C1-18

Why型質問応答システムNAZEQA Web版

Web NAZEQA, a Web-based Why-Question Answering System

磯崎 秀樹* 東中 竜一郎*
Hideki Isozaki Ryuichiro Higashinaka

1 はじめに

我々は最近、「なぜ … なのか？」という質問に答えるWhy型質問応答システムNAZEQAの研究を行っている[Higashinaka 08b, Higashinaka 08a, 磯崎 08]。今回は、言語処理学会で発表したシステムをさらに改良し、Webデータで予備的な実験を開始したので、それについて報告する。

Webを扱う意義としては、少なくとも以下の２つが考えられる。

- ユーザはしばしば、古い情報よりも最新の情報に興味があり、また、新聞だけではカバーできないような多様なトピックに興味がある。
- 新聞データで有効な手法が大規模なWebデータで有効とは限らず、手法の有効性を確認する必要がある。

一方、新聞データによる実験にも意義がある。まず、文字化け対策やフォーマット変換など、本来の研究目的とは違うところで無駄に時間を費やすことがない。実験の再現性という面でも有利である。

最新のWebのデータを利用する場合、同じURLでも、内容が変わっていたり、つながらなかったりする。商用サーチエンジンを利用するのは簡単であるが、検索アルゴリズム・検索対象文書・検索ログなどの更新により、出力は時々刻々と変化する。検索を実行したこと自体が、ネットワークのあちこちにあるキャッシュを更新したり、文書検索のスコアに影響を与えたりする可能性がある。したがって、実験の再現性や異なる手法の比較という面では問題が大きい。

我々は、NAZEQAの実現にあたり、1998年から2001年の毎日新聞CD-ROM版に基づき、以下のような1000問の「なぜ」型質問を用意した。そして、その10分割交差検定によってシステムの性能を測定、改良を行っている。

質問：○○氏が首相を辞めたのはなぜか？

正解：980730015.10 首相は[自民党が惨敗した参院選]翌日の１３日、同党総裁を退陣する意向を表明した記者会見で、敗因や経済政策の失敗について一切言及しなかった。

980730015.12 自民党単独政権復帰にこぎつけ、中央省庁再編や日露関係改善などに功績はあったものの、[金融不況による経済危機で足もとをすくわれた]。

980730015.16 しかし、[肝心の改革は失速、一方で経済危機、失業、教育荒廃、高齢不安など国民の痛みばかりが浮き彫りになり、退場を余儀なくされた]。

先頭の数字は、毎日新聞の記事番号と文番号である。正解は人手で探し出したものであり、一般に複数ある。[ ]内は、その文の中でもとくに正解として重要なポイントである。自動採点の都合上、各正解は１文とした。

言語処理学会の原稿を書いた時点での性能は、１文を回答として返すシステムで、１位正解率が0.153、５位以内正解率が0.445、５位MRRは0.257であった。

その後、利用する素性を見直した結果、現在では１位正解率が0.205、５位以内正解率が0.493、５位MRRが0.309に向上している。この改良については、別の機会に報告する。

2 NAZEQAの構成

NAZEQAの設計にあたり、我々は、正解が以下のような性質で特徴づけられると仮定した。

1. 理由表現の存在：理由を述べる文には、特徴的なパターンがあるであろう。
2. 類似性：正解の近辺の文章で使用されている語彙は、質問文に出現する語彙をほとんど含むだろう。

*日本電信電話株式会社, Nippon Telegraph and Telephone Corporation

まず、1 であるが、過去の研究 [乾 05] では、理由を述べる表現は多様かつ曖昧であり、検出が難しいことが知られている。したがって「ので」「ため」「理由」「原因」などの表現で候補を絞ってしまうと、かなりの数の正解を逃してしまう。

そこで NAZEQA では、EDR コーパスの意味解析結果を利用し、機械学習によって、理由表現のパターンをマイニングすることを試みた。この意味解析結果で cause タグを含む文を正例、含まない文を負例として BACT[Kudo 04a] で学習することにより、パターンをマイニングした。

まず、EDR コーパスの各文を CaboCha[工藤 04b] で解析しなおし、固有表現・品詞・日本語語彙大系の意味カテゴリなどの情報を付与した木を作成した。この木に正例・負例のラベルを付与して BACT にかけた。(なお、正例の数が負例の数に比べて圧倒的に少ないため、この学習結果を分類器として利用しても、性能はよくない。)

次に、回答候補の中から正解を選び出せる採点関数を学習するため、前述の 1000 問のうち 900 問でトレーニング、残りの 100 問でテストを行う 10 分割交差検定を行って、有効な素性を調べていった。この学習には、Ranking SVM[Joachims 02] の考え方を利用したが、効率面から、実際には OCAS[Franc 08] を利用している。

この学習にあたっては、上記で得られた理由パターンや、2 の類似度を利用している。

3 Web 対応で行ったシステムの変更

Web の検索エンジンを用いた Why 型質問応答の先行研究としては、田村ら [田村 08] によるものがあるが、これは子供向けの自然科学に関する５０問を対象としたクローズドな実験である。

我々は新聞記事でトレーニングした NAZEQA の検索部分だけを Web サーチエンジンで置き換えたシステムを作成し、これを以下の実験で利用した。

古い新聞記事を対象とした実験では、処理の対象となる文書が限られているため、形態素解析・係り受け解析・固有表現抽出・理由表現検出など、必要な処理をあらかじめ行っておくことができた。

Web を対象としてユーザに自由に使ってもらうシステムにするため、ユーザから質問を受け取って、検索エンジン goo に投げ、その結果を見て上位 20 文書を取得、html ファイルをプレインテキストに変換し、これらの処理をするように変更した。PDF など、他の形式のファイルは利用しなかった。

なお、Web 特有の様々な問題があることは容易に想像される。たとえば、信憑性の低い情報や広告などである。信憑性の低いサイトを排除するためにブラックリストを用意したり、有力なポータルサイトには、広告などを除く処理を用意することが考えられる。

しかし、そのようなアドホックな要素の導入は、システムを脆弱にし、メンテナンスが大変になるので、今回は行わなかった。

今回の実験の目標は、我々がこれまでに作成した NAZEQA が、Web においてどれほどの性能を持っているかを検証することである。

新しく作成した Web 対応部分のデバッグのため、Q&A サイトから、質問のタイトルに「なぜ」を含み、その１文だけで質問の意味がわかる問題を抜き出して利用した。これは、NAZEQA の質問解析が複数文対応になっていないためである。

この設定では、サイトに特化したシステムを作るのが一番簡単であるが、NAZEQA の性能を検証するという本実験の目的には合わないので、goo の検索結果から URL を抽出する部分以外、サイトに特化した処理はしていない。

また、goo には自然文検索の機能があるが、本研究と競合する機能なので利用せず、通常の検索語列挙による検索を用いる。

この実験結果を見て、以下のような、サイトに依存しない処理を追加した。

- 検索語としては、原形ではなく活用した形を利用する。たとえば、質問文に「読まない」という表現があった場合、「読む」で検索するよりも、「読ま」で検索する方がよい結果が得られる。
- 新聞に疑問文はほとんど出現しないが、Web では多数出現する。しかし、これらはほとんどの場合、質問の正解にはならない。そこで、疑問文は回答候補として扱わない。
- Web には、非常に長いファイルが存在する。これを処理するには時間がかかるので、grep により、検索語を含む文の近辺の文（前後５文）だけを回答候補として採用した。
- Web には、文の切れ目のはっきりしない文章がたくさんあり、改行コード
や line feed は文の途中にも出現する。

  そこで、行を連結して文を復元したいが、句点なしのテキストデータが並んでいるときには、１文が非

常に長くなってしまう。そこで、１文の長さがある一定の文字数を超えるときには、連結せず、文を切ることにした。

- NAZEQA では、引用符の中の句点を無視して１文を決定しているが、Web の文書の中には、引用符の対応が正しくないものがある。そこで、句点の位置ですでにある文字数を超えている場合は、引用符の中でも切る。
- 複数のサイトから完全に同じ文が回答候補として得られることがあるが、冗長なので最上位のもの以外は削除した。

4　実験

実験 (open test) には、NTCIR formal run の問題のうち、「なぜ」と言い換えられる問題 33 問を用いた。

その結果は図 1 のようにまとめることができる。N は Web NAZEQA の出力における正解の最高順位である。g は、その際に取得した goo のスニペットにおける正解の最高順位である。g+は、「理由」「原因」「ため」「ので」のいずれかを含む、という条件を付加した場合の goo のスニペットにおける性能である。いずれも、５位までを採点の対象とした。

なお、これらの問題を入力したときに、NTCIR QAC の関連文書が上位にランキングされて成績に影響を与える可能性がある。そこで、URL に ntcir という文字列を含む文書は無視した。

スニペットの MRR を計算するさい、テストデータそのものが上位に来ていると、スニペットの成績は不利である。確認してみると、上位スニペットに正解がある場合、NTCIR の関連文書が上位に来ているケースは１件だけであり、成績への影響はほとんどなかった。

今回の実験は、わずか３３問と小規模なものなので、NAZEQA の優位性を主張するには、もう少し大規模な実験が必要である。

しかし、Q&A サイトでの実験から、以下のような傾向がわかった。

- 質問のテーマが IT や語学などの場合、専門のサイトが多数あり、Q&A サイトの対応するページは上位に来にくい。その場合、正解と判断できる文を抜き出せることは少ない。
- それ以外の場合、検索結果の上位に Q&A サイトが上位に来る可能性が高い。この場合、そのページから正解と判断できる１文がしばしば得られる。

QAC4	N	g	g+
Q0002	-	-	-
Q0005	-	-	-
Q0006	1	2	3
Q0007	3	3	2
Q0008	3	1	4
Q0010	5	-	-
Q0012	-	4	3
Q0015	1	1	1
Q0026	3	3	2
Q0029	1	1	1
Q0030	1	1	1
Q0031	-	-	-
Q0035	-	1	-
Q0037	1	1	1
Q0040	5	3	3
Q0041	1	-	-
Q0051	-	-	-
Q0053	3	-	-
Q0061	1	2	1
Q0062	1	-	-
Q0064	1	-	-
Q0065	-	3	3
Q0068	-	-	-
Q0070	-	-	-
Q0071	-	-	-
Q0073	1	1	2
Q0079	-	-	5
Q0087	-	-	-
Q0090	-	-	5
Q0092	1	1	1
Q0094	1	5	-
Q0098	4	2	-
Q0099	-	-	4
1 位	12	6	6
2 位	0	3	3
3 位	4	4	4
4 位	1	1	2
5 位	2	1	2
5 位 MRR	0.424	0.281	0.295
5 位以内	0.576	0.455	0.515

図 1: goo の文書検索順位と NAZEQA での正解順位

- NAZEQA は文を単位として処理を行うため、その性能は文区切りに非常に敏感である。

また、QAC4 の問題を使った実験から、以下のような傾向がわかった。

- 「ミール」のような多義語では、文書検索の性能が低下し、正解しにくい。
- 新聞報道とは異なる観点での回答が見つかる。たとえば、質問あるいは新聞報道を否定するようなものがある。たとえば「なぜ自然エネルギーは環境に良いとされているのですか？」という質問に対する上位の回答は、いずれも、自然エネルギーはよくない、という文である。
- NAZEQA は周囲の文を考慮しているため、ブログのように、話題が頻繁に変わる場合に失敗することがある。

5 まとめ

「なぜ」に答える質問応答システム NAZEQA を Web 対応にしたバージョンをつくり、その性能の検証を始めた。

今回の実験では、サイトに依存した処理を一切加えなかった場合の性能を調べた。goo のスニペットとの成績を比較して、良好な成績であることが判明した。

参考文献

[Franc 08] Franc, V. and Sonnenburg, S.: OCAS optimized cutting plane algorithm for support vector machines, in *Proceedings of the International Machine Learning Conference* (2008)

[Higashinaka 08a] Higashinaka, R. and Isozaki, H.: Automatically Acquiring Causal Expression Patterns from Relation-annotated Corpora to Improve Question Answering for why-Questions, *ACM Transactions on Asian Language Information Processing*, Vol. 7, No. 2 (2008)

[Higashinaka 08b] Higashinaka, R. and Isozaki, H.: Corpus-based Question Answering for why-Questions, in *Proceedings of the Third International Joint Conference on Natural Language Processing*, pp. 418–425 (2008)

[乾 05] 乾 孝司, 奥村 学：文書内に現れる因果関係の出現特徴調査, 情報処理学会自然言語研究会 NL-167 (2005)

[磯崎 08] 磯崎 秀樹, 東中 竜一郎：パターンマイニングを用いて「なぜ」に答えるシステム, 言語処理学会第１４回年次大会発表論文集 (2008)

[Joachims 02] Joachims, T.: Optimizing Search Engines using Clickthrough Data, in *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining* (2002)

[Kudo 04a] Kudo, T. and Matsumoto, Y.: A boosting algorithm for classification of semi-structured text, in *Proceedings of EMNLP*, pp. 301–308 (2004), http://chasen.org/˜taku/software/bact/

[工藤 04b] 工藤 拓, 松本 裕治：カーネル法を用いた言語解析における高速化手法, 情報処理学会論文誌, Vol. 45, No. 9, pp. 2177–2185 (2004), http://chasen.org/˜taku/software/cabocha/

[田村 08] 田村 元秀, 村上 仁一, 徳久 雅人, 池原 悟：Web 検索エンジンを用いた Why 型質問応答システム, 言語処理学会第１４回年次大会発表論文集 (2008)