1D-7

# Web 掲示板を利用した Q&A データの自動生成

早川 晃央† 土方 聡子† 前橋 孝弘† 韓 東力†

日本大学文理学部 情報システム解析学科†

## 1. はじめに

インターネットの普及に伴い、Web 上のデータは急激に増加し、現在では様々な情報が不規則な形式で氾濫している。そこで我々は情報抽出の一環として、Web 上にあるテキスト情報を整理し、自動的に質問回答対(Q&A)を生成するシステムを構築しようと考えた。

既存研究には投稿記事があらかじめ質問と回答に分類されている QA サイト利用して質問回答対を作成するものがある[1,2]。

本研究では、質問記事と回答記事が混在している Web 掲示板を用いた質問回答対の作成と整理を試みる。

システムでは Web 掲示板よりユーザが指定した期間の記事から記事データとして[投稿日時]、[投稿者名]と[本文]の 3 要素を収集したのち、記事の分類、質問回答対の作成の順で処理を行い、図 1 に示されるように TreeView で質問回答対を表示する（投稿者 ID は伏せている）。記事の分類、質問回答対の生成・表示についてはそれぞれ第 2 章と第 3 章で述べ、第 4 章以降では評価と結論を述べる。

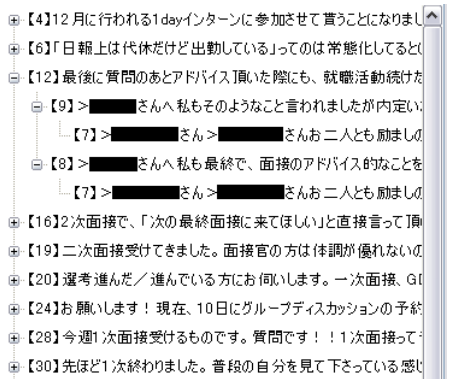


図 1 TreeView による質問回答対の表示

## 2. 記事の分類

質問回答対を生成するためには、収集した記事を質問記事と回答記事に分類する必要がある。投稿記事があらかじめ質問と回答に分類されている QA サイトを利用した既存研究と異なり、本研究では質問と回答の対応関係が明確でない Web掲示板サイトである「みんなの就職活動日記」[1]を研究対象として記事分類を行う。

### 2.1. 分類基準

Web 掲示板には質問と回答が混在している記事や、独り言など意味を成さない記事が大量に存在するため、質問記事と回答記事に明確に分類することが難しい。本研究では 2 通りの分類基準を用いて記事分類を試みる。

・基準 1：質問記事と質問以外の記事で分類
・基準 2：回答記事と回答以外の記事で分類

基準 1 では、返信が存在する質問記事を Q 記事とし、それ以外の記事を!Q と定め、基準 2 では、質問記事との間に明確な回答関係が認められており、かつユーザにとって有意義な情報を含んでいる記事を A 記事とし、それ以外の記事を!A と定めた。以下では基準 1 を「Q&!Q」分類、基準 2 を「A&!A」分類という。

サポートベクタマシーン(SVM) [3]を用いて、2 種類の分類基準に基づき、記事分類を試みる。

### 2.2. 分類実験

記事本文をJuman[2]で形態素解析し、結果として得られた記事本文中に存在する全ての形態素をSVMの素性とする。ただし、同一の形態素が複数存在する場合でも重みは常に 1 とし、存在しない場合は重みを 0 とする。

任意に選定した 7 社の記事データ 1200 件に対して人手でそれぞれ Q と!Q および A と!A を付与したのち、全データを A グループ、B グループ、C グループのそれぞれ 400 件に分ける。そのうちの 2 グループを学習データに、残りの 1 グループをテストデータとして扱い、交差検定法で分類実験を行う。実験結果は表 1 の通りである。

表 1 分類実験の結果

| 分類基準 | 形態素 | 助詞を除いた形態素 |
|---|---|---|
| Q&!Q | 79.18% | 79.18% |
| A&!A | 91.63% | 91.96% |

実験結果によって、Q&!Q 分類より A&!A 分類の方が精度が高いことが確認できた。後続の質問回答対の作成処理をより高い精度で実行する

Automated Generation of Q&A Data from Web Message Board

†Akio Hayakawa, Satoko Hijikata, Takahiro Maehashi, Dongli Han

†Department of Computer Science and System Analysis, the College of Humanities and Sciences, Nihon University

[1] http://www.nikki.ne.jp/

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/ juman.html

ため、実際のシステムでは A&!A 分類を採用している。

## 3. 質問回答対の生成と表示

Web 掲示板から収集した記事を A と!A に分類したのち、A と判定された記事を回答記事として、それに対応する質問記事を探索する。掲示板の性質から異なる A 記事から同一の Q 記事に辿り着く場合もある。

「みんなの就職活動日記」の Web 掲示板では、投稿記事に対して返信記事を書く場合、自動的に“>[名前]さんへ”という宛名表現が本文の先頭に挿入される。[名前]には、返信先の記事の投稿者名が入る。7 社各 600 件、合計 4200 件の記事の中から A と判定された回答記事 2246 件に対して宛名表現の有無を調査したところ、宛名表現が存在する記事が 2224 件と回答記事全体の 99%を占めていることが分かった。そこで我々は、質問記事の探索に宛名表現を利用できると判断した。ただし、この調査では“>[名前]さんへ”以外の表記も宛名表現と見なしている。

宛名表現を利用するにあたり、宛名表現が存在する回答記事 2224 件を確認したところ、“>[名前]さんへ”以外の表記法が 436 件見つかった。“>[名前]さんへ”以外の表記では、投稿者が“[名前]さん>”などと独自に宛名表現を改編している場合や、“>[名前]さん・[名前]さんへ”などと宛名が 2 つ以上存在している場合が見られた。

これら全てのパターンに対応するため、本研究では全記事の投稿者名を保存した「名前のデータベース」を考案した。宛名表現を調査した結果、いかなる表記法であっても[名前]には“さん”が付いていることが確認された。そこで、名前のデータベースに格納された投稿者名に“さん”を付けたものを回答記事の本文と照合することで、本文から宛名表現を抽出し、質問記事の探索に利用する。

名前のデータベースを利用して本文から抽出された宛名表現から、回答記事から直近の、宛名と一致する投稿者名の記事を探索し、該当する記事が存在した場合はその記事を質問記事として質問回答対を作成する。

ここまでの処理が全て終了したのち、TreeView に出力された質問回答対をユーザに提示する。

## 4. 評価

本手法により自動作成された質問回答対の精度を評価する。ランダムで選んだ 3 社各 120 件の記事から質問回答対を作成したところ、合計 64 対の質問回答対が得られた。それらを人手で作成した質問回答対と比較しながら、以下の 3 段階の基準で評価した結果を表 2 に示す。

【Level 1】 Q 記事も A 記事の集合も完全に一致する。

【Level 2】 質問回答対の Q 記事が一致するが、それに対応する A 記事の集合が不完全である。

【Level 3】 人手による A 記事が全て見つかっているが、正解でない A 記事も含まれている。

表 2 質問回答対生成の精度評価

| | Level 1 | Level 2 | Level 3 |
|---|---|---|---|
| 再現率 | 70.77% | 80.00% | 86.15% |
| 適合率 | 80.70% | 91.23% | 98.25% |
| F 値 | 75.41% | 85.25% | 91.80% |

失敗の主な原因として、質問回答対を作成する際に、回答記事から直近の記事を質問記事として採用する手法にあると考えられる。この問題を解決するためには、回答記事と質問記事と思われる記事との間に関連性があるかを判定した上で、対応する質問記事として扱う必要があると思われる。また、システムを実際に利用したユーザに対して行ったアンケートではシステムの有効性が確認できた。

## 5. 結論

本研究では、質問記事と回答記事が混在している Web 掲示板に着目し、回答記事から質問記事を探索することにより、質問回答対の作成を試みた。SVM による記事分類の実験では A&!A 基準の有効性を確認できたが、質問回答対の作成に関しては手法の問題点も浮上している。今後は以上の問題点を踏まえ、質問回答対の作成方法を改善しながら、アンケートの結果に基づき、よりユーザにとって利用しやすいようインターフェースに機能を追加する予定である。

## 参考文献

[1] 池上敬明, 竹内孔一: “Web 上の QA データの構造の抽出と利用”, 言語処理学会第 11 回年次大会, C2-7, 2005.

[2] 鈴木佑輔, 横田隼, 酒井浩之, 増山繁: “Web 掲示板からの質問・回答対応の自動抽出”, 人工知能学論文誌 25 巻 1 号 SP-Q, pp 168 - 173 , 2010.

[3] T. Joachims“Making large-Scale SVM Learning Practical". Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press. 1999.