LE-10

統計的機械学習を用いた質問タイプ同定

Question Type Classification using Statistical Machine Learning

鈴木 潤† 佐々木 裕† 前田 英作†
Jun Suzuki Yutaka Sasaki Eisaku Maeda

1 はじめに

質問応答は，自然文で与えられた質問に対し，大量文書から適切な解答を導き出す技術であり，TREC(Text REtrieval Conference) QA-Track, QAC(Question and Answering Challenge) が開催されるなど，新しい情報アクセス技術として世界的に注目を集めている．

人間の場合でも，相手が何を訊いているのか正しく理解できなければ，相手の質問に答えることはできない．それと同様に，計算機による質問応答においても，与えられた質問の意図を分析する「質問解析」が正しくできなければならない．特に，質問応答では，質問文が何を訊いてのかを示す「質問タイプ」(地名，人名，人数など数十種) を正しく同定することが非常に重要である．また，この質問タイプ同定問題は，より一般的に，自然文による質問の意図を判別する問題と捉えることもでき，本稿で提案する手法は，情報検索/抽出，対話など自然文を扱う問題に広く適用可能な技術である．

TREC QA-Track 参加システムなどの代表的な質問応答システムでは，意味的制約 [1]，ルールベース [2]，パターンマッチング [3, 4] などの方法を用いて質問タイプを同定していた．しかしながら，こうした方法では，限られた種類の質問文に対してのみ有効であったり，人手によるルール作成コストが大きいなどの問題があった．

これを解決する一つの手法が統計的機械学習手法を導入する方法である．既に，これまでの質問応答システムにおいて，決定木学習 [5] や最大エントロピー法 [6] などの統計的機械学習手法を用いた質問タイプ同定手法が提案されている．しかし，多様な言語表現に対応し，かつ，高精度な質問タイプ同定を実現するには，意味情報や質問文の構造といった，より多くの情報を素性として考慮しなければならない．

サンプル数に対して素性空間の次元数が非常に高くなる上記のような問題に対して，高精度な分類を可能にする統計的機械学習手法として Support Vector Machine(SVM)[7, 8] が知られている．

そこで，本稿では，質問タイプ同定に適した素性抽出手法と，統計的機械学習手法 SVM を組み合わせた，新しい質問タイプ同定法の提案を行う．

2 SVM を用いた質問タイプ同定

質問応答技術では，最初の処理となる質問解析を誤ると後の解析全てに悪影響を及ぼす．よって，与えられた質問が何について聞いているのか判断する質問タイプ同定は，質問応答技術の精度をあげる上で極めて重要な処理の一つである．

英語質問応答システムでは，TREC QA-Track 参加システムが様々な質問タイプセットを定義している．本稿では，日本語固有表現抽出タスク IREX により定義された固有表現タイプ 8 種を基礎として 35 種類の質問タイプを設定する．

2.1 質問タイプ同定に適した素性抽出法

質問タイプを特徴付ける素性として，文献 [3, 4] から，質問文中の意味情報と意味情報の構造が有効であると考えられる．また，質問タイプ同定には，質問タイプ「人名」に対する「～は誰ですか」のような典型的な表現が存在する．よって，質問タイプ同定には，単語あるいは意味情報と，それらが質問文中でどのような関係で使われているかが極めて重要であると考えられる．そこで，本稿では単語属性 N−gram を用いて質問タイプを特徴付ける素性を抽出する手法を提案する．

2.1.1 単語属性 N−gram

本稿では単語，品詞，意味情報を要素とした連鎖には，質問タイプを効果的に特徴付ける特徴が含まれていると仮定し，これらを質問タイプ同定問題の素性として用いる手法を提案する．ここで，本稿では単語，品詞，意味情報を単語属性と呼び，単語属性の連鎖を単語属性 N−gram と呼ぶ ($N = 1, 2, 3, \ldots$)．

このように単語，品詞，意味情報といった各種の単語属性を全て用いることにより，質問タイプを特徴付ける表層的な定型表現や意味的な構造を網羅的に考慮した素性集合を作成することができる．

図 1 に，単語属性 N−gram の具体的な抽出例を示す．単語属性 N−gram の要素となる単語，品詞，意味情報の抽出法については以下に述べる．

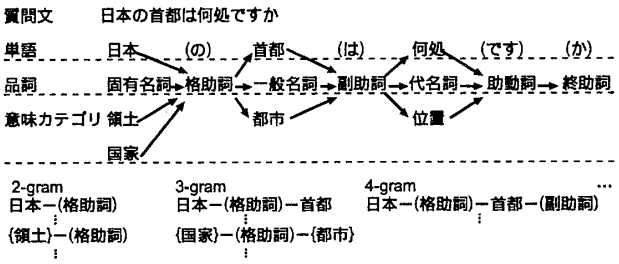


図 1: 単語属性 N−gram の概略及び抽出例

各素性は質問文中に出現したか，しないかの 1 または 0 を値としてとる．

2.1.2 単語及び品詞

質問文内に出現する単語及びその単語の品詞を単語属性 N−gram の要素として用いる．ただし，単語自身は質問文中に出現する全単語ではなく，自立語のみを用いる．

各単語の品詞を判定するために，形態素解析器として ALTJAWS[1]を使用する．

2.1.3 意味カテゴリ

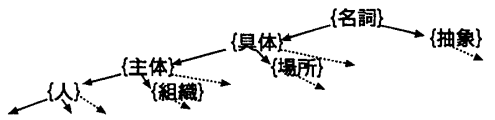
前述の形態素解析器 ALTJAWS は，各単語に意味情報として意味カテゴリも同時に付与する．

意味カテゴリとは，日本語語彙大系 [9] 内に記述されている各単語の意味を表記するためのカテゴリであり，全 2715 カテゴリ存在し，木構造により記述されている．

意味カテゴリから素性を抽出する際には，ある意味カテゴリに単語が属する場合は対象カテゴリより上位のカテゴリにもその単語は属する，また，ある単語が複数の意味カテゴリに属する場合はその全てのカテゴリに属する，と考えて抽出を行う．図 2 に，具体的な意味カテゴリからの素性抽出例を示す．

† 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

[1] http://www.kecl.ntt.co.jp/icl/mtg/resources/altjaws2-j.html



(a)木構造で定義された意味カテゴリ

例：{人}に属する単語が出現した時

名詞	具体	抽象	主体	場所	人	組織	人間	…	2715
1	1	0	1	0	1	0	0		0

例：{人}と{場所}に属する単語が同時に出現した場合

名詞	具体	抽象	主体	場所	人	組織	人間	…	2715
1	1	0	1	1	1	0	0		0

(b)意味カテゴリから抽出される素性ベクトル例

図 2: 意味カテゴリに関する素性抽出方法の概略図

2.2 Support Vector Machine

Support Vector Machine (SVM) は Vapnik らによって提唱されたノンパラメトリックな機械学習手法であり，正例及び負例の二クラス間の距離 (マージン) を最大にするという基準で識別平面を決定する Large Margin Classifier の一つである．

SVM の特徴として，サンプル数に対して素性空間の次元が高い場合でも高精度な識別が行える点，カーネル関数を用いることにより非線形識別関数を容易に扱うことができる点が挙げられ，質問タイプ同定に用いるのに適していると考えられる．

3 実験

本稿では，人手作成ルール (RULE)[2]、決定木学習 (DT)、最大エントロピー法 (ME) と SVM を用いた質問タイプ同定実験を行った．

質問応答テストコレクション 10000 問 [2] を用い，5fold の交差検定による精度を比較した (訓練 4，評価 1)．実験には 16 質問タイプとそれ以外の質問タイプをまとめた「その他 (OTHER)」という計 17 質問タイプを設定した．本稿で提案する素性抽出手法を適用した際に得られる素性ベクトルは 23528 次元であった．

SVM のカーネル関数には，1 次および 2 次の多項式型カーネルを用いた．また，多クラス分類への適用手法としては，one vs. rest 法を使用した．

4 実験結果及び考察

表 1 に結果を示す．表中の SVM1 は多項式カーネル 1 次の SVM，SVM2 は多項式カーネル 2 次の SVM を用いた実験の結果を表す．macro はマクロ平均を表し，各質問タイプの F 値の平均であり，micro はマイクロ平均を表し，全サンプルの F 値の平均である．

人手作成ルールの手法が良い精度を得られなかった一番の理由として，質問文の多様な表現に対応できなかったことが考えられる．逆に，質問タイプ「人名」に対する「～は誰ですか」のような質問タイプを特徴付ける典型的な表現の質問には，ほぼ正解することができていた．

機械学習手法間の精度比較では，SVM を用いた提案手法が最も高い精度を得た．決定木学習，最大エントロピー法では，サンプル数の少ない質問タイプの精度が低くなっている．これは，本稿で用いた素性空間がサンプル数に対して次元数の高いものであることから，サンプル数の少ない質問タイプでは，その影響は更に大きくなり，汎化能力の高い分類器を学習できなかったことが原因と考えられる．

この結果から，サンプル数の少ない質問タイプでも，SVM は他の機械学習手法より効率良く高精度な分類器を作成していることがわかった．

比較手法と比べて SVM が統計的有意差をもって高精度であったことから，本稿で用いた素性抽出手法と SVM の組合せによる質問タイプ同定手法は有効であることが示せた．

表 1: 提案手法と既存手法の性能比較

		RULE	DT	ME	SVM1	SVM2
タイプ	質問数	F measure				
年齢	130	0.784	0.878	0.708	**0.893**	0.855
日付	1885	0.832	0.924	0.927	**0.959**	0.957
事柄	165	0.545	0.296	0.518	**0.622**	0.620
場所	1530	0.616	0.575	0.714	0.773	**0.779**
値段	250	0.734	0.810	0.571	**0.834**	0.789
組織名数	140	**0.746**	0.654	0.632	0.705	0.733
人数	365	0.853	0.834	0.835	**0.869**	0.864
組織名	1605	0.618	0.541	0.697	0.729	**0.736**
割合	190	0.817	0.765	0.705	**0.819**	0.814
期間	260	0.439	**0.734**	0.626	0.721	0.728
人名	1615	0.816	0.707	0.851	**0.897**	0.897
製品名	135	0.402	0.185	0.275	0.507	**0.537**
役職名	270	0.790	0.751	0.821	**0.903**	0.879
物質	130	0.498	0.387	0.407	**0.630**	0.606
時間	125	0.718	0.778	0.742	0.803	**0.826**
作品名	150	0.316	0.299	0.270	**0.479**	0.433
その他	1055	0.434	0.575	0.609	0.666	**0.679**
Macro		0.670	0.629	0.642	**0.754**	0.749
Micro	10000	0.683	0.670	0.754	0.811	**0.812**

5 まとめ

本稿では，質問文中に現れる単語属性とその連鎖が質問タイプを特徴付けるのに有効であると考え，単語属性 N−gram を用いた素性抽出手法を提案した．また，サンプル数に対して素性空間が高次元でも高精度な分類が行える機械学習手法である SVM と組み合わせて質問タイプ同定実験を行い，本稿で提案した素性抽出及び SVM を用いた質問タイプ同定手法の有効性を示した．

参考文献

[1] 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, 情報処理学会 自然言語処理研究会 NL-135, pp. 181–188 (2000).

[2] 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, 情報処理学会 情報学基礎研究会 FI-64, pp. 77–82 (2001).

[3] Harabagiu, S., Pasca, M. and Maiorano, S.: Experiments with Open-Domain Textual Question Answering, *Proc. of COLING-2000* (2000).

[4] Hovy, E., Hermjakob, U. and Lin, C.-Y.: The Use of External Knowledge of Factoid QA, *Proc. of TREC 2001*, NIST (2001).

[5] Zukerman, I. and Horvitz, E.: Toward Understanding WH-Questions: A Statistical Analysis, *Proc. of Association for Computational Linguistics (ACL-2001)*, ACL (2001).

[6] Ittycheriah, A., Franz, M., Zhu, W. and Ratnaparkhi, A.: Question Answering Using Maximum-Entropy Components, *Proc. of NAACL 2001*, ACL, pp. 33–39 (2001).

[7] Cortes, C. and Vapnik, V. N.: Support Vector Networks, *Mahine Learning*, Vol. 20, pp. 273–297 (1995).

[8] 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).

[9] 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, 岩波書店 (1997).