社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2006－NL－171（7）
2006／1／12

Web 文書を対象とした用語説明文抽出手法における抽出範囲の特定

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

本稿では，ある用語に関する説明文をウェブ上に存在する文書から抽出する際の，抽出範囲を特定する手法を提案する．本手法では，ウェブ上に存在する用語説明文に特有の表現を利用し，説明文抽出のためのテンプレートを半自動的に生成する．さらに，生成したテンプレートによって抽出された用語説明文 1 文を起点とし，文同士の関連度計算による抽出箇所の拡張と特定を行う．提案した抽出範囲特定手法を用語説明文抽出システムに実装し，その抽出性能評価実験を行った結果，本システムの用語説明文抽出精度は 39.0%であり，既存の用語説明文抽出システムの精度が 17.8%であったのと比較し，本システムは有効性があることを確認した．

Specification of the range of extraction in term explanation extraction method for Web document

Junichi TSUCHIHASHI，Kenji ARAKI
Graduate School of Information Science and Technology，HOKKAIDO University
{jun1，araki}@media.eng.hokudai.ac.jp

In this paper, we propose the method for specifying the range of the extraction in term explanation extraction from the document that exists on the web. In this method, we generate some templates using a peculiar expression to the explanation that exists on the web for the explanation extraction. And, from one term explanation sentence extracted with the template, we extend and specify the range of extraction by use of relevance ratio between sentences. As a result of the evaluation experiment of performance, the accuracy of our system was 39.0% while the accuracy of an existing system was 17.8%.

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多様かつ膨大な量の情報を獲得できるようになった．このことに伴い，我々は日常的に生じる疑問に対して，ウェブ上の情報からその回答を得る機会が増えている．また，近年ではこれらウェブ上の情報をコンピュータにおける大規模な知識情報として，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].

質問応答システムにおける典型的な質問例として，「～とは何ですか？」のように，ある用語（被定義語）の定義や説明を求めるものが挙げられる．ウェブ上からある用語の説明を獲得するためには，ウェブ上の文書中から，その用語についての説明であると考えられる箇所を特定し，抽出する必要がある．説明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んでいる文を文書中から抜き出す手法がある[2]．この説明文に特有の表現を複数組み合わせたものは「テンプレート」と呼ばれ，これまでの我々の研究[3]や，類似研究[4]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[4]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表現が限定されてしまう．これに対し，ウェブ上に存在する用語説明文は，その作成者によって表現に多様な揺れがあり，それらを用語辞典のみ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用語説明文ではなく，ウェブ上の用語説明文をサンプル文としてテンプレートの生成を行った．

また，用語説明文を抽出するにあたり，今回生成したテンプレートを使用するが，我々の予備調査の結果，ウェブ上に存在する用語説明文は複数の文から構成されている場合がほとんどであることが明らかとなった．ある用語200語に関して，それぞれ説明文が記述されているウェブ上のサイトを1件ずつ用意し（計200件），その説明文数を調査した結果を表1に示す．ある用語についての説明文が記述されているサイト200件中，説明が2文以上で構成されているサイトは197件であった．このことに対し，テンプレートによる文抽出は1文単位で行われるため，獲得される情報としては不十分であると考えられる．そこで我々は1文単位の抽出ではなく，文書中の用語説明の抽出範囲を特定し，その範囲内に含まれる全ての文を抽出することを考えた．

本稿では，ウェブ上に実在する用語説明文を構成する表現を利用し，テンプレートを半自動的に生成する手法を提案する．また，生成されたテンプレートによって抽出される説明文1文をもとに，それに隣接する文同士の関連度を算出し，抽出範囲を特定する手法を提案する．最後に提案した手法に関して性能評価実験を行い，その結果と考察について報告する．

表1　説明文数によるサイト件数の分布

説明文数	1	2	3	4	5	6	7以上
サイト件数	3	15	23	64	53	33	9

2　テンプレートの半自動生成

2．1　用語説明文の構造的特徴

日本語で記述されている一般的な用語説明文に見られる表現の特徴として，「とは」や「は」などのように被定義語の直後に出現する表現（以降，これを「文中表現」と呼ぶ）と，「です」や「である」などのように文末に出現する表現（以降，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種類の表現を構成要素として含んでおり，またこの特徴はウェブ上の用語説明文にも当てはまることが明らかにされている[2]．本研究では，このような用語説明文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ることによってテンプレートを生成する．

今回，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「imidas 2004[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「imidas 2004」中に記述されている説明文と意味内容の一致する文をウェブから検索し，1文単位で抽出する．以上の方法により，300のサンプル文を収集した．

2．2　文中表現の獲得

収集した各サンプル文を，形態素解析システム「茶筅[6]」によって形態素解析する．ただし，被定義語については1語とみなし形態素による分割を行わない．次に，被定義語の直後に続いて出現する5つの形態素に関して，それぞれ連続する形態素組の共起頻度を計算し，共起頻度の高いN種類の形態素組を文中表現とする．ここで，「連続する形態素組」とは，隣り合う複数個（2〜5個）の形態素のまとまりを意味する．

2．3　文末表現の獲得

2．2と同様に，収集した各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つの形態素に関して，それぞれ連続する形態素組の共起頻度を計算し，共起頻度の高いM種類の

形態素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．これら全ての文中表現と文末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 2 に示す．

表 2　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　用語説明抽出範囲の特定

３．１　抽出範囲特定の流れ

我々の提案する抽出範囲特定手法の流れを以下に示す．

1) ２．４で生成したテンプレートによって説明文であると考えられる1文を抽出し，「説明文集合」のひとつとして包含する．ここで，「説明文集合」とは本システムが最終的に抽出する説明文の集合である．
2) 説明文集合に含まれる全ての文それぞれと，説明文集合の直後（または直前）に位置する1文（以降，これを「被検査文」と呼ぶ）との「文間関連度」を計算し，各文間関連度の平均値を算出する．ここで，「文間関連度」とは任意の2つ文における文同士の関連度のことであり，具体的な算出法法については３．２にて後述する．
3) 被検査文中に「この」や「それ」などのような指示語が存在する場合は，2）において算出した文間関連度平均値を1.3倍し，これを被検査文の「関連度」とする．
4) 被検査文の関連度がある閾値を超えた場合，その被検査文を説明文集合に包含し，2）に戻る．関連度が閾値を超えない場合，その被検査文の直前までを説明範囲とみなし，それまでに説明文集合に包含されている全ての文を抽出する．ただし，関連度が閾値を超えず，かつ，被検査文を構成する形態素数が6以下であった場合は，説明文集合に包含せず，その被検査文の直後（または直前）に存在する文を新たに被検査文とみなして2）に戻り，同様の処理を行う．

３．２　文間関連度の計算

文間関連度の計算にはベクトル空間モデル[7]を利用する．ベクトル空間モデルは情報検索の分野で幅広く利用されている[8]．出現する単語にもとづいて文書を1つのベクトルで表現し，その向きによって内容を判断する手法である．仮に2つの文書に同一単語が多く出現する場合，その2つの文書の内容は類似していると考えられる．このことから，文書をベクトルで表現することによって，内容の近い文書ベクトル同士は近くに位置し，内容の異なる文書ベクトル同士は遠くに位置するような空間がつくられる．

本手法では，このベクトル空間モデルを文単位で適用する．各文をベクトルで表現し，比較する2文の文間関連度を2つのベクトルの内積によって求める．ここで，生成されるベクトルは，比較する2文にそれぞれ存在する全種類の形態素を要素として構成されている．つまり比較する2文を通じて存在する形態素の種類がn種類ならば，2つのベクトルはn次元で構成されることになる．ベクトルの各要素の値は，文中の出現頻度に品詞情報により重み付けをしたものである．品詞情報による重み値を表3に示す．各品詞の重み値は，名詞・動詞・形容詞・副詞の重み値を固定した場合に，助詞と助動詞の重み値を変化させ，最適な値をとったものである．ただし被定義語に関しては，それを1語とみなし，品詞による分割は行わずに，特別に高い重み値を与えている．これは，被定義語を含む文では，その被定義語に関する何らかの記述がなされていると考えられるためである．

表 3　各品詞および被定義語の重み値

品詞（または被定義語）	重み値
被定義語	2.2
名詞	2.0
動詞	1.0
形容詞	1.0
副詞	1.0
助詞	0.5
助動詞	0.5

４　性能評価実験

４．１　テンプレートによる用語説明文抽出精度についての評価

ウェブ上に実在する用語説明文から半自動的に生成したテンプレートの用語説明文抽出精度について評価する．具体的には，テスト用語（被定義語）を 500 語用意し，「テスト用語 500 語に関して，テンプレート A が抽出した文のうち，どれだけ適切な説明文を抽出できたか」を調べる．

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[9][10][11]より無作為にテスト用語 500 語を選択し，検索エンジンの入力用語とする．このとき，各入力用語に対して，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．また，本実験では検索エンジンとして Google[12]を利用した．

次に，検索エンジンによって入力用語が含まれているページを検索し，検索結果のうち上位 200 件のページ中から，入力用語と，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しているため，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一切考慮しない．また，その文の文字数や長さについても考慮しない．

テンプレートによって抽出された文を，それぞれの用語に対する正解文と比較し，意味内容の一致していたものを正解，一致していなかったものを不正解とする．専門用語辞典に記述されている用語説明は，複数文にわたり記述されている場合がほとんどであるが，その場合は，記述されている全ての文を入力用語に対する正解文とし，テンプレートによって抽出された文が，複数ある正解文のうちいずれかと意味内容が一致していれば正解とする．なお比較の際の，意味内容の一致・不一致は，各文中の単語の一致・不一致をもとに第一著者が判断する．また，抽出された文が重文や複文のような場合においても，少なくとも一部が正解文と一致していれば正解とする．このとき，あるテンプレート A の用語説明文抽出精度は以下の式(1)によって定義される．

$$P_A = \frac{\text{テンプレートAの正解数}}{\text{テンプレートAによる抽出文の総数}} \quad (1)$$

本実験の結果および，各テンプレートを構成する文中表現・文末表現の出現頻度の一部を表 4 に示す．

４．２　抽出範囲特定による用語説明文抽出性能についての評価

次に，提案した抽出範囲特定手法を実装した用語説明文抽出システムを用いて，用語説明文を実在するウェブ文書から抽出し，その抽出性能について評価する．本システムにおける入力は用語であり，出力はシステムがウェブ文書中から抽出する用語説明文の集合である．

本実験では，本システムと類似したシステムとして，Cyclone[13]に同じ実験データを入力した際の性能比較を行う．Cyclone は文献[4]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，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，用語説明文を複数文にわたり抽出する．

ある用語についての説明文というものは，補足事項や追加事項などの存在，また，それらの事項に関しての，妥当性の検証の問題もあり，その用語の全ての説明文を完全に網羅することは不可能である．実際，同じ用語においても事典・辞書によっては，その説明内容が異なっている場合は少なくない．また，用語説明文抽出システムを利用

するユーザの立場から見た場合，様々な観点や尺度によってユーザ満足度は大きく変化するものと考えられる．これらのことから，用語説明文抽出システムを評価するための絶対的な評価基準を設定することは困難である．よって今回は，我々が独自に評価基準を設定し評価を行った．

具体的な実験方法は次の通りである．用語事典「imidas 2006[14]」より無作為に用語 300 語を選択し，システムの入力用語とする．この際，選択した各用語に関して「imidas 2006」に記載されている用語説明文を本実験での正解文とする．各入力用語について，それぞれのシステムにより用語説明文抽出を行い，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とする．意味内容の一致・不一致の判別基準は**4．1**にて記述したものと同様である．このとき，システムの精度(P)および再現率(R)は，それぞれ以下の式(2)，(3)のように定義される．

$$P = \frac{\text{システムが抽出できた 正解文数}}{\text{システムによる全抽出 文数}} \quad (2)$$

$$R = \frac{\text{システムが抽出できた 正解文数}}{\text{全正解文数}} \quad (3)$$

なお，今回の実験において本システムでは，生成したテンプレートのうち，予備実験により，システムの抽出精度が最も高くなるテンプレートを選択し利用する．実際には，**4．1**で抽出精度が 50%以上のテンプレートのみを選択し，使用した．また，検索エンジンには Google を利用し，2005 年 11 月 14 日から 2005 年 11 月 17 日の時点で Google のキャッシュとして存在している文書を対象とし実験を行った．今回は仮に，抽出された文のうち，抽出元となるページの検索順位上位 30 件までのものを最終的な出力とした．

各用語における精度ならびに再現率の平均値を本実験の実験結果として表 5 に示す．

5　考察

4．1において，最高で約 77%の抽出精度をもつテンプレートを生成することができた．しかし，全テンプレート 96 種類中，11 種類のテンプレートは抽出精度が 30%未満であった．このことより，多様な表現によるテンプレートを多数生成することができたが，各テンプレートの抽出精度には大きなばらつきがあることを確認した．どのテンプレートをシステムに利用するかによって，システム全体としての精度や再現率は大きく変化することが予想されるため，今後，テンプレートの選択・利用について十分な検討が必要である．

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，正解数が少なかったため，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．

4．2において，本システムは Cyclone とほぼ同程度の再現率であり，70%以上のものとなった．

表 4　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%	76.7%
～とは …である	69% 38%	75.2%
～の定義は …である	10% 38%	71.8%
～とは …を指す	69% 9%	71.1%
～とは …のことです	69% 12%	69.2%

表 5　システムの抽出精度および再現率

システム	精度	再現率
本システム	39.0%	71.1%
Cyclone	17.8%	78.8%

一方，精度においては Cyclone が 17.8%であったのに対し，本システムは 39.0%の精度となり，本システムが 21.2%上回った．精度が上回った主な原因として，抽出範囲の特定手法による違いが考えられる．Cyclone が，抽出対象となる文書に関して，説明文の抽出範囲を一定の文数または箇所と定めているのに対し，本手法は抽出範囲を文書に応じて柔軟に決定することができる．このことから本手法は，今回設定した正解文に対して，より適切な説明文が抽出されたと考えられる．よって今回の実験結果より，本手法が有効であることが確認された．また，今回の実験において両システムともに精度が絶対値から見ると低いものとなっているが，これは出力される説明文に対して，正解文数が非常に少ないことが原因であると考えられる．よって，今回は 1 つの事典に記述されている説明文のみを正解文としたが，複数の事典を用いるなど，正解文の追加を行うことによってシステムの精度は向上するものと考えられる．

また，**４．２**にて述べたように，このようなシステムを評価するための基準はひとつではない．システムを評価するための尺度の一つとしてユーザ満足度が挙げられる．しかし，例えばあるユーザが，用語についての概要のみ知りたいのか，もしくは用語についての詳細で正確な内容を知りたいのか，などの観点の違いによってユーザ満足度は大きく変化すると考えられる．よって今後は，このような観点の違いによるユーザ満足度についての実験と調査を行う予定である．

６　おわりに

本稿では，ウェブ上に存在する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によってテンプレートを半自動的に生成する手法を提案した．また，テンプレートによって抽出される説明文をもとに文同士の関連度を計算し，説明文の抽出範囲を特定する手法を提案した．

生成した各テンプレートについて抽出精度評価実験を行った結果，最高で約 77%の抽出精度をもつテンプレートを生成できることを確認した．

また，提案した抽出範囲特定手法を実装したシステムの抽出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比較し，精度において 21.2%上回り，その有効性が確認された．

今後は，最適なテンプレートの決定と関連度計算における適切な重み値の設定を行い，システムの抽出性能の向上に努める予定である．また，ユーザ満足度を始め，他の様々な評価基準による追加実験を行う予定である．

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] 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．

[3] 土橋　惇一，荒木　健治，“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”，言語処理学会第 11 回年次大会論文集，pp.791-794，2005

[4] 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．

[5] imidas 2004，集英社，2003．

[6] 茶筅，http://chasen.aist-nara.ac.jp

[7] G. Salton, A. Wong, and C. S. Yang, “A Vector Space Model for Automatic Indexing”, Communications of the ACM, Vol.18, No.11, pp.613-620, 1975

[8] 大谷　紀子，“情報検索におけるベクトル空間モデルの応用”，武蔵工業大学環境情報学部紀要第五号，http://www.yc.musashi-tech.ac.jp/~kiyou/no5/P099-109.pdf

[9] 北川　高嗣，須藤　修，他，情報学事典，弘文堂，2002．

[10] 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．

[11] 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．

[12] Google，http://www.google.co.jp/

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

[14] Imidas 2006，集英社，2005