テキストの内容を表す記述要素の自動生成手法の検討

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

検索エンジンを用いてWebページを探す時、ユーザは目的とするページに含まれるであろう言葉（クエリ）を予想し、実際にクエリを文中に含むページに絞って調べることで目的のページを探し出そうとする。

しかし単にクエリを含むページを提示するだけでは、ユーザの期待したページを提示するとは限らない。何故ならそのページ中のテキストがクエリを含んでいたとしても、クエリに関する説明をしているとは限らないからである。例えば「個人情報保護法」について検索しても、それは個人情報保護法の「定義」の話、「施行」に関する話、「社会に及ぼした影響」の話など様々な内容があり得る。またそれだけではなく、あるフォーラムで個人情報保護法の話題が上がったというだけの記事まで提示しており、個人情報保護法に関する説明がない場合もあり、ユーザの目的とは異なるページが多数提示され得る。

このような事態に対処するため、検索ではSnippetの提示や関連語の提示、タイトルの生成といった形で、ページ中にある情報をユーザに提示する研究が行われている[1][2]。またウェブディレクトリを用いて事前に話題ごとにページをまとめておく事で、目的のページにたどり着きやすくする方法もある[3]。

しかしここで取り上げる最も重要な問題は、クエリを含んでいるからといってページ本文がクエリについて説明しているとは限らないという点である。

我々はこの問題に対処するため、検索によって提示されたページが、クエリに関してどのような説明をしているのか、その内容を記述要素という形で簡潔に表す事で問題の解決を試みた[4]。記述要素は「定義」「施行」「影響」などといった語で表される。先行研究ではWikipediaから記述要素を表し得る一般的な特徴を抽出し、テキストへの付与を行った。しかしこの時の実験では、Wikipediaから生成した記述要素の特徴は検索結果となるWebテキストと合致しなかったため、オープンテストにおいて記述要素を正しく付与する事ができなかった。

そこで本稿では、Web検索を用いて集めたテキストとテキストの共起情報を用いて記述要素の付与を試みる。特にここでは記述要素となる言葉が本文中に含まれていない文への付与を主な目的とした。

2　提案手法

2.1　全体像

本研究での目標は、検索で得られたページに対して、そのページ本文において「クエリに関するどのような内容を説明しているか」を記述要素で表し、付与する事である。これは図**??**のように表される。

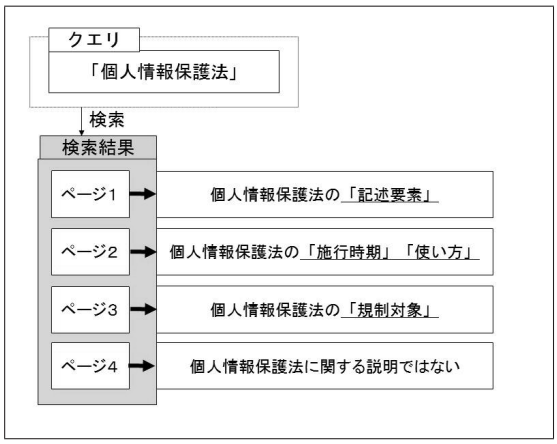

図 1: システム全体図

以下では記述要素の定義や具体的な取得手法と、入力文に対する記述要素の付与手法について説明を行う。

2.2　記述要素の定義と取得

記述要素とは、クエリに関して説明する文章が、具体的にどのような説明をしているかを表す語である。例えば「個人情報保護法では5000人以上の個人情報

をもつ事業者はすべて規制の対象になるが、この場合の個人情報とは個人名を含む。」のような文章が入力されたときは、「規制対象」といった語がこれに該当する。

これだけの条件だと記述要素となり得る語は無限に存在する。しかし実際に用いられる記述要素はクエリによって異なる。「個人情報保護法」では用いられる「施行時期」という記述要素も、「カーネルトリック」というクエリを入れられたときに用いられることはない。これは実際に用いられる記述要素はクエリに依存するという事を意味する。

そこで我々は記述要素の付与を行う前に、クエリに対して用いられ得る記述要素を集めることにした。ここでは記述要素を名詞・未知語の連続と設定して、以下の手順で記述要素を取得した。

1. 「(クエリ) の**」で Web 検索を行う
2. 「の」に接続される名詞・未知語を取得

次に記述要素の付与について説明を行う。

2.3 記述要素の付与

本手法では、記述要素が付与できるような文には、その中に記述要素を決定するような特定のキーワードを含んでいると仮定し、これらのキーワードの有無によって記述要素の付与を試みた。なお、ここではこれらのキーワード、またはキーワードの集合を指してトリガと呼ぶ。

以下にトリガの生成方法を説明する。先に述べた考え方に沿えば、同じクエリ・記述要素の文の中では、共通のトリガを保持していると考える事ができる。そこで以下の手順でトリガの生成を行った。

1. 「クエリの記述要素」で Web 検索を行い、このフレーズを含む段落を集める
2. 集めた段落の集合を形態素解析し、内容語を抽出する
3. 得られた段落の集合中で共起しやすい単語集合を取得し、トリガとする

ただし、全ての内容語を取り扱っていては組み合わせ数は膨大になってしまう。その全てのパタンだけトリガを作るのは現実的ではないため、ここではトリガになり得る語を制限した。設定条件は以下の通りである。

1. 「クエリの記述要素」で集めた段落集合において、全段落数の 10%以下でしか出現しない形態素
2. 段落を集めるときに用いた「クエリの記述要素」の記述要素名と一致する形態素

なお 2 つ目の条件は、トリガ生成に用いる学習データの集め方の影響で全ての段落には必ず「クエリの記述要素」というフレーズが入ってしまうために設定した。こうすることで、トリガ生成上の強力なバイアスになる事を避ける。

3 実験と結果

本研究での提案手法に対して評価実験を行った。実験に使用するデータは Google 検索により得られたページを対象とし、形態素解析には MeCab を使用した (1)。また入力するクエリは「個人情報保護法」と設定した。

3.1 記述要素の取得実験

2.2 節の方法で記述要素の取得を行った。Google で「個人情報保護法」をクエリにして検索して得られたページのうち、上位 1 万件を対象に記述要素を抽出した結果、異なり数で 366 個の記述要素を得た。その一部を以下に示す。

> ガイドライン／違反／運用／影響／解釈／解説／改正／概要／完全施行／観点／関連／規制／規定／義務／義務規定／教育／見直し／施行／施行状況／...

しかしこの中には類似した記述要素、意味の曖昧な記述要素も多数存在した。その例を以下に示す。

> 類似した記述要素
> （施行以来／施行後）（過剰／過剰反応）（対応／対策）
> 曖昧な記述要素
> 精神／ポイント／意味／基礎／基礎知識／基本／基本理念／関係

またここで得られた記述要素の中には、「個人情報保護法の記述要素」で検索するとほとんどページが得られないような語も多かった。そこで、付与の際に扱う記述要素は曖昧なものなどを人手で省いた 54 個に限定した。

3.2 記述要素の付与実験

3.1 節で取得した記述要素ごとにトリガの生成を行った。本実験で生成するトリガは、形態素解析で名詞・未知語・動詞・形容詞に分類された語とする。またトリガは 1 形態素のもの、2 形態素からなるもの、3 形態素からなるものの 3 種類を作る。例えば「運用」という記述要素の 2 形態素からなるトリガ (以下 2 トリガと記述する) なら、「対応, 過剰」のような 2 つの語を 1 つのトリガとしている。

ここで生成したトリガを用いてオープンテストを行った。その際に用いた正解データは、Web からクエリ「個人情報保護法」を含む文をランダムに 100 文集めたものに、人手で記述要素の付与をおこなったものである。なお、記述要素が当てはまるか判断に迷ったものは、本実験では当てはまらないものとしてカウントしている。

表 1: 正解セット 1

文数	100
適合率	0.08
再現率	0.61
あてはまる記述要素の無い文	21

適合率の算出は、1 文に対して現在候補として保有している 54 個の記述要素を全て付与した場合を計算している。また 1 文に対しては複数個の記述要素が付与されることがある。

次にオープンテストの結果を以下に示す。なお、入力した文に対してあらゆるトリガが該当しない場合、該当する記述要素は存在しないものとしてシステムは出力を行う。また再現率の括弧の中の値は、正解セットの上限値と比較しての値である (つまり正解セットと同じ値であれば 100%となる)。

表 2: オープンテストの結果

	再現率	適合率	F 値	平均候補数
1 トリガ	0.61(100%)	0.09	0.16	48.9
2 トリガ	0.59(98%)	0.10	0.17	43.1
3 トリガ	0.61(100%)	0.12	0.20	35.5

再現率は高い精度で得られたものの、適合率が低いという結果になった。これは 1 文あたりの平均候補数が高いためだと考えられる。事実、表 2 において、1 トリガよりも平均候補数の少ない 3 トリガは精度が高くなっている。

そこで次に、使用するトリガに制約をかけ、1 つの記述要素で用いるトリガの数を減らし、実験を行った。

3.3 トリガに制約をかけた記述要素の付与実験

3.2 節で用いたトリガの中には、他の記述要素でも用いられるトリガや、「する」「ある」のようなどのような文でも出現しやすい語がトリガに含まれていた。トリガは一つの記述要素を明確に示していた方が都合が良く、このような性質は都合が悪い。

そこで記述要素ごとにトリガを差別化するため以下の制約を与えた上で、新たな正解セットを用いて実験を行った。

(1) 先の実験 (実験 1) で 1 度以上使われたトリガを使用
(2) 実験 1 で間違った抽出を 2 度以上したトリガを使用しない
(3) そのトリガが 3 個以上の異なる記述要素で使われていたら使用しない

実験は (1)、条件 (1)(2) を組み合わせたもの、条件 (1)(2)(3) を組み合わせた 3 パタンで行った。この実験で新たに作った正解セットと、実験結果を以下に示す。

表 3: 正解セット 2

文数	100
適合率	0.06
再現率	0.72
あてはまる記述要素の無い文	19

表 4: トリガ制限をしたオープンテストの結果

	再現率	適合率	F 値	平均候補数
1 トリガ (1)	0.7	0.07	0.13	41.4
2 トリガ (1)	0.7	0.08	0.14	36.45
3 トリガ (1)	0.62	0.09	0.16	27.31
1 トリガ (1)(2)	0.42	0.15	0.22	5.9
2 トリガ (1)(2)	0.54	0.10	0.17	20.87
3 トリガ (1)(2)	0.55	0.10	0.16	21.81
1 トリガ (1)(2)(3)	0.37	0.16	0.22	3.39
2 トリガ (1)(2)(3)	0.52	0.10	0.17	18.45
3 トリガ (1)(2)(3)	0.55	0.10	0.17	20.31

表 4 で最も高い精度を出したのは、1 トリガ (1)(2)(3) の 0.16 だった。このときの平均候補数は 3.39 と少ない。しかし、再現率は 0.37 となっており、正解データのとり得る最大値 0.72 と比較すれば約半分の 51%に低下している。

結果として、トリガの制約を厳しくしても適合率が大きく向上することは無かった。

4 考察

4.1 トリガを用いた記述要素の付与

本実験の結果より、記述要素の付与は再現率は高いが適合率は低く、使用するトリガに制約を与えても適

合率が大きく向上することは無いことが分かった。その理由は、トリガを用いた手法では、システムが 1 文に対して付与する記述要素の数が多く、その中に多くの不正解を含んでいたためである。

本実験では、この問題に対応するため 3.3 節で使用するトリガに制約を与え、実際に使用するトリガを記述要素ごとに差別化するアプローチを取った。しかしその結果も適合率は最大で 0.16 と低い値をとっている。また再現率も最大値に比較して 51%に低下していた。

注目したいのは、この時はトリガに与えた制約が最も厳しいこともあって、平均候補数も 3.39 と低い値をとっている事である。これは 3.3 節で目的とした「記述要素ごとにトリガを差別化したい」という狙いは少なくとも達成しているという事を意味する。その上で適合率が上がらないという事は、トリガを用いた手法では高い適合率で記述要素を割り当てる事はできないと考えられる。これは、本実験で最初に仮定した「記述要素が付与できるような文には、その中に記述要素を決定するような特定のキーワードを含んでいる」が必ず成立するとは限らないといえる。

この条件が成立しない理由は、本実験で「記述要素となる言葉が本文中に含まれていないような文を主な対象とした」ためだと考えられる。実際、正解セットを作る際も年月日などの明確な情報があれば「施行時期」の説明だと判断することができたが、そのような特徴的な語が人が見る限り発見できない文も多数見られた。

従って、トリガを用いた手法について以下の特徴があると考えられる。

> 再現率は高い。しかしその理由は一つの強力なトリガがあったからではなく、トリガの異なり数によって幅広い文をカバーしていたため
> 「記述要素となる言葉が本文中に含まれていないような文」を対象として記述要素の自動付与を行う場合、トリガを用いた手法は適合性の面で有効ではない

4.2 問題の再設定

4.1 節より、記述要素と一致する語を直接含まないような文にトリガを用いた記述要素の付与を行う場合、トリガを用いた手法では高い精度を得る事は難しい事が分かった。そこで今後は、記述要素付与において精度を保証するような仕組みを考えるために問題を再設定し、「ある記述要素集合と入力文があった時、そのペアは正しいか否かを判定する」タスクに取り組む。

5 おわりに

本稿では検索支援のため、検索では見つけにくい、記述要素と一致する語を直接含まないような文を対象として記述要素の付与を試みた。

具体的には Web から「クエリの記述要素」を含む文を集め、そこから共起頻度の高い語を含んでいるものを記述要素を決定するトリガとし、トリガを含む文に対して該当する記述要素を付与した。そして実験の結果、トリガを用いた手法は再現率は保証しても適合率は保証できないことがわかった。

今後はシステムの適合率を向上させるために、タスクを「ある記述要素集合と入力文があった時、そのペアは正しいか否かを判定する」問題と再設定し、記述要素の付与を試みる。

使用した言語資源及びツール

(1) 形態素解析器 MeCab,Ver.0.9.1, 京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト, http://mecab.sourceforge.jp/

(2) IPA 品詞体系日本語辞書「IPADIC」, Ver.2.7.0, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/stable/ipadic/

参考文献

[1] 野田武史, 大島裕明, 手塚太郎, 小山聡, 田島敬史, 田中克己. 主題語からの話題語自動抽出とこれに基づく Web 情報検索. 電子情報通信学会技術研究報告, pp.955-958,2006.

[2] 長安 義夫, 山本 和英. タイトルパタンによる文書の一文概要生成. 言語処理学会第 13 回年次大会,pp.684-687,2007.

[3] 隅田飛鳥, 後河内脩平, 三浦二三高, 相川昌裕, 鳥澤健太郎. WWW 文書集合から自動抽出した意味的関係を用いた大規模な検索用ディレクトリの試作. 言語処理学会第 13 回年次大会, pp.1121-1124,2007.

[4] 久保木武承, 山本和英.Web ページ検索結果の絞込みのための記述要素の提示. 言語処理学会第 16 回年次大会, pp.278-281,2010.