# 常識を用いた質問応答のための正解判定用表現パターンの自動獲得

**村田祐一　秋葉友良**

豊橋技術科学大学　情報工学系
e-mail: murata@cl.ics.tut.ac.jp

## 1　はじめに

質問応答技術は，自然言語の質問文による検索質問について，組織化されていないオープンドメイン文書集合から直接の答えとなる部分のみを抽出する精度重視の情報検索技術であり，米国 NIST の TREC や日本の NTCIR において，大規模な評価が行われている．

質問応答では，質問文中に，正解の上位概念を表す語が特定のパターンで現れることが多い [2]．例えば,「夏目漱石の有名な作品は何ですか」という質問は,「坊ちゃん」等の正解を持つが，この時，質問文中に現れた「作品」と正解「坊ちゃん」は，上位下位関係にある．この上位語（質問焦点）を利用し，回答候補との上位下位関係の有無を調べることで，正解選択の手掛かりとすることが出来る．

上位下位関係の判定には，EDR 電子化辞書の概念体系辞書や WordNet 等のシソーラスを用いる方法がある．しかし，質問応答システムはオープンドメイン環境での動作が求められるため，既存のシソーラスでは規模が小さく回答候補をカバーでない．
また，上位下位関係の判定には，予め文書集合から自動獲得した上位語下位語ペアを用いる方法もある．上位語下位語ペアの自動獲得は，文献 [3] により，”A such as B”といった表現を用いた自動抽出として，最初に提案された．文献 [4] では，品詞パターンを用いて同格の語のペアを自動抽出する手法に付いて述べられている．文献 [5] では，上位下位関係を表す代表的な表現パターンについて，下位語として表れる名詞の用法が分析されている．

しかし，上位語下位語ペアの自動獲得では，通常，短い単語のペアの獲得に制限し,“ 情報技術関連企業 ”等の長い複合名詞や“ 夏目漱石の作品 ”などの句は獲得の対象としない．単語以上に複雑な表現まで考慮に入れると獲得対象文字列を特定することが困難であり，また計算および空間コストが大幅に増大するためである．特に，質問応答では，対象文書中のあらゆる表現が候補となりえるため，あらかじめ全てを獲得することは困難である．

そこで我々は，大規模コーパスをそのまま知識源として使用し，語間の上位下位関係を判定する手法を提案している．語間の上位下位関係の有無は，コーパス中の“ 夏目漱石の作品「坊ちゃん」”のような表現の出現の有無で検査できる[1]．文献 [2] では，この上位下位関係判定機能を組み込むことによる質問応答システムの性能の向上が報告されている．先行研究では判定用の表現パターンを人手で作成したが，本稿では，表現パターンをコーパスから自動抽出する手法について述べる[2]．

## 2　表現パターンの自動獲得

我々の表現パターン獲得の手法は，以下のステップから成る．

1. 種となる上位語下位語ペアを用い，コーパスより表現パターン候補を複数抽出する．
2. 得られた複数の表現パターン候補の性能を正例と負例を含む上位語下位語ペア（開発用テストセット）を用いて評価し，表現パターンを選択する．

### 2.1　表現パターンの抽出

#### 2.1.1　種子セットの抽出

EDR 概念体系辞書 [1] より，表現パターン抽出の種子セットとなる上位語下位語ペアの抽出を行った．EDR 概念体系辞書は，約 41 万の概念間の上位下位関係を規

[1]現時点では QAC 等の評価型ワークショップの対象外ではあるが，本手法を用いれば「滋賀県の草津」などの下位語（回答候補）が句になる場合も扱う事が出来る．

[2]文献 [3] で表現パターンを用いて上位語下位語ペアの獲得を行っているが，本稿では表現パターンの方を自動獲得する点が異なることに注意されたい．

表 1: 自動獲得した表現パターン例

| 表現パターン | 判定できた例 | 精度 |
|---|---|---|
| [下位] 以外の [上位] | [バナナ] 以外の [果物] | 0.5690 |
| [下位] など [上位] | [土星] など [惑星] | 0.6207 |
| [下位] という [上位] | [卵子] という [生殖細胞] | 0.4504 |
| [下位] は [上位] | [犬] は [動物] | 0.2500 |
| [上位]・[下位] | [眼鏡]・[老眼鏡] | 0.2167 |

定したものである．EDR 概念体系辞書のルートから深さ 4 以上の階層を持つ概念より，1 階層分の距離を持つ上位語下位語ペアを抽出した．ペアに含まれる語は品詞の判定が行われ，名詞と判定されなかった語は除外した．

#### 2.1.2 表現パターン候補の抽出

上位下位関係を表す表現パターンを含む文例を得るため，抽出された上位語下位語ペアが共起する文を新聞記事コーパスより抽出した．抽出された文は，形態素解析器茶筌により形態素解析を行い，タグ付けを行った．上位語下位語にあたる名詞には，それぞれ [上位], [下位] タグを，それ以外の名詞には [名詞] タグ，名詞以外の形態素は，茶筌の品詞 ID をタグとした．

タグ付けを行った文は，最初に出現した [上位] 又は [下位] のタグを持つ形態素を始端，もう片方のタグの二つ後ろの形態素を終端として切り出し，表現パターン候補とした．表現パターン候補は，更に以下のように処理された．[上位][下位] タグ間に 3 以上の [名詞] タグを持つパターンは除かれた．[上位][下位] タグに囲まれない [名詞] タグは除かれた．[上位][下位] タグに囲まれたパターンが同じで，その外の品詞 ID が異なる表現パターン候補が 4 種類以上あれば，それらは [上位][下位] タグを両端に持つ表現パターン候補として一つにまとめられた．出現頻度が 1 以下の表現パターン候補は除かれた．最後に，表現パターンの [上位][下位][名詞] タグに対応する部分については表層表現を削除しタグだけを残した．以上の処理により，57 の表現パターンが抽出された．得られた表現パターンの一部を表 1 に示す．

### 2.2 表現パターンの評価と選択

自動抽出された表現パターンには，様々な性能の表現パターンが混在している．例えば“ [下位] のような [上位] ”という表現パターンは，“ [犬] のような [動物] ”等の表現に用いられ，精度の高い例となっている．“ [上位] の [下位] ”などの表現パターンは，“ [香水] の [オーデコロン] ”等の上位下位関係も表すが，それ以外にも“ [少女] の [帽子] ”といった所有関係など，様々な関係を表すため，精度が低いパターンとなっている．表現パターンによる上位下位関係の判定を質問応答システムに適用するには，性能の良い表現パターンの集合を作成する必要がある．表現パターンの性能を評価する開発用テストセットを作成し，選択の指標とした．

#### 2.2.1 開発用テストセットの作成

NTCIR-3 の QAC1 テストコレクション [6] を用いて表現パターンの開発用テストセットを作成した．

まず質問セットとして QAC1 ドライランおよびリファレンスランの全 1013 問から質問焦点を含み名称を問う 527 問を選択し，各質問文から質問焦点を人手により抽出した．質問応答システムにより，質問セットの各質問に対する回答候補をそれぞれ 20 ずつ取得した．使用した質問応答システムは文献 [2] によるもので，1 節で述べた文書集合をそのまま知識源とし質問の質問焦点と回答候補の上位下位関係を判定する機能を持つが，候補抽出の際にはこの機能は使わなかった．(以後，この上位下位関係判定機能を用いない質問応答システムをベースライン QA システム，上位下位関係判定機能を用いたシステムをベースライン+上位下位関係判定機能 QA システムと呼ぶ.）質問焦点を上位語，質問の正解と回答候補の和集合の各要素を下位語候補として，上位語と下位語候補のペアを作成した．各ペアについて質問焦点と正解のペアを正例とし，それ以外のペアを負例とした．これにより，986 の正例と 8286 の負例を持つ 9272 ペアのテストセットが作成された[3]．

作成した開発用テストセットの正例は質問応答の正解だけを含む．すなわち負例には質問応答の正解ではないが上位下位関係にあるペアも含まれる．この意味でテストセットは粗い近似である．しかし，開発用テストセット作成における上位下位関係の人手判定コストを避けることを考え，また多くの誤りは含まないであろうとの予測から，今回はこの手法を採用した．

#### 2.2.2 表現パターンの選択

表現パターンの性能を開発用テストセットにより評価し選択の基準とした．開発用テストセットのペアそれぞれについて，コーパスからペアが共起する文集合

[3]回答候補数が 20 に満たない問題が存在するため，総ペア数は質問数 (527) の 20 倍より少なくなっている．

表 2: 評価テーブル

| | 判定結果 | |
|---|---|---|
| | true | false |
| 正例 | CA | CD |
| 負例 | IA | ID |

を抽出した．文集合の中にパターンにマッチする[4]部分文字列を含む文が一つ以上存在する場合，ペアは上位下位関係であると判定，どの文ともマッチしない場合，上位下位関係でないと判定した．結果は表現パターンごとに，表 2 のように，１．正例を上位下位関係と判定した (CA)　２．正例を上位下位関係で無いと判定した (CD)　３．負例を上位下位関係と判定した (IA)　４．負例を上位下位関係でないと判定した (ID)　の４つに分けて頻度を計算し，パターン選択の指標とした．

本稿では，表 2 の値を用いて計算した精度 $\frac{CA}{CA+IA}$ を選択基準とし，表現パターンの選択を行った．選択基準の閾値は 0.0 から 0.5 まで 0.1 刻みで変化させ，閾値以上の精度を持つ表現パターンを使用した．パターンとその精度の例を表 1 に示す．

# 3　評価実験

獲得した表現パターンの性能を調べるため，評価実験を行った．

## 3.1　上位下位関係判定の評価

NTCIR QAC Task1 の formalrun テストコレクションを用いて評価用テストセットを作成した．200 問の質問中，質問焦点を持ち名称を問う 121 問を用い，ベースライン QA システムにより各質問に対してそれぞれ 20 の回答候補を下位語候補として取得した．下位語候補は各質問から人手で抽出した質問焦点とのペアを作り，それが上位下位関係であるかを人手で判定し，上位下位関係であれば正例，それ以外を負例とした．この操作は各質問に対し，回答候補のスコア順に 5 の負例が出るまで繰り返された．また，各質問の質問焦点と正解のペアを正例として加えた．これにより，正例 963 組と負例 581 組を含む，1544 組の評価用テストセットが作成された．[5]

[4]パターン中の [上位][下位] タグは，それぞれペア中の対応する語とマッチする必要があることに注意

[5]解答の不正解ペアは，質問数 (121) の 5 倍に満たない．これは回答候補中に 5 以上の不正解ペアが存在しない質問があったためである．

表 3: 精度により選択した表現パターンの集合による上位下位関係判定の性能

| - | 閾値 | パターン数 | 精度 | 再現率 |
|---|---|---|---|---|
| 人手作成パターン | - | 10 | 0.9057 | 0.2492 |
| 自動獲得パターン | 0.5 | 7 | 0.9444 | 0.0353 |
| | 0.4 | 10 | 0.9200 | 0.0955 |
| | 0.3 | 15 | 0.9113 | 0.2669 |
| | 0.2 | 24 | 0.8297 | 0.3593 |
| | 0.1 | 26 | 0.7722 | 0.4611 |
| | 0.0 | 57 | 0.7728 | 0.4663 |

この評価用テストセットを用いて，2.2 節で選択した表現パターンの集合を評価した．各ペアについて，表現パターン集合中のいずれか一つの表現パターンがコーパスに現れた場合に，上位下位関係にあると判定した．判定結果は，表現パタンの集合ごとに表 2 の評価テーブルとして得られる．この表から精度と再現率を計算し，評価指標とした．本稿では，精度，再現率を次式のように定義した．

$$\text{精度 } (P) = \frac{CA}{CA + IA} \tag{1}$$

$$\text{再現率（}R\text{）} = \frac{CA}{CA + CD} \tag{2}$$

比較対象として文献 [2] で手作業で作成された表現パターン集合を用いた．評価結果を表 3 に示した．閾値 0.3 の結果では，精度と再現率の両方で人手作成した表現パターンの集合を上回った．

## 3.2　質問応答システムによる評価

自動獲得した表現パターンをベースライン+上位下位関係判定機能 QA システムに適用し，質問応答の性能を調べた．評価は，NTCIR QAC1 の formalrun テストセット全 200 問から質問焦点を持ち名称を問う質問 121 問が用いられた．121 問の質問セットは，全部で 181 の正解を持つ．質問応答システムは各質問ごとに 5 つずつ，全部で 605 の回答候補を出力し，その性能を評価した．比較対象としてベースライン QA システム，および人手で作成した表現パターン集合をベースライン+上位下位関係判定機能 QA システムに適応したものを用いた．人手で作成した表現パターンはその信頼度に応じてスコアが与えられている．一方，自動獲得された表現パターンは全て同じ値のスコアを与えた．結果

表 4: 精度を閾値とする表現パターン集合による QA の性能

| 表現パターン | 閾値 | 正解数 | 改善数 | 改悪数 | MRR |
|---|---|---|---|---|---|
| baseline | - | 89 | - | - | 0.431 |
| 人手作成 | - | 92 | 25 | 9 | 0.524 |
| 自動獲得 | 0.5 | 89 | 5 | 2 | 0.431 |
| | 0.4 | 89 | 6 | 4 | 0.453 |
| | 0.3 | 95 | 21 | 12 | 0.482 |
| | 0.2 | 92 | 24 | 13 | 0.477 |

表 5: 人手作成パターンに自動獲得パターンを追加した結果

| 表現パターン | 閾値 | 正解数 | 改善数 | 改悪数 | MRR |
|---|---|---|---|---|---|
| baseline | - | 89 | - | - | 0.431 |
| 人手作成 | - | 92 | 25 | 9 | 0.524 |
| 自動獲得 | 0.5 | 92 | 26 | 10 | 0.528 |
| | 0.4 | 92 | 26 | 9 | 0.530 |
| | 0.3 | 90 | 24 | 12 | 0.496 |
| | 0.2 | 90 | 24 | 14 | 0.477 |

を表 4 に示した.

列“ 正解数 ”は，回答候補中に含まれる正解の総数を，列“ MRR ”は，QAC の Task1 の評価指標である Mean Reciprocal Rank の結果を表す．また，改善数，改悪数は，ベースラインと比較して MRR が改善した問題数，改悪した問題数を表す.

自動獲得した表現パターン集合を用いた結果では，ベースラインと比較して閾値 0.4 以下で性能の上昇が確認された．また，閾値 0.3 は，正解数が最も多いが，MRR 値では人手作成したパターンを適用した QA システムより低くなっている.

次に，人手で作成した表現パターンと自動獲得した表現パターンを組み合わせて用いた場合の結果を表 5 に示す．この評価では，自動獲得した表現パターンには人手で与えたスコアの最小値が与えられた．閾値 0.4 以上にて人手作成パターンを用いた QA システムの MRR 値を改善した.

# 4 考察

自動獲得したパタンは，上位下位関係判定において人手作成のパターンの性能を上回った．一方，質問応答システムに単独で適用した場合，ベースラインシステムの性能は改善したが，人手によるパターンを用いた場合には及ばなかった．この原因として，人手作成表現パターンが信頼度に従いスコアが与えられているのに対し，自動獲得表現パターンは全て同じスコアであったことが考えられる．今後，自動獲得した表現パターンにその信頼度に応じたスコアを与えることで，性能が改善できると考える．また，現在は各パターンを個別に用いているが，複数のパターンにあてはまった回答候補はより信頼度が高いと考えられる．複数のパターンに与えられたスコアを組み合わせる手法について検討したい.

# 5 まとめ

上位下位関係語を判定する表現パターンを自動獲得する手法を述べた．NTCIR QAC1 テストコレクションを用いた評価実験により，適切なパターンの獲得と，QA への適用による性能向上を確認した．本手法は上位下位関係だけでなく，同義語関係，類義語関係など様々な関係を判定する表現パターンの獲得に応用できる．語や句の間の様々な関係を表す表現パターンを獲得し，質問応答の正解判定に用いる予定である.

# 参考文献


[1] EDR 電子化辞書 http://www.iijnet.or.jp/edr/

[2] 秋葉 友良, 伊藤 克亘, 藤井 敦. 質問応答における常識的な解の選択と期待効用に基づく回答群の決定. 情報処理学会研究報告, 2004-NL-163, pp.131-138, Sep. 2004.

[3] Marti Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora. In proceedings of the Fourteenth International Conference on Computational Linguistics, Nantes, Flance,1992

[4] Michael Fleischman, Eduard Hovy, Abdessamad Echihabi. Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked, In proceedings of the 41st Annual Meeting of the Association for Conputational Linguistics, pp. 1-7, July 2003

[5] 安藤 まや, 関根 聡. 上位語・下位語を含む連体修飾表現の言語的分析. 言語処理学会, 第 10 年次大会, pp. 205-208, 2004

[6] J.Fukumoto, T.Kato, F.Masui. Question Answering Challange (QAC-1) Question answering evaluation at NTCIR Workshop Meetion Part IV, QAC, National Institute of Informatics, pp.1-10, 2002