

## 文章構造を用いた Why 型質問応答システム

## Why-question answering using document structure

車 智修\*1 鍋島 啓太\*2 水野 淳太\*3 岡崎 直観\*2\*4 乾 健太郎\*2  
 Jisoo Cha Keita Nabeshima Junta Mizuno Naoaki Okazaki Kentaro Inui

\*1 東北大学工学部情報知能システム総合学科

Department of Information and Intelligent Systems. School of Engineering. TOHOKU University

\*2 東北大学 大学院情報科学研究科

Graduate School of Information Sciences. TOHOKU University

\*3 情報通信研究機構

National Institute of Information and Communications Technology (NICT)

\*4 科学技術振興機構 さきがけ

Japan Science and Technology Agency (JST)

This paper presets an approach to answer why questions by using document structure. Unlike the factoid questions, why questions require elaborated sentences rather than a few words. Therefore, we choose core sentences that are relevant to a why-question and sentences related to the core sentences. We formalize this task as a personalized PageRank, where nodes represent sentences with seed weights and edges represent similarity between sentences. The experiments demonstrate the superiority of the proposed method on *Yahoo! Chiebukuro* Q&A pairs.

## 1. はじめに

ウェブ上の膨大な文書から、必要な情報のみを適切に抽出する技術は重要である。質問応答 (QA) システムは、自然文で入力した質問に対して、適切な回答部分だけを抽出する技術であり、Answers.com\*1 といった実用化が進められている他、IBM の開発した Watson がクイズ番組で勝利するなど、開発が進められてきた。

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 QA は、Factoid 型と呼ばれ、TREC[Dang 07] や NTCIR[Kato 05] といった評価型ワークショップを通じて研究が進められてきた。一方で、方法や理由を問い主に文章で回答する Non-factoid 型については、質問回答コミュニティへの投稿データをコーパスとして、入力された質問文とコーパス中の質問文の対応を取り、その回答を出力するアプローチが研究されてきた。本研究でも、回答データには質問回答コミュニティの一つである Yahoo! 知恵袋\*2 を用いる。本研究は理由を問う Why 型質問について、回答部分の周辺文の情報に着目するために Personalized Pagerank を用いて回答候補をランキングする手法を提案する。

Why 型質問応答システムに関して、様々な手法が提案されてきた。渋沢らは、「から、原因」といった原因や理由を表す表現と、「それ、以下」といった指示表現を手がかりとして、人手で作成したパターン辞書を元に理由文を抽出する手法を提案した [渋沢 07]。石下らは、質問のタイプに依存しない質問応答システムを提案し、タイプごとの質問と回答の特徴を Yahoo! 知恵袋から学習し、回答の特徴と質問文との関連度を用いて Web 文書から回答部分を抽出した [石下 09]。呉らは「ネガティブ (ポジティブ) な事象の理由はネガティブ (ポジティブ) な事象である事が多い」という意味的極性の傾向と、質問と回答の単語間の意味的關係を機械学習による回答抽出に用いることで、Why 型質問応答システムの性能を向上させた [呉 13]。本研究では、質問文と回答文との関係だけでなく、回答文とその周辺

文との関係を利用することにより、回答抽出の性能改善を図る。

## 2. 提案手法

提案手法は図 1 に示すように、回答候補検索と、回答候補のランキングの 2 つのステップからなる。まず 1 文からなる Why 型の質問を入力として与える。回答候補検索では、与えられた質問文に対して、コーパスから回答候補文書を検索する。次に、検索された文書から各文を回答候補として抽出して、ランキングを行う。最終的にランキングされた回答候補のリストを出力する。本研究の主眼は、ランキングを行う際に文書構造を用いて、その性能を向上することにある。

**回答候補検索:** 回答候補検索を行うために、まず入力で与えられる Why 型質問文に対し形態素解析を行い、質問文に含まれる内容語 (名詞, 動詞, 形容詞) を文書検索のクエリとして、コーパスから検索を行う。検索対象となるコーパスには、2004 年 4 月 1 日から 2009 年 4 月 7 日の間に「Yahoo! 知恵袋\*3」に投稿された回答文書約 50 万件を用いた。検索して得られた回答文書に、文書のうちクエリの内容語をすべてを含む 1 文が存在しない場合は、関連性が低いと考え回答文書から除く。得られた回答文書の各文を、回答候補とする。

**Personalized PageRank** による回答候補のランキング:

Why 型質問の回答として適切な説明文は、周辺文で補足して説明されていたり、回答候補文書中で重要な文となっていることが多い。そこで本研究では、文書の重要文抽出の手法としてよく用いられる Personalized PageRank[Jeh 03] を使い、各回答候補文をノードとした回答候補のランキングを行う。Personalized PageRank とは、PageRank の Random Surfer Model に優先すべきノードにジャンプするテレポーターション確率を導入したものである。ノード  $X$  の PageRank スコア  $P(X)$  は、以下の式 1 を反復的に処理することにより求められる。

$$P(X) = (1 - d) * V(X) + d * \sum_{Y \in G} w(Y, X) * P(Y) \quad (1)$$

連絡先: 車 智修, 東北大学, 宮城県仙台市青葉区荒巻字青葉  
 6-3-09, jisoo.cha419@gmail.com

\*1 <http://wiki.answers.com/>

\*2 <http://chiebukuro.yahoo.co.jp/>

\*3 <http://chiebukuro.yahoo.co.jp/>



図 1: システムの構造

$G$  はノードの集合,  $V(X)$  はノードの重み,  $w(Y, X)$  は  $Y$  から  $X$  方向のエッジの重みを表す. ここで  $d$  はテレポーション確率をあらわすパラメータである. 実験では最も性能が向上した値  $d=0.05$  を用いた.

ノードの重みとしては, 質問文との類似度を用いた. 2 文間  $X, Y$  の類似度の測定には以下の式を用いた.

$$\text{sim}(X, Y) = \frac{1}{|X|} \sum_{x \in X} \max\{\text{dice}(x, y), y \in Y\} \quad (2)$$

$x, y$  は文  $X, Y$  の内容語 (名詞, 動詞, 形容詞),  $\text{dice}(x, y)$  は 2 単語間  $x, y$  の dice 係数を表す. dice 係数は, 高度言語情報融合フォーラム (ALAGIN)\*4 で公開されている単語共起頻度データベース (Version 1.1) を利用した. 算出方法は以下のとおりである.

$$\text{dice}(x, y) = \begin{cases} 1 & (x = y) \\ \frac{2 * \text{cnt}(x, y)}{\text{cnt}(x) + \text{cnt}(y)} & (\text{otherwise}) \end{cases} \quad (3)$$

ここで  $\text{cnt}(x)$  は ALAGIN コーパスの文書集中での単語  $x$  の頻度,  $\text{cnt}(x, y)$  は単語  $x, y$  が 4 単語以内で共起した回数を表す.

次に同じ回答候補文書内の文同士でエッジを貼る. エッジの重みとしては, 文同士の類似度を用いて式 2 のスコアで求めた.

### 3. 実験

#### 3.1 実験設定

Why 型質問応答における文章構造の有効性を測るため, 評価実験を行った. この実験の目的は本論文の提案手法で抽出した回答文の中に正解回答が上位に含まれているかの評価である. Personalized PageRank の効果を調べるため, 比較対象として通常の PageRank を用いた場合と, 文章構造を使わない, つまり式 1 のパラメータ  $d$  の値を 0 にした場合の 2 つと比較した. また既存手法との比較として, 渋沢らの [渋沢 07] 手法を用いた. この手法は「ので, 原因」などの特徴語を手がかりとし, 各文にスコアリングを行う.

正解データの作成にはまず, コーパス中に正解の回答文が存在し, かつ回答候補が 10 文以上得られる 5 個の Why 型の質問を手で用意した. 次に各質問文のクエリによりコーパスから抽出した回答候補文書の各文を回答候補とし, 各文が質問に対する理由の説明になっているかどうかを判定し, 理由になっているものを正解データとした.

1 つの質問あたりの平均回答候補数は約 133 件, 平均正解数は約 9 件であった. 評価方法として, 上位  $N$  個の全体的な精度を表すための  $P@N$  (Precision at the top- $N$  answer) と, 順位を考慮した  $MRR$  (Mean Reciprocal Rank) で性能を評価した.

#### 3.2 実験結果

実験結果を表 1 に示す. まず,  $d=0.0$  の場合に比べ,  $d=0.05$  の方が性能が向上した. このことから文同士の類似度を組み込むことが有効であると言える. 通常の PageRank を用いた場合

手法	MRR	P@3	P@5	P@7
既存手法 [渋沢 07]	0.232	0.467	0.400	0.300
Personalized PageRank ( $d=0.05$ )	0.254	0.600	0.440	0.360
Personalized PageRank ( $d=0.00$ )	0.171	0.533	0.440	0.340
PageRank	0.147	0.267	0.160	0.100

と比較すると, Personalized PageRank の方が性能が高く, 単に文同士の類似度を入れるだけでは性能は向上しないことが分かる.

また既存手法に比べ, Personalized PageRank を用いた手法の方が性能が向上した. 既存手法と比較し, 提案手法のみが正解していた回答としては, 例えば「ウサギの耳はなぜ長い?」という質問に対し, 「ウサギの耳は, 音を集めるアンテナの役目をしています」のように, 回答候補に「原因, なので」のような特徴語を含んでいないものも上位に含まれていた.

#### 4. まとめ

本研究では, Personalized PageRank を用いて Why 型質問応答システムの性能の向上を行う手法を提案した. 実験の結果, 既存手法より性能が向上し, 回答文書構造を使うことによる Why 型質問応答システムの性能向上の可能性を示した. 今後は, 回答として 1 文を取り出すのではなく, 回答文書から適切な長さの回答を抽出する Why 型質問応答システムを構築する予定である.

#### 謝辞

「Yahoo!知恵袋」データを提供して頂いたヤフー株式会社と国立情報学研究所の皆様へ感謝いたします. 本研究は, 東北大学工学部情報知能システム総合学科「Step-QI スクール」の支援を受けた.

#### 参考文献

- [Dang 07] Dang, H. T., Lin, J., and Kelly, D.: Overview of the TREC 2007 question answering track, in *Proceedings of TREC*, p. 3 (2007)
- [Jeh 03] Jeh, G. and Widom, J.: Scaling personalized web search, in *Proceedings of the 12th international conference on World Wide Web*, pp. 271–279 ACM (2003)
- [Kato 05] Kato, T., Fukumoto, J., and Masui, F.: An overview of NTCIR-5 QAC3, in *Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 361–372 (2005)
- [呉 13] 呉 鍾勲, 鳥澤 健太郎, 橋本 力, 川田 拓也, Saeger, S. D., 風間 淳一, 王 軼諷: 意味的極性と単語クラスを用いた Why 型質問応答の改善, 情報処理学会論文誌, Vol. 54, No. 7 (2013)
- [渋沢 07] 渋沢 潮, 林 貴宏, 尾内 理紀夫: Why 型質問の回答文を WEB から自動抽出するシステムの開発と評価, 情報処理学会論文誌, Vol. 48, No. 3, pp. 1512–1523 (2007)
- [石下 09] 石下 円香, 佐藤 充, 森 辰則: Web 文書を対象とした質問の型に依らない質問応答手法, 人工知能学会論文誌, Vol. 24, No. 4, pp. 339–350 (2009)

\*4 <http://alagin.jp/>