾値を設定し，その閾値を超える精度を持つテンプレートのみを実装するのが望ましいと考えられる．

５　おわりに

本論文では，ウェブ上の用語説明文より特徴表現を獲得し，その表現の組み合わせよりテンプレートを生成する手法を提案した．

生成したテンプレートの性能評価実験を行った結果，提案手法は最高で約 77%の抽出精度を持つテンプレートを自動生成できることを確認した．

生成した全てのテンプレートを実装したシステムの性能評価実験において，本システムは既存の説明文抽出システムとほぼ同等の精度を持つことを確認した．また，本システムの不正解抽出文における補足候補文の割合は 45%であり，既存の説明文抽出システムが 16%であったのと比べ有効性が確認された．

今後は，テンプレート実装の際の最適な閾値を設定し, システム全体の抽出精度の向上に努める．また，サンプル数を増やし，より精度の高いテンプレートの生成を試みる予定である．

参考文献

[1] 桜井　裕，佐藤　理史，“ワールドワイドウェブを利用した用語説明の自動生成”, 情報処理学会論文誌, Vol.43, No.5, pp.1470-1480, 2002.

[2] 土橋　惇一，荒木　健治，“WWW 検索エンジンを利用した用語解説文抽出システム”，北海道情報処理シンポジウム 2004 講演論文集，pp.26-27，2004.

[3] 藤井　敦，石川　徹也，“World Wide Web を用いた事典知識情報の抽出と組織化”，電子情報通信学会論文誌, Vol.J85-D-II，No.2, pp.300-307，2002.

[4] 西野　文人，橋本　三奈子，落谷　亮，“テキストからの用語とその定義文の抽出”，言語処理学会第 5 回年次大会発表論文集, pp.124-127，1999.

[5] Imidas 2004，集英社，2003.

[6] 茶筅，http://chasen.aist-nara.ac.jp

[7] Cyclone，http://cyclone.slis.tsukuba.ac.jp

[8] 北川　高嗣，須藤　修，他，情報学事典，弘文堂，2002.

[9] 伊藤　正男，井村　裕夫，高久　史麿，医学書院医学大辞典，医学書院，2003.

[10] 金森　久雄，荒憲　治郎，森口　親司，有斐閣経済辞典　第 4 版，有斐閣，2004.