SVMを用いた学習型質問応答システムSAIQA-II

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[†]

近年，大量の文書を用いて自然文によるユーザからの質問に答える質問応答（QA: Question Answering）システムに関する研究が注目を集めている．これまでいくつかのQAシステムが開発されてきたが，それらの多くは人手で作成されたルールや評価関数を用いて，質問の答えを大量の文書から抽出するアプローチをとっていた．これに対し，本論文では，機械学習技術を用いて，日本語QAシステムの主要なコンポーネントをそれぞれ学習データから構築することにより，QAシステム全体を構築する方法について述べる．具体的には，質問タイプや答えの判定を2クラス分類問題としてとらえ，質問文やその正解例から学習された分類器により，これらの機能を実現する．本アプローチのフィージビリティの確認のため，機械学習手法Support Vector Machine（SVM）を用いて学習型QAシステムSAIQA-IIを実装し，2,000問の質問・正解データによるシステム全体の5分割交差検定を行った．その結果，システムの性能として，MRR値で約0.4，5位以内正解率で約55%の正解率が得られることが明らかになった．

SAIQA-II: A Trainable Japanese QA System with SVM

Yutaka Sasaki,[†] Hideki Isozaki,[†] Jun Suzuki,[†]
Koji Kokuryou,[†,☆] Tsutomu Hirao,[†] Hideto Kazawa[†]
and Eisaku Maeda[†]

This paper describes a Japanese *Question-Answering* (*QA*) System, SAIQA-II. These years, researchers have been attracted to the study of developing Open-Domain QA systems that find answers to a natural language question given by a user. Most of conventional QA systems take an approach to manually constructing rules and evaluation functions to find answers to a question. This paper regards the specifications of main components of a QA system, question analysis and answer extraction, as 2-class classification problems. The question analysis determines the question type of a given question and the answer extraction selects answer candidates that match the question types. To confirm the feasibility of our approach, SAIQA-II was implemented using Support Vector Machines (SVMs). We conducted experiments on a QA test collection with 2,000 question-answer pairs based on 5-fold cross validation. Experimental results showed that the trained system achieved about 0.4 in MRR and about 55% in TOP5 accuracy.

1. はじめに

近年，自然文で書かれた質問文に対する解答を大量の文書から抽出する質問応答（QA: Question Answering）システムの研究が多くの研究者の注目を集めている．たとえば，本論文が対象としているQAシステムとは，「英国のビクトリア女王が即位したのは何年ですか？」というユーザからの質問文に対して，システムが知識源として持っている大量の新聞記事を参照することにより，「1837年」と答えるシステムである．

このようなQA研究への関心の高まりは，TREC（Text REtrieval Conference）において，1999年にQA Trackが新たに設定されたことが1つの契機となっている．1999年に開催された第1回のQA Trackの概要は，以下のようにまとめられる[22]．

- 約528,000記事を対象．
- 正解が短い事実情報となる質問文200問により評

† 日本電信電話株式会社NTTコミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation
☆ 現在，NTTコムウェア西日本株式会社
Presently with NTT COMWARE WEST Corporation

価する．

- 新聞記事から抜き出された，連続した文字列およびその文字列が現れる記事 ID のペアを，1 位〜5 位までランキングして出力する．
- 50 バイト以内の文字列を出力する課題と 250 バイト以内の文字列を出力する課題の 2 つの課題により評価する．

このような TREC の QA Track のタスク設定に沿った仕様による日本語 QA の研究としては，質問文や文書の構文情報を利用しながら短いパッセージを抽出する研究[11]や解答を含む 50 バイトの文字列を答える日本語 QA システムの比較・評価に関する研究[15]等がある．

さらに，質問の答えとして，50 バイトといった文字列ではなく，解答そのもの（exact answer）を答える研究も行われている[16]．2002 年に開催された第 4 回の TREC QA Track においても，50 バイトといった文字列ではなく，解答そのものを答えるタスク設定に変更された．

日本においても，国立情報学研究所主催の第 3 回 NTCIR ワークショップ[8]の新規課題として，Question Answering Challenge（QAC）[14]が採用されたことにより，最近，多くの研究者の注目を集めている．NTCIR QAC1 では，メイン（解答を 5 位までランキング），リスト（解答をすべて列挙），枝問（関連質問に答える）の 3 種類のタスクにおいて，解答そのものを答える性能の評価が行われた．

このような状況の中，日本語 QA システムを実現するため，以下の点を明らかにすることが重要な課題として浮上している．

- どのようなシステム構成が有効か．
- どのような技術が利用できるのか．

その 1 つの方向性として，機械学習技術を用いて，QA システムを構成するアプローチがある[6),7),12),13),17),23)]．質問文の解析への機械学習技術の応用には，決定木学習用いた研究[23]や SVM [21]を用いた研究[18]がある．解答抽出・選択に関しては，パーセプトロンを用いた研究[13]や決定木を用いた研究[12]，SVM を用いた研究[17]が行われている．

これらの研究は，QA システムの一部のコンポーネントにおいて機械学習技術を採用した研究であり，システム全体を機械学習技術を用いて構成および評価した研究はない．Ittycheriah ら[6),7)]は，最大エントロピー法を質問解析，固有表現抽出，解答選択に適用しているが，50 バイトのパッセージを答える英語質問応答システムを対象としており，本論文が対象とする解答そのものを答えるシステムの評価は行われていない．

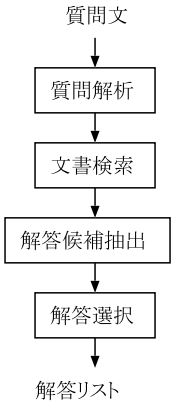

図 1 従来の QA システムの構成
Fig. 1 Block diagram of conventional QA systems.

そこで，1 つの疑問として，

- 従来，人手により作成されたルールや評価関数を用いて構成されていた QA システムの主要コンポーネントに機械学習技術を適用することにより，解答そのものを答える QA システム全体を構成することは可能か？ また，可能であればどの程度の性能が得られるか？

ということがあげられる．

そこで，本論文では，

(1) 学習技術を用いたシステム全体の構成法を明らかにする，
(2) 個々のコンポーネントの評価ではなく，システム全体を通して，テストセットを用いた評価を行う，

という 2 点に絞って述べる．

本論文では，以下のような構成でこれらの点について述べる．2 章では，従来の典型的な QA システムの概要を述べる．3 章では，学習型 QA システムについて述べるための準備を行う．続いて，4 章において，学習型 QA システム SAIQA-II を実装する方法を述べる．5 章では，本構成のフィージビリティ確認と評価を行う．学習型 QA システムの利点について 6 章で考察し，最後に 7 章において，結論を述べ本論文を締めくくる．

2. 従来の QA システムの概要と問題点

この章では，従来の QA システムの構成について述べる．ほとんどの QA システムは，**図 1** のようなモジュールにより質問応答を実現している．

質問解析モジュール 質問文を解析し，質問タイプを同定する．

表 1 固有表現
Table 1 Named entities.

固有表現	説明	例
PERSON	人名	小泉，ブッシュ
LOCATION	地名	北アメリカ，米国，名古屋
ORGANIZATION	組織名	アメリカ政府，NTT，関西国際空港会社
ARTIFACT	製品名，作品のタイトル	カローラ，「徹子の部屋」
DATE	日付	1 月，クリスマス，5 月 4〜7 日
TIME	時間	午後 3 時，7:30AM
MONEY	金額	$5，1,000 円，20 ペソ
PERCENT	割合	5%，半分，3 分の 1

文書検索モジュール 解答抽出の高速化のために対象文書を制限する．質問文から取り出された検索語により文書やパラグラフをスコアリングし，スコアの上位 N 件を取り出す．

解答候補抽出モジュール 文書から質問タイプに合った解答候補を取り出す．

解答選択モジュール 質問タイプ，検索語等の情報を利用して，解答候補を順位付けする．

質問タイプは，解答として何を選択するかを特定するための重要な手がかりである．質問タイプの分類は，システムによって様々であるが，人名，地名，組織名，日付，時間等が共通に用いられている．これらの固有名詞や数値表現に関する具体的な定義は IREX [4] の固有表現抽出タスクで詳細に規定されている．

これらのモジュールの中で，質問タイプの同定と解答候補の選択が QA システムにおいて重要なポイントである．この 2 つが重要な理由を以下に述べる．

質問タイプの同定 質問文の質問タイプの同定を間違えると，その質問に正解することはほとんど不可能になる．たとえば，「プリウスを発表した会社はどこですか」という質問が会社名を聞いているにもかかわらず，質問タイプを間違って LOCATION と同定すると，「東京」等の地名を答える可能性が非常に高くなる．

解答候補の選択 TREC QA Track の仕様や QAC のメインのタスクでは，1 つの質問に対して，5 件の解答を 1 位から 5 位までランキングして出力する．質問に応じて選び出された数十〜数百件の解答候補から，できるだけ正解を上位に含むように 5 件を選び出せるかどうかが性能を大きく左右する．

従来これらの機能は，人手により経験的に作成されたルールやパターン，評価関数により実現されていた．しかし，未知の質問文に対しての正解率も高くなるように予測しながら，様々なパラメータやパターンを調整する過程は，多大な時間と労力を必要とし，かつ非常にセンシティブな作業である．たとえば，ある質問の正解が得られるようにわずかな変更を行うことで，他の質問に対する質問タイプや解答候補の判定が変わり，全体の正解率が大きく低下するという現象が生じる．

3. 準 備

この章では，学習型 QA システムの構成について述べるための準備として，固有表現のためタグ集合および QA システムの各コンポーネントの学習に用いるためのデータセット，学習アルゴリズム SVM の概要について述べる．

3.1 タグ集合

システムを通じて学習の基盤となるのがタグ集合である．本構成法においては，タグ集合は IREX の固有表現タグを採用しており，

- 質問タイプ
- 解答候補のタイプ

の両方に共通して使用する．本論文で対象とする固有表現を**表 1** に示す．一般的にはタグ集合は固有表現のタグに限る必要はなく，明確な基準の下に自由に定義されたタグを用いることもできる．たとえば，BIRD（鳥），FISH（魚）等，クラスを表すタグを用いることもできる．ただし，本論文では，タグの上位下位関係は対象としていない．

定義 1 （タグ集合 T）タグ集合 T は，英数字からなる文字列の有限集合であり，文書中の単語，語（term）や名詞句の種類を表現する．T の要素をタグ名と呼ぶ．

定義 2 （タグ付け）ある文書中の文字列の前後に，タグ集合 T の要素 t_i に対応する開始タグ $\langle t_i \rangle$ と終了タグ $\langle /t_i \rangle$ を挿入することをタグ付けと呼ぶ．ただし，開始タグと終了タグの間にはタグを含んではならない．

たとえば，「日本の小泉首相が...」にタグ付けした結果は，「⟨LOCATION⟩ 日本 ⟨/LOCATION⟩ の

表 2 2,000 問の分類
Table 2 Question types of 2,000 questions.

質問タイプ	質問数
PERSON	323
ORGANIZATION	319
LOCATION	308
ARTIFACT	141
DATE	377
TIME	25
MONEY	50
PERCENT	38
上記以外	419
合計	2,000

⟨PERSON⟩ 小泉 ⟨/PERSON⟩ 首相が...」となる．

3.2 学習データ

タグ集合とともに，学習型 QA システムの実現に必要な要素として学習データがある．本研究で用いた学習データは以下のとおりである．

タグ集合 固有表現タグ集合 T={PERSON，ORGANIZATION，LOCATION，ARTIFACT，DATE，TIME，MONEY，PERCENT}．

タグ付き文書集合 毎日新聞 98，99 年の新聞記事に T の固有表現タグを付与した文書集合．文書への固有表現タグの付与には，Isozaki ら[5]の固有表現抽出ツールを利用した．

質問文集合 人手で作成した 2,000 問の質問文[16]を利用．この質問セットは，無作為に記事を選び，その記事の中に現れる固有表現が正解となるような質問文を作成されたものである．各質問文に対してその質問タイプが付与されている．質問文 2,000 問の内訳を**表 2** に示す．これらの 2,000 問のうち，固有表現タグ集合に含まれる質問タイプの質問 1,581 問だけを利用する．

正解集合 各質問文の正解とその正解が現れる文書の ID の集合を作成．

なお，タグ付き文書集合は，学習データの作成のために用いられるとともに，評価実験において，評価用の質問の答えを抽出するための知識源としても利用される．

3.3 Support Vector Machine による学習

この節では，学習型 QA システムの説明のための準備として，機械学習アルゴリズム Support Vector Machine（SVM）に与えるために，文字データをどのように数値化するか，また，SVM がどのように学習を行うかについて簡単に紹介する．

3.3.1 素性ベクトル

従来の QA システムで行われてきた，質問文の分類や解答の選択を機械学習により実現するためには，質問文集合に含まれる質問文とその質問タイプや，文書中での正解の現れ方に関する特徴を数値化した数値ベクトルに変換する必要がある．原データの持つ特徴は素性（そせい）（feature）と呼ばれ，素性の値をベクトル化した数値ベクトル $\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$ は，素性ベクトルと呼ばれる．すなわち，$\boldsymbol{x}$ の i 番目の値 x_i は，i 番目の素性の値を表している．

たとえば，ある生徒が特徴として「身長 125cm，体重 35kg，兄弟なし，好きな色は黄色」という特徴を持っていたとする．この場合，素性ベクトルの 1 番目の素性を身長，2 番目を体重，3 番目を兄弟の有無 {0,1}，4〜6 番目の素性を，好きな色がそれぞれ赤，青，黄色であるか否か {0,1} で表現すると，この生徒の素性ベクトルは，$\boldsymbol{x}$=(125,35,0,0,0,1) となる．

3.3.2 正例，負例

学習データは，素性ベクトル $\boldsymbol{x}_i$とそのラベル y_i のペアの集合である．学習アルゴリズムが n クラスへの分類器（classifier）を構成できるとすると，ラベルは n 種類となる．SVM は 2 クラス分類器を構成するアルゴリズムであるため，素性ベクトルには，正（positive）または負（negative）のラベルを与える．正のラベルを与えられた素性ベクトルは正例（positive example），負のラベルを与えられた素性ベクトルは負例（negative example）と呼ばれる．

3.3.3 SVM の概要

これまで SVM を自然言語処理に用いた研究が数多く行われており，また，いくつかの解説[9),19)]が存在するため，本項では，SVM についての最小限の説明を行う．

素性ベクトルの次元が n であるとすると，1 つの素性ベクトルは n 次元空間中の点として表すことができる．正例と負例をすべてこの n 次元空間に配置したとする．この空間は入力空間（input space）と呼ばれる．正例と負例を分ける 2 クラス分類問題は，正例と負例を分離する超平面（分離平面）を決める問題に帰着できる．SVM は，ノイズを許容しつつ，分離平面に最も近い正例と負例との間のマージンを最大化するような分離平面を求める．また，カーネル関数（kernel function）$K(\boldsymbol{x}_1,\boldsymbol{x}_2)$ により，入力されたデータを高次元の素性空間（feature space）に写像し，素性空間において分離平面を求めることにより，入力空間においては非線形となる分離も可能である．本論文では，2 次の多項式カーネル $K(\boldsymbol{x}_1,\boldsymbol{x}_2)=(\boldsymbol{x}_1\cdot\boldsymbol{x}_2+1)^2$ を用いた．

以上のようにして得られた分離平面を用いて，以下のように分類器が構成される．新たに与えられた素性

図 2 SAIQA-II の GUI
Fig. 2 GUI of SAIQA-II.

ベクトルに対して，分離平面の正例側をプラス，負例側をマイナスとし，分離平面からの距離を正規化した値を計算することにより，分類器は与えられたデータが正，負の 2 クラスのどちらに属するかを判定する．さらに，平面からの距離は，クラスに属する度合いを表す評価値としても用いられることがある．つまり，平面からの距離の値が大きいほど，そのクラスに強く属していると考えるのである．

4. 学習型 QA システム SAIQA-II の実装

この章では，QA システム SAIQA-II の実装について述べる．SAIQA-II は**図 2** ような GUI を持つ．左上が質問文を入力フレーム，左下が解答表示フレームである．右側のフレームには，質問解析結果，記事検索結果，要約が提示される．要約は解答の根拠を表しているが，本論文の対象外である☆．

学習型 QA システムは，従来の QA システムと同様の構成を持つが，質問解析，解答選択の 2 つの重要モジュールが，例から学習された分類器により実現される点が特徴である．

学習型質問解析モジュール 質問文を素性ベクトルに変換し，あらかじめ学習してある質問分類器により質問タイプへ分類する．

文書検索モジュール 質問文中の自立語を検索語として取り出し，検索語を含むパラグラフをスコアリングし，スコアが上位 N 件であるパラグラフを取り出す．検索の評価関数としては，一般的な TF·IDF [20] を用いた☆☆．

解答候補抽出モジュール タグ付き文書から質問タイプに合った固有表現を解答候補として取り出す．

学習型解答選択モジュール 各解答候補に対して，質問タイプ，検索語等の情報を基に，素性ベクトルを作成し，あらかじめ学習してある解答分類器により，解答として出力するかどうかを決定する．

以下，学習型質問解析モジュールと学習型解答選択

☆ 平尾ら [1] の「質問に適応した文書要約手法」を用いた．

☆☆ ここでは，TF·IDF は，検索語の出現回数（TF: Term Frequency）と検索語を含むパラグラフの数の逆数（IDF: Inverse Document Frequency）の積である．

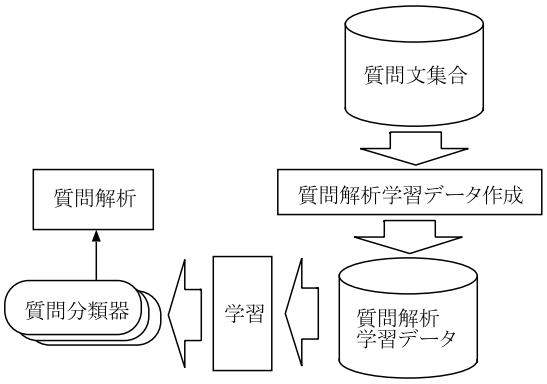

図 3 学習型質問解析の構成
Fig. 3 Block diagram of trainable question analysis.

モジュールの構成法について述べる．

4.1 学習型質問解析モジュール

質問解析は，典型的な分類問題と考えることができる．つまり，ユーザが入力した質問文を，PERSON, LOCATION 等の質問タイプに分類する問題ととらえることができる．たとえば，質問文「日本の首都はどこ？」は，地名を尋ねている質問であるから LOCATION に分類され,「プリウスを発売した会社はどこ？」は組織名を尋ねているので，ORGANIZATION に分類される．

質問文集合を学習データに変換し，これらのデータから質問分類器を作成し，質問分類器により質問解析を実現するまでの流れを**図 3** に示す．

4.1.1 学習フェーズ

質問解析は，1 つの質問文を複数のクラス（質問タイプ）に分類する問題であるが，本論文で用いる学習アルゴリズム SVM は 2 クラス分類器であるため，8 種類の質問タイプについて，それぞれその質問タイプに該当する/しないを分類する分類器を作成する．

たとえば，質問タイプが PERSON である質問文から作成された素性ベクトルを正例（正解例）とし，質問タイプが PERSON 以外である質問文から作成された素性ベクトルを負例（誤り例）として，正例と負例を分ける分類器を構成することにより，新たに与えられた質問文の質問タイプが PERSON であるか，PERSON でないかを決めることができる．

以下，このような質問分類器の学習を記号を用いながら正確に述べる．質問文データ D_Q は以下のような 3 つ組の集合とする．

$$\langle \text{質問 } ID, \text{質問文}, \text{質問タイプ} \rangle$$

ここで，質問タイプは固有表現タグ集合 T の要素であり，システム全体で共通に用いられるものである．

このデータ D_Q に基づいて，質問分類器の学習は以下のような手順で行われる．

(1) 質問文を素性ベクトルに変換する関数を F_Q とする．
(2) 固有表現タグ集合 T の各要素 t_i について，以下を行う．
  (a) 正例集合 $E^+_{t_i} = \{\}$.
  (b) 負例集合 $E^-_{t_i} = \{\}$.
  (c) 各質問タイプについて，以下のように正例と負例を作成する．各データ $\langle n_j, q_j, t_j \rangle \in D_Q$ について，
    - 質問文 q_j を素性ベクトル $\boldsymbol{v}_j = F_Q(q_j)$ に変換する．
    - $t_i = t_j$ ならば，$\boldsymbol{v}_j$を正例集合 $E^+_{t_i}$ に追加．
    - $t_i \neq t_j$ ならば，$\boldsymbol{v}_j$を負例集合 $E^-_{t_i}$ に追加．
  (d) 学習アルゴリズム SVM により，各 t_i に関する質問分類器 $g^{t_i}_Q = SVM(E^+_{t_i}, E^-_{t_i})$ を作成．

質問文を素性ベクトルに変換する関数 $F_Q(q_j)$ は，以下のような素性ベクトルを構成する．これらの素性は鈴木らの研究[18]に基づいて決定した．

質問文に含まれるすべての単語の意味カテゴリ すべての意味カテゴリについて，その意味カテゴリに含まれる単語が質問文 q_j に 1 回以上現れたかどうかを {0,1} で列挙する．

質問対象語の意味カテゴリ すべての意味カテゴリについて，質問対象語がその意味カテゴリに含まれる単語かどうかを {0,1} で列挙する．

質問文に含まれる疑問詞 学習データ D_Q の質問文に現れるすべての自立語について，質問文 q_j に現れたかどうかを {0,1} で列挙する．

単語の連鎖 学習データ D_Q に現れる質問文の単語，品詞，意味カテゴリの bigram（2 語連鎖）と trigram（3 語連鎖）について，質問文 q_j に現れたかどうかを {0,1} で列挙する．

この結果，次元が高くかつ 0 の多いスパースな素性ベクトルが作成されるが，SVM は数万次元の素性ベクトルを扱うことが可能である．なお，質問対象語（question focus[10]）とは，質問文中で解答が何であるかを限定している語である．たとえば,「〜食べ物は何ですか？」の「食べ物」が質問対象語である．

意味カテゴリとしては，日本語語彙大系の意味体系[3]を利用した．意味体系は日英機械翻訳システム ALT-J/E のために開発された概念の階層である．概念階層は木構造で表されており，各ノードは意味カテゴリを

表している．約 3000 の意味カテゴリを持ち，約 30 万語の単語を含む辞書において各単語に対して意味カテゴリが付与されている．

4.1.2 実行フェーズ

学習フェーズで構築した，各質問タイプごとの質問分類器 $g_Q^{t_i}$ は，質問文 q が与えられたとき，q から作成された素性ベクトルがその質問タイプに属するかどうかを分離平面からの距離により判定する．

以下，このような質問分類の実行を記号を用いながら正確に述べる．質問文 q が与えられたとき，q の質問タイプは以下のような手順で決定される．

(1) 質問タイプ集合 $QT = \{\}$.
(2) 質問文を素性ベクトルに変換する関数を構築で用いた F_Q とする．
(3) 質問文 q を素性ベクトル $\boldsymbol{v}$=$F_Q(q)$ に変換する．
(4) 固有表現タグ集合 T の各要素 t_i について，以下を行う．
  (a) 質問分類器 $g_Q^{t_i}(\boldsymbol{v})$ により，$\boldsymbol{v}$を正例，負例に分類する．
  (b) $\boldsymbol{v}$が正例であれば，質問タイプ集合 QT に t_i を追加．

複数の質問タイプが質問タイプ集合に追加された場合は，以下の解答抽出/選択の処理を簡単にするために，分類器が出力する値の一番大きな質問タイプを採用する．

4.2 学習型解答選択モジュール

解答分類器の学習の前処理として，解答選択の高速化のためにタグ付き文書集合に対して以下のような準備を行う．

(1) 文書を形態素解析し，単語，品詞，固有表現，意味を含んだ形態素情報の形式で格納する．
(2) 各単語には，パラグラフ番号，文番号，文節番号，単語番号を記事中で連番になるように付与する．

なお，本論文では簡単化のために形態素を単語と見なす．

定義 3 （形態素解析結果）1 つの形態素の形態素情報 m_i は次のようなリストである．

固有表現の場合

$$\langle n_p, n_s, n_b, n_w, \text{出現形，固有表現タグ} \rangle$$

固有表現ではない場合

$$\langle n_p, n_s, n_b, n_w, \text{出現形，よみ，原形，品詞，意味カテゴリリスト} \rangle$$

ただし，n_p，n_s，n_b，n_w は，それぞれその単語が現れたパラグラフ番号，文番号，文節番号，単語番号を表す．形態素解析結果は，形態素情報のリスト $\langle m_1, \ldots, m_n \rangle$ である．

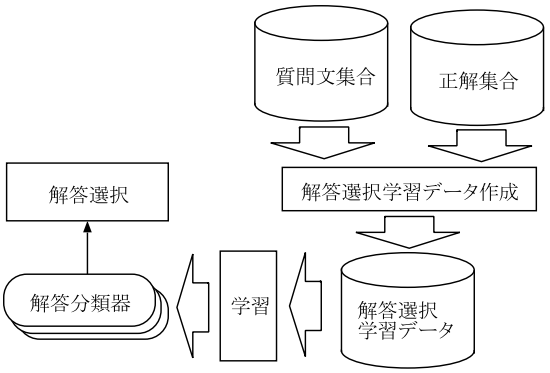

図 4 解答選択の学習の構成
Fig. 4 Block diagram of trainable answer selection.

例 1 「⟨DATE⟩10 月 1 日 ⟨/DATE⟩ に発売する」を形態素解析した結果を上記定義に従って表記したものを示す．

< < 0,0,0,0,10 月 1 日，DATE >,
< 0,0,0,1, に，に，に，格助詞，[] >,
< 0,0,1,2, 発売する，はつばいする，発売する，他動詞，[売り] > >

前処理により，上記の解析結果に変換しておくことで，2 つの単語が同じ文節に含まれるかどうかの判定や，ある単語の前後 n 文節内にある品詞が含まれるかどうかの判定が文節番号等を使うことで効率的に行える．

各解答候補に対して，質問タイプ，検索語等の情報を基に素性ベクトルを作成し，あらかじめ学習してある解答分類器により，解答としての確からしさを判定する．その概要を**図 4** に示す．

4.2.1 学習フェーズ

質問文データ D_Q に加えて，各質問文に対する正解のデータ D_A を準備する．正解のデータ D_A は以下の 3 つ組の集合である．

$$\langle \text{質問 } ID, \text{正解，記事 } ID \rangle$$

なお，ある質問 ID に対する正解は複数与えられてもよい．

このデータ D_Q および D_A に基づいて，解答分類器の学習は以下のような手順で行われる．

(1) 解答候補を素性ベクトルに変換する関数を F_A とする．
(2) 固有表現タグ集合 T の各要素 t_i について，以下を行う．
  (a) 正例集合 $E_{t_i}^{+} = \{\}$.
  (b) 負例集合 $E_{t_i}^{-} = \{\}$.
  (c) 各データ $\langle n_j, q_j, t_j \rangle \in D_Q$，$\langle n_j, a_j, d_j \rangle \in D_A$ について，

- d_j に含まれる解答候補 w_j を素性ベクトル $\boldsymbol{v}_j$=$F_A(w_j, d_j, q_j, t_j)$ に変換する．
- $w_j = a_j$ ならば，$\boldsymbol{v}_j$を正例集合 $E^+_{t_i}$ に追加．
- $w_j \neq a_j$ ならば，$\boldsymbol{v}_j$は負例集合 $E^-_{t_i}$ に追加．

(d) 学習アルゴリズム SVM により，各 t_i に関する解答分類器 $g^{t_i}_A = SVM(E^+_{t_i}, E^-_{t_i})$ を作成．

タグ付き文書集合に基づいて，解答候補 w_j を素性ベクトルに変換する関数 $F_A(w_j, d_j, q_j, t_j)$ により作成される素性ベクトルの各素性は以下のとおりである．これらの素性は鈴木らの研究[17]に基づいて決定した．ここでは，質問文に含まれる自立語を特に重要語と呼ぶ．

- 解答候補 w_j が質問文 q_j に含まれるか否か．
- w_j が質問対象語の意味カテゴリに含まれるか否か．
- w_j の直前の名詞が質問対象語の意味カテゴリに含まれるか否か．
- w_j と最も近い重要語との単語距離☆．
- w_j と重要語との平均単語距離．
- w_j の前後 n 単語，n 文節，n 文，n パラグラフにそれぞれ現れる重要語の割合．
- w_j の前後 n 単語，n 文節，n 文，n パラグラフにそれぞれ現れる重要語の品詞の割合．
- w_j の前後 n 単語，n 文節，n 文，n パラグラフにそれぞれ現れる重要語の意味カテゴリの割合．
- 学習データ D_A に現れるすべての句読点・接辞について，w_j の直前/直後の単語と一致したかどうかを，それぞれ {0,1} で列挙．
- すべての意味カテゴリについて，w_j がその意味カテゴリに属するかどうかを {0,1} で列挙する．
- 学習データ D_A に現れるすべての単語について，w_j に最も近い単語と一致するかどうかを {0,1} で列挙する．

なお，上記の n は $1, 2, \ldots, 5$ とする．先に述べた形式で，形態素解析結果を保存しておくことにより，同じパラグラフ，文，文節に入っている重要語の数や単語間の距離を即座に決定できる．また，D_A に現れる単語にユニークな単語番号をあらかじめ付与しておくことにより，単語番号 k の単語と w_j の一致は，該当する素性が素性ベクトルの i 番目の素性から始まるとすると，素性ベクトルの $i + k$ 番目の値を 1 にするだけ素性ベクトルに反映できる．これと同様の手法は，意味カテゴリや品詞についても使える．

正解データの形式から分かるように，質問に対する正解は，パラグラフ位置や記事中の文字位置までは特定されていない．これはデータ作成の作業における工数の制約による．1 つの記事には正解が複数含まれる可能性があるため，記事中で，正解と同一の固有表現すべてについて上記の素性ベクトルを作成する．

4.2.2 実行フェーズ

解答候補 w が与えられたとき，w を解答とするかは以下のような手順で決定される．ただし，解答候補を抽出した記事を d，質問文を q，質問タイプを t とする．

(1) 解答集合 $AS = \{\}$.
(2) 解答候補を素性ベクトルに変換する関数を構築で用いた F_A とする．
(3) 解答候補 w を素性ベクトル $\boldsymbol{v}$=$F_A(w, d, q, t)$ に変換する．
(4) 固有表現タグ集合 T の各要素 t_i について，以下を行う．
  (a) 解答分類器 $g^{t_i}_A(\boldsymbol{v})$ により，$\boldsymbol{v}$を正例，負例に分類する．
  (b) $\boldsymbol{v}$が正例であれば，解答集合 AS に w を追加．

本論文での評価は，上位 5 位まで解答をランキングして出力するタスクであるので，分類器が出力する値により解答候補をランキングし，その上位 5 件を解答として出力する．

以上のように，学習型質問解析モジュールと学習型解答選択モジュールの作成法が明らかになり，システム全体を学習型として構築する方法が示された．

5. 評　　価

フィージビリティの確認のために，システム全体の評価を行った．1,581 問の質問・正解セットを各質問タイプが均一になるように，5 セットに分割し，システム全体での 5 分割交差検定（5-fold cross validation）を行った．すなわち，4 つのセットのデータを用いて，質問解析，解答選択の両方のモジュールを学習し，残りの 1 セットで評価を行う．この処理を評価用セットを変えながら 5 つのすべてのセットについて行い，結果を平均する．

システムの最終的な出力結果として得られた解答を，標準的に用いられる次の 2 つの評価値により評価した．

Top5 スコア $Top5 \stackrel{\text{def}}{=} R/|\,Qs\,|$. ただし，$|\,Qs\,|$ は質問数，R は 5 位以内に正解が含まれた質問数で

☆ 記事 d_j 中の 2 つの単語間の単語数+1.

表 3 出力の評価結果
Table 3 Evaluation results of outputs.

質問タイプ	1 位	2 位	3 位	4 位	5 位	1〜5 位	総計	TOP5(%)	MRR
PERSON	129	48	24	11	6	218	323	67.5	0.511
ORGANIZATION	83	32	18	15	8	156	319	48.9	0.346
LOCATION	65	38	22	18	9	152	308	49.3	0.317
ARTIFACT	26	7	3	3	0	39	141	27.9	0.222
DATE	123	54	34	28	18	257	377	68.2	0.456
TIME	8	3	2	2	0	15	25	60.0	0.427
MONEY	13	8	2	0	0	23	50	46.0	0.353
PERCENT	13	4	1	2	0	20	38	52.6	0.417
計	440	165	106	59	41	880	1,581	55.7	0.393

ある．

MRR（Mean Reciprocal Rank）は各質問について，ランクの 1 位から 5 位まで順に正解かどうかをチェックしていき，最初に正解と判定されたランク n のポイント $1/n$ を与え，質問数で平均したもの．

さらに，システムの出力が不正解となった原因を探るため，途中段階の質問解析，文書検索，解答候補抽出の精度も評価した．

質問解析： システムが判定した質問タイプが正しいかどうかを評価．

文書検索： 検索結果の上位 10 件に正解を含む記事が含まれるかどうかを評価．

解答候補抽出： 10 件の記事から取り出した解答候補に中に，正解が含まれるかどうかを評価．

全体の評価結果を**表 3** に示す．表 3 は，各質問タイプについての，1 位〜5 位での正解数，TOP5 値，MRR 値を表している．本論文で述べた構成法により，全体で MRR=0.393，TOP5=55.7%の質問応答が実現できることが確認された．

6. 考　　察

システム全体の性能としては，MRR = 0.393，TOP5 = 55.7%が得られた．この値は人手作成システム SAIQA [16]により，同じ 2,000 問を用いて評価した結果（MRR = 0.383，TOP5 = 54.4%）と同等な結果を示している．

さらに詳細に個々のモジュール別に分析してみると，システム全体での 5 分割交差検定における質問タイプ判定の精度は**表 4** のとおりであった．質問タイプにより，76〜98%と幅はあるが，平均すると，質問解析には 88%という高い正解率で成功している．次に，検索の正解率を**表 5** に示す．検索は，検索語をもとに関連記事を取り出すため，質問のタイプには影響されない．検索も平均 92%という高い正解率で成功している．これは，記事を参考に質問文を作成したため，質問文の中の検索語で絞り込むと上位 10 件に候補が入りやすかったことが一因であると考えられる．また，解答候補抽出の精度は平均 74.3%であった（**表 6**）．解答抽出の精度は固有表現の抽出精度を表しているが，

表 4 質問解析の結果
Table 4 Results of question analysis.

質問タイプ	正解数	質問数	正解率 (%)
PERSON	307	323	95.0
ORGANIZATION	243	319	76.2
LOCATION	264	308	85.7
ARTIFACT	109	141	77.3
DATE	370	377	98.1
TIME	23	25	92.0
MONEY	44	50	88.0
PERCENT	32	38	84.2
合計	1,392	1,581	88.0

表 5 検索の評価結果
Table 5 Evaluation results of text retrieval.

質問タイプ	正解数	質問数	正解率 (%)
PERSON	302	323	93.5
ORGANIZATION	305	319	95.6
LOCATION	280	308	90.9
ARTIFACT	129	141	92.1
DATE	342	377	90.7
TIME	20	25	80.0
MONEY	45	50	90.0
PERCENT	30	38	78.9
合計	1,453	1,581	91.9

表 6 解答候補抽出の評価結果
Table 6 Evaluation results of answer candidate extraction.

質問タイプ	正解数	質問数	正解率 (%)
PERSON	286	323	88.5
ORGANIZATION	222	319	69.6
LOCATION	194	308	63.0
ARTIFACT	54	141	38.6
DATE	327	377	86.7
TIME	19	25	76.0
MONEY	44	50	88.0
PERCENT	29	38	76.3
合計	1,175	1,581	74.3

Isozaki らの固有表現抽出[5]の精度と比べて，ORGANIZATION や LOCATION，ARTIFACT の抽出の精度が低いのは，質問文作成者が質問の対象として選択した固有表現が複雑な複合語となる傾向があったためだと考えられる．

以上のような結果から，学習型 QA システムにおいて，システム全体の精度を向上させるためには，全体的なモジュールの精度向上にあわせて，特に解答選択の学習の部分を精度を向上させる必要があるという知見が得られた．

システム全体の評価において，特に，ARTIFACT の正解精度が相対的に低かったのは，質問解析，解答候補の抽出において次のような困難な点があったためである．まず，ARTIFACT を尋ねる質問は,「〜は何」のような質問がほとんどであるが,「何」を使った質問は，人名，地名，組織名，割合に関する質問でも使われているため，ARTIFACT に関する質問文と他の質問文との特徴的な差が小さく，質問タイプの分類の時点での精度が低くなっている．さらに，あらかじめ文書集合にタグ付けされた ARTIFACT の精度が低いため，解答候補の抽出精度が下がっていた．固有表現抽出に関する評価型ワークショップ IREX においても，ARTIFACT の精度は他の固有表現よりも低いことが報告されている．その直観的な理由は，製品名等は自由に名前を付けられるため，普通名詞との区別が困難であったり，長い複合語の場合は，記事中の固有表現の開始と終了の位置の判定を誤りやすかったりするという点にある．

また，本論文は 5 分割交差検定によるシステム全体の評価を対象としているが，参考のために，QAC のフォーマルランの質問文 200 問に対する正解率を**表 7** に示す☆．QAC の質問文は，本論文が対象としている固有表現以外にも植物の名前等のクラス名や高さ等，数値表現を尋ねる質問を多く含むため，大枠の比較しかできないが，MRR=0.323 は QAC 参加 15 システム中の中位に位置する．我々の質問タイプの分類に沿って QAC の質問を分類すると，質問数が極端に少ない質問タイプを除けば，MRR=0.4 程度となり，5 分割交差検定と同様な結果が得られた．なお,「その他」に正解があるのは，質問タイプの選択と解答候補の抽出の双方において偶然同じ固有表現に間違えたためである．

さらに，学習技術を利用するメリットを性能以外の面について考えると，大きく分けて次の 2 つのメリットが考えられる．

- システム拡張が容易になる．
- 対象文書の変更や質問タイプの分類の変更が容易にできる．

まず，質問文とその正解の例を加えていくことにより，その質問自体に正解するように学習が行われるとともに，さらに，未知の質問文に対する精度の向上も期待できる．次に，固有表現タイプおよび質問タイプの分類を変更する必要がある場合，システム本体の変更は必要なく，例を差し替えるだけで，この変更に対応することができる．ただし，本論文の構成では，例や質問タイプの変更を反映するには，学習データ全体を用いて再度学習をする必要があるため，構築と実行のフェーズが分かれている必要がある．

☆ 5 分割検定で作成した 1 つのシステムにより評価．

表 7 QAC データによる評価
Table 7 Evaluation results on the QAC testset.

質問タイプ	質問数	TOP5(%)	MRR
PERSON	44	56.8	0.444
ORGANIZATION	23	60.9	0.431
LOCATION	32	65.6	0.460
ARTIFACT	29	48.3	0.372
DATE	14	50.0	0.357
TIME	2	0	0
MONEY	3	33.3	0.333
PERCENT	4	0	0
小計	151	54.3	0.404
その他	49	6.1	0.061
総計	200	42.5	0.323

7. おわりに

本論文は，質問文およびその正解例から，質問に対する解答法を学習する QA システム全体の構築法について述べた．この中で，質問解析，解答選択の 2 つのモジュールを分類問題としてとらえ，各分類器を質問文とその正解例から学習する方法，および学習した分類器の利用方法を明らかにした．我々の知る限り，学習技術により，解答そのものを答える日本語 QA システム全体を構成する方法はこれまで示されておらず，1 つの構成法を与えることは重要なステップである．また，SVM を用いたシステムを実装し，システム全体を通して，学習データにより学習させたときの性能を交差検定により測定した．その結果，人名，地名等，主要な 8 種類の質問タイプについて，平均約 55%の 5 位以内正解率が得られることが判明した．

1,581 問の質問文セットに対する 5 分割交差検定の際，構築フェーズに総計約 20 時間，実行フェーズに総計約 7 時間を要した．個々のモジュールの単体テストでは，さらに大きなデータセットでの実験も可能であ

るが，システム全体の交差検定を現在より大きなデータセットで行うためには，学習および分類の高速化が必要である．また，テストセットの拡張等により対象質問タイプを広げていくことと，システム全体の性能向上も今後の課題である．また，Hirao ら[2)]が SVM を用いた要約手法を研究を行っており，これらの成果を解答の根拠を提示するモジュールに採用し，要約の有用性も評価していきたい．

参 考 文 献

1) 平尾 努，佐々木裕，磯崎秀樹：質問に適応した文書要約手法とその評価，情報処理学会論文誌，Vol.42, No.9, pp.2259–2269 (2001).
2) Hirao, T., Isozaki, H., Maeda, E. and Matsumoto, Y.: Extracting Important Sentences with Support Vector Machines, *Proc. COLING-2002*, pp.342–348 (2002).
3) 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).
4) Information Retieval and Extraction Exercise (IREX) (1999). `http://cs.nyu.edu/cs/projects/proteus/irex/`
5) Isozaki, H. and Kazawa, H.: Efficient Support Vector Classifiers for Named Entity Recognition, *Proc. COLING-2002*, pp.390–396 (2002).
6) Ittycheriah, A., Franz, M., Zhu, W-J. and Ratnaparkhi, A.: Question Answering Using Maximum-Entropy Components, *Proc. NAACL* (2001).
7) Ittycheriah, A., Franz, M., Zhu, W-J. and Ratnaparkhi, A.: IBM's Statistical Question Answering System — TREC-10, *Proc. TREC-10* (2001).
8) 神門典子：NTCIR とその背景，人工知能学会誌，Vol.17, No.3, pp.296–300 (2002).
9) 前田英作：痛快！サポートベクトルマシン—古くて新しいパターン認識手法，情報処理学会誌，Vol.42, No.7, pp.676–683 (2001).
10) Moldovan, D. and Harabagiu, S.M.: Lasso: A Tool for Surfing the Answer Net, *Proc. 8th Text Retrieval Conference* (*TREC-8*), pp.175–184 (1999).
11) 村田真樹，内山将夫，井佐原均：類似度に基づく推論を用いた質問応答システム，自然言語処理研究会，2000-NL-135 (2000).
12) Ng, H.T., Kwan, J.L.P. and Xia, Y.: Question Answering Using a Large Text Database: A Machine Learning Approace, *Proc. Empirical Methods in Natural Language Processing* (*EMNLP-2001*), pp.67–73 (2001).
13) Pasca, M.A. and Harabagiu, S.M.: High Performance Quesiton/Answering, *Proc. SIGIR-2001*, pp.366–374 (2001).
14) Question Answering Challenge (QAC) (2002). http://www.nlp.cs.ritsumei.ac.jp/qac/
15) 佐々木裕，磯崎秀樹，平 博順，廣田啓一，賀沢秀人，平尾 努，中島浩之，加藤恒昭：質問応答システムの比較と評価，信学技報，NLC-2000-10, pp.17–24 (2000).
16) 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，情報学基礎研究会，No.064-012, 情報処理学会 (2001).
17) Suzuki, J., Sasaki, Y. and Maeda, E.: SVM Answer Selection for Open-Domain Question Answering, *Proc. Coling-2002*, pp.974–980 (2002).
18) 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).
19) 高須淳宏：Support Vector Machine による分類，発見科学とデータマイニング，第 12 章，pp.118–127, 共立出版 (2001).
20) 徳永健伸：情報検索と言語処理，東京大学出版会 (1999).
21) Vapnik, V.N.: *The Nature of Statistical Learnig Theory*, Springer (1995).
22) Voorhees, E.M. and Tice, D.M.: Building a Question-Answering Test Collection, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in the Information Retrieval*, pp.192–199 (2000).
23) Zukerman, I. and Horvitz, E.: Using Machine Learning Techniques to Interpret WH-Questions, *Proc. Association for Computational Linguistics* (*ACL-2001*), Toulouse, France, pp.547–554 (2001).

(平成 15 年 8 月 18 日受付)
(平成 15 年 11 月 4 日採録)

佐々木　裕（正会員）

1986 年筑波大学第三学群情報学類卒業．1988 年同大学院修士課程理工学研究科修了．同年日本電信電話株式会社入社．現在，NTT コミュニケーション科学基礎研究所に所属．博士（工学）．1995 年〜1996 年サイモン・フレーザー大学（カナダ）客員研究員．主として自然言語処理，機械学習に関する研究に従事．人工知能学会，言語処理学会，ACL 等各会員．

磯崎　秀樹（正会員）

1983 年東京大学工学部計数工学科卒業．1986 年同工学系大学院修士課程修了．同年日本電信電話株式会社入社．1990 年〜1991 年スタンフォード大学ロボティクス研究所客員研究員．現在，NTT コミュニケーション科学基礎研究所特別研究員．博士（工学）．人工知能・自然言語処理の研究に従事．電子情報通信学会，人工知能学会，言語処理学会，AAAI，ACL 各会員．

鈴木　　潤（正会員）

1999 年慶應義塾大学理工学部数理科学科卒業．2001 年同大学院理工学研究科計算機科学専攻修士課程修了．同年日本電信電話株式会社入社．現在，NTT コミュニケーション科学基礎研究所に所属．2003 年 10 月より奈良先端科学技術大学院大学博士後期課程在学．主として自然言語処理，機械学習に関する研究に従事．ACL，言語処理学会各会員．

国領　弘治

1984 年彦根工業高校卒業．同年日本電信電話株式会社入社．1997 年〜2000 年 NTT コムウェアに所属．2001 年〜2002 年 NTT コミュニケーション科学基礎研究所に所属．現在，NTT コムウェア西日本株式会社に所属．主として自然言語処理に関する研究に従事．

平尾　　努（正会員）

1995 年関西大学工学部電気工学科卒業．1997 年奈良先端科学技術大学院大学博士前期課程修了．同年，NTT データ通信株式会社（現,（株）NTT データ）入社．2000 年より，日本電信電話株式会社 NTT コミュニケーション科学基礎研究所に所属．博士（工学）．自然言語処理の研究に従事．言語処理学会，ACL 各会員．

賀沢　秀人（正会員）

1995 年東京大学理学部物理学科卒業．1997 年同大学院理学系研究科修士課程修了．同年日本電信電話株式会社入社．現在，NTT コミュニケーション科学基礎研究所に所属．主として自然言語処理，機械学習に関する研究に従事．ACL，IEEE 各会員．

前田　英作（正会員）

1984 年東京大学理学部卒業．1986 年同大学院修士課程理学系研究科修了．同年日本電信電話株式会社入社．現在，NTT コミュニケーション科学基礎研究所．知能情報研究部知識処理研究グループ リーダ．工学博士．1995 年〜1996 年ケンブリッジ大学（英国）客員研究員．主としてパターン認識，統計的機械学習，生物情報処理の研究に従事．IEEE，ACL，電子情報通信学会，日本バイオインフォマティックス学会会員．