

意味的知識を用いた Why 型質問応答の改善

呉 鍾勲 鳥澤 健太郎 橋本 力 川田 拓也 De Saeger, Stijn 風間 淳一 王 軼謨
情報通信研究機構 情報分析研究室

1 はじめに

質問応答の研究において、factoid 型質問に対する質問応答技術は広く研究されてきたが、Why 型質問、How-to 型質問を含む non-factoid 型質問に対する質問応答技術への研究は比較的少ない。また、最高レベルの non-factoid 型質問応答システム [4, 5, 6] の精度 (例えば、Why 型質問に対する上位 150 個の結果で 34% の MRR [6]) は最高レベルの factoid 型質問応答システムの精度 (上位 1 個の結果で 85% の精度 [1]) よりも非常に低いというのが実状である。

本稿では、このように困難なタスクであると認識されてきた non-factoid 型質問応答、特に Why 型質問応答の精度を向上するための手法を提案する。本研究は、「否定的な事象の理由は否定的な事象であることが多い」、そして「肯定的な事象の理由は肯定的な事象であることが多い」という意味的極性に関わるパターンが Why 型質問とその回答によく現れるという観察を出発点とし、このような意味的極性のパターンを機械学習で学習し Why 型質問回答の性能改善を試みる。より具体的には、例えば、以下のように「なぜガンになりますか?」という「否定的な事象」の理由を求める質問 Q1 に対して、ガンリスクを高めるという「否定的な事象」を説明する A1-1 と、ガンを予防するための「望ましいこと」を説明する A1-2 が回答候補として出てきたとするならば、ここで、「否定的な事象の原因は否定的な事象であることが多い」という意味的極性のパターンから A1-1 を Q1 の正しい回答として選べる。

Q1: なぜガンになりますか?

A1-1: ニトロソアミンなどの発がん因は細胞のもつ遺伝子を変化させ、ガンの リスクを高める。

A1-2: 健康的な体重を維持することはガンの リスクを下げる。

こうした意味的極性のパターンは non-factoid 型質問応答において、我々の知る限りこれまでに検討されることがない。また、本研究のもう一つの基本的アイデアは、Why 型質問が含む単語とその回答が含む単語間の意味的な相関関係を用いて性能の向上を図る

ということである。例えば、Q1 のように「病気」の原因を求める質問の回答は「有害物質」(A1-1 のように)、「ウィルス」、「身体の部位」などを表す単語を含む場合が多い。質問と回答からこのような「病気」と「有害物質」間の相関関係を把握し、質問応答における回答抽出に応用することによって Why 型質問応答の性能向上が期待できる。このため、提案手法では単語クラスタリング手法 [3] を用いて大規模なウェブ文書から単語の意味的クラス (意味的に類似する単語の集合) を自動獲得し、機械学習の素性として活用する。

Why 型質問と回答における意味的極性のパターンを Why 型質問応答に適用する場合、意味的極性を持つ言語表現、すなわち評価表現の内容も考慮する必要がある。これは、回答に意味的極性が異なる複数の表現が存在すると (本研究の評価データでは約 33% の正しい回答がこのような特徴を持っていた。)、意味的極性のみでは Why 型質問応答の性能改善が期待できないためである。例えば、A1-2 が以下のように肯定的な表現と否定的な表現を持つと仮定すると、Q1、A1-1、A1-2 が同様に否定的な表現を持つことになり、これらの質問と回答における意味的な極性のパターンの効果が期待できない。

“がんの良いとされる食品を食べ過ぎるのはガンの予防に 効果的ではない が、健康的な体重を維持することはガンの リスクを下げる”

このような問題を解決するため、提案手法では Why 型質問と回答候補における評価表現の極性と共に評価表現の内容 (評価表現を構成している単語、係り受け関係など) も考慮した。そして、この際のデータの過疎性を回避するため、評価表現の内容を単語と共に単語の意味的クラスで表現した。

最終的に以上のようなアイデアは回答候補のランキングを行う教師あり学習で使われた。人手で作成した 362 個の Why 型質問と Web から抽出されたその回答候補 (各 20 個からなる) で評価実験を行った結果、提案手法による約 11% の性能改善を確認した。

なお、non-factoid 型質問応答において、構文情報 (係り受け関係など)、意味情報 (WordNet 情報など)、統計情報 (頻度など) などを素性として学習した分類

器を回答のリランキングに用いた先行研究 [2, 6, 5] があったが、提案手法は、単語クラスと意味的極性という意味的知識を用いた新たな素性を提案し、Why 型質問応答におけるその有効性を示した点が先行研究と大きく異なる。

2 提案手法

提案手法は以下に述べる**回答検索**と**回答のリランキング**の2ステップからなる。本研究の目的は意味的知識による**回答のリランキング**の性能向上である。

回答検索: NTCIR-6 の non-factoid 型質問応答のタスクにおいて、最高性能を示した Murata ら [4] を実装して回答検索を行った。まず、Why 型質問の単語を情報検索ツール Solr¹の入力として与え、6 億件のウェブ文書から上位 300 個の文書を検索する。そして、検索結果の文書から接続の回答候補が2文を共有するように一連の5文を回答候補として抽出した。これは、回答候補の抽出の誤りによって正しい回答が取れない可能性を防ぐためである。質問 q に対して抽出された回答候補 ac は式 (1) によってランク付けられ、上位 20 個の回答候補を次の段階である回答のリランキングの入力として与える。また、Murata ら [4] と同様に Why 型質問応答の手がかりとなる単語（「理由」、「原因」、「要因」）を質問の単語として追加した。

$$S(q, ac) = \max_{t_1 \in T} \sum_{t_2 \in T} \phi \times \log(ts(t_1, t_2)) \quad (1)$$

$$ts(t_1, t_2) = \frac{N}{2 \times \text{dist}(t_1, t_2) \times df(t_2)}$$

ここで、 T は q と ac の共に現れる単語（名詞、動詞、形容詞を含む）の集合を表す。 N は文書の数（6 億）、 $\text{dist}(t_1, t_2)$ は ac においての t_1 と t_2 間の距離、 $df(t)$ は t が現れる文書の頻度、 $\phi \in \{0, 1\}$ は $ts(t_1, t_2) > 1$ であるか否かを示す indicator である。（1 if $ts(t_1, t_2) > 1$, 0 otherwise）。

回答のリランキング: このプロセスは教師あり学習によって作られた分類器（本稿では SVMs）のスコアによって行われる。実験では、362 個の Why 型質問に対する上位 20 個の回答候補を手でチェックしたデータを基にした 10-fold cross validation 方法で提案手法の有効性を検証した。

3 回答のリランキングのための素性

本節では回答のリランキングのための素性を説明する。具体的には、質問と回答候補のテキストおよび表

現を以下のように抽出変換した上で、SVMs の教師あり学習の素性として使用する。

1. 質問と回答候補のテキストに対して JUMAN による形態素解析と KNP による構文解析を行い、解析結果から形態素、文節、係り受けの n -gram を抽出する。そして、これらを基にして素性を作成する（表 1 の MSA1～MSA4）。これらの素性は既存研究においても良く用いられているものである [5, 6]。
2. 質問と回答候補の形態素、文節、係り受けの n -gram に現れる単語を単語のクラスに変換し、変換した n -gram のうち単語クラスを含むものののみを取り出す。これらを単語クラス n -gram と呼び、Why 型質問が持つ単語とその回答が持つ単語間の意味的な相関関係を示すための素性作成に用いる（表 1 の SWC1 と SWC2）。単語クラスは、6 億件のウェブ文書から取り出した名詞間の係り受け関係、名詞と動詞間の係り受け関係を名詞の文脈情報として用いて、類似する文脈を持つ名詞をクラスタリング [3] することにより獲得された。名詞 n の単語クラスは $c = \text{argmax}_{c^*} p(c^*|n)$ により判定され、合計 550 万名詞に対する 500 個の単語集合を単語クラスとして使用した。
3. 質問と回答候補の形態素、文節、係り受けの n -gram に現れる単語を単語の極性辞書（「意見（評価表現）抽出ツール」²の辞書を使用）によって極性（ポジティブとネガティブ）に変換し、変換した n -gram のうち単語の極性を含むものののみを取り出す。これらを単語極性 n -gram と呼び、表 1 の SA@W1 と SA@W2 の作成に用いる。そして、単語クラスと単語極性の組み合わせで同様に形態素、文節、係り受けの n -gram を変換し、変換した n -gram のうち単語クラスと単語極性の組み合わせを含むものののみを取り出す。これらを単語クラス／極性 n -gram と呼び、表 1 の SA@W3 と SA@W4 を作成するために使う。
4. 「意見（評価表現）抽出ツール」を使い、質問と回答候補のテキストから極性を持つ評価表現を抽出し、これらの評価表現から形態素、文節、係り受けの n -gram、単語クラス n -gram、単語クラス／極性 n -gram を取り出す。そして、取り出した n -gram と評価表現の極性を合わせを基にして表 1 の SA@P1～SA@P10 を作成する。

特に、3 と 4 の極性は性能向上において重要であるが、論文冒頭で述べたように、これは多くの場合「良いと

¹ <http://lucene.apache.org/solr>

² <http://alaginrc.nict.go.jp/opinion/index.html>

MSA1	質問や回答候補のテキストに現れる形態素、文節、係り受けの n -gram。質問からと回答候補からの n -gram は区別される。
MSA2	回答候補から取り出した MSA1 の n -gram のうち、質問の単語を含むもの。
MSA3	MSA1 の n -gram のうち、手がかりとなる単語（理由、原因、要因）を含むもの。質問からと回答候補からの n -gram は区別される。
MSA4	質問の単語のうち、回答候補に現れたものの比率。
SWC1	MSA1 の n -gram に現れる単語を単語クラスに置き換えた n -gram のうち、単語クラスを含むもの。これらを単語クラス n -gram と呼ぶ。質問からの単語クラス n -gram と回答候補からの単語クラス n -gram は区別される。
SWC2	回答候補からの単語クラス n -gram のうち、単語クラスの元になる単語が質問の単語である n -gram。
SA@W1	MSA1 の n -gram に現れる単語を単語の極性辞書（意見抽出ツールの辞書を使用）によって極性（ポジティブとネガティブ）に置き換えた n -gram のうち、置き換えた極性を持つもの。これらを単語極性 n -gram と呼ぶ。質問からの単語極性 n -gram と回答候補からの単語極性 n -gram は区別される。
SA@W2	回答候補からの単語極性 n -gram のうち、極性の元になる単語が質問の単語である n -gram。
SA@W3	MSA1 の n -gram に現れる単語を単語クラスと単語の極性の組み合わせに置き換えた n -gram のうち、置き換えた組み合わせを持つもの。これらを単語クラス／極性 n -gram と呼ぶ。質問からの単語クラス／極性 n -gram と回答候補からの単語クラス／極性 n -gram は区別される。
SA@W4	回答候補からの単語クラス／極性 n -gram のうち、単語クラス／極性の元になる単語が質問の単語である n -gram。
SA@P1	質問の評価表現の極性と回答候補の評価表現の極性が一致するか否かを示す指示変数。一致する対があると 1 を持つ。評価表現は意見（評価表現）抽出ツールを用いて抽出する。
SA@P2	SA@P1 が 1 になった際の極性、ポジティブとネガティブ。
SA@P3	評価表現に現れる形態素 n -gram、文節 n -gram、係り受け n -gram と評価表現が持つ極性の組み合わせ。質問の評価表現からの n -gram と回答候補の評価表現からの n -gram は区別される。
SA@P4	回答候補の評価表現から取り出した SA@P3 の n -gram のうち、質問の単語を含むもの。
SA@P5	質問の単語のうち、回答候補の評価表現を含む文に現れたものの比率。
SA@P6	評価表現の単語クラス n -gram と評価表現が持つ極性の組み合わせ。質問の評価表現からと回答候補の評価表現からのものは区別される。
SA@P7	回答候補の評価表現からの単語クラス n -gram と評価表現が持つ極性の組み合わせのうち、単語クラスの元になる単語が質問の単語であるもの。
SA@P8	評価表現の単語クラス／極性 n -gram と評価表現が持つ極性の組み合わせ。質問の評価表現からの単語クラス／極性 n -gram と回答候補の評価表現からの単語クラス／極性 n -gram は区別される。
SA@P9	回答候補の評価表現からの単語クラス／極性 n -gram と評価表現が持つ極性の組み合わせのうち、単語クラス／極性の元になる単語が質問の単語である n -gram。
SA@P10	質問からの SA@P6 の n -gram と回答候補からの SA@P6 の n -gram の組み合わせ。それぞれの n -gram の元になる評価表現の極性が一致する場合のみ（ SA@P1 の指示変数が 1 である評価表現間のみ）、 n -gram と評価表現の極性が素性として用いられる。

表 1: 教師あり学習に用いられた素性。 n -gram の n は $n \in \{1, 2, 3\}$ 。

される出来事の原因は良いとされる事象である」「悪いとされる出来事の原因は悪いとされる事象である」という傾向があるからである。なお、こうした評価極性の利用は 2 の単語クラスによるデータの過疎性の回避があつて、より有効になる。表 1 は提案手法の回答のリランキングのために用いられた素性を示している。

4 実験

評価実験では人手で作成した評価データを用いて提案手法の評価を行った。評価データは質問作成と回答候補の判定の二段階で作成された。

質問作成の目的は、対象文書に正しい回答があると保証される多様な Why 型質問を作成することであつた。当然ながら、このような設定には「実世界のユーザは検索対象になる文書に求める回答があるか否かを気にせず、自分が知りたい事象の理由や原因に関わる Why 型質問をする。」という現実とはズレがあり、このようなデータを用いた質問応答システムの評価結果は実世界におけるユーザの質問に対する評価結果の上

限値と考えられる。

質問作成のため、連続する 3 文で構成されたパッセージからある事象の理由や原因を説明する部分を探し、その部分が回答になる Why 型質問を著者以外のアノテータが人手で作成した。しかし、もととなるパッセージを対象文書からランダムに選択すると、そこにある事象の理由や原因を表す説明が含まれる可能性は非常に低いため、パッセージは「理由」、「原因」、「要因」という手がかりとなる単語を含む連続 3 文に限定した。その結果、手がかりとなる単語を含む 12,000 個のパッセージからは 362 個の Why 型質問が作成できた。このように作られた質問の内容はパッセージ抽出の手がかりとなった「理由」、「原因」、「要因」に偏りがあると考えられるが、上述した質問作成の目的のための実用的な妥協点と言える。

次に、作成された 362 個の質問を入力とした回答検索の上位 20 個の結果を 3 名のアノテータ（著者以外）が判定した。与えられた回答候補が質問の正しい回答か否かについて 3 名が判定を行い、3 名の判定結果における多数決によって最終判定結果を得た。3 名の判

定結果は相当な一致率 (Fleiss の kappa 値で 0.611) を示した。判定結果を見ると 223 個の質問 (362 個の 61.6%) に対する上位 20 個の結果に正しい回答が含まれており、これらの質問に対する正しい回答は平均 4.1 個であった。

4.1 実験結果

実験では 10-fold cross validation 方法で提案手法の有効性を検証した。そして、上位 1 個の回答の精度を示すための P@1 と上位 n 個の回答候補における全体的な精度を示すための MAP (Mean Average Precision) で性能評価を行った。提案手法の回答候補のリランキングには TinySVM の線形カーネルで学習した SVMs を用いた。

表 2 は実験結果を示している。B-QA は提案手法の回答検索のみの結果を、B-Ranker は表 1 の MSA1~MSA4 のみで学習した SVMs をランク付けに用いた結果を、Proposed は提案手法の評価結果を示している。UpperBound は回答検索結果に正しい n 個の回答がある場合、これらをいつも上位 n にランク付けするシステムの評価結果を示し、この実験の上限値を表す。この上限値に対する B-QA、B-Rank、Proposed の相対 P@1 値と相対 MAP 値を括弧で示した。また、回答検索結果に質問作成の基となったパッセージを回答候補に追加し、これらを Proposed に用いられた SVMs によってランク付けした結果を Retrieval-Oracle に示した。この結果は回答検索結果にいつも一つ以上の正しい回答が含まれると提案手法が高精度で回答を出せるという可能性を示している。

System	P@1	MAP
B-QA	0.254 (0.412)	0.288 (0.467)
B-Ranker	0.306 (0.497)	0.341 (0.553)
Proposed	0.370 (0.600)	0.370 (0.600)
UpperBound	0.616 (1)	0.616 (1)
Retrieval-Oracle	0.702	0.712

表 2: 実験結果

B-QA と B-Ranker に比べ、提案手法は最も高い性能を示している。B-QA との 11.6% の性能差 (P@1) は回答検索に対する性能向上を、B-Ranker との 6.4% の性能差 (P@1) は本稿で提案した素性の有効性を示している。なお、スペースがないために詳細は省くが、意味的極性に関わる素性 (表 1 の SA@W と SA@P)、もしくは単語クラスに関わる素性 (表 1 の SWC) を取り除いた場合、いずれも性能低下 (P@1 で 1.7%~5% の性能低下) を示し、これら全部を取り除いた場合 (表 2 の B-Ranker) は P@1 で 6.4% の性能低下がある

ことを確認した。これらの結果は、意味的極性、単語クラス各々が独立に性能向上に貢献しているが、両者の組み合わせがより有効であることを示している。

5 まとめ

本稿では、「否定的な事象の原因は否定的な事象であることが多い」、そして「肯定的な事象の原因は肯定的な事象であることが多い」という意味的極性に関わるパターンと質問が持つ単語 (例えば「病名」とその回答が持つ単語間 (例えば「有害物質」、「ウィルス」、「体の部位」) の意味的な相関関係を意味的知識として用いて Why 型質問に対する質問応答システムの精度を向上するための手法を提案した。362 個の Why 型質問に対する評価実験により、提案手法による約 11% の性能改善を確認し、意味的知識の有効性を示した。

参考文献

- [1] D. A. Ferrucci, E. W. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. M. Prager, N. Schlaefer, and C. A. Welty. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, Vol. 31, No. 3, pp. 59–79, 2010.
- [2] R. Higashinaka and H. Isozaki. Corpus-based Question Answering for Why-questions. In *Proc. of IJCNLP*, pp. 418–425, 2008.
- [3] J. Kazama and K. Torisawa. Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations. In *Proc. of ACL-08: HLT*, pp. 407–415, 2008.
- [4] M. Murata, S. Tsukawaki, T. Kanamaru, Q. Ma, and H. Isahara. A System for Answering Non-Factoid Japanese Questions by Using Passage Retrieval Weighted Based on Type of Answer. In *Proc. of NTCIR-6*, 2007.
- [5] M. Surdeanu, M. Ciaramita, and H. Zaragoza. Learning to Rank Answers to Non-Factoid Questions from Web Collections. *Computational Linguistics*, Vol. 37, No. 2, pp. 351–383, 2011.
- [6] S. Verberne, L. Boves, N. Oostdijk, and P.-A. Coppen. What is not in the Bag of Words for Why-QA? *Computational Linguistics*, Vol. 36, No. 2, pp. 229–245, 2010.