様式 C-19

科学研究費補助金研究成果報告書

平成　２２年　３月　　３１日現在

研究種目：基盤研究(C)
研究期間：2007～2009
課題番号：19500121
研究課題名（和文）
テキストマイニングとの融合による高度オープンドメイン質問応答の実現
研究課題名（英文）
Development of an advanced open-domain question answering system by incorporating text mining methods
研究代表者
秋葉　友良（AKIBA TOMOYOSHI）
豊橋技術科学大学・工学部・准教授
研究者番号：00356436

研究成果の概要（和文）：大規模なドキュメントから人の情報要求に適合する情報を効率よく獲得するための技術として、オープンドメイン質問応答が期待されている。本研究では、従来の手法では対象外であった発掘型質問について、実際の QA データベースの調査を行い、そこで得た知見を利用して質問応答システムの開発を行った。また、質問応答の精度向上のための手法や、他言語で記述されたテキストや音声データを対象とするための手法を開発した。

研究成果の概要（英文）：
Open domain question answering is a promising technology for obtaining the information relevant to one's information needs from large-scale documents. In this research, we focused on ``mining type questions'', which had not been dealt with the existing QA system. We firstly investigated the mining type questions appeared in the real-world question database, then developed a QA system targeting them by exploiting the findings. In addition, we developed the methods to improve the accuracy of the traditional question answering, the QA methods targeting a foreign language documents, and the methods targeting speech data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,400,000	420,000	1,820,000
2008 年度	1,100,000	330,000	1,430,000
2009 年度	900,000	270,000	1,170,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：自然言語処理、音声情報処理
科研費の分科・細目：知能情報学
キーワード：オープンドメイン質問応答，テキストマイニング，情報検索，音声ドキュメント処理

様式 C-19

科学研究費補助金研究成果報告書

１．研究開始当初の背景

テキスト情報源の電子化およびインターネットなどの共有可能なテキスト情報の急増を背景に、大規模なテキストから人の情報要求に適合する情報を効率よく獲得するための情報アクセス技術が重要な研究課題となっている。情報検索を高精度化する技術として、自然言語で表された質問を入力とし大規模な検索対象から該当する答の部分だけを抽出するオープンドメイン質問応答技術が有望である。質問応答は従来、事実型質問と呼ばれる名称や数量を問う事実型質問(factoid question)を中心に研究が行われてきた。しかし、定義を問う質問(定義型) や、理由を問う質問(WHY 型)などの非事実型質問(non-factoid question)を対象とした質問応答の研究が活発になってきた。

２．研究の目的

従来の質問応答では対象外であった、文書の複数の部分の組合わせにより答が明らかになるような質問（発掘型質問）を対象とした質問応答システムを開発する。また従来の factoid 型質問応答システム、non-factoid 型質問応答システムの性能改善や機能向上を行うとともに、全てのシステムを統合し如何なる質問にも回答する究極の質問応答システムを実現する。

３．研究の方法

(1) 以前の研究で構築した「質問データベース」および新聞記事などの既存テキストコーパスをベースに、実世界質問を対象とした発掘型質問データベースを構築する。収集した質問を分類・分析し、システム構築に必要な知識を獲得する。

(2) 質問データベースの分析を元に、システムの構成要素となるモジュールを実装する。質問解析器は、回答抽出に必要な処理の分類結果を元に入力質問を分類し、質問文から回答抽出に必要な情報を取り出す。回答抽出器は、抽出した情報から検索質問を構成し、検索対象文書のマイニングにより回答候補を抽出する。これらを統合して発掘型質問応答システムを構築する。

(3) 従来型の質問応答システムについて、回答タイプ判定性能、回答候補抽出性能、回答判定性能などの基本性能の向上および、有用な回答抽出範囲の検討、他言語で記述された対象文書への対応、音声データを対象とした質問応答などの質問応答機能の向上を行う。これらの質問応答システム全てを統合したユニバーサル質問応答システムの実装を行なう。

４．研究成果

(1) 本研究で対象とする発掘型質問の調査のため、Web 上の質問サイトに投稿された質問を調査した。技術、社会、学問、生活、趣味などの質問カテゴリそれぞれからランダムに質問を調べ、一つの文書に答がそのまま掲載されているのではなく、複数の文書に記述された情報を統合することで答が導きだせるという基準に基づき、質問を選択し分類を行った。その結果、(1)比較・最上級を問う質問、(2)初出を問う質問、(3)経緯・軌跡を問う質問、(4) ２つ以上の事項の関連性を問う質問、の４つに分類できることが分かった。これらの質問文を意味表現に構造化し、これらの質問に答えるために必要な要素を特定した。

(2) 調査結果に基づき、入力質問文字列から各タイプの意味表現を抽出する質問文解析器を構築した。解析器は、正規表現パターンによる人手で作成した規則に基づいて実装を行った。また、ルールベースの手法では多様な入力質問文字列に対応するには限界があると考え、質問データベースの類似質問との類似度に基づく事例ベースの質問解析手法も実装した。また、新聞記事を対象とした回答抽出法を具体化した回答タイプ毎の回答抽出ルールを作成した。これを質問文解析器と組み合わせて、質問応答システムの実装を行った。ただし、抽出結果には誤りが多く、精度の向上に課題が残されている。

(3)従来型の質問応答システムについて、以下に挙げる性能向上手法および機能向上手法の開発を行なった。これらの factoid 型、non-factoid 型、発掘型の質問応答の質問入力部を統合し、すべての回答タイプに対応する質問応答システムの仮実装を行った。

① factoid 型質問応答の発展型として、従来型の「短い答」ではなく、回答周辺に存在する情報を付加し、かつその回答自身が言語として適切な表現となっているような「詳解」を抽出するシステムの検討を行った。まず、既存の QA テストコレクションを調査し、詳解の分類を行なった。これらの詳解を抽出するため、構文解析機を使った手法を実装した。この手法で誤って抽出した詳解をさらに分析することで、詳解抽出の手がかりとなる情報を明らかにした。

② 質問文とは異なる言語で記述された文書から答を抽出する言語横断質問応答として、統計的機械翻訳に基づくパッセージ検索手

法を検討した。特に、本手法で利用する翻訳モデルの改善を行った。翻訳モデルの改善手法として、日英方向のユニグラム線形補間翻訳モデル、英日方向のユニグラム線形補間翻訳モデル、IDF 重み付け英日翻訳モデル、確率的潜在意味解析による英日翻訳モデルを検討した。また、翻訳モデルの学習法として、内容語だけを考慮した対訳コーパスからの学習手法を検討した。評価実験により、検索対象言語から検索質問言語への方向の翻訳モデルが逆側の翻訳モデルより優れていること、単語重要度の利用が有効であること、内容語のみを考慮した学習手法が有効であること、が分かった。

③ non-factoid 型の質問応答の拡張・性能改善を行った。QA コーパスを知識源として用いる、質問タイプに依存しないシステムを実装した。このシステムをベースに、入力質問と回答候補の間の、予測される回答タイプの一致の判定に利用する特徴量抽出法について検討を行った。従来は特徴量として機能語と品詞化した内容語を利用していたが、さらに質問に現れる内容語を他の内容語と区別して抽象化した特徴量を考案し、その利用による性能の向上を確認した。また、検討した特徴量を用いたオンライン機械学習による回答抽出法の検討を行った。

④ ニュース音声・学会講演音声などの音声データを音声認識して得られる音声ドキュメントを対象とする質問応答では，認識誤りへの対応が問題になる。特に、質問に対する正解自体が誤認識される可能性は高く、その場合正解を見つけることは極めて困難になる。この問題に対処するため、固有表現の出現位置を正確に求める従来の固有表現抽出の代わりに、ある音声区間に特定のタイプの固有表現が存在するか否かだけを判定する固有表現検出を利用する手法を開発した。固有表現検出は、より簡単な問題設定であることと、正解周辺のより広いコンテキストを利用することで、音声認識誤りを含むテキストに対しても頑健に動作する。音声ドキュメントを対象とした質問応答での評価実験により、提案法は従来法に比べ高い性能を示すこと、正解が誤認識される場合でも頑健に動作することが分かった。

⑤ 音声ドキュメントを対象とする質問応答の性能向上には、フロントエンドの検索性能の向上も必要である。その際に問題となるのは、音声認識誤りの影響による自動書き起しテキストの劣化の問題である。この問題に対し、本研究では、音声認識による自動書き起しと人手書き起しの間の差異を「翻訳」によって補完する検索手法を開発した。評価実験の結果、提案手法は、特に小さいサイズの文書を検索対象とするタスクにおいて良い性能を示すことが分かった。さらに、文脈を利用した文書拡張手法、言語モデルに基づく検索手法を利用することを試み、それらの併用による検索性能の向上を確認した。

⑥ 音声ドキュメントでは段落や文のような意味的まとまりを表す区切りが明示されていないため、可変長区間を対象とした検索手法が必要になる。可変長音声区間を対象とする内容検索手法として、固定区間に分割して中心を検索する手法、固定区間の検索の後に区間内での２段階目の検索を行なう手法、区間を設定せずに検索の後近傍の検索結果を事後的にまとめる手法を検討し、評価実験により事後的にまとめる手法が有効であることを明らかにした。

⑦ 音声ドキュメントを対象とする検索では、音声データから検索語の出現位置を検出する技術が必要となる。この音声からの検索語検出（Spoken Term Detection）問題に対し、音素と音声ドキュメント中の位置の間の距離で定義される距離空間上の索引を用いる全く新しい手法を開発した。提案法は、検索の際にしきい値を設定する必要がなく、類似した候補から順番に検索結果を出力することが大きな特長である。また、認識の複数候補を扱う自然な拡張法が存在する。講演音声を対象とした予備実験の結果、高速な検索語検出が可能であることを確認した。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計８件）

① Kouichiro Honda, Tomoyosi Akiba, Language Modeling Approach for Retrieving Passages in Lecture Audio Data, Proceedings of International Conference on Language Resources and Evaluation (LREC 2010), 査読有, (掲載決定)、2010.
② Tomoyosi Akiba, 他(全９名， １番目)、Developing an SDR Test Collection from Japanese Lecture Audio Data, Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009), 査読有, pp.324-330, 2009.
③ 秋葉友良、横田悠右、認識候補から正解テキストへの翻訳に基づく講演音声ドキュメントのアドホック検索、情報処理学会論文誌、査読有、Vol.50, No.2,

pp. 514-523, 2009.
④ Tomoyosi Akiba, 他（全９名、１番目)、Construction of a Test Collection for Spoken Document Retrieval from Lecture Audio Data、情報処理学会論文誌、査読有、Vol. 50, No. 2, pp. 501-513, 2009.
⑤ Tomoyosi Akiba, Yusuke Yokota, Spoken Document Retrieval by Translating Recognition Candidate into Correct Transcriptions, Proceedings of International Conference on Speech Communication and Technology, 査読有, pp. 2166-2169, 2008.
⑥ Tomoyosi Akiba, Kei Shimizu, Atsushi Fujii, Statistical Machine Translation based Passage Retrieval for Cross-Lingual Question Answering, Proceedings of International Joint Conference on Natural Language Processing, 査読有, pp. 751-756, 2008.
⑦ Tomoyosi Akiba, Hirofumi Tsujimura, Error-Tolerant Question Answering for Spoken Documents, Proceedings of International Conference on Speech Communication and Technology, 査読有, pp. 326-329, 2007.

〔学会発表〕(計 22 件)
① 金子泰輔、秋葉友良、ハフ変換に基づく音声ドキュメントの高速検索語検出法、日本音響学会春季研究発表会、2010 年 3 月 10 日、電気通信大学（東京）
② 本田耕一郎、秋葉友良、講演音声を対象とした部分音声区間の内容検索タスクの設定とその検索手法の検討、2010 年 3 月 8 日、電気通信大学（東京）
③ 兵藤達浩、秋葉友良、統計翻訳を用いた言語横断質問応答における翻訳モデルの改善、第８回情報科学技術フォーラム、2009 年 9 月 3 日、東北工業大学（仙台）
④ 小田貴博、秋葉友良、Non-factoid 型質問応答システムにおける質問タイプ判別法の改善、第８回情報科学技術フォーラム、2009 年 9 月 3 日、東北工業大学（仙台）
⑤ 秋葉友良、本田耕一郎、翻訳モデルを用いた講演音声ドキュメントの内容検索―文脈情報の利用と言語モデリング検索手法の適用、第 3 回音声ドキュメント処理ワークショップ、2009 年 2 月 27 日、豊橋技術科学大学（豊橋）
⑥ 秋葉友良、音声言語研究関連分野の１０年の歩み「音声検索」、第１０回音声言語シンポジウム、2008 年 12 月 9 日、早稲田大学（東京）
⑦ 伊藤雄、秋葉友良、事実型オープンドメイン質問応答システムにおける周辺情報を考慮した詳解の抽出、言語処理学会第 14 回年次大会、2008 年 3 月 18 日、東京大学駒場キャンパス（東京）
⑧ 秋葉友良、辻村裕史、固有表現検出を用いた認識誤りに頑健な音声ドキュメント質問応答、第２回音声ドキュメント処理ワークショップ、2008 年 2 月 29 日、豊橋技術科学大学（豊橋）
⑨ 秋葉友良、横田祐右、翻訳モデルに基づく講演音声ドキュメントのアドホック検索、日本音響学会秋季研究発表会、2007 年 9 月 19 日、山梨大学（甲府）
⑩ Junta Mizuno, Tomoyosi Akiba , Non-factoid Question Answering Experiments at NTCIR-6: Towards Answer Type Detection for Realworld Questions, Sixth NTCIR Workshop, 2007 年 5 月 17 日、東京

〔産業財産権〕
○出願状況（計１件）

名称：系列信号検索装置および系列信号検索方法
発明者：秋葉友良, 金子泰輔
権利者：豊橋技術科学大学
種類：特許
番号：特願 2009-286883
出願年月日：平成 21 年 12 月 17 日
国内外の別：国内

６．研究組織
(1)研究代表者
秋葉 友良（AKIBA TOMOYOSHI）
豊橋技術科学大学・工学部
研究者番号：00356346

(2)研究分担者
中川 聖一（NAKAGAWA SEIICHI）
豊橋技術科学大学・工学部
研究者番号：20115893