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

2007－NL－180 (17)
2007／7／25

WWW検索エンジンを用いた質問文内の用語特定手法

北條 奈緒美[1] 鞠々堀 正幹[2] 北 研二[3]

[1]徳島大学 大学院 先端技術科学教育部 システム創生工学専攻
[2]徳島大学 大学院 ソシオテクノサイエンス研究部
[3]徳島大学 高度情報化基盤センター

近年，ユーザが入力した質問文に対して大量の知識源から回答を得る質問応答システムの研究が注目されている．特に，インターネットの普及により，Googleに代表されるWWW検索エンジンを用いて，WWW空間から回答を探す技術が研究されている．これらのシステムでは，質問文内から抽出されたキーワードをWWW検索エンジンに入力し，検索結果から回答を出力する．本稿では，質問文からキーワードを抽出する際に起こる，用語の過分割問題に着目し，WWW検索エンジンを用いた質問文内の用語特定手法を提案する．本手法では，学習用質問文内の各用語に対して，WWW検索エンジンの検索結果(サマリ)から継続度，品詞，文字種などの特徴量を抽出し，Support Vector Machine(SVM)を用いて用語判定モデルを作成する．そして，解析用質問文内の各用語候補に対しても同様の方法で特徴量を抽出した後，用語判定モデルを用いて用語を特定する．実際に，NTCIR4-QAC2の質問文に対して用語特定を行った結果，従来手法に比べて約55%の質問文に対して用語特定による精度向上が認められた．

A method to specify terms in a question sentence by WWW search engine

Naomi Hojo[1] Masami Shishibori[2] Kenji Kita[3]

[1]College of Systems Innovation Engineering,Systems Innovation Engineering,
Graduate School of Advanced Technology and Science, The University of Tokushima
[2]Institute of Technology and Science, The University of Tokushima
[3]Center for Advanced Information Technology, The University of Tokushima

Recently, Question-Answer(QA) systems using the WWW search engine have been developed by the spread of the Internet technology. These systems extract keywords from the question sentence and search the answer from retrieval results obtained by keywords on the WWW search engine. In this paper, We pay attention to the overdivision of the term when keywords are extracted from the question sentence, and propose the method to specify the term in the question sentence by WWW search engine. On this method, the concatenation level that uses the summary, the part of speech, and the character kind for each term candidate are extracted as the feature. And, This method specifies terms of term candidates by Support Vector Machine(SVM). Actually, From experimental resalts using 140 question sentences of NTCIR-QAC2, it was found that the accuracy of the term specific improved compared with the conventional method.

1 はじめに

近年，インターネットの普及により，Googleに代表されるWWW検索エンジンを用いて，WWW空間から回答を探す質問応答システムに関する研究が活発に行われている [1]-[3]．WWW空間を対象とした一般的な質問応答システムは，まず質問文を解析して質問の内容を的確に表記したキーワードを抽出する．次に，既存のWWW検索エンジンに抽出したキーワードを入力して文書検索を行う．最後に，得られた文書集合から回答候補を絞りこみ，ユーザに提示する．

本研究では，質問文からキーワードを抽出するフェーズに着目し，より適切なキーワードを抽出することで質問応答システムの回答精度を向上させることを目的とする．質問文を解析する際に，質問文内の用語 (意味のつながりの強い複合語やフレーズ等) が形態素辞書に未登録の場合には，形態素解析の結果，用語が過分割される問題が生じる．ここで，回答を導くために決定的なキーワードとなる用語が過分割されると，次段の文書検索において検索結果内に含まれる正解単語数が減少し，システム全体の精度に大きな影響を与えてしまう．よって，質問文中の用語部分をあらかじめ特定しておくことが必要になる．

従来の用語特定技術として，パターンマッチによる手法 [4] や出現頻度を用いた手法 [5] がある．パターンマッチによる手法は，人手によるデータ収集やルール作成が必要であるため，多大な時間と労力の消費が問題となる．また，出現頻度情報による手法は，名詞の連続からなる用語を対象に，コーパスでの出現頻度を取得して用語の特定を行うため，名詞以外を含む用語の抽出が困難である．このように従来手法では，質問応答システムへの適用は難しいのが現状である．

そこで本稿では，WWW検索エンジンの検索結果，特にサマリを用いた用語特定手法を提案する．本手法では，学習用質問文内の各用語に対して，WWW検索エンジンの検索結果 (サマリ) から継続度，品詞，文字種などの特徴量を抽出し，Support Vector Machine(SVM)[6] を用いて用語判定モデルを作成する．そして，解析用質問文内の各用語候補に対しても同様の方法で特徴量を抽出した後，用語判定モデルを用いて用語を特定する．従来手法 [5] がコーパス全体での頻度情報を使用しているのと異なり，本手法では検索結果上位のサマリを使用して特徴量を生成している．そのため，既存のWWW検索エンジンでの順位付け (信頼性) が反映されており，より精度の高い用語特定を行えることが期待できる．

2 質問応答システムにおける用語特定の必要性

質問応答システムでは，入力された質問文を解析し，質問タイプを同定するとともに特徴的なキーワードを抽出する (質問文解析)．次に，抽出したキーワードをWWW検索エンジンの入力として与え，適合する文書集合をWWW空間から検索する (文書検索)．そして，絞りこんだ文書から質問タイプに合致する回答候補を見つけ出し，キーワードとの単語間距離などを用いて回答を得る (回答候補絞りこみ)．

特に，質問文解析において，用語 (意味のまとまりの強い複合語やフレーズ等) を分割することなく抽出することが重要になる．次段の文書検索部において，過分割された用語をAND検索した場合，意図に反した検索結果が得られ，検索結果から正解回答を絞りこむことが困難になる．本研究で対象とする用語は，映画やテレビドラマの作品名，著名人の人名，歴史上のできごと，政治経済の用語など，Web上のフリーの百科事典ウィキペディア [7] に登録されている語句とする．

ここで，実際に既存のWWW検索エンジン [8] を用いて質問文を解析した結果を例示する．質問文は，用語として宮崎駿監督のジブリ作品「紅の豚」を含むものを用いる．まず，品詞を特定するために形態素解析を行った結果，用語部分である「紅の豚」は「紅 (名詞) / の (助詞) / 豚 (名詞)」に過分割され，名詞である「紅」と「豚」がキーワードとして抽出される．次に，これらのキーワードを検索語として，WWW検索エンジンによりAND検索を行うと，映画「紅の豚」に関する情報は，検索結果の下位にいくつか出現するものの，沖縄の「紅 豚」の情報や，「豚 しゃぶ専門店 紅 月」という店舗の情報が上位を占めた．このように，映画「紅の豚」に関する文書がほとんど得られず，目的とは違った文書が得られた．しかし，「紅の豚」を用語として検索を行うと，映画「紅の豚」に関す

る情報が上位を占めた．このように，質問文解析部においてあらかじめ用語部分を特定しておくことにより，次段の文書検索でより目的に近い文書を得ることができる．

3 従来の用語特定技術

従来の用語特定手法として，パターンマッチングによる手法 [4] および出現頻度情報による手法 [5] について説明する．

パターンマッチングによる手法は，特定する用語のパターンを作成し，そのパターンにマッチする単語の連続部分を用語と判定する．パターンマッチングによる手法は，分野が限定されている場合，その分野で用いられる表現のみをカバーできるパターンを用意すれば良いため，非常に有効となる．しかし，新しい分野など，それまでのパターンでカバーされない分野では全く機能しない場合が考えられる．また，これらのパターンは，人手により経験的に作成する必要があるため，多大な時間と労力を必要とする．

出現頻度情報を用いた手法は，名詞の連続からなる用語を対象としたもので，連接する名詞の出現頻度を用いて用語の特定を行う．この手法は，まず特定のコーパスから単名詞の連接で作られる単名詞バイグラムを作成する．そして，作成した単名詞バイグラムのコーパスでの出現頻度を集計しスコアを計算する．また，用語は複数の単名詞から生成される複合名詞も含まれるため，複合名詞に拡張しスコアを計算する．最後に，拡張された複数名詞のみでの出現頻度を集計しスコアを補正する．このスコアの高い用語候補から順にソートを行い，上位 K 件を用語として特定する．出現頻度情報による手法では，正解語を含む長めの語であれば大部分をカバーできる．しかし，名詞の連続からなる用語のみを対象としているため，名詞以外からなる用語の抽出が困難である．また，出現頻度情報を特定のコーパスから取得しているため，分野が限定されてしまうという問題も生じる．

4 WWW 検索エンジンを用いた用語特定手法

4.1 用語特定手法の概要

本稿では，WWW 検索エンジンを用いた用語特定手法を提案する．図 1 に提案する用語特定手法の流れを示し，手順を説明する．本手法は，学習フェーズと用語特定フェーズから構成され，学習および用語特定のために SVM を用いている．なお，両フェーズで共通する用語候補生成，特徴量抽出については，4.2 および 4.3 で詳しく述べる．

学習フェーズ

手順 1: 学習用入力文に対し，用語候補を抽出する (形態素解析，*N*-gram 文字列抽出)．

手順 2: WWW 検索エンジンを用いて，各用語候補の特徴量を抽出する．

手順 3: 人手で作成した正解データをもとに，SVM を用いて用語判別モデルを作成する．

用語特定フェーズ

手順 1: 解析用質問文に対し，用語候補を抽出する (形態素解析，*N*-gram 文字列抽出)．

手順 2: WWW 検索エンジンを用いて，各用語候補の特徴量を抽出する．

手順 3: 用語判別モデルを参照し，SVM により用語特定を行う．

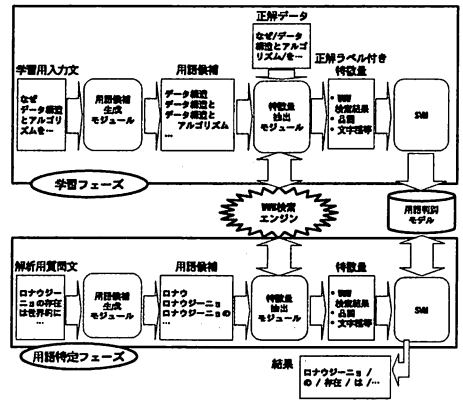

図 1: 用語特定手法の流れ

4.2 用語候補生成処理

用語特定の前処理として用語候補を生成する．以下に用語候補生成の手順を示す．

手順 1: 形態素解析

与えられた質問文 Q に対して，形態素解析を行い，形態素 $M_i(1 \leq i \leq n)$ を得る．

手順 2: N-gram 文字列の生成

手順 1 で得た形態素 M_i から N-gram 文字列 X_j を生成する．このとき，X_j は以下のように表される．

$X_j : M_i * M_{i+1} * \cdots * M_{i+N}; (j = N, *$は連結$)$

また，X_j は以下の条件に従って生成される．

条件 1: $1 \leq N \leq 5$

条件 2: M_i は名詞もしくは未知語

条件 3: M_{i-1} は名詞もしくは未知語以外

条件 4: M_{i+N} が句読点もしくは記号であれば，N-gram 文字列の生成終了

上記の手順に従い，質問文 Q「なぜデータ構造とアルゴリズムを学のか」に対して用語候補抽出を行うと，まず Q は「なぜ/データ/構造/と/アルゴリズム/を…」のように形態素解析される．次に，上記の条件に従い X_j を生成すると，「データ構造」「データ構造と」「データ構造とアルゴリズム」…のように N-gram 文字列が得られ，これらを用語候補とする．

4.3 特徴量の抽出

4.2 節で生成した N-gram 文字列 X_j に対して，特徴量ベクトル $Sig(X_j)$ を作成する．$Sig(X_j)$ は，1) 継続度，2) 現在の形態素 M_{i+N} の品詞，3) 直前の形態素 M_{i+N-1} の品詞，4) 現在の形態素 M_{i+N} の文字種，5) 直前の形態素 M_{i+N-1} の文字種，6) 現在の N の値，以上 6 つの特徴量からなる．

継続度とは，任意の前後の形態素がどの程度継続しているかを示す値である．以下に，継続度を計算する手順を示す．

手順 1: 検索結果サマリの取得

4.2 節で抽出した用語候補 X_j を入力とし，直前の用語候補 X_{j-1} の検索結果のサマリ $S(X_{j-1})$ を取得する．取得するサマリの件数は，検索結果の上位 200 件とする．

手順 2: 用語候補 X_{j-1} の頻度を取得

手順 1 で取得したサマリ $S(X_{j-1})$ における，用語候補 X_{j-1} の頻度を計算する．

手順 3: 用語候補 X_j の頻度を取得

手順 2 と同様に，サマリ $S(X_{j-1})$ における，用語候補 X_j の頻度を計算する．

手順 4: 継続度の計算

手順 2 および手順 3 で得た頻度をもとに，式 (1) により継続度を計算する．

$$\text{継続度} = \frac{\text{サマリ } S(X_{j-1}) \text{ 中の用語候補 } X_j \text{の頻度}}{\text{サマリ } S(X_{j-1}) \text{ 中の用語候補 } X_{j-1} \text{の頻度}} \quad (1)$$

上記の手順に従い，用語候補 X_3「データ構造とアルゴリズム」の継続度を求めた例を示す．まず X_2「データ構造と」を入力とし，検索結果上位 200 件のサマリ $S(X_2)$ を取得する．次に，$S(X_2)$ 中の X_2「データ構造と」の頻度，および X_3「データ構造とアルゴリズム」の頻度を求める．それぞれの頻度が 156，87 であった場合，式 (1) より継続度は 0.56 となる．

5 評価

5.1 実験条件

本手法の有効性を確かめるために，NTCIR ワークショップ [9] の QAC Task で実際に使用された質問文 (NTCIR4-QAC2 から抜粋した 140 文*) に対して評価を行った．学習用データは，WWW 検索エンジン Google[8] や，フリー百科事典ウィキペディア [7] から手作業で収集した 500 文を用いた．これらに形態素解析 [10] を施し，過分割された用語部分に対して正解ラベルを付与した．ラベル付けされた形態素を用い，4.2 節と同様に N-gram 文字列を抽出し，正解ラベルが連続する N-gram 文字列を正事例 (用語)，それ以外を負事例 (用語でない) とした．以下に評価の手順を示す．

手順 1: キーワード抽出

質問文 140 文から，以下の 2 種類の方法によりキーワードを抽出する．

抽出方法 1: 形態素解析結果から，名詞および形容詞を抽出する．この抽出方法をキーワード抽出の従来手法とする．

* 提案手法と形態素解析との結果を比較し，変化のあった 140 文を抜粋した．

表 1: 正解率別の質問数および割合

	質問数	割合 (%)
正解率＋	76	55
正解率－	42	30
正解率＝	22	15

表 2: 抽出方法別の平均正解率

	平均正解率 (抽出方法 1)	平均正解率 (抽出方法 2)
正解率＋	0.2479	0.3791
正解率－	0.3004	0.1890
正解率＝	0.1533	0.1533

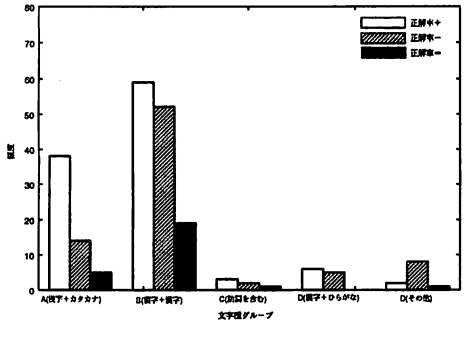

図 2: 字種グループ別の頻度

抽出方法 2: 本手法により用語特定された語句および抽出方法 1 におけるキーワードを抽出する.

手順 2: AND 検索を行い，サマリを取得

手順 1 で抽出したキーワードで AND 検索を行い，検索結果 100 件のサマリを取得する．なお，検索結果が 100 件に満たない場合は，全検索結果を用いる.

手順 3: サマリに含まれる解答の頻度を取得

手順 2 で取得したサマリ中に含まれる，質問に対する正解単語の頻度を人手で求める.

手順 4: 正解率を計算

手順 3 で得た頻度をもとに，式 (2) により正解率を計算する.

$$\text{正解率} = \frac{\text{サマリ中の正解単語の頻度}}{\text{検索件数 (100 もしくはそれ以下の件数)}} \quad (2)$$

5.2 実験結果

表 1 に正解率別の質問数および割合を示し，表 2 に抽出方法別の平均正解率を示す．また，図 2 に各質問文に含まれる用語の字種グループ別の頻度をまとめた結果を示す.

表 1 および表 2 における正解率は，上記の手順 1 で述べた 2 種類のキーワード抽出方法を比較し，本手法を用いた抽出方法 2 において正解率が上昇したもの，下降したもの，変化がなかったものの 3 種類に分類した．表 1 より，評価に用いた質問文 140 文のうち約 55%において，正解率が上昇していることが分かる．さらに，表 2 に示す抽出方法別の平均正解率では，その値が大幅に上昇していることが確認できる．よって，本手法を用いたキーワード抽出方法 2 により，より適切なキーワードが抽出され，質問応答システム全体の回答精度の向上につながったと言える．正解率上昇の理由として，用語部分をあらかじめ特定しておくことで，用語が過分割されることなく AND 検索ができるため，回答をより多く含む検索結果を絞りこむことができたと考えられる．従来のキーワード抽出手法 1 では，用語部分の過分割によりキーワードが分散してしまうため，期待しない検索結果がノイズとなり，意図する検索結果が得られなくなる.

同様に表 1 および表 2 より，質問文の約 30% において正解率が低下しているのが確認できる．低下の理由として，用語としてまとめられたキーワードにより，過度に検索内容を絞りこんでしまい，検索件数自体が減少してしまったことが考えられる．また，一般的に検索キーワードが多い場合，検索件数が少なくなるため，さらに本手法により用語部分をまとめてしまうと，検索内容はさらに絞りこまれ，キーワードに一致するページが見つけられない場合があった.

次に図 2 では，各質問文に含まれる用語の字種に着目し，字種グループ別に頻度をまとめた．字種グループは以下に示す表 3 のように分類した．まず字種グループ A では，正解率が上昇したものについての頻度が高いため，漢字+カタカナからなる用語について，用語特定の効果が高いことが分かる．例えば，「チベット自治区には主にどの民族が暮らしていますか.」という質問文において，特に「チベット自治区」に着目する．従来手法によりキーワードを抽出すると，「チベット自治区」は「チベット」「自治」「区」に過分割される．これらを AND 検索すると，様々な自治区に関する検索結果が得られ，そのうちの一つとしてチベットに関する話題が挙げられる．そのため「チベット自治

表 3: 字種グループ

	組合せ字種	例
A	漢字 + カタカナ	ハンカチ王子，イラク大使
B	漢字 + 漢字	湾岸戦争，阪神大震災
C	助詞を含む	関ヶ原の戦い，徹子の部屋
D	漢字 + ひらがな	うつ病，もののけ姫
E	その他	NHK 大河ドラマ，ウィンドウズ 98

区」に関する詳しい話題は少なく，サマリ内に含まれる回答は少なくなる．それに対し，本手法により用語特定を行い「チベット自治区」を用語として検索を行うと，それに限定した結果が得られ，サマリ内に回答が多く出現する．よって，漢字+カタカナからなる用語については，用語特定の効果が認められ，質問応答システムの精度向上につながると言える．

次に字種グループ B では，正解率が下降したものについての頻度が高くなっていることから，漢字の連続からなる用語について，用語特定の精度が低いことが分かる．例えば，「"接吻" という芸術作品を作ったのは誰ですか．」という質問文に対して，漢字の連続部分である「芸術作品」に着目する．従来手法でのキーワードは「接吻」，「芸術」，「作品」となり，AND 検索の結果，サマリ内には多くの回答が含まれた．しかし，本手法により「芸術作品」を用語としてまとめ，同様に検索を行うと，キーワードとして「芸術作品」が強くなり，一般的な意味の「接吻」を含む様々な「芸術作品」に関する検索結果が得られた．よって，このような漢字の連続からなる用語については，用語特定を行うことによって逆の効果になる場合があった．

6 まとめ

本稿では，質問文からキーワードを抽出する際に起こる，用語の過分割問題に着目し，WWW 検索エンジンを用いた質問文内の用語特定手法を提案した．本手法では，学習用質問文内の各用語に対して，WWW 検索エンジンの検索結果 (サマリ) から継続度，品詞，文字種などの特徴量を抽出し，Support Vector Machine(SVM) を用いて用語判定モデルを作成した．そして，解析用質問文内の各用語候補に対しても同様の方法で特徴量を抽出した後，用語判定モデルを用いて用語特定を行った．評価実験では，NTCIR4-QAC2 の質問文に対して用語特定を行い，55%の質問文に対して用語特定による精度向上が認められ，質問応答システムにおける用語特定の有効性を示すことができた．今後は，従来の用語特定手法を質問応答システムに組み込み，本手法との比較実験を試みたい．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (B)(17300036)，科学研究費補助金基盤研究 (C)(17500644) を受けて行われた．

参考文献

[1] 福本淳一，伴井文人：質問応答技術 －大量のデータをもとに任意の質問に答える－，情報処理，45 巻 6 号，2004 年

[2] 佐々木裕，磯崎秀樹，平博順，平尾努，賀沢秀人，鈴木潤，国領弘治，前田英作：SAIQA －大量文書に基づく質問応答システム－，自然言語処理，145-12，2001 年

[3] 佐々木裕，磯崎秀樹，鈴木潤，国領弘治，平尾努，賀沢秀人，前田英作：SVM を用いた学習型質問応答システム SAIQA-II，情報処理学会論文誌，Vol.45，No.2，2004 年

[4] 小川泰嗣，望主雅子，別所礼子：複合語キーワードの自動抽出，自然言語処理，97-15，1993.

[5] 湯元紘彰，森辰則，中川裕志：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理，145-17，2001.

[6] 大北剛：サポートベクターマシン入門，共立出版 2005.

[7] ウィキペディア (Wikipedia) http://ja.wikipedia.org/wiki/

[8] Google http://www.google.co.jp/

[9] NTCIR Workshop http://research.nii.ac.jp/ntcir/index-j.html.

[10] MeCab http://mecab.sourceforge.net/