Vol.2014-NL-216 No.21
Vol.2014-SLP-101 No.21
2014/5/23

質問応答システムにおける詳細な質問タイプの分類

若山 龍太[1,a)]　白井 清昭[1,b)]

概要：質問タイプの分類はファクトイド型質問応答システムにおける重要な要素技術である．従来手法では，あらかじめ定義されている質問タイプの粒度が粗いため，実用的な質問応答システムに用いるには不十分であるという問題があった．本研究では，関根の拡張固有表現階層に基づく詳細な質問タイプを定義し，質問文の質問タイプを Support Vector Machine (SVM) ならびに k-NN 法を用いて自動分類することを試みる．また，分類器の訓練データとして，正解の質問タイプが付与された質問文のコーパスに加えて，固有表現タグ付きコーパスを併用する手法を提案する．実験の結果，質問タイプ分類の正解率は 60.3%となった．学習素性の有効性を検証した結果，自立語，疑問詞の素性が質問タイプの分類に有効であること，訓練データの量が多いときには単語 bi-gram も有効な素性であることがわかった．一方，訓練データとして固有表現タグ付きコーパスを併用することの効果は確認できなかった．

1. はじめに

質問応答システムとは，自然言語で表現された質問文を入力として受け付けて，文書集合から回答を抽出し，ユーザに提示するシステムである．質問応答の技術は以前から様々な研究機関で研究されてきた．日本では国立情報学研究所による NTCIR(NII Testbeds and Community for Information access Research) [1]，海外では MUC(Message Understanding Conference) [2]，TREC(Text REtrieval Conference) [3] 等といった評価型ワークショップが行われ，それらが提供するタスク設定およびベンチマークデータを利用して，参加者が実装したシステムの性能評価が行われている．

質問応答システムは幾つかのサブシステムから構成され，質問文解析・質問タイプ分類・文書検索・回答候補抽出・回答選択といった処理が行われる．本研究では，質問応答システムの構成要素のうち，質問タイプの分類に着目する．質問タイプとは，質問文が尋ねている事柄の分類である．例えば「メリッサというコンピュータウィルスを作ったのは誰ですか？」という質問文に対する回答は「デービッド・スミス」であるが，この質問は人名を尋ねているので，質問タイプは「PERSON」となる．他にも，地名が問われていることを示す「LOCATION」や，企業名が問われていることを示す「COMPANY」等の質問タイプが用意される．

一般的なファクトイド型の質問応答システムでは，回答候補を抽出する際，その前段の処理としてまず質問タイプを推定し，その質問タイプに適合する回答候補の中から回答を選択する．したがって，質問文のタイプを正しく分類できないと，正しい回答を得ることは難しい．例えば「iPad を開発したのはどこですか？」という質問文は企業名を尋ねており，質問タイプとしては「COMPANY」と分類されるべきである．しかし，誤って異なる質問タイプ（例えば「LOCATION」）に分類されてしまった場合，回答候補抽出機能の性能がいくら高くても正しい回答を選択することはできない．そのため，質問応答システムにおける質問タイプ分類は重要な問題である．一方，質問タイプの分類はそれほど簡単ではない．先に挙げた「iPad を開発したのはどこですか?」という質問は，「どこ」というキーワードは場所を尋ねる質問であることを示唆するが，実際の質問タイプは「LOCATION」ではなく「COMPANY」である．

質問応答システムに関する研究の初期の段階では，質問タイプの分類は人手によって経験的に作成したルールやパターンマッチングによって行われることが多かった．しかし，未知の質問文に対しては質問タイプの分類精度が低下してしまうことや，ルール作成の作業コストが高い等の問題点があった．これらの問題を克服するために，機械学習を用いた質問タイプ分類手法が提案されている [4, 5]．

1.1 研究の目的

本研究では，実用的な質問応答システムで使用することを前提とした質問タイプ分類モジュールを実装することを目的とする．本研究では特に以下の 3 点に特徴がある．

(1) 詳細な質問タイプを分類する．

[1] 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
[a)] ryota.wakayama@jaist.ac.jp
[b)] kshirai@jaist.ac.jp

従来は IREX [6] の固有表現タグに基づく 8 個程度の質問タイプが用いられることが多かった．これに対し，本研究では関根の拡張固有表現階層 [7] を質問タイプの定義として用いる．同階層はおよそ 200 個の詳細な固有表現タグから構成されているため，様々な質問に対して適切な質問タイプを割り当てることができる．

(2) 機械学習に基づく質問タイプ分類手法を実装し，学習素性の有効性を評価する．
多くの先行研究と同様に，本研究でも機械学習の手法を用いて質問タイプを分類する．また，関根の拡張固有表現階層のような詳細な質問タイプを用いたとき，どのような学習素性が有効であるかを実験により評価する．

(3) 異なる 2 種類の訓練データを利用し，正解率の向上を試みる．
質問タイプを分類するモデルを教師あり機械学習するには，正しい質問タイプが付与された質問文を集めたコーパスが必要である．しかし，そのような質問文を集めたコーパス，特に関根の拡張固有表現階層に基づく質問タイプが付与されたコーパスの整備は進んでいない．一方，新聞記事などの一般的なテキストに固有表現タグが付与されたコーパス（固有表現タグ付きコーパス）は整備が進んでおり，現時点でも比較的大規模なコーパスが利用可能である．本研究では，(1) 質問タイプが付与された質問文のコーパス，(2) 固有表現タグ付きコーパスの 2 種類の訓練データを利用する．本来，質問タイプの分類のためには (1) のコーパスが使われるが，大規模なコーパスは存在せず，データスパースネス問題を生じやすい．しかし，(2) のコーパス，すなわち質問タイプに対応する固有表現タグが付与された平叙文からも，質問文のタイプの分類に有用な情報を得ることができると考えられる．また，(2) のコーパスを併用することで訓練データの量を増やすことができる．

2. 関連研究

前節で述べたように，質問タイプは質問応答システムにおいて重要な役割を果たす．多くの質問応答システムでは，質問タイプはあらかじめ人手によって定義される．特に，回答候補（固有名詞）の絞り込みに用いるため，質問タイプは固有表現の種類と同じように定義されることが多い．

従来手法の多くで採用されている質問タイプの定義は粗いものが多く，実用的な質問応答システムでは不十分と考えられる．質問タイプ分類に関する先行研究において，佐々木らの手法 [5] で用いられている質問タイプは，IREX [6] の固有表現タグに基づくもので 8 種類 (PERSON:人名，LOCATION:地名，ORGANIZATION:組織名，ARTIFACT:製品名/作品のタイトル，DATE:日付，TIME:時間，MONEY:金額，PERCENT:割合) しかない．また，鈴木らの手法 [4] で用いられている質問タイプでは 17 種類 (AGE:年齢，DATE:日付，EVENT:事柄，LOCATION:場所, MONEY:値段, NORGANIZATION:組織名数, NPERSON:人数，ORGANIZATION:組織名，PERCENT:割合，PERIOD:期間, PERSON:人名, PRODUCT:製品名, PTITLE:役職名，SUBSTANCE:物質名，TIME:時間，TITLE:作品名, OTHER:その他) である．質問タイプの種類が少ないと，与えられた質問に対する回答候補を適切に絞り込めないという問題が生じる．例えば「夏目漱石が生まれた町はどこですか?」と「世界で一番長い川は何ですか?」という質問文は，従来の粗い質問タイプではともに LOCATION に分類されるが，実際に尋ねている内容は，前者が市区町村名，後者は河川名と大きく異なっており，単純に質問タイプを LOCATION と分類してしまうと，正しい回答を選択することは難しい．

一方，遠藤らは詳細化された質問タイプを提案している [8]．彼らは，IREX の固有表現タグの下位分類を新たに定義し，およそ 60 個からなる質問タイプの集合を作成した．質問タイプの分類は，ヒューリスティクスならびにコーパスから得られた動詞と固有表現の共起頻度に基づいて行う．ただし，質問タイプの分類には詳細に分類された固有表現辞書を必要とするが，これは人手で作成されているため，機械学習に基づく手法とは異なり，多様な質問文への対応が困難であるという問題点がある．

本研究では，200 種類のカテゴリーを持つ関根の拡張固有表現階層をもとに，遠藤らの手法よりも詳細化した質問タイプを用意し，機械学習に基づく手法で質問タイプを分類するシステムを構築する．我々の知る限り，これまで機械学習手法を適用して関根の拡張固有表現タグに基づく詳細な質問タイプの分類が試みられたことはない．

3. 提案手法

3.1 質問タイプの定義

本研究では，質問タイプの定義として関根の拡張固有表現階層 [7] を利用する．その一部を**図 1** に示す．

関根の拡張固有表現階層は名前を中心とした単語の意味の分類で，固有表現の種類毎に階層的に分類されており，200 種類の固有表現のタイプが定義されている．

3.2 質問タイプの判定

本研究では，質問タイプ分類モジュールを構築するための機械学習アルゴリズムとして，Support Vector Machine(SVM) と k-NN 法の 2 つを用いる．まず，個々の訓練データを素性ベクトルで表現する．ここでの素性ベクトルは，後述する学習素性を次元とし，素性が存在するときに重みを 1 とする二値ベクトルである．SVM の学習には

ENE			ENE英語表記	例
名前_その他			Name_Other	たま，ポチ，オグリキャップ，トントン
人名			Person	岡本文弥，カーン，長門美保，フォスター，武帝
神名			God	アテネ，インドラ，ゼウス，大国主命，帝釈天
組織名(Organization)	組織名_その他		Organizaton_Other	総務課，孔門の十哲，向田ファミリー，精華町町内会，第二工学部
	国際組織名		International_Organization	国際連盟，イスラム諸国会議機構，南太平洋フォーラム，東南アジア条約機構
	家系名		Family	久我氏，清水家，近衛家，伏見宮家
	民族名(Ethnic_Group)	民族名_その他	Ethnic_Group_Other	ケルト人，モンゴロイド，トラジャ(人)，チェコ人，アフリカーナー
		国籍名	Nationality	イスラエル人，アメリカ人，日本国籍
地名(Location)	地名_その他		Location_Other	タイムズ・スクエア，グランド・ゼロ，日本三景，天国，エデンの園
	地形名(Geological_Region)	地形名_その他	Geological_Region_Other	アルタミラ洞窟，野島断層，秋芳洞，阿波の土柱，利根川構造線
		山地名	Mountain	富士山，間ノ岳，青崩峠，中央アルプス，木曽駒ケ岳
		島名	Island	ラクシャドウィープ諸島，友ヶ島，大スンダ列島，西表島，沖縄諸島
		河川名	River	早出川，アーレ川，マージー川，千種川，ダニューブ川
		湖沼名	Lake	大浪池，グレート湖，シルヤン湖，丸沼，サロマ湖
		海洋名	Sea	日本海，バルト海，周防灘，関門海峡，ホルムズ海峡
		湾名	Bay	シェレホフ湾，浦戸湾，九十九湾，ピョートル大帝湾，ベンガル湾

図 1 関根の拡張固有表現階層 (抜粋)
Fig. 1 Sekine's Extended Named Entity Hierarchy(excerpt)

liblinear [*1] を利用した．一方，k-NN 法では，データ間の距離あるいは類似度を測る尺度が重要な役割を担う．本研究では，素性ベクトルを素性の集合とみなし，Dice 係数によってデータ間の類似度を測った．SVM, k-NN 法はともに教師あり機械学習であり，訓練データとして正しい分類クラスが付与されたデータの集合が必要である．訓練データの詳細は 3.4 項で述べる．

3.3 学習素性

質問タイプを分類するモデルを学習するための素性として，自立語，単語 bi-gram，疑問詞，係り受け関係の 4 つを用いた．これらの素性を得るために，訓練データの文に対して形態素解析や文節の係り受け解析を行う．形態素解析には MeCab [9] を，文節の係り受け解析には CaboCha [10] を利用した．以下，それぞれの学習素性について説明する．

(1) 自立語

質問文中に出現する自立語を素性とする．例えば「エアロスミスのデビュー作は何ですか.」という質問文の場合は，以下のような単語を素性として抽出する．

エアロ スミス デビュー 何

あらかじめ自立語に相当する品詞のリストを用意し，形態素解析の結果得られた品詞がそのリストに存在する単語の基本形を素性として抽出する．ただし，「する」「れる」「いる」「ある」「なる」「いう」の語の品詞は動詞であるが，例外的に自立語の素性としない．

[*1] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 1 学習素性として利用する疑問詞の一覧
Table 1 List of interrogatives used as features

何, どこ, 何処, どちら, どなた, いつ, 何時, いくつ, 幾つ, だれ, 誰, どう, どの, なぜ, 何故

(2) 単語 bi-gram

単語 bi-gram とは，N-gram において N が 2 である単語列，すなわち連続する 2 単語の列を表す．単語 bi-gram は，単語列に対して 2 単語単位で 1 単語ずつずらして単語列を抽出して得られる学習素性である．例えば「夏目漱石の名作は何ですか.」という質問文の場合は，以下のような 2 単語ずつの単語列を抽出する．

夏目+漱石 漱石+の の+名作 名作+は は+何 ...

(3) 疑問詞

「何」,「いつ」,「誰」等の疑問詞を学習素性として利用する．疑問詞は，質問文が尋ねている内容や意味を類推する上で重要な手がかりの一つとなる．本研究では，**表 1** に示す疑問詞を学習素性とした．

(4) 係り受け関係

文節の係り受け解析を行い，係り受け関係にある語を抽出して学習素性とする．例えば「夏目漱石の名作は何ですか.」という質問文から**図 2** に示すような文節の係り受け解析結果が得られる．図 2 において，/は単語の境界を，矢印は文節の係り受け関係を表す．文節の係り受け関係から，それぞれの文節の主辞（太字で示された単語）を取り出し，係り受け関係にある語のペアとして抽出する．したがって，図 2 の質問文か

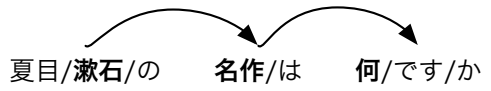

図 2 文節の係り受け解析の例
Fig. 2 Example of dependency analysis between *bunsetsu*

夏目漱石の名作は何ですか．,Book
彼の長男の職業は何ですか．,Position_Vocation
日本三大祭りはなんですか．,Occasion_Other
エアロスミスがデビューしたのはいつですか．,Date

図 3 QAC 質問文コーパス (抜粋)
Fig. 3 QAC corpus(excerpt)

らは以下のような係り受け関係が学習素性として抽出される．

漱石→名作 名作→何

3.4 訓練データ

本研究では訓練データとして，以下の 2 つのコーパスを利用する．

(1) Question Answering Challenge–1(QAC–1) [11] より公開されている質問文のデータセット（以下，「QAC 質問文コーパス」と称す）
(2) 毎日新聞のテキストに対して関根の拡張固有表現階層に基づくタグが付与された拡張固有表現タグ付きコーパス [12] （以下，「新聞コーパス」と称す）

QAC 質問文コーパスは QAC–1 より提供されているデータであり，質問文とその回答の集合である．このデータは質問応答システムの評価に用いられる．ただし，本研究では質問文の質問タイプを分類することを目的としているため，QAC–1 のテストコレクションのうち質問文のみを利用する．次に，各質問文に対し，その正しい質問タイプを人手で付与した．関根の拡張固有表現階層の中から，その質問の回答に該当する固有表現のタイプをひとつ選択し，質問タイプとして付与した．QAC 質問文コーパスの一部を**図 3** に示す．Book, Occasion_Other, Date などが人手で付与された質問タイプである．

現在利用可能な QAC 質問文コーパスは，質問文数が 1,218 個と少なく，これのみを訓練データとするのは不十分であると予想された．そのため，QAC 質問文コーパスに加えて，毎日新聞の拡張固有表現タグ付きコーパスを利用することによって訓練データの増加を図る．

本研究で利用する新聞コーパスは，新聞記事に対して約 29 万個の固有表現タグが付与されたデータである．新聞コーパスにおいては，多くの文は質問文ではなく平叙文であるので，本来は質問文の質問タイプを分類するモデルの学習データとして利用することはできない．しかしながら，利用する新聞コーパスの文には固有表現タグが付与されており，この固有表現タグを利用することにより質問タイプ（固有表現）と関連の深い素性を学習できる可能性がある．例えば，文学名 (Book) という固有表現の周囲に「名作」という単語が頻出することが学習できれば，「夏目漱石の名作は何ですか」の質問タイプを Book と正しく分類できる．ここでは，文中に固有表現が一つ含まれている文について，その固有表現タグを文の仮想的な質問タイプとみなし，文と質問タイプの組を獲得する．これらを質問タイプ分類モデルを機械学習するための訓練データとする．ただし，平叙文には疑問詞は出現しないので，新聞コーパスを訓練データとするときは，3.3 項で述べた 4 つの素性のうち疑問詞の素性は用いない．

4. 評価実験

本節では，3 節で述べた質問タイプの分類手法の評価実験について報告する．

4.1 実験方法

本実験では QAC 質問文コーパスをテストデータとして使用した．QAC 質問文コーパス，新聞コーパス，両者の併用の 3 種類の訓練データを用いて，SVM ならびに k-NN 法による分類器を学習し，正解率 (分類器の出力する質問タイプが正解と一致する割合) を測った．QAC 質問文コーパスを訓練データとしたときは，5 分割交差検定で質問タイプ分類の正解率を評価した．QAC 質問文コーパスと新聞コーパスを併用したときも 5 分割交差検定を行ったが，訓練データは 80%の質問文コーパスと新聞コーパスをあわせたものを用いた．また，学習素性の有効性を評価するために，素性集合を変えて分類器の学習を行い，正解率を比較した．

4.2 実験結果

SVM による質問タイプ分類の実験結果を**表 2** に示す．この表における「訓練コーパス」の列は SVM の学習に用いた訓練コーパスを示している．「QAC」「新聞」の列の●はそれぞれ QAC 質問文コーパス，新聞コーパスを用いたことを表す．「学習素性」の列は SVM で用いた素性を示している．「自立語」「単語 bi-gram」「疑問詞」「係り受け関係」の列の●はそれぞれの素性を用いたことを表す．素性の組み合わせとしては，すべての素性を使うか，あるいは 1 つの素性を除くかのいずれかであり，素性の組み合わせの説明を「内容」の列に記した．なお，平叙文の集合である新聞コーパスを使うときは疑問詞の素性は常に使わない．最後に「平均正解率」は質問タイプ分類の正解率を示しており，5 分割交差検定における 5 回の試行の平均である．

また，k-NN 法による実験結果を**表 3** に示す．「訓練データ」「学習素性」「平均正解率」は，表 2 と同じく，使用した訓練データと学習素性ならびにそのときの正解率を示している．「k の値」は k-NN 法における k，すなわち質問タ

表 2　SVM による質問タイプ分類の実験結果
Table 2　Results of question type identification by SVM

訓練コーパス		学習素性				内容	平均正解率
QAC	新聞	自立語	単語bi-gram	疑問詞	係り受け関係		
●		●	●	●	●	全素性を使用	59.0%
			●	●	●	自立語なし	58.5%
		●		●	●	単語bi-gramなし	60.3%
		●	●		●	疑問詞なし	58.6%
		●	●	●		係り受け関係なし	59.9%
	●	●	●	−	●	全素性を使用	18.3%
			●	−	●	自立語なし	17.7%
		●		−	●	単語bi-gramなし	15.4%
		●	●	−		係り受け関係なし	18.1%
●	●	●	●	−	●	全素性を使用	56.1%
			●	−	●	自立語なし	55.3%
		●		−	●	単語bi-gramなし	54.4%
		●	●	−		係り受け関係なし	55.7%

イプの判定に用いる最近傍データの数を表す．今回の実験では $k = 1, 3, 5$ とした．

4.3　考察

本項では，質問タイプの違い，学習アルゴリズムによる違い，新聞コーパスを訓練データとして用いることの効果，学習素性の有効性などの観点から実験結果について考察する．

4.3.1　質問タイプによる違い

前項で報告した実験結果のうち，最高の正解率を得たのは，QAC 質問文コーパスを訓練データとし，提案する 4 つの素性のうち単語 bi-gram を除いた 3 つを学習素性として利用し，学習アルゴリズムとして SVM を用いたときで，その正解率は 60.3%(表 2 より) であった．ただし，60.3%という正解率自体は先行研究と比べてかなり低い．例えば，佐々木らは SVM を用いた質問タイプの分類システムを実装しており，その正解率を 88.0%と報告している [5]．ただし，佐々木らの研究では質問タイプの種類は 8 種類であるのに対し，本研究では関根の拡張固有表現階層に基づく 200 種類の質問タイプを使用している点が異なる．質問タイプの数が増えれば増えるほど質問タイプの自動判定は難しくなるため，質問タイプの数が少ない先行研究に比べて，詳細な質問タイプを用いた今回の実験の正解率が低いことは自然な結果である．ただし，実用的な観点から言えば，60.3%という正解率は十分ではなく，大幅な改善が必要である．

4.3.2　学習アルゴリズムによる違い

本実験では，機械学習アルゴリズムとして SVM と k-NN 法の 2 つを採用した．QAC 質問文コーパスを訓練データとし，学習素性として自立語・疑問詞・係り受け関係の 3 つの素性を用いたとき，SVM の正解率は 60.3% (表 2 より)，k-NN 法の正解率は 52.0%(表 3 より，ただし $k = 5$ のとき) であった．また，QAC 質問文コーパスと新聞コーパスの両方を訓練データとし，疑問詞以外の素性を用いたときには，SVM の正解率は 56.1%(表 2 より)，k-NN 法の正解率は 51.3%(表 3 より，ただし $k = 5$ のとき) であった．これらの結果から，今回の実験では，SVM は k-NN 法より質問タイプを分類するための手法として適していることがわかる．

また，表 3 の結果を見ると，k-NN 法で k の値を 1,3,5 と変化させたとき，$k = 5$ のときが正解率が一番高くなる傾向が見られる．今回の実験では 3 種類の k についてしか実験を行わなかったが，k を 5 より大きく設定したときの正解率は調べる価値がある．

4.3.3　新聞コーパスを訓練データとして用いることの効果

3.4 項で述べたように，新聞コーパスの使用は，関根の拡張固有表現階層に基づく詳細な質問タイプが付与された質問文のコーパスの量が少ないという問題に対し，固有表現を含む新聞記事中の平叙文を訓練データとして流用するという考えに基づいている．そこで，訓練データとして新聞コーパスを併用することの効果を検証する．

SVM の場合，表 2 より，QAC 質問文コーパスを訓練データとしたときの正解率は最高で 60.3%であるのに対し，QAC 質問文コーパスと新聞コーパスの両方を訓練データとしたときの正解率は 56.1%であった．したがって，新聞コーパスを訓練データとして使用することの効果は見られなかった．ただし，前者は疑問詞の素性を用いているのに対し，後者では使用していない．疑問詞の素性を用いていないことが，後者が前者の正解率より劣る原因になっている可能性もある．そこで，同じ素性集合 (疑問詞の素性を除いた素性集合) で比較すると，QAC 質問文コーパスを訓練データとしたときの正解率は 58.6%となり，2 種類のコーパスを訓練データとして利用したときの正解率 (56.1%) はこれよりも低い．したがって，同じ素性集合で比較しても新聞コーパスを併用することの有効性は確認できなかった．

k-NN 法の場合，表 3 より，QAC 質問文コーパスを訓練データとしたときの正解率は 51.6%であるのに対し，QAC 質問文コーパスと新聞コーパスの両方を訓練データとした

表 3 k-NN 法による質問タイプ分類の実験結果
Table 3 Results of question type identification by k-NN

訓練コーパス		学習素性				内容	kの値	平均正解率
QAC	新聞	自立語	単語bi-gram	疑問詞	係り受け関係			
●		●	●	●	●	全素性を使用	1	49.0%
							3	51.2%
							5	51.6%
			●	●	●	自立語なし	1	48.5%
							3	50.8%
							5	50.9%
		●		●	●	単語bi-gramなし	1	49.4%
							3	51.8%
							5	52.0%
		●	●		●	疑問詞なし	1	48.3%
							3	49.4%
							5	49.9%
		●	●	●		係り受け関係なし	1	48.3%
							3	50.6%
							5	51.2%
	●	●	●	–	●	全素性を使用	1	13.7%
							3	15.0%
							5	15.3%
			●	–	●	自立語なし	1	13.0%
							3	13.7%
							5	14.2%
		●		–	●	単語bi-gramなし	1	11.3%
							3	11.5%
							5	12.3%
		●	●	–		係り受け関係なし	1	14.0%
							3	14.4%
							5	15.1%
●	●	●	●	–	●	全素性を使用	1	48.2%
							3	50.3%
							5	51.3%
			●	–	●	自立語なし	1	46.9%
							3	48.8%
							5	49.6%
		●		–	●	単語bi-gramなし	1	46.0%
							3	49.5%
							5	50.4%
		●	●	–		係り受け関係なし	1	46.6%
							3	49.5%
							5	50.5%

ときの正解率は 51.3%であった (いずれも $k = 5$ の場合). 正解率の差は SVM のときよりは大きくないものの，新聞コーパスを併用したときの正解率は QAC 質問文コーパスのみを訓練データとしたときよりも劣っている．また，疑問詞の素性を除いた素性集合で比較すると，QAC 質問文コーパスを訓練データとしたときの正解率は 49.9%であるのに対し，QAC 質問文コーパスと新聞コーパスの両方を訓練データとしたときの正解率は 51.3%となり (いずれも $k = 5$ の場合)，新聞コーパスを併用することで若干の改善が見られた．この実験結果からは，固有表現タグの付与された平叙文を質問タイプ分類モデルの学習に使うことの有効性が確認できる．ただし，質問タイプの分類に疑問詞の素性が有効であることはある程度自明であり，質問文のコーパスのみを使うときには当然疑問詞の素性を利用するべきである．また，k-NN 法の正解率は SVM よりも低い．したがって，k-NN 法において疑問詞の素性を使わないときに新聞コーパスを併用することの有効性が確認できたとはいえ，この結果は実用的な観点からはあまり意味がない．

新聞コーパスの使用が質問タイプ分類の正解率向上に貢献しない理由を考察する．QAC 質問文コーパスから得られる 4 種類の素性の異なり数は 12,098 個であるのに対し，新聞コーパスから得られる素性の数は 178,667 個であった．素性の数は大幅に増えているのにも関わらず正解率が向上しないのは，両者の素性集合に隔たりがあることが原因の一つと考えられた．そこで，QAC 質問文コーパスから獲得された素性集合 (12,098 個) の素性のうち，新聞コーパスから獲得された素性集合 (178,667 個) にも含まれるものの割合を調べたところ，2.4%しかなく，両者の素性集合にほとんど重なりがないことがわかった．特に，k-NN 法では，重複する素性の数が少ないことから，テスト文と訓練データ中の文の類似度を Dice 係数で求めても，類似度が 0 となる文がほとんどであり，テスト文と似ている文を検索できなかったために正解率が低かった．なお，本実験で用いた 4 種類の学習素性だけでは，QAC 質問文コーパスと新聞コーパスの素性集合の重なりが小さかったが，両コーパスに共通して出現し，かつ質問タイプの分類にも有効な別の素性が発見できれば，新聞コーパスが質問タイプの正解率向上に貢献する可能性がある．

4.3.4 学習素性の有効性の検証

次に，自立語，単語 bi-gram，疑問詞，係り受け関係の 4

種類の学習素性の有効性について検証する．ここでは，全素性集合と 1 つの素性を除いた素性集合を用いたときの正解率を比較し，後者が前者に比べて正解率が大きく低下するときに，その素性は有効性が高いとみなす．

まず，QAC 質問文コーパスを訓練データとしたときについて考察する．表 2 から，SVM の場合，有効な素性は自立語と疑問詞で，両者の貢献度はほとんど差がない．一方，単語 bi-gram，係り受け関係の素性は，これを除いた素性集合を用いたときの正解率が全素性を用いたときよりも高くなり，悪影響を及ぼすことがわかった．一方，k-NN 法 ($k = 5$) の場合，一番有効に働く素性は疑問詞で，次に自立語，係り受け関係の順となる．単語 bi-gram の素性は悪影響を及ぼすことがわかった．単語 bi-gram や係り受け関係は素性の種類が多く，訓練データの量が少ないときは過学習を起こしやすい．QAC 質問文コーパスの量は 1,218 文と少ないため，単語 bi-gram や係り受け関係の素性が有効に働かなかったと考えられる．

次に，新聞コーパスを訓練データとしたときの素性の有効性について考察する．表 2 から，SVM の場合，一番有効に働く素性は単語 bi-gram であり，次いで自立語，係り受け関係となる．一方，表 3 から，k-NN 法 ($k = 5$) の場合，同様に一番有効に働く素性は単語 bi-gram で，次いで自立語，係り受け関係となる．QAC 質問文コーパスと比べて新聞コーパスははるかに量が多いため，素性の種類が多い単語 bi-gram でも有効に働いたと考えられる．

最後に，QAC 質問文コーパスと新聞コーパスの両方を訓練データとしたときの素性の有効性を考察する．表 2 から，SVM の場合，質問タイプの判定の正解率向上に大きく寄与する素性は単語 bi-gram，自立語，係り受け関係の順となる．一方，表 3 から，k-NN 法 ($k = 5$) の場合，一番有効に働く素性は自立語であった．単語 bi-gram と係り受け関係の素性はほとんど差がない．

訓練データの違いによって有効な素性が異なるので一概には言えないが，全体の傾向としては，質問タイプの分類に有効なのは疑問詞，自立語の素性である．また，訓練データの量が大きいときは単語 bi-gram も有効に働く．

疑問詞の素性についてさらに検証してみよう．表 2 より，学習アルゴリズムとして SVM，訓練データとして QAC 質問文コーパスを用いたとき，全素性を使ったときの正解率は 59.0%なのに対し，疑問詞の素性を除いたときの正解率は 58.6%とあまり変わらなかった．さらに，このときの両者における素性の数を調べると，全素性を用いたときの素性数は 12,098 個，疑問詞を含めなかったときは 12,088 個であり，ほとんど差がない．つまり，訓練データに出現する疑問詞の素性 (疑問詞の種類) は 10 個しかなかった．ただし，SVM では効果が薄かったが，k-NN 法では 4 つの素性の中で最も有効性が高かった．また，直観的にも，質問タイプの分類に「誰」「どこ」などの疑問詞は有効に働くと考えられる．

5. おわりに

本研究では，質問応答システムにおける質問タイプ分類の性能の向上を図るために，質問タイプとして新たに関根の拡張固有表現階層を利用し，機械学習に基づく手法で実装した．自立語，単語 bi-gram，疑問詞，係り受け関係の４種類の学習素性を用いて，SVM・k-NN 法の２種類の機械学習アルゴリズムにより分類モデルを学習し，質問タイプ分類の正解率を測定した．その結果，機械学習アルゴリズムとして SVM を利用した場合では，質問タイプ分類の正解率は 60.3%という結果が得られた（訓練データとして QAC 質問文コーパスのみを利用し，学習素性を自立語・疑問詞・係り受け関係の 3 種類を用いた場合）．

また，4 種類の学習素性の有効性を検証した結果，QAC 質問文コーパスを訓練データとした場合，質問文中に出現する自立語や疑問詞が学習素性として有効であることがわかった．一方，新聞コーパスを訓練データとして用いる場合，つまり訓練データの量が十分に多い場合には，単語 bi-gram も有効な学習素性であることがわかった．

質問タイプの分類モデルを学習するための訓練データとして，２種類のコーパス（QAC 質問文コーパスおよび新聞コーパス）を利用する手法について検証した．しかしながら，質問文のコーパスに加えて平叙文からなる新聞コーパスを併用することで，質問タイプ分類の正解率を向上させることはできなかった．質問タイプの分類に有効で，かつ質問文と平叙文に共通して頻出する素性の発見が今後の課題となる．

参考文献

[1] NTCIR: NII Testbeds and Community for Information access Research Project homepage, http://research.nii.ac.jp/ntcir/index-ja.html.
[2] Ralph, G. and Sundheim, B.: Message Understanding Conference - 6: A Brief History, *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, pp. 466–471 (1996).
[3] TREC: Text REtrieval Conference homepage, http://trec.nist.gov/.
[4] 鈴木　潤，佐々木裕，前田英作：単語属性 N-gram と統計的機械学習による質問タイプ同定，情報処理学会論文誌，Vol. 44, No. 11, pp. 2839–2853 (2003).
[5] 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).
[6] IREX: Information Retrieval and Extraction Exercise NE homepage, http://nlp.cs.nyu.edu/irex/NE/.
[7] Sekine, S. and Nobata, C.: Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy, *Proceedings of LREC*, pp. 1977–1980 (2004).
[8] 遠藤哲哉，福本淳一：詳細化された質問タイプによる質問応答システム，情報処理学会研究報告，Vol. 2004-NL-159, No. 1, pp. 25–30 (2004).

[9] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.
[10] 工藤　拓，松本裕治：チャンキングの段階適用による日本語係り受け解析，情報処理学会論文誌，Vol. 43, No. 6, pp. 1834–1842 (2002).
[11] Fukumoto, J. and Kato, T.: An overview of Question and Answering Challenge (QAC) of the next NTCIR workshop, *Proceedings of the Second NTCIR Workshop Meeting*, pp. 375–377 (2001).
[12] 橋本泰一, 乾　孝司, 村上浩司：拡張固有表現タグ付きコーパスの構築，情報処理学会研究報告，Vol. 2008, No. 113, pp. 113–120 (2008).