課題研究報告書

質問応答システムにおける詳細な質問タイプの同定手法の実装と評価

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

若山 龍太

2014 年 3 月

課題研究報告書

質問応答システムにおける詳細な質問タイプの同定手法の実装と評価

指導教官　白井 清昭 准教授

審査委員主査　白井 清昭 准教授
審査委員　島津 明 教授
審査委員　飯田 弘之 教授

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

1110702 若山 龍太

提出年月: 2014 年 2 月

概 要

本課題研究では、質問応答システムにおける質問タイプの同定タスクに関する新たな手法を提案し、その実装および評価を行った。従来の質問タイプ同定タスクでは、人手で経験的なルールにより質問タイプを作成していたり、質問タイプの定義が粗いため実用的な質問応答システムとして不十分であること等の問題点があった。

本研究ではこれら従来手法の問題点を改善するため、質問タイプとして関根の拡張固有表現階層 [1] を利用し、質問タイプの数を、従来手法でよく用いられた 8 種類から 200 種類へ拡張した。併せて、教師あり機械学習に基づく手法を用いることにより、未知の質問文に遭遇した場合における質問タイプ同定性能の向上を試みた。また、本研究において利用可能であった質問文コーパスは 1,218 文と数が少なく、これのみを機械学習のための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、質問文コーパスに加えて、毎日新聞の固有表現タグ付きコーパスを用いて訓練データの増強を図った。

これらの施策に基づき実験を行った結果、質問タイプの正解率は、機械学習アルゴリズムとして Support Vector Machine(SVM) を利用した場合では、平均 60.3%（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・単語 bi-gram・係り受け関係の 3 種類としたとき）という結果が得られた。

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　研究の背景

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、文書集合から回答候補を抽出し、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。日本では国立情報科学研究所による NTCIR(NII Testbeds and Community for Information avccess Research)[2]、海外では MUC(Message Understanding Conference)[3]、TREC(Text REtrieval Conference)[4] 等といった評価型会議が行われ、各会議の提供するタスク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・質問タイプ同定・回答候補抽出・回答選択といった処理が行われる。質問応答システムを構成する要素技術は、自然言語処理や大規模データ解析等、様々な分野に応用することが可能であり、情報化社会において重要性の高い技術である。本課題研究では、質問応答システムの構成要素のうち、質問タイプの同定に焦点を当てている。質問タイプとは、質問文が問うている事柄を分類したものである。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、人名を問うているので、この場合の質問タイプは「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同定し、その質問タイプに関連付けられている回答候補の中から回答を選択する。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」と同定されるべきである。しかし、誤って異なる質問タイプ（例えば「LOCATION」）が同定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同定問題は極めて重要な要素の一つである。一方、質問タイプの同定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問う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。

従来、質問応答システムにおける質問タイプの同定は、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同定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同定手法が提案されている [6]。

1.2 研究の目的

本課題研究では、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同定モジュールを実装することを目的とする。本課題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同定する。

   従来は IREX[5] 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、本課題研究では関根の拡張固有表現階層 [1] 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同定手法を実装し、学習素性の有効性を評価する。

   多くの先行研究と同様に、本課題研究でも機械学習の手法を用いて質問タイプを同定する。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同定する際、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」と呼ばれる）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。本課題研究では、(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、質問タイプの同定のためには (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、質問文のタイプの同定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。このような考えに基づき、質問タイプ同定の正解率を向上させるために (1) と (2) の 2 種類の訓練データを利用する。

1.3　報告書の構成

本課題研究報告書の構成は以下の通りである。2 章では本課題研究の関連研究について述べる。3 章では質問タイプの同定モジュールの実装について述べる。4 章では実装したモジュールの評価実験について報告する。最後に 5 章で本課題研究のまとめと今後の課題を述べる。

第2章　関連研究

情報検索の評価型ワークショップである TREC に質問応答タスクが導入されて以来、汎用ドメイン質問応答システムに対する関心は高まっているが、自然言語による質問応答システムは新しい研究分野ではなく、1950 年代から研究が始まり、今日まで様々なシステムが提案されてきた。本章では、これまでに考案された幾つかの質問応答システムを取り上げて概観するとともに、一般的な質問応答システムの構成について説明する。最後に質問タイプの同定に関する先行研究をいくつか紹介するとともに、その問題点を論じる。

2.1　様々な質問応答システム

2.1.1　ELIZA

ELIZA[9] は、人と機械の自然言語による会話を可能とすることを目的として、1966 年に Joseph Weizenbaum により開発されたシステムである。

ELIZA は、対話で参照されている世界に関する知識を持たず、(1) キーワードの同定、(2) 最小の文脈の発見、(3) 適切な質問文変換の選択、(4) キーワードが欠落している場合の応答生成、(5) 会話の終了判定 等の機能によって、ユーザが ELIZA に対して行った発話の一部を利用しながら会話を継続する。

2.1.2　LUNAR

LUNAR[10] は、1973 年に Bolt Beranek and Newman Inc. (BBN) で開発されたシステムで、地質学者が月面で採取された岩石や地表物質のデータベースに自然言語（英語）でアクセスし、科学分析データを容易に比較、評価できるようにすることを目標としていた。ユーザは、データベースアクセス用の専用コマンドを入力する必要はなく、“Give me all lunar samples with Magnetite” や “How many samples contain Titanium” といった自然言語で表現された質問文によりデータベースへの問い合わせを行う。これを実現するために、LUNAR は質問文の構文解析器、規則を基にした意味解釈コンポーネント、 データベース検索、推論モジュールから構成されている。

2.1.3 IBM Watson

IBM Watson [11] は、IBM 社のグランドチャレンジ (成功する保証のない、技術的に困難な課題への取り組み) の一環として、IBM Research が 2005 年から開発に着手した質問応答システムである。2012 年 2 月に、米国クイズ番組「Jeopardy!」において、同番組の人間のクイズチャンピオン 2 人と対戦して勝利する結果を収めている。Jeopardy!は、1964 年から続いているクイズ番組で、既に 9,000 回以上放映されている。クイズは、歴史・科学・スポーツ等の幅広い分野から出題され、知識が問われる通常の問題に加え言葉遊びのような問題が提示される場合もある。クイズの方式としては、司会者が問題を読み上げると同時に問題文が提示され、回答者は早押しで回答する。各問題には事前に賞金が設定され、回答が正しければ回答者にその賞金が与えられ、不正解であればその賞金と同じ金額を失い、他の回答者に回答権が移る。例えば次のような問題が出題される。

> 「WHILE MALTESE BORROWS MANY WORDS FROM ITALIAN, IT DEVELOPED FROM A DIALECT OF THIS SEMITIC LANGUAGE」(マルタ語はイタリア語から多くの語彙を借りているが、それはこのセム語系言語の方言から発展した)

この場合の正解は「アラビア語」であるのだが、この問題文で本質的に問われている内容を解析し、また代名詞の適切な照応解析を行うことは困難である。また、問題文の前半部分が回答に直接関係していないことも回答をさらに困難にしており、情報の適切な取捨選択も必要となっている。

このような出題に対し精度の高い回答を応答するために、IBM Watson では、意味を考慮したマッチングにより正答を探索する機能を導入している。すなわち、回答候補の中から唯一の回答を抽出する際に、その回答の正しさの根拠をスコアリングし、回答＋確信度のリストから閾値以上の一番高い確信度をもつ答えを選出する、という処理を行っている。一方、IBM Watson では以下のような機能は実装されていない。

1. 音声認識
2. 画像認識
3. インターネット上のデータの探索（対戦時）
4. 人間のような感情や直感

IBM 社では、IBM Watson で培った技術を、医療診断支援、潜在的な薬物相互作用検査、裁判における判例参照等の分野で今後活用する。

2.2　質問応答システム概要

IBM Watson は高度な質問応答性能を有しているが、訓練データの収集、学習アルゴリズムの検討等、多大なコストをかけて実現している。また、当該クイズ番組以外の出題形式への対応は発展途上であり、どのような質問にも万能的に回答できるわけではない。

本節では、従来の一般的な質問応答システムの概要について説明する。一般的な質問応答システムの処理フローを図 2.1 に示す。

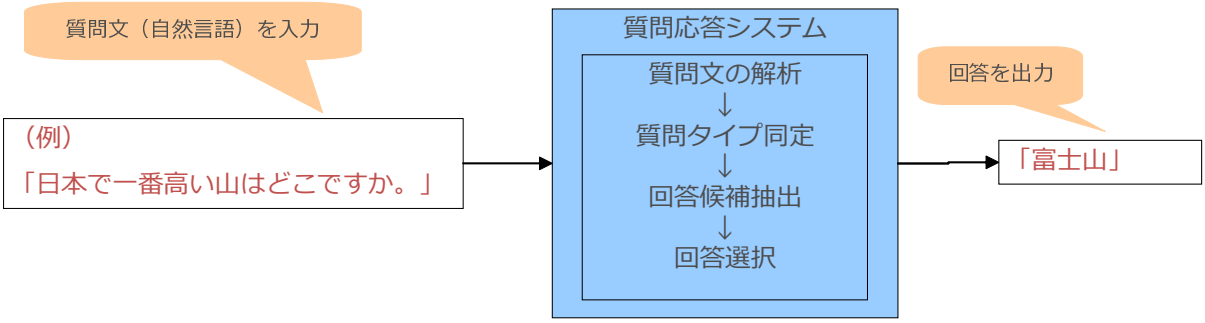

図 2.1: 一般的な質問応答システムの処理フロー

2.2.1　質問文の解析

質問応答システムは、入力として受け付けた質問文を解析する。一般に、形態素、構文、意味解析といった様々な深度の解析が行われる。例えば、形態素解析器を使って質問文を単語（形態素）単位に分解して品詞を付与する。

質問文を形態素に分解した後、文書集合から回答を含む文書を検索するための検索キーワードが抽出される。一般には質問文に含まれる自立語を検索キーワードとする。例えば、佐々木らが提案した質問応答システム SAIQA II[7] では、質問文から自立語を検索キーワードとして抽出し、そのキーワードを含むパラグラフをスコアリングし、そのスコアの上位 N 件のパラグラフを獲得する方式を採用している。

同時に、後述する質問タイプ同定の処理に機械学習の手法が適用される場合は、学習素性ベクトルも抽出される。ここでの学習素性ベクトルとは、質問文の特徴を表わすベクトルであり、質問タイプを同定するための手がかりとなるものである。学習素性ベクトルとしてどのような情報を抽出するかは重要な研究テーマであるが、従来は自立語、単語 N-gram、疑問詞等がよく用いられてきた。

2.2.2　質問タイプの同定

機械学習による質問タイプの同定においては、質問文をあらかじめ定義された質問タイプに分類するための訓練データをあらかじめ構築する必要がある。訓練データとは、学習

素性ベクトルとその正解ラベルのペアの集合である。この訓練データから各種機械学習アルゴリズムにより質問タイプを分類するモデルを自動的に学習する。そして、システムに質問文が入力されたとき、学習された分類モデルを用いて質問タイプを同定する。

2.2.3 回答候補の抽出

質問タイプを同定した後に、回答候補の抽出を行う。まず、質問文解析によって得られた検索キーワードをクエリとして、知識源として用意された文書集合から質問文との関連が深い文書を検索する。これには従来の情報検索技術が適用される。次に、検索された文書の中から、回答の候補となる単語を抽出する。従来のファクトイド型の QA システムでは、回答の多くは固有名詞である。そのため、テキストから固有表現を抽出し、回答候補とする場合が多い。さらに、質問タイプに適合しない固有表現は回答候補から除外する。例えば、質問タイプが「LOCATION」なら、場所を表す固有表現のみが回答候補となる。

2.2.4 回答の選択

抽出したそれぞれの回答候補に対しスコアリングを行い、最も高いスコアを獲得した回答を選択する。回答候補のスコアは、回答候補が出現する文書における回答候補と検索キーワードとの距離等を基に算出される。距離によるスコアリングは、質問文中のキーワードの近くに出現する回答候補は正しい回答である可能性が高いという考え方に基づく。

2.3 質問タイプの定義

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。機械学習に基づいた質問タイプ同定に関する先行研究において、佐々木らの手法 [7] で用いられている質問タイプは、IREX [5] 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 [6] 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と同定してしまうと、正しい回答を選択することはできない。

一方、遠藤らは詳細化した質問タイプを提案した[8]が、質問タイプの同定がルールベースで行われており、未知の質問文への対応が困難になる問題点がある。

本課題研究では、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[8]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同定するシステムを構築する。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同定が試みられたことはない。

第3章　実験方法

本章では、本課題研究の目的である、詳細な質問タイプの同定、新たに提案した質問タイプ同定手法における学習素性の有効性の評価、および 1.2 節で述べた 2 種類の訓練データによる正解率向上の検証を行うために実施した実験方法について述べる。

3.1　質問タイプの定義

3.1.1　関根の拡張固有表現階層

本課題研究では、質問タイプの定義として関根の拡張固有表現階層 [1] を利用する。関根の拡張固有表現階層を一部抜粋したものを図 3.1 に示す。

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 3.1: 関根の拡張固有表現階層 (一部抜粋)

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人工的な分類で、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 機械学習アルゴリズム

本課題研究では、質問タイプ同定モデ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 つを利用する。これら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。後述するように、本課題研究では、QAC 質問文のセットおよび毎日新聞のテキストに固有表現タグを付与したコーパスという 2 つのデータを訓練データとする。

3.2.1 Support Vector Machine

SVM[12] は、1995 年に AT&T の V.Vapnik によって統計的学習理論の枠組みで提案された、ノンパラメトリックな教師あり学習手法で、局所解に収束しないという特徴をもっている。SVM は「マージン最大化基準」と呼ばれる、特徴空間における識別境界の位置決定基準をもっており、訓練データの中で最も他の特徴クラスと近い位置にいるもの（これをサポートベクトルと呼ぶ）を基準として、そのサポートベクトルと識別境界とのユークリッド距離（マージン）が最大になるような位置に識別境界を設定する。

上記のような特徴をもつ SVM を用いた場合、従来の機械学習手法を用いた場合と比較して良好な識別結果が得られることが、多くの先行研究においても報告されている [6, 7]。

3.2.2 k-NN 法

k-NN 法（k-nearest neighbor method、k 近傍法）は最も単純な機械学習アルゴリズムの一つである。k-NN 法では、データを多次元空間上のベクトルとして表現する。訓練データは、正しいクラスラベルが付与された特徴ベクトルの集合となる。いま、未知のデータが与えられたとき、そのデータの特徴ベクトルと、訓練データにおける個々の特徴ベクトルとの距離を計算し、k 個の最近傍標本を選択する。そして、選択された k 個の標本の中から、最も多く存在しているクラスラベルを探索し、それを正解として選択する。

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。特徴ベクトル間の類似度評価尺度は幾つか考案されており、一般にはユークリッド距離が用いられるが、集合に対する類似度の計算では、Jaccard 係数、Dice 係数、Simpson 係数が用いられる。ある 2 つの集合を X, Y とすると、それぞれの係数は次式により定義される。

1. Jaccard 係数

$$\frac{|X \cap Y|}{|X \cup Y|} \tag{3.1}$$

2. Dice 係数

$$\frac{2 \times |X \cap Y|}{|X| + |Y|} \tag{3.2}$$

3. Simpson 係数

$$\frac{|X \cap Y|}{min(|X|, |Y|)} \tag{3.3}$$

それぞれの係数の定義式において、分子はほぼ同じであるが分母に違いがある。Jaccard 係数は 2 つの集合の和集合の要素数、Dice 係数は 2 つの集合の要素数の和、Simpson 係数は小さい方の集合の要素数が分母となっている。

本課題研究における k-NN 法を用いた質問タイプ同定実験では、Dice 係数を類似度の評価尺度として採用している。

3.3 学習素性

実験では、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、質問文および後述する毎日新聞の固有表現タグ付きコーパス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[13] を、文節の係り受け解析には CaboCha[14] を利用した。以下、それぞれの学習素性について説明する。

3.3.1 自立語

単語には大きく分けて自立語と付属語がある。自立語とはそれ自身が何らかの意味を持つ単語であり、一方付属語とは特に意味を持たずに文法的な役割を表す単語である。例えば、名詞、動詞、形容詞等は自立語であり、助詞、助動詞、句読点等は付属語である。自立語は従来の質問応答システムの研究においてもよく利用されてきた素性である。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

自立語は以下のように抽出する。まず、形態素解析を行い、文を単語に分割し、同時に個々の単語の品詞を決める。次に、品詞を基準として自立語を素性として抽出する。あらかじめ自立語に該当する品詞のリストを用意し、そのリストに該当する品詞の語を自立語と判定する。但し、抽出した自立語の表記ゆれによる学習精度の低下を防ぐために、以下の処理を行う。

- 動詞等の活用語は基本形に変換する（例：「走っ」→「走る」）
- 表 3.1 に示した単語はストップワード (質問タイプ同定の正解率の向上に寄与しない可能性が高い語) と見なし学習素性から除外する。これらの単語の品詞は「動詞」であり、品詞に従って分類すれば自立語になるが、実際には付属語であると考えられるためである。

表 3.1: ストップワード一覧

する, れる, いる, ある, なる, いう

3.3.2 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 何+です です+か

単語 bi-gram の学習素性によって、質問文を構成する単語列の並びから質問文の表現、意味を学習することが期待できる。

3.3.3 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。本課題研究では、表 3.2 に示す疑問詞を学習素性の抽出対象とした。

表 3.2: 学習素性として利用する疑問詞の一覧

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

3.3.4 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 3.2 に示すような文節の係り受け解析結果が得られる。図 3.2 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、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

図 3.2: 文節の係り受け解析の例

3.4 訓練データ

この節では、質問タイプを分類するモデルを教師あり学習によって構築するための訓練データについて述べる。本課題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) [15] 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. 毎日新聞の固有表現タグ付きコーパス（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。本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同定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 3.3 に示す。また、付与した質問タイプの頻度分布を表 3.3 に示す。ただし、表 3.3 では頻度 10 以上の質問タイプの頻度のみを示した。

```
QAC0-10001-01,0,QAC0-20001-01,QAC0-30001-01,d,, 夏目漱石の名作は何ですか。,0,,,,,, ○,,,Book,
QAC0-10001-02,0,QAC0-20001-02,QAC0-30001-02,d,, 彼の長男の職業は何ですか。,0,,,,,,,,,Position_Vocation,
QAC0-10002-01,0,QAC0-20002-01,QAC0-30003-01,d, 日本三大祭りはなんですか。, 日本三大祭りは何ですか。,0,,,,,, ○,,,Occasion_Other,
QAC0-10003-01,0,QAC0-20003-01,QAC0-30002-01,d,, エアロスミスがデビューしたのはいつですか。,0,,,,,,,,,Date,
QAC0-10003-02,0,QAC0-20003-02,QAC0-30002-02,d,, エアロスミスのデビュー作は何ですか。,0,,,,,,,,,Music,1
```

図 3.3: QAC 質問文コーパス (抜粋)

通常の質問応答システムでは、QAC 質問文コーパスのように質問形式の文を入力とする。しかしながら、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。そのため本課題研究では、QAC 質問文コーパスに加えて、毎日新聞の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補強を図る。

本課題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

表 3.3: QAC 質問文コーパスにおける質問タイプの頻度分布 (抜粋)

質問タイプ	出現頻度
Person	255
Date	118
Company	86
City	58
Product_Other	54
Money	37
Country	34
Physical_Extent	21
Province	20
N_Person	19
Contx_Other	17
Percent	14
GOE_Other	14
Food_Other	14
Movie	13
Age	13
Numex_Other	11
Position_Vocation	10
Music	10

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）と関係する単語を学習できると期待できる。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。なお、文中に固有表現が複数出現する場合は、文の仮想的な質問タイプを一意に決定することができないため、訓練データとはしなかった。新聞コーパスに含まれる文の数は合計 89,817 件であり、この中から一つの文に複数の固有表現タグが付与されているものを除いた 21,385 件を訓練データとした。以下、それぞれのコーパスに基づく訓練データについて説明する。

3.4.1 訓練データ作成の手続き

本項では、訓練データ作成の詳細な手続きについて説明する。まず、QAC 質問文コーパスに基づく訓練データは次のように作成する。本研究では、関根の拡張固有表現階層に基づく固有表現を質問タイプとみなすので、200 種類ある質問タイプにあらかじめ ID を

付与する。次に、すべての QAC 質問文を走査して学習素性から構成される特徴ベクトルを抽出する。特徴ベクトルは、自立語、単語 bi-gram、疑問詞そして係り受け関係の４種類の学習素性の列となる。特徴ベクトルの要素、すなわち学習素性には、全体で一意となる ID を付与する。つまり、一つの質問文を特徴ベクトルの ID 列に変換する。ひとつの QAC 質問文から、それに付与されている正解の質問タイプの ID および特徴ベクトルの ID 列を訓練データとしてファイルに出力する。

本システムの訓練データ出力処理によって生成される訓練データのファイルフォーマットを以下に示す。

表 3.4: 訓練データのファイルフォーマット

＜質問タイプの ID ＞ ＜特徴ベクトル ID(1) ＞:1 ＜特徴ベクトル ID(2) ＞:1 ... ＜改行＞

訓練データの実際の出力例は以下のようになる。

表 3.5: 訓練データの例

102 1899:1 3122:1 4056:1 6622:1 7483:1 7848:1 13921:1
119 1899:1 3661:1 3828:1 4056:1 6622:1 8582:1 10755:1 11717:1 20318:1
125 1899:1 4056:1 6106:1 6622:1
26 797:1 1193:1 1899:1 2125:1 2781:1 3613:1 3962:1
158 205:1 498:1 1033:1 1231:1 1542:1 1899:1 2781:1 3911:1

表 3.4 のフォーマットは、SVM の学習ツール LIBSVM のフォーマットに準拠している。また、k-NN 法は今回の実験のために独自に実装したが、このフォーマットを基に２つの文の類似度（Dice 係数）を計算するプログラムを作成した。

固有表現タグ付き毎日新聞コーパスに基づく訓練データは、ファイルフォーマットは表 3.4 で示したものと最終的に同じになる。ただし、新聞コーパスの元のテキストデータファイルのフォーマットは、QAC 質問文コーパスのファイルフォーマットと異なるため、別のプログラムを実装して学習素性を抽出した。なお、新聞コーパスの文は平叙文がほとんどであるので、疑問詞の学習素性は抽出しない。

第4章　実験結果

3章で説明した実験方法に基き、SVM・k-NN法それぞれの機械学習アルゴリズムによって構築した学習モデルで質問タイプの同定実験を行った。また、自立語・単語bi-gram・疑問詞・係り受け関係の４つの学習素性が、それぞれどの程度質問タイプ同定の性能の向上に寄与するのかを検証した。なお、実験は5分割交差検定により行った。すなわち、訓練データを5分割し、そのうちの１つをテストデータ、他の４つを訓練データとして質問タイプ同定の正解率を測る実験を行った。また、テストデータを変えながらこの試行を5回繰り返した。

4.1　SVMの実験結果

SVMによる質問タイプ同定の実験結果を表4.1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。また、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」は質問タイプ同定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。

表 4.1: SVMによる質問タイプ同定の実験結果

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%

それぞれの学習素性が質問タイプの同定にどの程度貢献するかを調べるために、全素性の集合、ならびに１つの素性のみを除いた素性集合を用いて質問タイプの判定をしたときの正解率を比較する。結果を図 4.1 に示す。図 4.1 の正解率は表 4.1 に示した正解率と同じだが、比較のためわかりやすくグラフで表示している。

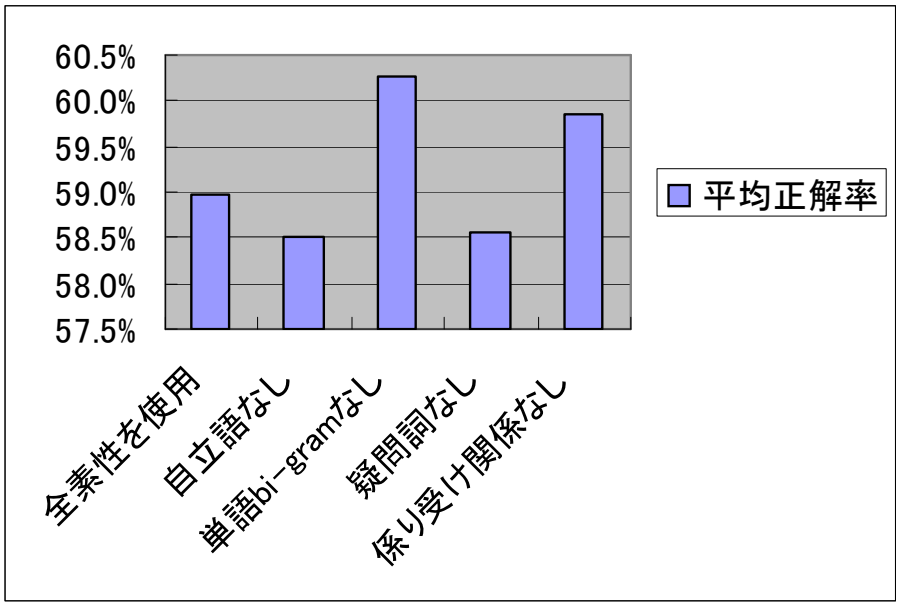

図 4.1: 質問タイプ同定の実験結果 (SVM, QAC)

同様に、新聞コーパスのみを訓練データとしたときの素性集合の平均正解率を比較したグラフを図 4.2 に示す。ここでは疑問詞の素性は常に使わないことに注意されたい。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たときの結果を図 4.3 に示す。

QAC 質問文コーパスと新聞コーパスの両方を使ったときの正解率 (図 4.3) は、QAC 質問文コーパスのみを使ったときの正解率 (図 4.1) よりも低くなった。これは、質問文ではない新聞コーパスから得られる素性が質問タイプの判定に悪影響を与えたためと考えられる。そこで、出現頻度による簡単な素性選択の手法を試した。出現頻度の低い素性はノイズになると考え、これを除外した。訓練データとして QAC 質問文コーパスと新聞コーパス両方を用い、出現頻度が 1 以下、2 以下、5 以下、10 以下の素性を除外したときの実験結果を表 4.2、表 4.3、表 4.4、表 4.5 にそれぞれ示す。

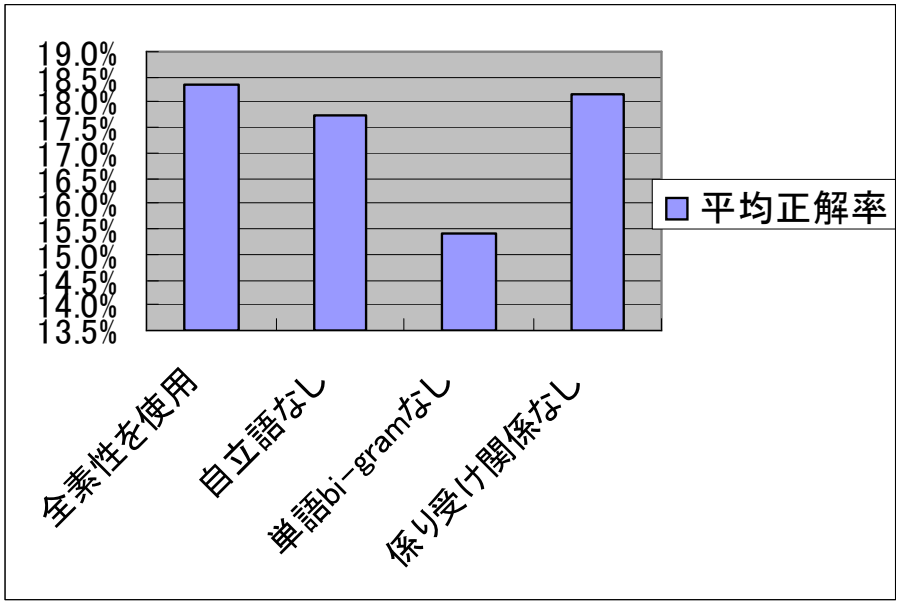

図 4.2: 質問タイプ同定の実験結果 (SVM, 新聞)

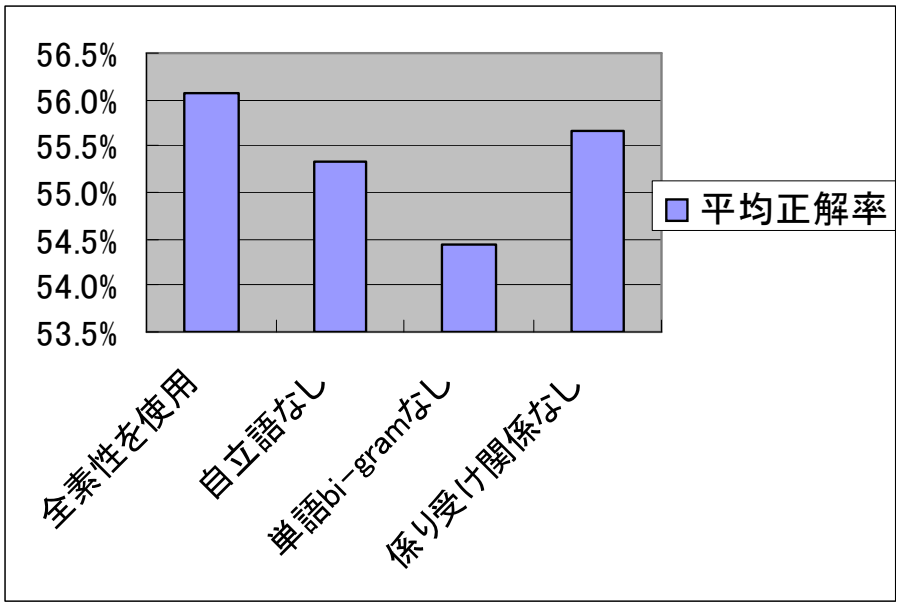

図 4.3: 質問タイプ同定の実験結果 (SVM, QAC+新聞)

表 4.2: 質問タイプ同定の正解率 (SVM, QAC+新聞, 頻度 1 以下の素性なし)

交差検定	正解率 [%]
1	56.6
2	55.7
3	52.8
4	61.7
5	52.4
平均	55.8

表 4.3: 質問タイプ同定の正解率 (SVM, QAC+新聞, 頻度 2 以下の素性なし)

交差検定	正解率 [%]
1	55.3
2	58.3
3	56.2
4	54.0
5	54.1
平均	55.6

表 4.4: 質問タイプ同定の正解率 (SVM, QAC+新聞, 頻度 5 以下の素性なし)

交差検定	正解率 [%]
1	52.8
2	52.3
3	56.2
4	55.7
5	60.5
平均	55.5

表 4.5: 質問タイプ同定の正解率 (SVM, QAC+新聞, 頻度 10 以下の素性なし)

交差検定	正解率 [%]
1	55.3
2	59.1
3	59.1
4	57.0
5	48.9
平均	55.9

4.2 k-NN 法の実験結果

k-NN 法による実験結果を表 4.6 に示す。この表における「訓練コーパス」の列は k-NN 法における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 k-NN 法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。また、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。「k の値」は k-NN 法における k、すなわち質問タ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。最後に「平均正解率」は質問タイプ同定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。

表 4.6: k-NN 法による質問タイプ同定の実験結果

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	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	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	1	46.6%
							3	49.5%
							5	50.5%

それぞれの学習素性が質問タイプの同定にどの程度貢献するかを調べるために、SVM のときと同様に、全素性の集合、ならびに１つの素性のみを除いた素性集合を用いて質問タイプの判定をした結果を比較する。図 4.4 は、学習アルゴリズムとして k-NN 法を用い、$k = 5$ としたとき、それぞれの素性集合の平均正解率を比較したものである。

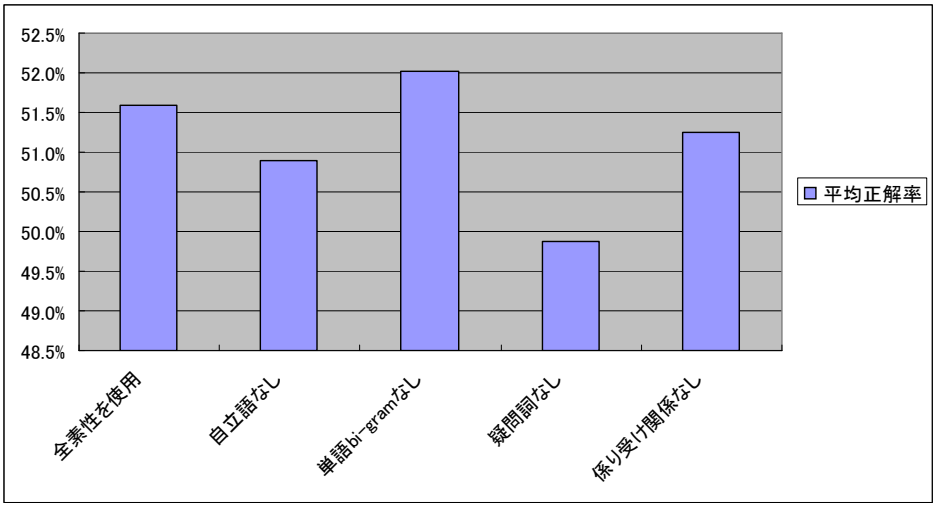

図 4.4: 質問タイプ同定の実験結果 (k-NN, QAC)

同様に、新聞コーパスのみを訓練コーパスとし、$k = 5$ の k-NN 法を用いたときの素性集合の平均正解率を比較したグラフを図 4.5 に示す。ここでは疑問詞の素性は常に使わないことに注意されたい。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たときの結果を図 4.6 に示す。

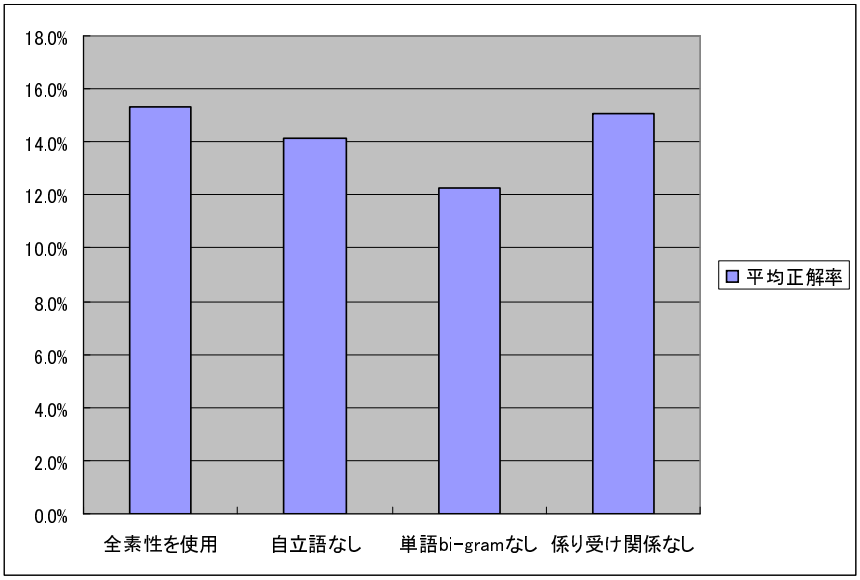

図 4.5: 質問タイプ同定の実験結果 (k-NN, 新聞)

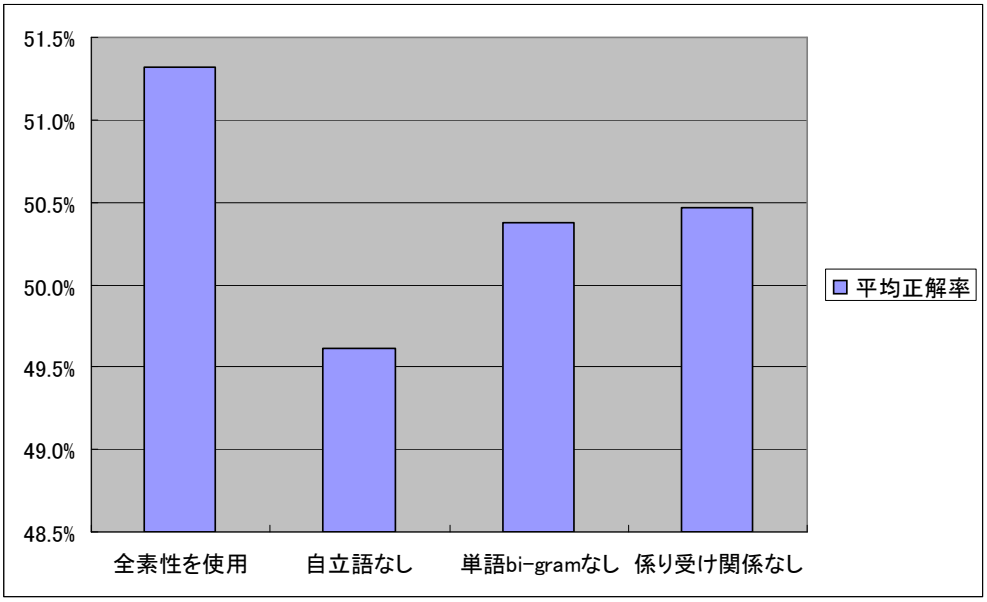

図 4.6: 質問タイプ同定の実験結果 (k-NN, QAC+新聞)

4.3 考察

本節では、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

前節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 4.1 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同定システムを実装しており、その正解率は 88.0%と報告している [7]。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と考えられるため、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る提案手法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 4.1 より)、k-NN 法の正解率は 52.0%(表 4.6 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 4.1 より)、k-NN 法の正解率は 51.3%(表 4.6 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同定するための手法として適していることがわかる。

また、表 4.6 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 節で述べたように、本課題研究では QAC 質問文コーパスと新聞コーパスの 2 種類の訓練データを使用する。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、QAC 質問文コーパスを訓練データとしたときの正解率は 60.3%(表 4.1 より) 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%(表 4.1 より) 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%(表 4.1 より) となり、2 種類の訓練データを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 4.6 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同定のモデルの学習に使うことの有効性が確認できる。ただし、質問タイプの同定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同定の正解率に貢献しない理由を考察するために、素性の数を調べた。QAC 質問文コーパスから得られる 4 種類の素性の総数は 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、質問タイプの同定に無関係な素性が大量に抽出され、それがノイズとなって悪影響を与えていると考えられる。そこで、出現頻度の低い素性はノイズになる可能性が高いと考え、低頻度の素性を除いて SVM を学習する実験を行った。訓練データは QAC 質問文コーパスと新聞コーパスの両方を用いた。表 4.2〜表 4.5 より、出現頻度が 1、2、5、10 以下の素性を除いたときの正解率はそれぞれ 55.8%、55.6%、55.5%、55.9%であった。これらはいずれも全素性を用いたときの正解率 56.1%(表 4.1) よりも低い。出現頻度の低い素性を削除することは頻度に基づく簡単な素性選択を行っているとみなせるが、今回の実験では素性選択の有効性は確認できなかった。

また、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ると、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。今回の実験では 5 分割交差検定によって QAC 質問文コーパスの文をテスト文としている。つまり、テスト文の素性と新聞コーパス中の素性にほとんど重なりがなく、このことも新聞コーパスの使用が正解率の向上に寄与しない理由のひとつと考えられる。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、かつ質問タイプの同定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。実験では、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た。もし、後者の正解率が前者の正解率と比べて大きく低下するなら、除いた素性は質問タイプ同定の正解率の向上に大きく貢献するといえる。

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。図 4.1 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引き起こしやすい。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。図 4.2 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、図 4.5 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、単語 bi-gram が有効に働いたと考えられる。また、質問タイプ同定のタスクにおいては、主に平叙文から構成される新聞コーパスを訓練データとするとき、自立語や係り受け関係よりも単語の出現順序がその文の意味を抽出する情報として重要であることを示唆している。

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。図 4.3 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、図 4.6 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 4.1 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、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、質問タイプの同定に「誰」「どこ」などの疑問詞は有効に働くと考えられる。

第5章　まとめ

5.1　結論

本課題研究では、質問応答システムにおける質問タイプ同定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、質問タイプ同定の正解率を測定するとともに、それぞれの学習素性の有効性について検証した。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、質問タイプ同定の正解率は 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、質問タイプ同定の正解率を向上させることはできず、今後の研究課題となった。

5.2　今後の課題

本課題研究では、質問タイプとして新たに関根の拡張固有表現階層を導入したが、これにより実際の質問応答システムの回答精度の向上にどの程度寄与するかは、今回実装した質問タイプ同定モジュールを質問応答システムに組込んで検証を行うことが必要である。

新聞コーパスを質問タイプ同定の正解率向上に役立てるために必要な学習素性・機械学習手法についての知見は得られておらず、今後の研究課題である。但し、質問文のコーパスに正解ラベル（質問タイプ）を付与する作業コストは非常に大きく、本課題研究で用いた新聞コーパスのように、平叙文から有効な学習モデルを構築する手法を確立することができれば質問応答システムの性能向上に大きく貢献するはずであり、依然として重要な研究課題であると考えられる。

参考文献

[1] Satoshi Sekine et al., “Definition, dictionaries and tagger for Extended Named Entity Hierarchy”, Proceedings of LREC, pp.1977-1980, 2004.

[2] NTCIR(NII Testbeds and Community for Information access Research) Project homepage http://research.nii.ac.jp/ntcir/index-ja.html

[3] Ralph Grishman, Beth Sundheim, “Message Understanding Conference - 6: A Brief History” In Proceedings of the 16th International Conference on Computational Linguistics (COLING), pp.466-471, 1996.

[4] TREC(Text REtrieval Conference) homepage http://trec.nist.gov/

[5] IREX(Information Retrieval and Extraction Exercise) NE homepage http://nlp.cs.nyu.edu/irex/NE/

[6] 鈴木 潤ほか, “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.

[7] 佐々木 裕ほか, “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.

[8] 遠藤 哲哉ほか, “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, 2004-NL-159, pp.25-30, 2004.

[9] Joseph Weizenbaum, “ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine”, Communications of the ACM, Vol.9, No.1, pp.36-45, 1966.

[10] W. A. Woods, “Progress in natural language understanding-An application to lunar geology ”, In Proceedings of AFIPS National Conference, pp.441-450, 1973.

[11] Rob High, “The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works”, IBM, Redbooks, REDP-4955-00, pp.3-10, 2012.

[12] Vladimir N. Vapnik, “The Nature of Statistical Learning Theory”, Springer, 1995.

[13] MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

[14] 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.

[15] J. Fukumoto and T. Kato, “An overview of Question and Answering Challenge (QAC) of the next NTCIR workshop”, In Proceedings of the Second NTCIR Workshop Meeting, pages 375-377, 2001.