用語説明抽出に基づくWeb文書の事典的利用

藤井 敦　石川 徹也
図書館情報大学
{fujii,ishikawa}@ulis.ac.jp

1 はじめに

自然言語処理に基づく知識抽出や情報検索の対象としてWorld Wide Webが注目を集めて久しい．最近では，主に機械翻訳への利用を目的として，対訳関係にあるWebページを自動抽出する研究が行われている [3, 4].

著者らは，検索エンジンを使って検索したWebページから，検索キーワードの説明文を抽出し，Webを事典的に利用することを可能にした [6]．具体的には，説明文に使われやすい言語表現をテンプレート化し，キーワード検索された文書から，テンプレートに一致する文を抽出する．しかし，文単位の言語表現では捉えられない説明への対応や，評価実験に基づく性能分析が今後の課題として残されていた．

本論文では，言語表現に基づく用語説明抽出法について再説し，これを拡張するために，Webページに含まれるHTMLタグに基づく抽出法を提案する．さらに，テスト入力用語を用いた評価実験について説明する．

2 用語説明抽出法

2.1 言語表現に基づく手法

用語説明文は,「とは」や「は」を含むことが多い．事実，これらの表現は新聞記事や辞典を対象とした用語説明抽出に用いられている [5, 9, 12].

著者らの先行研究 [6] では,「CD-ROM 世界大百科事典」[11] から用語説明のテンプレートを半自動的に収集し，人手によるテンプレート作成のコストを削減した．予備調査の結果，用語説明文は「（用語）とは（定義）である」のような特徴的な2文節で構成されることが分かった．そこで，百科事典の説明中に頻繁に共起する文節に基づいてテンプレートを作成した．

百科事典では様々な種類の用語が説明されており，人名や地名の説明は専門用語の説明とは異なる．そこで，専門用語の説明のみを対象とするために，EDR 専門用語 [10] に登録されている見出し語とそれらの説明を収集した．次に，形態素解析器「茶筌」[13] を用いて説明文を単語に分割し，品詞情報に基づいて文節にまとめた．ここで，茶筌の品詞情報と文節に関する規則を独自に作成して利用した．さらに，説明文中の見出し語を共通の変数に置換し，この変数を含む文節とそれに共起する文節を統計情報に基づいてソートした．最後に，ソートした文節対のうち上位100件を人手で確認あるいは修正した．現在，テンプレートは20以上あり，なお人手による分析・拡張を行っている．

上記手法で作成したテンプレートを用いることで，用語説明を文単位で抽出できる．しかし，文より大きな単位での説明には対応できない．例えば,「これを～と定義する」のように照応表現が使われている文には，説明の実体は含まれない．そこで，今回新たに「これ，これら」のような指示語を用意し，抽出した文が指示語を含む場合には，先行する N 文も同時に抽出した．ここで，N はパラメタであり，現在は $N = 3$ としている．

2.2 HTMLに基づく手法

WebページはHTMLのフォーマットに基づいて記述されており，様々な種類のHTMLタグを含んでいる．これらのタグのいくつかは，テキスト情報に一定の構造を与えるためのものである．

そこで，HTMLタグで規定された文書構造に基づいて，文よりも大きな単位での用語説明抽出を試みた．用語説明を含むWebページに特徴的に見られるHTMLタグの使用を調査し，2つの使用法に着目した．

ひとつは,「<H>」や「<B>」タグなどを使って用語を見出し化し，後続する段落で説明を行う方法である．説明には，文章や箇条書が使われたり，両者が併用されることもある．もうひとつは,「<A>」タグを用いて，用語説明をリンク先で行う方法である．リンク先としては，同一ページ内の箇所や別のWebページがある．これらのHTMLタグを用いることで，言語表現テンプレートでは抽出できない用語説明の抽出が期待できる．

しかし，ここで問題となるのは，用語説明を含んでいるWebページが特定できても，用語説明として抽出する範囲が必ずしも明確ではないことである．そこで，本研究では比較的簡単な手法を用いた．すなわち，見出し語あるいはリンクで指定された箇所から，K文を用語説明として抽出する．ここで，K はパラメタであり，経験的に $K = 3$ としている．

用語説明抽出の例を図1に示す．この図では，「データマイニング」の用語説明が提示されており，最上位と最下位が言語表現テンプレートを用いて抽出した説明文，それ以外はHTMLタグに基づいて抽出した説明である．各見出しからは，用語説明を抽出した元のWebページへのリンクがはられており，抽出された説明では不十分な場合や，その用語についてさらに知りたい場合に対処している．

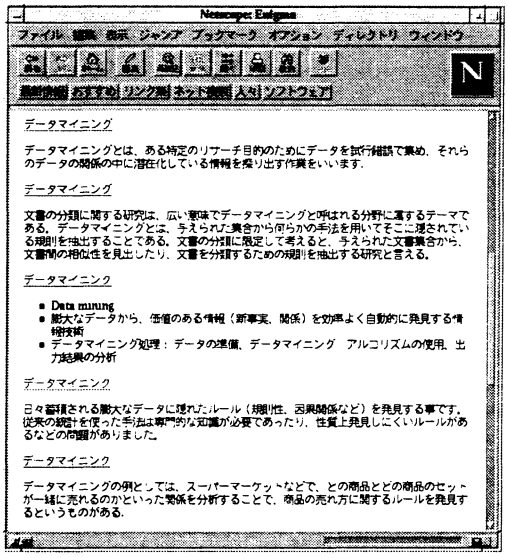

図 1:「データマイニング」の用語説明抽出例

3 評価実験

3.1 方法

本研究で提案した用語説明抽出法の抽出精度を評価するためには，テスト入力用語を作成し，抽出された説明に対する正解判定を行う必要がある．テスト入力の作成には，専門用語辞典などから用語を収集する方法がある．しかし他方において，正解判定は人手のコストが高いので，現実の情報要求を反映するような比較的少数の用語を選択することが好ましい．

そこで，情報検索の評価用に作成された「NACSISコレクション」[2]の検索課題からテスト入力を収集した．このコレクションは65学会の論文から収録した約33万件の日英抄録，日本語の検索課題（予備版21件，公式版39件），各検索課題に対する正解文書リストからなる．検索課題には「タイトル」「検索要求」「検索要求説明」などのフィールドがあり，このうち「タイトル」は検索要求を簡潔に表現する専門用語ひとつで構成されることが多い．NACSISコレクションの検索課題は，出版された論文を検索することを目的として作成されているため，現実の情報要求をある程度反映しており，本研究の評価に適していると考えた．

具体的には，予備版の検索課題21件から，一般語である「新聞記事」（課題番号0023）を除外して，残り20件の専門用語をテスト入力として用いた．また，「ロボット」（課題番号0001）と「マイニング手法」（課題番号0012）は比較的多義なので，情報要求フィールドに基づいて，それぞれ「自律移動ロボット」「データマイニング」に変更した．

上記20件のキーワードに対して用語説明抽出を行い，提示された順番に上位20の用語説明に対して正解判定を行った．特許検索のように網羅性が重視される処理とは異なり，本評価では，たとえ一つの用語説明でもユーザにとって十分な場合がある．そこで，再現率（recall）よりも精度（accuracy）を重視した．用語説明の抽出は，検索エンジン「goo[1]」で検索されたWebページの順位に基づいて行った．

そして，抽出された用語説明に対して，「正解（A）」「元のWebページが用語説明を含んでいる（B）」「元のWebページにも用語説明がなく不正解（C）」のいずれかを判定した．

ここで，A判定とB判定は，自動要約の研究における「informative」，「indicative」要約にそれぞれ対応する点に注意が必要である．すなわち，B判定された用語説明は，それ単体では説明として不十分でも，元の文書を読む価値があるかどうかを判断するためには十分である．

3.2 結果と考察

用語説明抽出の実験結果を表1に示す．表中で，「検索文書数」はキーワード検索によって得られた文書数を示し，「用語説明数」は，判定結果ごとに言語表現とHTMLに基づく手法で抽出された用語説明数をそれぞれ示している．用語説明が20よりも多く抽出されたのは「データマイニング」「機械翻訳」「ニューラルネットワーク」

[1] http://www.goo.ne.jp/

であり，これらに対しては上位20件についてのみ正解判定を行った．それ以外のキーワードに対しては全ての用語説明に対して正解判定を行った．

キーワード検索文書数が0でない17のキーワードのうち，正解判定結果に関わらず用語説明がひとつでも抽出できたものは，言語表現に基づく手法では9件，HTMLに基づく手法では8件，両者を併用すると10件あった．そこで，それぞれの抽出法の被覆率は52.9%と47.1%であり，合計で58.8%となった．

表1に基づいて計算した各手法の抽出精度を表2に示す．A判定の用語説明については，言語表現とHTMLに基づく手法の精度は比較的匹敵しており，B判定の用語説明まで含めると，両者の差異が顕在化した．また，両手法を併用すると平均的な精度となった．A判定については45.3%の抽出精度であり，抽出された用語説明を少なくとも3つ読めば必要な情報が得られることが分かった．また，A判定とB判定の両方については，69.5%の抽出精度であった．言い替えれば，用語説明を含むWebページを約70%の精度で特定できた．

表1: 用語説明抽出の実験結果

キーワード	検索文書数	用語説明数（言語表現/HTML） A	B	C	合計
（自律移動）ロボット	822	0/0	0/0	0/0	0/0
文書画像理解	52	1/3	0/0	0/0	1/3
特徴次元リダクション	0	0/0	0/0	0/0	0/0
知的エージェント	331	2/2	5/1	4/2	11/5
associative rule	1,544	0/0	0/0	0/0	0/0
キーワード自動抽出	35	1/0	1/0	0/0	2/0
（データ）マイニング	3,205	6/5	6/0	3/0	15/5
ルーブ領域解析	0	0/0	0/0	0/0	0/0
故障診断	1,439	0/0	3/0	0/0	3/0
コロケーション	504	1/0	0/0	0/1	1/1
最大共通部分グラフ	0	0/0	0/0	0/0	0/0
通信品質保証	27	0/0	0/0	0/0	0/0
係り受け解析	147	0/0	0/0	0/0	0/0
カタカナ外来語	34	0/0	0/0	0/5	0/5
知識抽出	1,732	0/0	1/0	0/2	1/2
機械翻訳	2,853	1/8	5/1	3/2	9/11
LFG	786	0/0	0/0	0/0	0/0
語彙機能文法	28	0/0	0/0	0/0	0/0
ニューラルネットワーク	9,642	10/3	0/0	2/5	12/8
位置計測	718	0/0	0/0	0/0	0/0
合計	23,899	22/21	21/2	12/17	55/40

表2: 各手法の抽出精度（%）

言語表現		HTML		併用	
A	A+B	A	A+B	A	A+B
40.0	78.2	52.5	57.5	45.3	69.5

4 さらなる拡張

本研究で提案した用語説明抽出法をさらに拡張するための方法について議論する．

まず，検索エンジンを使って検索したWebページから用語説明を動的に抽出するのは時間効率が悪い．そこで，用語説明を静的に索引付けすることが好ましい．そのためには，既存の用語辞典や新聞記事から新語を自動的に抽出し，それらに対して定期的に用語説明を更新する必要がある．

次に，抽出された用語説明数が多い場合には，尤度が高いものからユーザに提示する必要がある．抽出結果の中にある一定数の正しい用語説明が含まれれば，それらは何らかの単語を共有しやすい．そこで，抽出結果に頻出する単語を特定し，それらを含む用語説明から順番に提示する方法がある．

最後に，日本語を母国語としないユーザへの対応がある．ひとつの方法は，言語表現テンプレートを多言語化して，外国語ページからも用語説明を抽出することである．それとは別に，ユーザが入力したキーワードを一旦日本語に翻訳して日本語ページから用語説明を抽出し，さらに抽出された用語説明をユーザ言語に逆翻訳する方法もある．著者らは，言語横断検索の研究の一環として，検索キーワードや文書の翻訳に関する手法を提案してきた[1, 7, 8]．そこで，今後これらの要素技術を統合することを計画している．

5 おわりに

言語表現とHTMLタグに基づいてWebページから用語説明を抽出して，Webを事典的に利用するための手法を提案した．テスト入力用語20件を用いた評価実験の結果，正しい用語説明を約50%の精度で抽出し，抽出元のWebページが用語説明が含むかどうかを約70%の精度で特定した．また，今後さらに拡張するための方法として，時間効率の改善，統計情報に基づく用語説明の順位付け，日本語以外の言語への対応について議論した．

謝辞

NACSISコレクションは学術情報センターの許諾を，CD-ROM世界大百科事典は日立デジタル平凡社の許諾を得て使用させて頂きました．

参考文献

[1] Atsushi Fujii and Tetsuya Ishikawa. Cross-language information retrieval for technical documents. In *Proceedings of the Joint ACL SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 29–37, 1999.

[2] Noriko Kando, Kazuko Kuriyama, and Toshihiko Nozue. NACSIS test collection workshop (NTCIR-1). In *Proceedings of the 22nd Annual International ACM*

SIGIR Conference on Research and Development in Information Retrieval, pp. 299–300, 1999.

[3] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 74–81, 1999.

[4] Philip Resnik. Mining the Web for bilingual texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pp. 527–534, 1999.

[5] 黒橋禎夫, 長尾真, 佐藤理史, 村上雅彦. 専門用語辞典の自動的ハイパーテキスト化の方法. 人工知能学会誌, Vol. 7, No. 2, pp. 336–345, 1992.

[6] 藤井敦, 石川徹也. 言語横断検索システム Quest. 言語処理学会第 5 回年次大会発表論文集, pp. 353–356, 1999.

[7] 藤井敦, 石川徹也. 技術文書を対象とした言語横断情報検索のための複合語翻訳. 情報処理学会論文誌, 2000. （掲載予定）.

[8] 藤井敦, 石川徹也. 言語横断検索における機械翻訳の利用 – 文書翻訳に基づく順位付けの精密化 –. 言語処理学会第 6 回年次大会発表論文集, 2000. （掲載予定）.

[9] 木田敦子, 乾裕子, 落谷亮, 西野文人. 新聞記事からの用語集作成のためのテキスト分析. 情報処理学会 自然言語処理研究会報告, Vol. 99, No. 95, pp. 85–92, 1999.

[10] 日本電子化辞書研究所. 専門用語辞書（情報処理）, 1995.

[11] 日立デジタル平凡社. CD-ROM 世界大百科事典プロフェッショナル版, 1998.

[12] 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.

[13] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明. 日本語形態素解析システム『茶筌』version 1.5 使用説明書. Technical Report NAIST-IS-TR97007, 奈良先端科学技術大学院大学, 1997.