話題語の詳細を表す記述要素の特定要因の分析

久保木 武承　　山本 和英
長岡技術科学大学 電気系
{kuboki, yamamoto}@jnlp.org

1　はじめに

増加するWebページや日々のやりとりに用いられるメール、ローカル上の個人的なファイルまで、コンピュータ上で利用されるテキストは増加の一途を辿っている。そこで把握しきれない情報の中から目的の情報を見つけて利用する事が求められている。

そこで目的のテキストを即座に見つけるための支援として、本研究では「ある話題語について検索したとき、得られた文章が話題語についてのどのような詳細を説明しているかを推定する」という問題を設定した[5]。具体例としては「ローパスフィルタ」について説明しているテキストに対し、説明内容にあわせて「機能」「構成」「種類」「副作用」などの語を付与する事を想定している。これらの語をここでは記述要素と呼ぶ。

本稿では文に対する記述要素の付与を試みることで、記述要素の自動生成システムの設計と課題を明らかにした。実験の結果、精度はオープンテストで42%から80%となった。誤りの大部分は入力文が話題語を詳細化する文では無かった事に起因しており、システムの精度向上にはこれを排除する必要がある。その中で特定のキーワードを用いたパタンマッチは50%から94%の精度をだしており、単純なキーワードマッチも、話題語ごとにキーワードを決定し、入力テキストが話題語を詳細化する文であるという前提が成立する限り効果があることがわかった。

2　関連文献

本研究のテーマは「目的のテキストを見つける」という観点から見ると検索技術の問題と捉えられる。検索そのものの精度を上げるという点で基本的な研究は多数存在する。直接的にはPageRankなどを用いた並び替えや、検索結果のフィルタリングや分類による閲覧性向上などがある[3][4]。入力されたキーワードの意味的な解釈や分類より精確な検索を試みるアプローチもある[1][2][6]。

しかし本来目的のテキストを見つけるには「このテキストは何についてどのような事が書かれているか」という問いに解答する必要がある。先に挙げたようなアプローチは、結果として検索結果の閲覧性を向上させても、テキストを取り扱う上で中心となる課題であるこの問題に答えない。そこで本研究では、「あるキーワードについて検索したとき、得られた文章がキーワードについてのどのような詳細を説明しているかを推定する」という課題に取り組む。

3　記述要素の付与

3.1　記述要素の定義

「このテキストは何についてどのような事が書かれているか」という課題を設定したが、以下、本稿では検索クエリにあたる「何について」を話題語、「どのような事が」を指す語を記述要素と呼ぶ。

記述要素は「話題語＋の」に接続される名詞であると仮定する。しかしこのような基準だけでは記述要素の可否判定は困難である。例えば「ローパスフィルタ」を話題語とした場合「機能」「影響」「問題」「構成」といった語とともに「カットオフ周波数」「掃除」「通過帯域」といった性質の異なる語も出現する。そこで本研究では検索時の検索者の意図を「検索者は話題に関して詳細化する説明が欲しい」と設定することで、記述要素として使用可能な名詞を限定した。具体的には以下のように記述要素とならない語を設定し、話題語を詳細化する名詞 (例:施行, 適用, メンバー, 機能, 影響, ...) となるように記述要素を限定した。

- 名詞が独立して話題語になる 例:カットオフ周波数, ガイドライン, ...
- 名詞が話題語を詳細化しない 例:関係法案, 清掃, 新曲, 詐欺疑惑, 最新版, ...
- 話題の説明の大部分を意味する 重例:要性, 要点, 解説, ...

これらの d。

3.2 記述要素の取得

記述要素として用いられる語は話題語ごとに異なる。本稿では先に話題語に対して記述要素のリストを作り、その後に文に対する記述要素の付与を試みた。

記述要素の候補は、任意に選んだ「話題語」について、「話題語+の+記述要素候補+は or が or を」で Google で検索し、MeCab によるスニペットの形態素解析データから取得する。但し記述要素として抽出する名詞は「非自立、記号、副詞可能、接尾以外から始まる」「数字、記号のみの形態素を含まない」「等、達を含まない」の条件を満たすとする。これらの条件に当てはまる物は話題語の詳細を表す事が少ないためである。また、話題語は記述要素の推定時に分野に偏った傾向が出る事を防ぐため「ローパスフィルタ」「個人情報保護法」「AKB48」の 3 語を用いた。ここで得られた記述要素候補語を人手で正解、不正解の判定をした結果を以下に示す。

表 1: 記述要素語の抽出結果

	ローパスフィルタ	個人情報保護法	AKB48
候補語数	234	60	193
記述要素	39(17%)	27(45%)	15(8%)
その他の語	195(83%)	33(55%)	178(92%)

記述要素候補語の取得に当たっては、先の予想通り「ローパスフィルタ」を話題語とした場合の「カットオフ周波数」「掃除」「通過帯域」のような話題を詳細化しない語が多数含まれ、誤りとなった。

3.3 記述要素の付与

ここでは入力文が話題を詳細化する文か否かを人手判定し、結果の考察時に記述要素の推定誤りとあわせて検証する。実験は次の手順で進める。まず話題語を含む 1 文を集め、3.2 節で集めた候補から記述要素を人手で割り当てる。その際、その記述要素が正しいと考えた根拠となった文中の語、フレーズを記述要素のキーワードとして取得し、キーワードと記述要素の対応辞書を作成する。記述要素の付与は、入力文がキーワードを含む時、対応する記述要素を割り当てる。クローズドテストでは辞書の精査を行い、誤りを出した物を除外した。オープンテストではクローズドテストで作成した辞書を用いて実験し、有効性を確認した。

作成した辞書では、キーワードは以下に示す 3 種類に分類した。

話題語+の+記述要素 「話題語の記述要素」が文中にある場合、当該記述要素を付与

記述要素直接 文中に直接記述要素候補語を含む場合、当該記述要素を付与

その他のキーワード 正解セットから人手で作成したキーワードに当てはまる場合、当該記述要素を付与

「その他のキーワード」の例をあげると、ローパスフィルタでは「カット」が記述要素「働き」、個人情報保護法では「提供」が「規定」、AKB48 では、「誕生+する+た」が「歴史」を表す、といった形になっている。

実験結果を表 2、表 3 に示す。ただし精度の評価からは、話題語を詳細化しない文は除外した。「総合」は全てのパタンの結果をあわせて求めた。

クローズドテストにおける総合的な精度は最低値が個人情報保護法の 68%だった。なお事前に話題語の詳細を説明しない文を実験データから除外しなかった場合、総合的な精度は 55〜57%となり、全般的に精度が低下した。オープンテストでは 23%から 30%だった。

注目したいのはその他のキーワードにおける記述要素付与の精度の高さである。この部分に限定すれば精度 70%以上を維持している。オープンテストでも個人情報保護法を除いて 76%、80%と高い値を示しており、分野依存ではあるものの、他の文との比較からキーワードを検討するのではなく、1 文から記述要素を決定する要素を決めていく方法で記述要素の推定精度を高められる事がわかる。

4 誤り分析

記述要素の特定要因を調べるため、誤り分析を行う。誤りの分布を表 4 に示す。なお、表 2、表 3 では話題語を詳細化しない文 (表中では「話題語が異なる」) を除外したが、ここでは記述要素の推定という課題全体からみた誤りの分布を確認したかったため、これを含めた。

誤りの例を以下に示す (ただし「記述要素が事前取得した候補に無い」は、問題が明らかなため除外する)。

1. 文の話題語が異なる

   例文 2 次ローパスフィルタの特性は、各種のものが作れます

表 2: クローズドテストにおける記述要素付与の精度

	ローパスフィルタ		個人情報保護法		AKB48	
種類	抽出数	精度	抽出数	精度	抽出数	精度
総合	45	0.87	31	0.68	28	0.93
話題語+の+記述要素	3	0.67	3	0.33	1	0.00
記述要素直接	6	1.00	10	0.70	7	1.00
その他のキーワード	36	0.89	18	0.72	20	0.95

表 3: オープンテストにおける記述要素付与の精度

	ローパスフィルタ		個人情報保護法		AKB48	
種類	抽出数	精度	抽出数	精度	抽出数	精度
総合	10	0.80	66	0.42	33	0.76
話題語+の+記述要素	2	0.50	6	0.33	7	0.29
記述要素直接	5	1.00	40	0.40	8	0.75
その他のキーワード	3	0.67	20	0.50	18	0.94

表 4: 誤りの分布

	ローパスフィルタ		個人情報保護法		AKB48	
	closed	open	closed	open	closed	open
文の話題語が異なる	25	18	6	53	19	77
記述要素が事前取得した候補に無い	4	0	5	17	0	1
別パタンに照合	1	1	2	14	1	7
照合誤り	0	0	2	5	1	0
その他	0	0	1	2	0	0
合計	30	19	16	91	21	85

理由 「ローパスフィルタ」ではなく「2 次ローパスフィルタ」の話題

2. 照合誤り

例文 個人情報保護法に基づき、県においても、同法についての情報提供、苦情処理のあっせんなどの取組みを行っています
選出 規定 その他のキーワード_提供
正解 対応 その他のキーワード_取組みを行う

3. その他 (話題は同じだが詳細化する説明ではない)

例文 個人情報を集めた時点が、個人情報保護法の施行前である場合は、利用方法等をどのように説明すればいいのか。
理由 疑問文であり、話題語を詳細化していない

4.1 話題語が異なる文

オープンテストの総合的な精度をクローズドテストのそれと比較した時、「話題語が異なる」を除外したものではオープンテストは総合評価が 7〜25%低下している。オープンテストの誤り分布とクローズドテストの誤り分布を比較すると、オープンテストでは「話題語が異なる」に分類されるものがローパスフィルタで 7 個、個人情報保護法 47 個、AKB48 で 58 個増加しており、その割合は誤り全体における 58%から 95%を占めている。精度を下げる主要な要因であり、当該文が事前に設定した話題語の詳細を説明する文であるか否かという判定が記述要素の付与において重要であり、記述要素を特定する要因の一つ素成る。よって今後記述要素の推定を行う際には、事前に「入力文が、ある語を詳細化する説明か否か」というタスクを解かなくてはならない。

4.2 別パタンに照合

「別パタンに照合」とはその他のキーワードで判定されるべきものが「話題語+の+記述要素」で判定されたような、本来用いられるべき方法と異なる方法で記述要素の付与が行われ、誤りとなったものである。誤りを観察したところ、クローズドテストでは 1 つを除いていずれも記述要素となる語を直接含んでいるために誤った付与を行っていた。具体的には本来は「その他のキーワード」で判定されるべきところを「話題語+の+記述要素」か「記述要素直接」によって割り当ててしまった事が原因だった。すなわちこの問題は「話題語+の+記述要素」や「記述要素直接」のパタン固有の問題であり、入力文がこのパタンに当てはまるときは、パタンを含むか否か以外の要因を持つ事にな

る。これは 4.1 節で挙げた問題に等しい。よってこれらのパタンを使う場合は、同時に当該語を詳細化する説明が後に続いているか否かを判定する必要がある。

4.3 照合誤り

照合誤りとは、文中の語が複数のキーワードを含み、誤ったものを付与してしまった場合を指す。

クローズドテストで発生下 2 件はそもそもキーワードが人手で限定できなかったために発生したものであり、それは「個人情報保護法の理念」「AKB48 のコンセプト」など抽象性の高い語だった。抽象性の高い語は記述要素固有のキーワードの有無と異なり、事前に見た別の文章「... が AKB48 のコンセプトである」のような文との類似性で判断していた。この事から、一部の記述要素はそれを表すキーワードだけでなく、内容に共通する名詞等の対応関係もとる必要がある。オープンテストでは、誤り 5 件中 4 件が記述要素:運用であり、キーワードが「利用」だった。上記の文ではいずれも利用は複合名詞の一部として出現しており、照合誤りの原因は実質的には形態素解析の問題にあり、これが誤った記述要素を付与する要因となった。

次に正しい記述要素が付与されなかった理由を検討する。データではいずれも誤って選出した際に用いたキーワードが、正しい記述要素のキーワードよりも文末に近い位置に出現している。今回用いたシステムではより文末側のキーワードが優先されるため、キーワードマッチの考え方のミスとは言いがたい。そこでそこで辞書と比較してみると、この中に辞書に登録されているキーワードは存在しないことがわかった。よってこれは単に辞書を拡張する事で解決する問題である。

結論として、単純な方法ではあるが、キーワードマッチは話題語を詳細化する文に対して行うという前提下で充分な効果を発揮する、記述要素を特定する要因である。

5 結論

目的のテキストを即座に見つけるための支援として、「ある話題語について検索したとき、得られた文章が話題語についてのどのような詳細を説明しているかを推定する」タスクを設定した。今回の実験を経て、特に記述要素の付与にあたっては「文がユーザに与えられた話題語について詳細化している事」が重要であり、この条件が成立する限りにおいては単純なキーワードマッチで高い精度が得られる事がわかった。ただし記述要素となる語が直接含まれる場合をキーワードにした場合は誤りとなる事が多く、対策が必要である。

使用した言語資源及びツール

(1) 検索エンジン “Google”. http://google.co.jp

(2) 形態素解析器「MeCab」,Ver.0.98, http://mecab.sourceforge.net/

参考文献

[1] Cory Barr, Rosie Jones, and Moira Regelson. The linguistic structure of English web-search queries. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 1021–1030, Honolulu, Hawaii, October 2008. Association for Computational Linguistics.

[2] Mehdi Manshadi and Xiao Li. Semantic tagging of web search queries. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 861–869, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[3] BRIN Sergey. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, Vol. 30, pp. 107–117, 1998.

[4] 村松亮介. 分類階層を利用した検索エンジンの検索結果の構造化とその提示方法の改良. *DEWS2008*, 2008.

[5] 久保木武承, 山本和英. 説明文と記述要素の関係要因の調査　～そこにクエリの「何」が書かれているのか ～. 第 1 回 テキストマイニング・シンポジウム, NLC2011-14, pp. 73–78, 2011.

[6] 若木裕美, 正田備也, 高須淳宏, 安達淳. 検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング. 情報処理学会論文誌. データベース, pp. 72–85, 2006.