多様な質問と文書を対象としたオープンドメイン質問応答

秋葉友良 (メディア科学リサーチセンター, セマンティックアーカイブコア)

1 はじめに

インターネットをはじめとする情報通信網の急速な発展を背景に、近年、テキストデータだけでなく、音声や動画像などのマルチメディアデータが氾濫するに至っている。このような多様な大規模データから効率的に必要な情報へと到達することを可能にする情報検索技術の開発が、情報爆発時代の重要課題といえる。一方、従来の文書検索技術の発展として、自然言語による質問文から文書中の解答個所を直接特定し回答する、オープンドメイン質問応答技術が注目されている。本稿では、本研究室におけるオープンドメイン質問応答研究について報告を行う。

2 オープンドメイン質問応答とは

オープンドメイン質問応答 (以下、質問応答) とは、入力された自然言語で表現された質問文の答を、新聞記事や Web 上の文書などの大規模なテキストから抽出する技術である。質問応答は、整理されたデータベースから質問の答を探す一昔前の自然言語理解の研究と比べ、前処理や組織化が行われていないプレーンなテキストを対象とする点で、対象分野によらない (オープンドメイン) 頑健で応用範囲の広い技術と言える。また、Web 検索エンジンに代表される検索キーワードを入力に関連文書を検索する文書検索に対し、自然言語による詳細な質問 (情報要求) から文書中の答の部分を特定する、高精度な情報検索技術と位置づけることができる。

質問応答は、米国 NIST 主催の評価型ワークショップ TREC(Text Retrieval Evaluation Conference) にて 1999 年から今日まで大規模な評価が行われてきており、現在も活発な研究分野である。日本においては、国立情報学研究所主催の評価型ワークショップである NTCIR のにおいて QAC(Question Answering Challenge) として、2002 年から現在まで 4 回の評価が行われている。また、2005 年からは言語横断質問応答 (CLQA; Cross Lingual Question Answering) の評価も行われている。

3 音声入力質問応答 [2]

質問応答は、自然言語で表現された質問を入力とする。そのため、人が質問を発する時に多用するモダリティである音声をそのまま質問応答の入力とすることは、質問応答の利便性を大きく向上させることにつながる。

音声入力への対応は、解答抽出対象の文書コレクションに言語モデルを適応した大語彙連続音声認識 (LVCSR) をフロントエンドに用いることで実現できるが、音声質問の認識誤により質問応答の性能が低下する。この問題に対し、質問応答の主要な構成要素であるパッセージ検索を利用することで、認識結果の N-Best 候補から質問応答にとって適切な認識候補を選択する手法を開発した。

4 質問応答の対話的利用 [1]

質問応答を実際に利用する場合、利用者がある検索トピックに関連する複数質問を連続して投げかけるという使い方が想定される。これを情報アクセス対話 (Information Access Dialog) と呼び、NTCIR の QAC-3 にて評価が行われた。

情報アクセス対話では、システムは以前に入力された質問とそれに対する回答を利用して、ユーザ意図 (情報要求) を適切に理解する必要がある。これは、質問文の省略補間、あるいは照応解析の問題として捉えることがでる。すなわち、補間する語句をコンテキストから選択する多義性解消の問題として定式化することにより、音声入力質問応答の場合と同様に前述のパッセージ検索を用いた多義性解消手法が適用できる。

5 音声ドキュメントを対象とした質問応答 [5]

従来のラジオやテレビでの放送に加えて、ポッドキャスティングなど Web 上での音声・動画像配信も爆発的に広まりつつある。これらのマルチメディアデータに埋もれる情報を検索する技術を目指し、音声ドキュメントを対象とした質問応答の研究に取組んでいる。

音声ドキュメントに対して質問応答を行う直接的な方法の一つは、大語彙音声認識を利用して得た音声文書の書き起しに対して、テキストを対象とした質問応答を適用するというものである。この方法の問題点は、大語彙音声認識の利用で生じる音声認識誤り、および認識器の辞書外の語の扱いにある。特に、質問は固有名詞を問うことが多いため、認識辞書に含まれない場合が多く、正解部分は誤認識が生じやすい。その場合、テキストを対象とした質問応答をそのまま適用しても正解を見つけることは難しい。

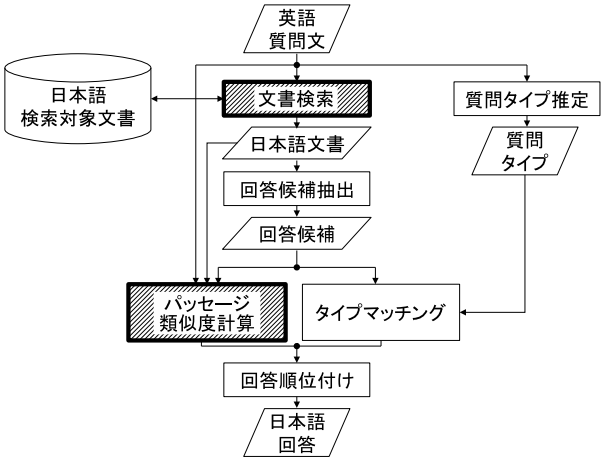

Fig. 1　CLQA system の概要

本研究では、テキストを対象とする質問応答で回答候補の抽出に利用する「固有表現抽出」の代わりに、回答候補周辺のより広い区間 (例えば、発話) に回答候補が含まれるかどうかを検出する「固有表現検出」を利用することで、誤認識に頑健な音声ドキュメントに対する質問応答を実現した。

6　他言語文書を対象とした言語横断質問応答 [4]

インターネット上の文書は多様な言語で記述されており、必要とする情報は質問とは異なる言語で記述されているかもしれない。質問文と検索対象文書の言語が異なる場合のオープンドメイン質問応答を、言語横断質問応答と呼ぶ。

言語横断質問応答には、処理中のいずれかの段階で翻訳が必要になる。前処理によって質問文と検索対象文書の言語を統一すれば、単一言語の質問応答の問題に帰着できる。翻訳に用いる手法によって、従来の手法は機械翻訳システムを用いる手法、対訳辞書を用いる手法の 2 つに分類できる。

筆者らは、統計的機械翻訳を利用した言語横断質問応答手法を開発した。提案手法は、前処理として翻訳を用いる従来手法とは異なり、統計翻訳モデルを質問応答プロセスに組み込むことで言語横断質問応答を実現する。図 1 に、英語から日本語の回答を探す英日言語横断質問応答システムの構成を示す[1]。システムは、入力が英語である点を除き、単言語質問応答システムとほぼ同じ構成を持つ。英語入力に対応するため、太字の文書検索モジュールおよびパッセージ類似度計算モジュールを統計的機械翻訳の翻訳モデルを用いて言語横断化している。この手法により、従来法より 1.5〜2 倍程度、性能を改善することができた。

[1]ただし、提案手法はどの言語対へも適用可能である。

7　実世界の任意の質問を対象としたユニバーサル質問応答 [3]

TREC や NTCIR における質問応答は、当初は事実型 (factoid) と呼ばれる一つの語句で答えられる質問を対象としてきた。例えば、人名や場所を尋ねる質問 (「誰?」「どこ?」) や時間や数量を尋ねる質問 (「いつ?」「いくら?」) が事実型質問に相当する。一方、最近では、理由を尋ねる質問 (「何故?」) や方法を尋ねる質問 (「どうすればいい?」) など、非事実型の質問を扱う質問応答の研究も活発になりつつある。筆者らは、人を相手にしたオープンドメイン質問応答ではどのような質問が入力されるかを網羅的に予測することは不可能であるとの考えから、実世界の任意の質問に回答することを目的としたユニバーサル・オープンドメイン質問応答の研究を行っている。

8　まとめ

研究室におけるオープンドメイン質問応答に関する研究を概観した。各研究についての詳細は、以下の発表論文を参照されたい。

発表論文

[1] T. Akiba. Exploiting dynamic passage retrieval for spoken question recognition and context processing towards speech-driven information access dialogue. In *Proceedings of International Conference on Language Resources and Evaluation*, pp. 1530–1535, 2006.

[2] T. Akiba and H. Abe. Exploiting passage retrieval for n-best rescoring of spoken questions. In *Proceedings of International Conference on Speech Communication and Technology (Eurospeech)*, pp. 65–68, 2005.

[3] 水野淳太, 秋葉友良. 任意の回答を対象とする質問応答のための実世界質問の分析と回答タイプ判定法の検討. 言語処理学会第 13 回年次大会, 2007.

[4] 清水慧, 秋葉友良, 藤井敦. 統計翻訳に基づくパッセージ検索の言語横断質問応答への適用. 言語処理学会第 13 回年次大会, 2007.

[5] 辻村裕史, 秋葉友良. 音声文書を対象とした質問応答のための固有表現検出法の検討. 日本音響学会秋季研究発表会, 2006.