WWW 検索エンジンを用いた質問文内の用語特定手法

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

北條奈緒美[1]　獅々堀正幹[2]　北研二[3]
N.Hojo　M.Shishibori　K.Kita
[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

1　はじめに

近年，インターネットの普及により，WWW 検索エンジンを用いた質問応答システムに関する研究が盛んに行われている [1][2]．一般的な WWW 質問応答システムは，まず質問文を解析して質問文の内容を示すキーワードを抽出した後，既存の WWW 検索エンジンで文書検索を行う．そして，得られた文書集合から回答候補を絞りこみ，ユーザに回答を提示する．質問文を解析する際，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，形態素解析によって用語が過分割される問題が生じる．ここで，回答を導くための決定的なキーワードとなる用語が過分割されると，文書検索結果に含まれる回答候補が減少し，システム全体の精度に大きな影響を与えてしまう．そこで本稿では，質問文内の用語部分をあらかじめ特定しておくことを目的とし，WWW 検索結果 (特にサマリ) を用いた用語特定手法を提案する．

2　WWW 検索エンジンを用いた用語特定手法

2.1　用語特定手法の概要

図 1 に，本稿で提案する用語特定手法の流れを示す．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に Support Vector Machine(SVM)[3] を用いる．

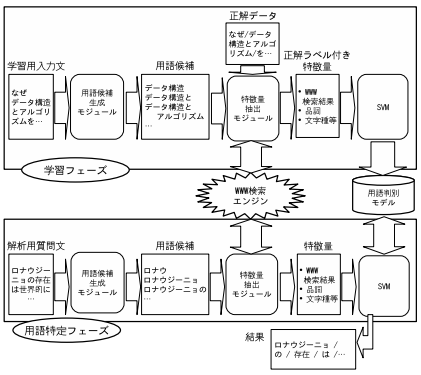

図 1: 用語特定手法の流れ

学習フェーズ

手順 1: 学習用入力文に対して，形態素解析および N-gram 形態素列の抽出を行い，用語候補語を生成する．

手順 2: WWW 検索エンジンを用いて，各用語候補語に対する特徴量を抽出する．

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て，形態素解析および N-gram 形態素列の抽出を行い，用語候補語を生成する．

手順 2: WWW 検索エンジンを用いて，各用語候補語に対する特徴量を抽出する．

手順 3: 用語判別モデルを参照し，SVM で用語特定を行う．

2.2　用語候補語の生成

以下に用語候補語生成の手順を示す．

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: 用語候補語の生成

手順 1 で得られた形態素 M_i から N-gram 形態素列を生成し，これらを用語候補語 X_j とする．このとき，X_j は以下のように表される．

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

また，X_j は以下の条件に従って生成される．

条件 1: $1 \leq N < 5$

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ならば生成終了

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.」に対して用語候補語を生成すると，まず質問文 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，上記の条件に従って「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 形態素列が生成され，これらが用語候補語となる．

2.3 特徴量の抽出

生成された用語候補語 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，継続度，現在の形態素 M_{i+N} の品詞，直前の形態素 M_{i+N-1} の品詞，現在の形態素 M_{i+N} の文字種，直前の形態素 M_{i+N-1} の文字種，用語候補語 X_j の長さの 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，生成された用語候補語 X_j に対する継続度を求める手順を示す．

手順 1: サマリの取得

直前の用語候補語 X_{j-1} の WWW 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: 用語候補語 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，用語候補語 X_{j-1} の頻度をとる．

手順 3: 用語候補語 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，用語候補語 X_j の頻度をとる．

手順 4: 継続度の計算

手順 2 および手順 3 で得た頻度を用いて，式 (1) で継続度を計算する．

$$\text{継続度} = \frac{\text{サマリ } S(X_{j-1}) \text{ における用語候補語 } X_j \text{の頻度}}{\text{サマリ } S(X_{j-1}) \text{ における用語候補語 } X_{j-1} \text{の頻度}} \quad (1)$$

上記の手順に従い，用語候補語 X_4「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_3「データ構造と」をキーワードに，検索結果上位 200 件のサマリ $S(X_3)$ を取得する．次に，$S(X_3)$ 中の X_3「データ構造と」の頻度，および X_4「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．これらの特徴量を用いて特徴量ベクトル $Sig(X_j)$ を作成し，SVM で学習および用語特定を行う．

3 評価実験

3.1 実験条件

本手法の有効性を示すため，質問応答システムにおける本手法および形態素解析法 (名詞，形容詞をキーワードとして用いる) の評価実験を行う．学習用データは，Web 上から手作業で 500 文収集し，形態素解析によって過分割された用語部分に正解ラベルを付与した．正解基準として，Web で公開されているフリーの百科事典「ウィキペディア」[4] に含まれている程度の語句を用語とした．また解析用質問文は，NTCIR ワークショップ [?] の QAC Task で実際に使用された質問文 140 文*を用いた．以下に実験の手順を示す．

*NTCIR4-QAC2 内の質問文に対し，提案手法と形態素解析法の結果を比較し，変化のあった 140 文を抜粋した．

手順 1: キーワード抽出

解析用質問文 140 文から，両手法により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 件未満の場合は，得られた検索結果を用いる．

手順 3: サマリに含まれる正解単語数をカウント

手順 2 で取得したサマリ内に含まれる，質問に対する正解単語数を人手で求める．

手順 4: 正解率を計算

手順 3 で得た正解単語数をもとに，式 (2) で正解率を計算する．

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

3.2 実験結果

表 1 に，各手法の平均正解率を示す．また，表 2 に形態素解析法との正解率の比較を示す．この表では，形態素解析法と比較して正解率が上昇したもの，低下したもの，変化がなかったものの 3 種類に分類し，それぞれの質問文数および割合を示した．さらに図 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示す．

表 1 より，形態素解析法と比較して本手法では全体の平均正解率が上昇していることが分かる．さらに，表 2 からは，評価に用いた質問文 140 文のうち半数以上 (55.0%) において正解率が上昇していることが分かる．よって，本手法を用いることで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考えられる．正解率上昇の要因として，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，より質問内容に合った検索結果を絞りこむことができたと考えられる．形態素解析法では，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．また同じく表 2 より，質問文の 30.0%において正解率が低下しているのが確認でき

表 1: 各手法の平均正解率

	平均正解率
形態素解析法	24.7%
提案手法	28.5%

表 2: 形態素解析法との正解率の比較

	質問文数	割合 (%)
上昇	76	55.0
低下	42	30.0
変化なし	22	15.0
計	140	100.0

る．低下の要因としては，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．一般的に検索キーワードが多い場合，検索件数が減少するため，さらに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る．

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の頻度が高いことから，漢字と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，「日本人でノーベル平和賞を受賞したことがあるのは誰ですか．」という質問文において，特に「ノーベル平和賞」に着目する．形態素解析法により「ノーベル平和賞」からキーワードを抽出すると，「ノーベル」「平和」「賞」に過分割される．これらのキーワードで AND 検索を行うと，さまざまな「ノーベル賞」に関する内容の文書が得られるため，「ノーベル平和賞」に関する詳しい内容は少なく，サマリ内に含まれる正解単語数も少なくなる．それに対し，本手法により用語特定を行い「ノーベル平和賞」をひとまとまりの用語として WWW 検索を行うと，「ノーベル平和賞」に限定した内容の文書が得られ，サマリ内に正解単語が多く出現する．よって，漢字と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回答精度の向上につながったと言える．

次に字種グループ B では，正解率が低下したものの頻度が高いことから，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，「インドネシアの歴代大統領は誰ですか．」という質問文に対して，漢字の連続部分である「歴代大統領」に着目する．形態素解析法でのキーワードは「インドネシア」「歴代」「大統領」となり，AND 検索の結果，サマリ内には多くの正解単語が含まれた．しかし，本手法により「歴代大統領」を用語としてまとめ，同様に検索を行うと，サマリ内の正解単語数は半分以下に減少した．これは「歴代」と「大統領」という語句は，「歴代大統領」とまとめて用いるよりも「歴代の大統領」のように付属語を含めて用いられることが多いためである．よって，このような漢字の連続からなる用語は，用語特定を行うことにより正解率が低下する場合があることが確認できた．

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

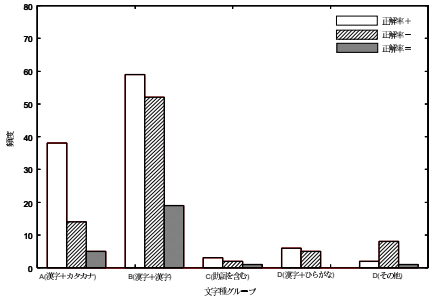

図 2: 字種グループ別の頻度

4 今後の改良点

4.1 漢字の過剰連結

今回提案した用語特定手法では，用語特定を行うことによって過度に検索内容を絞りこんで正解率を低下させてしまう場合があった．特に，これは漢字の連続からなる語句に多く見られた．その原因として，漢字の連続からなる語句は，他の字種からなる語句と比較して各形態素のつながりの強さに差があることが挙げられる．例えば，漢字の連続ではない「ハンカチ王子」や「イラク大使」，「うつ病」などは，1 つの単語として用いられることが多いのに対して，「歴代大統領」や「日本人技術者」などの漢字の連続からなる語句は，「歴代の大統領」や「日本人の技術者」のように，本来，形態素の間に付属語を含んで用いられるものが多い．このような漢字の連続からなる語句が用語特定によって過剰に連結されてしまったため，他の字種からなる語句に比べて正解率が低下したと判断できる．よって，漢字の連続からなる語句に対しては，本来付属語が入る箇所を特定し，「歴代 | 大統領」「日本人 | 技術者」のように，形態素の区切りを明確にする必要がある．

4.2 3 形態素の語句に対する区切り判定

漢字の連続からなる語句に対して，本来の適切な区切りを判定する．今回は特に，用語に多く見られる 3 形態素から構成される語句に限定した．3 つの形態素を $M_i(1 \leq i \leq 3)$，形態素の区切りを $P_j(j = 1, 2)$ とすると，3 つの形態素からなる語句は $M_1 \mid_{(P_1)} M_2 \mid_{(P_2)} M_3$ と表される．区切り判定には，区切り P_j をまたぐ形態素の継続度 (2.3 節) を用いる．提案手法では，前の形態素から後ろの形態素への継続度のみを用いたが，区切り判定ではその逆方向の継続度も用い，それぞれを前向き継続度，後向き継続度と表す．例えば，「†地球 | 温暖 | 化」において，区切り P_1 および P_2 をまたぐ，全ての形態素の組合せを考慮して継続度を求めると，表 4 および表 5 のようになる†．

表 4 および表 5 における上段は，区切りをまたぐ 2 形態素の継続度であり，分母が特徴的な形態素であれば，その

† $\frac{\text{†地球温暖}}{\text{†地球}}$ は $\frac{\text{サマリ } S(\text{†地球}) \text{ における「†地球温暖」の頻度}}{\text{サマリ } S(\text{†地球}) \text{ における「†地球」の頻度}}$ を表す

表 4: 区切り P_1 の継続度

	前向き継続度	後向き継続度	判
地球 | 温暖	$\frac{\text{地球温暖}}{\text{地球}} = 0.10$	$\frac{\text{地球温暖}}{\text{温暖}} = 0.65$	○
地球 | 温暖化	$\frac{\text{地球温暖化}}{\text{地球}} = 0.10$	$\frac{\text{地球温暖化}}{\text{温暖化}} = 0.67$	○

表 5: 区切り P_2 の継続度

	前向き継続度	後向き継続度	判
温暖 | 化	$\frac{\text{温暖化}}{\text{温暖}} = 0.95$	$\frac{\text{温暖化}}{\text{化}} = 0.09$	○
地球温暖 | 化	$\frac{\text{地球温暖化}}{\text{地球温暖}} = 0.98$	$\frac{\text{地球温暖化}}{\text{化}} = 0.07$	○

前後の形態素は限定されるため継続度は大きい値をとる．例えば，表 5 の「温暖 | 化」の前向き継続度では，「温暖」が特徴的な形態素であるため，その後の形態素は「化」に限定され，継続度の値は大きくなる．また両表の下段は，区切りをまたぐ 3 形態素の継続度であり，分母が 1 形態素の場合は，その形態素がかなり特徴的でない限り，前後の 2 形態素が限定されることは少なく，継続度は小さい値をとることが多い．さらに分母が 2 形態素の場合は，その前後の形態素が限定されるため，分母の 2 形態素にまとまりがあれば継続度は大きい値をとる．これらの値を用いて，形態素の区切りを判断するために以下の条件を設定した．

条件 1: 前向きおよび後向き継続度が 0.1 以下の場合，その区切りで形態素を切り離す

条件 2: 前向きおよび後向き継続度が 0.2 以上の場合，その区切りで形態素を結合する

条件 3: 前向きおよび後向き継続度の差が 0.2 以上の場合，その区切りで形態素を結合する

設定された条件から判定された結果を，同じく表 4 および表 5 に示す．表 4 では，2 形態素および 3 形態素の継続度ともに○ (結合) であるため，区切り P_1 は結合されて「地球温暖 | 化」となる．このとき，一方が○ (結合)，一方が× (切り離す) と判定された場合，その区切りは結合しても差し支えないが，両継続度ともに○ (結合) の場合と比較すると，そのつながりの強さは弱くなる．また表 4 の区切り P_2 も同様に結合されるため，「地球温暖化」となる．よって「地球 | 温暖 | 化」は，継続度による区切り判定によって「地球温暖化」に結合された．他の語句に対して区切り判定を行った結果を表 6〜表 8 に示す．

表 6 に示す「高齢 | 者 | 教室」に対する区切り判定では，区切り P_1 のみが結合され，「高齢者 | 教室」となった．また，表 7 に示す「火星 | 探査 | 機」は，区切り P_2 のみが結合され，「火星 | 探査機」となった．漢字の連続からなる 3 形態素の語句では，このような判定が多く見られ，他にも「常任 | 指揮者」「日本人 | 技術者」「長野 | 冬季五輪」などがあった．さらに，表 8 に示す「発売 | 開始 | 日」ように，全ての区切りで結合がない判定も多く見られ，「集団 | 暴行 | 事件」「生活 | 改善 | 薬」「事故 | 対策 | 本部」などがあった．

表 6: 「高齢 | 者 | 教室」に対する区切り判定

	前向き継続度	後向き継続度	判
高齢 | 者	$\frac{\text{高齢者}}{\text{高齢}} = 0.68$	$\frac{\text{高齢者}}{\text{者}} = 0.04$	○
高齢 | 者教室	$\frac{\text{高齢者教室}}{\text{高齢}} = 0$	$\frac{\text{高齢者教室}}{\text{者教室}} = 0.28$	○
者 | 教室	$\frac{\text{者教室}}{\text{者}} = 0$	$\frac{\text{者教室}}{\text{教室}} = 0$	×
高齢者 | 教室	$\frac{\text{高齢者教室}}{\text{高齢者}} = 0$	$\frac{\text{高齢者教室}}{\text{教室}} = 0$	×

表 7: 「火星 | 探査 | 機」に対する区切り判定

	前向き継続度	後向き継続度	判
火星 | 探査	$\frac{\text{火星探査}}{\text{火星}} = 0.12$	$\frac{\text{火星探査}}{\text{探査}} = 0.04$	×
火星 | 探査機	$\frac{\text{火星探査機}}{\text{火星}} = 0.07$	$\frac{\text{火星探査機}}{\text{探査機}} = 0.09$	×
探査 | 機	$\frac{\text{探査機}}{\text{探査}} = 0.28$	$\frac{\text{探査機}}{\text{機}} = 0$	○
火星探査 | 機	$\frac{\text{火星探査機}}{\text{火星探査}} = 0.41$	$\frac{\text{火星探査機}}{\text{機}} = 0$	○

表 8: 「発売 | 開始 | 日」に対する区切り判定

	前向き継続度	後向き継続度	判
発売 | 開始	$\frac{\text{発売開始}}{\text{発売}} = 0$	$\frac{\text{発売開始}}{\text{開始}} = 0.08$	×
発売 | 開始日	$\frac{\text{発売開始日}}{\text{発売}} = 0$	$\frac{\text{発売開始日}}{\text{開始日}} = 0.11$	×
開始 | 日	$\frac{\text{開始日}}{\text{開始}} = 0.14$	$\frac{\text{開始日}}{\text{日}} = 0.01$	×
発売開始 | 日	$\frac{\text{発売開始日}}{\text{発売開始}} = 0.11$	$\frac{\text{発売開始日}}{\text{日}} = 0$	×

5 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．評価実験では，評価に用いた質問文の約 55%におい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，漢字の連続からなる語句に対する用語特定精度の向上を目標とする．

6 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，桝井文人：質問応答技術 ー大量のデータをもとに任意の質問に答えるー，情報処理，45 巻 6 号，2004 年

[2] 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

[3] 大北剛：サポートベクターマシン入門，共立出版 2005.

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

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