複数文質問のタイプ同定

田村 晃裕[†]　高村 大也[††]　奥村 学[††]
[†]東京工業大学 工学部情報工学科　[††]東京工業大学 精密工学研究所
aki@lr.pi.titech.ac.jp　{takamura,oku}@pi.titech.ac.jp

1　はじめに

今日，Web に代表されるように電子化された情報が大量に存在している．この大量の電子化されたテキストから，効率よく情報を獲得する技術が重要になっている．そのような状況で，答えそのものを出力として想定している質問応答システムは，効率よく情報獲得できる有用なシステムといえる．

既存の質問応答システムの評価は，海外では評価型ワークショップ TREC による QA-TRACK，国内では QAC(Question & Answering Challenge) で行われている．これらのワークショップでシステムの入力として与えられる質問は，一文で構成される質問のみである．また，質問形式も「〜ですか？」や「〜の名前は？」のように，疑問形をしているものだけである．このことから，既存の質問応答システムの入力として想定されている質問は，疑問の形をした一文からなる質問のみと考えられる．

一方 Web には，質問をのせておくと他のユーザが答えてくれる Q&A サイトがある．そこによせられる質問には,「餃子の具が少し余ってしまいました．そのまま焼く他に何か調理法はあるでしょうか？」等のような複数文で構成される質問も数多くある．質問形式についても「〜を探しています」や「〜が分かりません」等のように疑問形をしていないものも存在する．

このような複数文からなる質問や疑問形をしていない質問は，既存のシステムでは想定外の入力で，既存の手法をそのまま適用できない．もしくは，適用できたとしても精度が極端に落ちてしまう．

そこで，本研究では，これらの質問にも対応するシステム構築の第一歩として，質問の文の数や形式に依らない質問タイプ同定手法を提案する．一般的な質問応答システムにおいて，質問タイプの同定は，システム全体の精度・効率に関わる重要な作業であるので，この作業を最初に扱うことにした．

これまで，質問タイプの同定に用いられてきた手法は色々あるが，今回想定するタイプ同定の入力は質問の形式を限定しないため，本研究では，パターンマッチなどよりも形式に依存しない統計的機械学習を用いる．機械学習技術にも色々あるが，データ数に対して多くの情報を素性として入れた時にも，精度の高い分類が可能である SVM(Support Vector Machines) を用いてタイプ同定を行う．SVM を用いたタイプ同定としては [1][2] で行われているが，一文の質問しか対象としていないので，本研究で複数文の質問に対しても有効なタイプ同定手法を提案する．

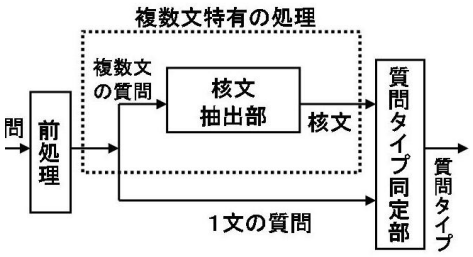

図 1: 質問タイプ同定の流れ

まず，2 節で提案するタイプ同定手法を述べ，3 節で提案手法の有効性を確認する実験とその結果の評価・考察を行う．最後に 4 節で，本研究のまとめと今後の課題について述べる．

2　提案手法

提案する質問タイプ同定手法の流れを図 1 に示す．

まず，質問が与えられたら，前処理として括弧部分を除く．除く理由としては，提案手法では，タイプ同定の際に構文解析を行うのだが，その際に括弧部分が悪影響を及ぼすと考えたからである．

そして，前処理で括弧部分を省かれた後は，質問が一文か複数文かで処理が分れる．質問が一文の場合，そのまま質問タイプ同定部によりタイプ同定される．

質問が複数文の場合，核文抽出部でタイプ同定の際に最も重要な一文（以降「核文」とよぶ）が決定される．この核文が質問タイプ同定部の入力となり，質問タイプ同定部でタイプ同定される．

核文抽出部が複数文質問特有の処理であり，先行研究にはない部分である．

以降，核文抽出部を 2.1 で，質問タイプ同定部を 2.2 で説明する．

2.1　核文抽出部

複数文からなる質問が入力として与えられた時に，核文を抽出する部分である．

複数文質問の場合，質問をそのまま SVM に適用してもノイズが多く，分類精度が落ちてしまう．そのた

め，以下の仮定を導入し質問の核を決めることで，ノイズを除去するという目的でこの処理を導入した．

仮定 1 複数文には，タイプ同定に必要な文とそうでない文がある

仮定 2 タイプ同定は，” タイプ同定に必要な一文” で同定が可能である

仮定 2 の” タイプ同定に必要な一文” を核文と呼び，核文抽出は，この文を抽出することを目的とする．核文抽出の手順は以下の通りである．

step1 文分割
入力として与えられた複数文質問を文に分割する．

step2 一文ずつ核文であるか判定
step1 で文分割された一文ごとに SVM を用いて核文であるか判定する．SVM は，TinySVM[1]を用いた．

step3 核文の決定
step2 で SVM による判定の際，SVM は分離平面からの距離を出力する．この分離平面からの距離を核文としてのふさわしさと捉え，SVM が出力した値が最大となった文を核文に決定する．

step2 の SVM で核文かどうか判定する際の素性とその組み込み方について述べる．素性は，形態素の unigram と bigram を用いた．文の形態素を抽出する際は，日本語形態素解析器 ChaSen[2]を用いた．

素性を組み込む際は，次の点を考慮する必要がある．ある文を核文かどうか判定する際は，判定する文の情報だけではその文が核文かどうか決められないという点である．具体例で，説明する．

質問 1 : よく効く花粉症の薬を教えてください. **頭痛と鼻づまりを治したいです．**よろしくおねがいします.
質問 2 : **頭痛と鼻づまりを治したいです．** よろしくおねがいします.

上の質問 1 と質問 2 の 2 つの質問を考える．それぞれの核文は，下線がひかれた文である．

ここで，質問 1 と質問 2 に太字で書かれた「頭痛と鼻づまりを治したいです.」という文が共に存在している．この太字の文は，質問 1 では核文としふさわしくないが，質問 2 では核文になる．質問 1 では「よく効く花粉症の薬を教えてください.」という文が前にあるため，太字の文が核文として適当でないのである．このことから，太字の文を核文かどうか判定する際，太字の文の情報だけでは核文かどうかを決めることができない．

つまり，核文かどうかを判定する際には，判定する文の前にどのような文があるのか，後にどのような文があるのかを考慮し，相対的に決定しなければならない．

この点を踏まえ，素性ベクトルは判定対象の前後の文の情報を含め，
（判定文より前にある文の unigram，判定文より前にある文の bigram，判定文の unigram，判定文の bigram，判定文より後にある文の unigram，判定文より後にある文の bigram）
の形にした．

この素性と素性の組み込み方の有効性については，3.2 で実験結果と共に示す．

2.2 質問タイプ同定部

一文からなる質問や核文抽出処理により抽出された核文に対してタイプ同定を行う部分である．1 節で述べたように SVM を用いてタイプ同定する．SVM は核文抽出部と同様 TinySVM を用いた．

素性は，形態素の unigram，bigram，名詞の意味カテゴリ，注目名詞，注目名詞の意味カテゴリの 5 つを用いた．これらの素性の有効性については，3.3 で実験結果と共に示す．

意味カテゴリは日本語語彙大系の意味体系 [4] を利用する．

注目名詞とは，先行研究 [1]，[3] の質問対象語（質問文中で回答が何であるかを限定している語）にあたるものである．先行研究では，質問対象語の特定に人手で記述したパターンを用いている．本研究では質問の形式を限定しないため，人手でパターンを作成するのは困難である．そこで，以下の手順により名詞を特定し，その名詞を質問対象の代用として素性に加えた．

step1 最後の動詞を含む文節・「？」で終る文節をみつける

step2 step1 でみつけた文節にかかる文節をみつける

step3 step2 でみつけた文節内の名詞・未知語を注目名詞とする

step2 で用いる係り受け関係は，日本語係り受け解析器 CaboCha[3]を使用した．

3 評価実験・結果と考察

2 節で提案した手法の有効性を確認する実験を行う．

3.1 実験データ

本研究の目的から，実験データとして，複数文の質問や疑問形をしていない質問を数多く含む Q&A サイトの質問を扱う．Q&A サイトは，はてな[4]と Yahoo!知恵袋[5]の 2 つを選び，2000 個ずつ，合計 4000 個の質問を抽出した．この 4000 個の質問に対して質問タイプと核文を人手でタグ付けした．

本研究で定めた質問タイプの種類と 4000 個の質問の分布を表 1 に示す．

[1] http://chasen.org/~taku/software/TinySVM/
[2] http://chasen.naist.jp/hiki/ChaseSen/
[3] http://chasen.org/~taku/software/cabocha/
[4] http://www.hatena.ne.jp/
[5] http://knowledge.yahoo.co.jp/

表 1: 質問タイプと質問の分布

質問の種類	個数	質問の種類	個数
回答が名詞	1139	回答がテキスト	1237
人名	64	理由	132
組織名	37	方法	500
地名	393	定義	73
施設名	139	叙述	228
製品名	238	意見	173
時間表現	108	その他のテキスト	131
数値表現	53	答えが Yes,No のみ	149
その他の名詞	107	2 つ以上尋ねる質問	808
		質問として不適	667

今回の実験では，表 1 の中の「答えが Yes,No のみ」「2 つ以上尋ねる質問」「質問として不適」以外の質問 2376 個をデータとして使用する．

3.2 核文抽出部の実験

質問 2376 個中，一文で構成される質問は 818 個，複数文で構成される質問は 1558 個であった．また，複数文質問の平均文数は 3.49 文であった．したがって，今回の核文抽出というタスクは，複数文質問 1558 個に対して，平均 3.49 文の中から核文を一つ決定することである．このタスクに対して 2.1 で説明した素性・素性の組み込み方の有効性を検証する．

評価は，システムが出力した核文と，人手でタグ付けした正解の核文との一致率で行う．2 分割交差検定によって実験を行った．

まず，提案手法の素性の組み込み方の有効性を確認するために，以下の 4 つのモデルを考え，核文抽出実験を行い比較した．尚，素性は，いずれのモデルも形態素の unigram と bigram を用いている．

判定文と前後全ての文全部：提案手法.
判定文のみ：判定文の素性だけで判定.
判定文と前後 1 文：前の文と後の文の素性として，それぞれ 1 文ずつを使用.
判定文と前後 2 文：前の文と後の文の素性として，それぞれ 2 文ずつを使用.

次に，提案手法の素性の有効性を確認するために,「素性に unigram だけを用いたモデル」,「素性に bigram だけを用いたモデル」に対して核文抽出実験を行い，比較した．尚，素性の組み込み方は，提案したように判定文とその前後の全文の情報を入れている．

以上の 2 つの実験結果を表 2 に示す.

表2 より，核文抽出の際の素性は，unigram と bigram 両方を用いるのが効果的であることが分かる．素性の組み込み方に関しては，判定文の情報だけではなく，提案したように前後の文の情報を組み込んだ方がよいことが分かる．また，前後の文の情報をいれる時は，いれる文の数を増やした方が結果がよい．これは，次の例のように，核文になりそうな文と核文が離れて存在する質問もあり，そのような質問に対しても，うまく

表 2: 核文抽出部の実験結果

素性の組み込み方	結果
判定文と前後全文	**90.9%**
判定文のみ	88.9%
判定文と前後 1 文	89.6%
判定文と前後 2 文	90.2%

素性	結果
unigram,bigram	**90.9%**
unigram のみ	88.3%
bigram のみ	90.3%

表 3: 核文抽出部の有効性の検証

質問タイプ	タイプ同定部の入力	核文抽出精度	タイプ同定結果
製品名	質問そのもの	なし	0.381
	核文抽出実験の結果	90.9%	0.467
	正しい核文	100%	0.480
方法	質問そのもの	なし	0.756
	核文抽出実験の結果	90.9%	0.778
	正しい核文	100%	0.798

核文決定ができるようになっていることを示している．

（例）
1 文目：インストール方法が分かりません．
（核文になりそうな文）
2 文目～：パソコンの環境など（説明）
最後の文: 方法の分かるサイトを教えて． （核文）

3.3 質問タイプ同定の実験

表 1 に示したように，質問タイプによってはデータ数が少ないものもある．そこで今回の実験では，同定する質問タイプとして，回答が名詞になるものから「製品名」，回答がテキストになるものから「方法」を選び，この 2 つに対してタイプ同定実験を行った．実験は，2 分割交差検定で行い，評価は F 値で行う．

3.3.1 核文抽出部の影響

核文抽出部の質問タイプ同定への影響力・有効性を調べる実験を行った．質問タイプ同定部の入力として以下の 3 つのモデルを考え，タイプ同定実験を行い，比較した．

核文抽出部の結果：3.2 の実験結果として出力された核文を用いる.（核文抽出処理の精度 1416/1558=90.9%）
質問そのもの：核文抽出処理をせずに，そのままの質問を使用.
正しい核文：人手でタグを付けておいた正しい核文を使用.（核文抽出処理の精度が 100%になった時の結果に等しい）

素性は，質問そのものをタイプ同定の入力とするモデルも考えるので，形態素の unigram,bigram, 意味カテゴリの 3 つを使用した.（注目名詞を特定する手法は，入力が一文であることを前提としているので，注目名詞に関する素性は使えない.）結果を表 3 に示す.

表 4: 素性の有効性の検証（タイプ同定部）

質問タイプ	素性	タイプ同定結果
製品名	all(5 つの素性)	**0.506**
	all-unigram	0.489
	all-bigram	0.579
	all-意味カテゴリ	0.483
	all-注目名詞	0.512
	all-注目名詞の意味カテゴリ	0.502
方法	all(5 つの素性)	**0.817**
	all-unigram	0.803
	all-bigram	0.787
	all-意味カテゴリ	0.820
	all-注目名詞	0.817
	all-注目名詞の意味カテゴリ	0.807

表3 より，核文抽出をすることで「製品名」では0.086,「方法」でも 0.022 上昇し，核文抽出処理がタイプ同定に有効であることが分かる．これより，2.1 で導入した複数文質問に対する仮定が，ある程度有効であったことがいえる．

また，核文抽出の精度が上がると最終的な質問タイプ同定の結果もよくなっている．

核文抽出の時点で間違っている質問が，正しくタイプ同定できた割合を調べてみると，

「製品名」: 2/16=13%「方法」7/23=30%

であった．このように，核文抽出の時点で間違えていると，質問タイプ同定で正解するのは難しい．タイプ同定の結果をよくするには，核文抽出の精度をあげる必要があることが分かる．

3.3.2 質問タイプ同定の素性の有効性

2.2 で説明した 5 つの素性 (形態素の unigram，bigram，意味カテゴリ，注目名詞，注目名詞の意味カテゴリ) から，一つずつ素性を抜いて同定実験を行った．F 値の変化によって，それぞれの素性の有効性を確認する．

尚，質問タイプ同定の入力は人手でタグ付けした正しい核文を使用した．結果を 4 に示す．

表 4 より，質問タイプによって有効な素性が異なることが分かる．最も有効な素性は,「製品名」に関しては意味カテゴリ,「方法」に関しては bigram である．しかし，bigram は「製品名」では結果を悪くしている．このことから，質問タイプにより同定する際の特徴が違うといえる．

具体的には,「製品名」は” 名詞” により特徴づけられていると思われる．例えば,「〜本を教えてください」では,「本」という” 名詞” がタイプ同定に重要な役割を示している．

一方,「方法」の質問をみてみると,「どのように〜？」などの” 表現” がカテゴリを特徴づけていると思われる．この違いが有効な素性の違いとして表れたと考えられる．

本研究では，注目名詞の特定方法を提案し，質問対象語の代用として素性に加えた．この有効性は，表 4 からはあまり感じられない．この理由は，素性として注目名詞と注目名詞の意味カテゴリの両方をいれたため，片方を素性からはずしても，もう片方が情報を補完したためと考えられる．表 3 に注目名詞の 2 つの情報を抜いた F 値が書いてあるので，その部分と比較すると,「製品名」では注目名詞を入れることにより 0.022 上昇，注目名詞の意味カテゴリについては 0.032 上昇,「方法」でも同様に上昇していることが分かり，有効性が確認できる．

4 おわりに

本研究では，質問の文の数や形式によらない質問タイプ同定手法を提案した．

複数文質問に対応するために，核文抽出処理を導入することを提案し，その有効性を確認した．また，実験より，核文抽出処理の精度の向上が最終的な質問タイプ同定の精度の向上につながることが分かった．

核文抽出の精度をあげるための素性の組み込み方も提案した．判定する文の情報だけではなく，その前後の文の情報も含めることを提案し，実験により有効性が確認できた．

質問の形式に依らないタイプ同定のため，素性として質問対象語の代りに注目名詞という素性をいれた．この注目名詞の特定方法を提案し，素性としての有効性を実験で確認した．ただ，今回提案した注目名詞の特定方法では，質問対象語として適さないものも含まれてしまう．この決定方法の改良は，今後の課題とする．

その他に今後として，今回は 1 つの質問に対して核文は 1 つに限定しているが，この仮定を変えることで，今回の実験では扱わなかった「2 つ以上の事柄をきく質問」などにも対応できるように拡張したい．

参考文献

[1] 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 , pp.635–646, 2004.

[2] 鈴木潤, 佐々木裕, 前田英作. “統計的機械学習を用いた質問タイプ同定”，情報技術レターズ, Vol.1, pp.89–90, 2002.

[3] 秋葉友良，藤井敦，伊藤克且. “質問応答における常識的な解の選択と期待効用に基づく回答群の決定”，情報処理学会, NL-163, pp.131–138, 2004.

[4] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦. “日本語語彙大系（1 意味体系）”，岩波書店, 1997.