# 任意の回答を対象とする質問応答のための実世界質問の分析と回答タイプ判定法の検討

**水野 淳太** † **秋葉 友良** †

† 豊橋技術科学大学 情報工学系
email: {jmizuno,akiba}@cl.ics.tut.ac.jp

## 1 はじめに

テキスト情報源の電子化およびインターネットなどの共有可能なテキスト情報の急増を背景に、大規模なテキストから必要な情報を効率よく獲得するための情報アクセス技術が重要な研究課題となっている。近年、情報検索を高精度化する技術として、自然言語で表された質問を入力とし大規模な検索対象から該当する答の部分 (単語や句) を抽出するオープンドメイン質問応答 (以下、質問応答) が注目を集めている。質問応答は、事実を問う質問に対して語や句などの短い表現で回答することを要求する事実 (factoid) 型質問について主に研究が進められきた。特に、米国 NIST の TREC(Text REtrieval Conference) や国内の NTCIR(NII-NACSIS Test Collection for IR System) プロジェクトなどで、評価型ワークショップおよびテストコレクションの構築が活発に行われている。

一方、事実型以外の質問を扱う質問応答研究としては、2006 年度の NTCIR から特定の質問型に限定する事無く、実世界に現れるような質問を対象とした評価が行われている。 筆者らはこれに対応した質問応答システムとしてユニバーサル質問応答システム (以下、UQA) の開発を行っている。本稿では UQA 実現のための第一段階として行った、実世界における質問の調査、回答タイプ判定手法の提案、質問サイトおよび NTCIR のテストセットを用いた UQA の予備実験の結果について報告する。

## 2 質問・回答の分析

### 2.1 WWW 質問サイトからの質問・回答の収集と分析

実世界にはどのような質問がありうるかを調べるために、WWW 上の質問に注目し、質問とその回答の収集を行った。今回は、ある利用者の書き込んだ質問に対し他の利用者が回答を登録する質問ポータルコミュニティサイトの一つである「教えて!goo」[1]を対象とした。利用者の一つの質問に対して、回答は複数の利用者によって複数の書き込みが行われる。この「質問」と「回答の集合」を一つの質問・回答ペアとして、1,187,873 ペアを収集した。

収集した質問・回答ペアを調べたところ、質問を見ただけでは質問型を特定するのが困難な場合が多く見られた。例えば、以下のような例がある。

質問：この世界で一番、壮大な自然を持っている国はどこなのでしょうか。

回答 1：直感的に私はロシアを思い浮かべます。もうひとつ私が挙げるのは日本です。流氷からサンゴ礁までを一国内で、しかもこれほどコンパクトにまとまって見られるところは他に例がありません。
回答 2：行った事があるのは、オーストラリアですね。あたり一面の地平線というのは、日本では経験の出来ない大自然だと思います。

この例の場合、質問だけを見ると、国名を答える事実型質問に見える。しかし回答を見ると、事実や経験に基づいた意見を述べている。単純に国名を答えるだけでは答えとして不足している典型的な例である。

そこで、回答の書き込みを調べることで、結果として各質問に対してどのような型の回答が行われたかという視点から質問・回答ペアの分類を行うことにした。これを回答タイプと呼ぶ。

回答タイプの種類は田村ら [1] の定義した質問タイプを参考に、回答タイプに適するように追加・変更を加えた。回答タイプの種類と、それらの WWW 質問サイト上での頻度分布を図 1 に示す。頻度分布は収集した質問・回答ペアのうちランダムに抽出した 2,064 ペアに対して調べた。

また、WWW 質問サイトでは質問の内容によって、コンピュータ、スポーツなどのカテゴリに分類されている。各カテゴリにおける回答タイプの分布を図 1 に示す。これから「コンピュータ」 では「方法」に関する質問が、「マネー」では「事実」に関する質問が多い事などが分かる。カテゴリによって回答タイプの分布は異なり、「趣味」カテゴリが、全体の分布に最も類似しているといえる。

### 2.2 NTCIR6-QAC4 の問題・回答の分析

情報アクセス技術に関する評価型ワークショップ NTCIR[2]では、オープンドメイン質問応答の評価タスク QAC(Question Answering Challenge) を実施している。2001 年の QAC1 から 2004 年の QAC3 までは、事実型の質問を対象としていた。現在評価の行われている NTCIR6-QAC4 では、非事実型の質問も対象とした QA の評価が行われている。例えば理由や説明などを尋ねるような、任意の回答を前提とした質問が対象となる。

[1] http://oshiete.goo.ne.jp/

[2] http://research.nii.ac.jp/ntcir/

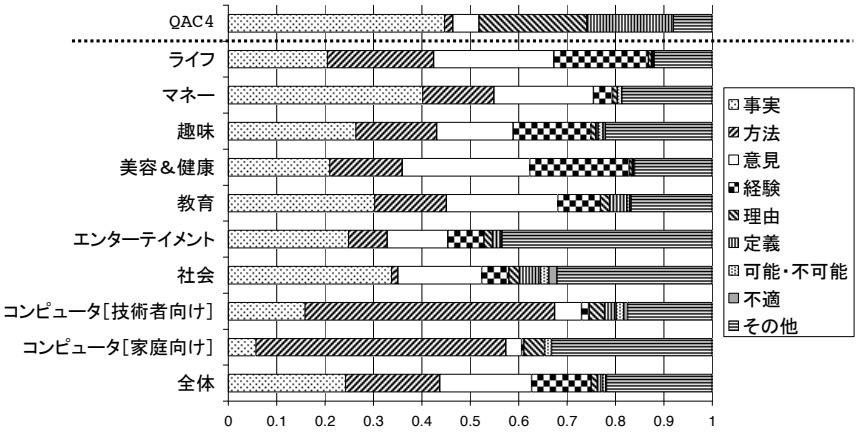


図 1: 各カテゴリにおける回答タイプの分布

表 1: 回答タイプとその割合

| 回答タイプ | 質問の表現例 | 割合 |
|---|---|---|
| 事実 | ～の場合、どうなるのでしょうか？ | 24.2% |
| 方法 | ～するにはどうすればよいでしょうか？ ～したいのですが、やり方が分かりません | 19.5% |
| 意見 | ～けど、どう思いますか？ ～と思いませんか？ | 18.9% |
| 経験 | ～された方いませんか？ ～なんて経験ありませんか？ | 12.3% |
| 理由 | ～となるのはなぜでしょうか？ ～となってしまうんですが、原因分かりますか？ | 1.3% |
| 定義 | ～とは何ですか？ | 1.1% |
| 可能・不可能 | ～できますか？ | 0.6% |
| 不適 | 質問として成り立っていないもの 質問者の不備などで回答の出来ないもの | 7.9% |
| 事実型 | ～という歌詞の曲知りませんか？ | 8.6% |
| 複数の事柄を答える回答 | | 4.9% |
| その他 | | 0.6% |

質問・回答の一例を以下に示す。

質問：NPO 法はどのような経緯を経て成立しましたか。

回答：１９９５年の阪神大震災でボランティア活動が広く社会に認知されたのをきっかけに、ＮＰＯに法人格を与えて活動を促進するため

QAC4 formalrun の質問は全部で 100 問である。これらに対し、WWW 質問サイトと同じ基準で回答タイプの付与を行った。正解は文書検索結果の上位 5 件を人手で調べる事によって作成した。WWW 質問サイトに比べ、QAC4 は質問の長さが比較的短く、また回答となる文書も新聞記事であるため、文章として整っている。そのため、回答タイプの付与は比較的容易であった。回答タイプの分布を図 1 の最上行に示す。WWW 質問サイトと比較すると、理由を答える回答が多く、経験を答える回答が少ないことが大きな違いである。回答タイプの分布という点では WWW 質問サイトとは大きく異なっていることが分かる。

## 3 UQA のアプローチ

まず、本論文では UQA の問題設定を単純化して「正解は回答候補の文書中の 1 段落」とする。QAC4 の問題設定では、回答部分の抽出や要約も想定をしているが、本論文では扱わない。

質問に対して正解となる回答は、質問と内容が類似しており、かつ「質問が予期する回答タイプ」と「回答の回答タイプ」が一致しているものであると考えた。そこで、質問に対して類似文書の検索、回答タイプの一致判定を別々に行い、その結果を統合する事で正解を得る、というアプローチを採用した (図 2)。

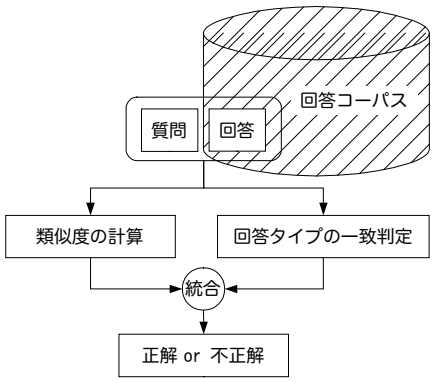


図 2: UQA のアプローチ

類似文書の検索では、文書検索で通常用いられる、

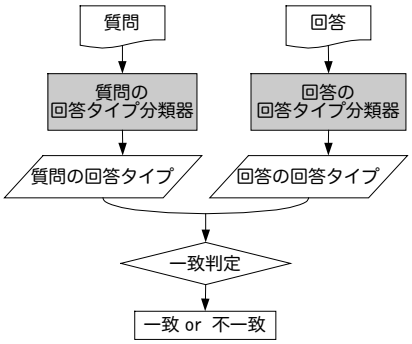


図 3: 手法 1

内容語をベースとした類似度尺度を用いた。実装には GETA[3]を用いた。類似尺度には TF-IDF を文書長で正規化した重み付け [2] を用いた。

回答タイプの一致判定には、機械学習による分類器を用いた 2 種類の手法を提案する。

**手法 1：質問と回答から回答タイプを個別に判定する**

質問と回答のそれぞれの回答タイプを求め、それらが一致するかどうかを判定する手法である。例えば質問に対して回答タイプが “理由” と分類された場合、回答も “理由” と分類されたものが正解となりうる。

図 3 にて判定の流れを説明する。質問と回答に対してそれぞれの回答タイプ分類器を用意し、それぞれの回答タイプを得る。最後に、得られた回答タイプが一致するかどうかを判定する。回答タイプ分類器は、“方法” についての分類器、“理由” についての分類器などのように各回答タイプごとに用意する。複数の回答タイプに分類された場合、そのいずれか一つでも一致していればよい、とした。

本手法では、人手で回答タイプ分類した結果を学習データとして用いる。

**手法 2：質問と回答の回答タイプの一致を直接判定する**

質問と回答の回答タイプが一致するかどうかを二値分類する手法である。手法 1 との大きな違いは、人手による回答タイプ分類を用いない点である。そのため、タイプ分類の体系や粒度、および人手による分類の揺れに、手法が影響を受けないという利点がある。

この手法を図 4 にて説明する。質問と回答の両方を入力にとり、回答タイプが一致するかどうかを二値分類器で判定する。この際、各々の回答タイプが求められる訳ではなく、一致するかどうかのみが分かる。

本手法では、学習データとして、質問と回答のペアをそのまま用いる。

[3]汎用連想計算エンジン http://geta.ex.nii.ac.jp

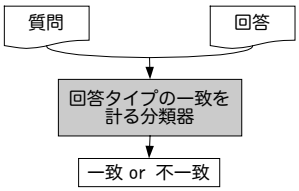


図 4: 手法 2

# 4 評価実験

QAC4 の formalrun100 問に対して評価実験を行った。QAC4 の回答抽出コーパスは毎日新聞 98〜01 年の記事である。回答は新聞記事における段落単位で行い、要約や回答部分の抽出などは行わない。QAC4 は現時点では評価中であり、正解セットが与えられていない。そのため正解判定は人手で行い、正解となる文字列が含まれていたら、その段落全体を正解であると定めた。評価尺度には上位 5 位の MRR(平均逆順位)を用いた。

$$\frac{1}{N}\sum_{k=1}^{N}\frac{1}{rank_k} \tag{1}$$

$rank_k \cdots$ 問題 $k$ における正解の最高順位
$N \cdots$ 問題数

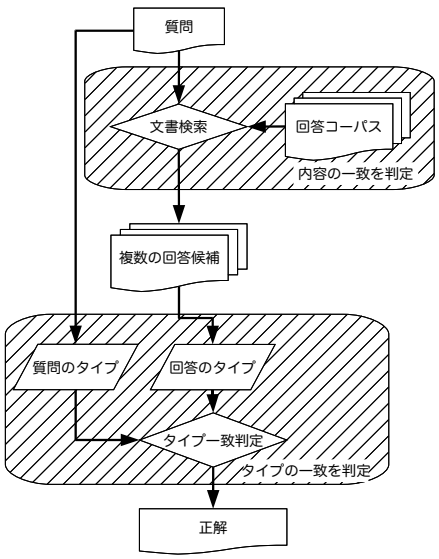


図 5: 作成したシステム

提案手法は図 5 に示すように実装した。この実装では文書検索時に得られる類似度のスコアを、回答タイプの一致を判定する際に反映させていないという点で、図 2 のアプローチの近似である。

処理の流れを以下に示す。

1. 質問文を入力として文書検索を行い、質問文と類似した文書を上位 5 件まで出力する

2. 得られた 5 つの文書を段落単位に分割する

3. 各段落に対し、回答タイプの一致判定を行い、SVM のスコア (分離平面からの距離) の高い順に並べ替える

4. 並べ替えた上位 5 件を回答として出力する

### 4.1 手法 1 の結果

QAC4 において質問と回答の回答タイプを個別に判定する手法 1 の適用を試みた。100 問を 90 問の学習データと、10 問のテストデータに分けて評価を行った。

機械学習には SVM を用いた。学習データの素性には、選択した品詞の形態素 uni-gram を用いた。例えば、名詞以外の形態素 uni-gram など、いくつかの組み合わせで評価を行った。その結果、質問に対する回答タイプ分類は 6 割程度の精度であった。しかし、回答に対する回答タイプ分類は 1 割程度の精度しか無く、この手法は有効でないと判断した。

### 4.2 手法 2 の結果

QAC4 において “質問と回答の回答タイプの一致” を二値分類する手法 2 の適用を試みた。10 問ずつ 10 セットに分割し、そのうち 90 問から分類器を作成し、残りの 10 問にてテストを行う。この操作を異なる分割にて 10 回繰り返す 10 fold cross validation を行った。

回答タイプ一致の分類器は SVM を用いて作成し、素性には品詞別の形態素 uni-gram と疑問表現、文末表現を文に出て来た表現そのままで用いた。形態素 uni-gram には機能語となる付属語として、接続詞、助詞、助動詞、副詞、連体詞、感動詞を用意した。文末表現は “最後の自立動詞から文末記号まで” と定義した。疑問表現は “なぜ”、“何の” など 36 種類の表現を用いた。以上の表現を、質問に現れた表現であるのか、回答に現れた表現であるのかを区別して素性として用いた。

SVM の学習データは以下の通りである。

**正例** 質問と正解段落とのペア。

**負例** 質問と、文書検索上位の文書の正解段落以外の全ての段落とのペア。

負例は、質問に対し内容的には一致しているが回答タイプは異なる回答とのペアである、ということが出来る。学習データのサイズは平均で正例が 263 個、負例が 3,205 個であった。

結果を表 2 に示す。ベースラインは文書検索の結果で第 1 位の文書の最初の 5 段落を出力した結果である。提案手法は、ベースラインに比べて改善出来ている事が分かる。特に素性として助詞、感動詞の組み合わせが有効であった。

表 2: QAC4 に対する手法 2 の実験結果

| 学習データ | | MRR |
|---|---|---|
| QAC4 | 助詞 | 0.244 |
| | 助詞、助動詞 | 0.247 |
| | 助詞、感動詞 | 0.282 |
| | 助詞、文末表現、疑問表現、副詞 | 0.120 |
| WWW 質問サイト | 助詞 | 0.322 |
| | 助詞、文末表現、疑問表現、副詞 | 0.329 |
| ベースライン | | 0.183 |

### 4.3 WWW 質問サイトの質問・回答を学習データに用いた手法 2 の結果

WWW 質問サイトの質問から回答タイプ一致判定のための正例、負例を作り、分類器を作成した。

SVM の学習データは QAC4 と同様の観点で、“趣味” カテゴリの質問・回答ペアを用いて作成した。正例と負例は以下の通りである。

**正例** 質問と正解のペア。

**負例** 質問と、それを入力とした文書検索の結果で第 1 位ではあるが正解ではなかった回答とのペア。このとき文書検索のためのコーパスには WWW 質問サイトのすべての回答を用いた。

正例は 67,260 個、負例は 66,481 個用意する事が出来た。SVM の学習に用いた素性は QAC4 を学習データとした際と同様の種類の形態素 uni-gram の組み合わせを用いた。

結果を表 2 に示す[4]。学習データを QAC4 とした場合と同様、提案手法によってベースラインを改善することが出来た。学習データをテストデータの分野とは異なるデータにした場合にも効果がある事が分かる。学習データのサイズの増大により、複雑な素性 (助詞、文末表現、疑問表現、副詞の組み合わせ) を用いた場合に性能の向上が確認出来た。

## 5 まとめと今後の課題

類似した文書の検索と回答タイプの一致判定を統合して UQA の実装を行った。QAC4 を対象とした評価実験の結果、質問と回答の回答タイプが一致するかどうかを直接二値分類する手法が有効である事が分かった。

今後は、本手法の回答タイプ一致判定において、より効果的な素性の調査を行いたい。また、WWW 質問サイトの質問を対象とした評価を行う予定である。

## 参考文献

[1] 田村昇裕 高村大也 奥村学, “複数分質問のタイプ同定” 言語処理学会 2005(D5-5) 2005.

[2] A.Singhal, C.Buckley, and M.Mitra, “Pivoted document length normalization” Proc. of SIGIR’96 1996.

[4] 一部の素性の取得に間違いがあった事が確認されたので、現在再実験中である