様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年５月２４日現在

機関番号：12701
研究種目：基盤研究(C)
研究期間：2010～2012
課題番号：22500124
研究課題名（和文）　利用者の納得に寄与する情報アクセス技術に関する研究

研究課題名（英文）　Study of information access methods that contribute to users' convincement

研究代表者
森　辰則（MORI TATSUNORI）
横浜国立大学・大学院環境情報研究院・教授
研究者番号：70212264

研究成果の概要（和文）：本研究では、情報アクセスシステムが提示する回答候補の採否を利用者が最終的に判断をしなければいけないことを当然の前提とし、利用者の納得に寄与する情報アクセス技術について検討を行った。具体的には、個々の回答候補の観点および回答候補間の関係の観点から、利用者の納得への寄与の検討をし、利用者からの納得に関する対話的フィードバックの検討を行った。

研究成果の概要（英文）：Supposing natural situations in which users have to determine the adoption or rejection of answer candidates that an information access system replies, we studied information access methods that contribute to users' convincement from the following viewpoints: individual answer candidates, relations among answer candidates, and interactive feedbacks from users.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	2,100,000	630,000	2,730,000
2011 年度	600,000	180,000	780,000
2012 年度	600,000	180,000	780,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：(1) ディレクトリ・情報検索, (2) 画像,文章,音声等認識, (3) ユーザインターフェース, (4) 自然言語処理, (5) 情報アクセス技術, (6) 質問応答システム

１．研究開始当初の背景

ICT技術の普及に伴い、膨大な量の文書情報がWeb等により入手可能となったが、利用者が欲する情報に容易に到達可能であるとは言い難い。例えば検索エンジンにより関連文書の効果的な絞込みが可能になりつつあるが、現在の検索エンジンは文書の順位付けを行い、文書の短い抜粋要約であるsnippetを提示するだけであるので、真に必要とする情報を得るためには文書の中身を精査する必要がある。そのため、情報を得るまでに利用者が読むべき文書を少なく抑え作業効率を向上させる情報アクセスに関する研究が行われている。例えば、文書から不要な部分を削りより短い文書を生成する自動要約や、「日本の首相は誰ですか。」といった利用者が入力した自然言語の質問文に対し情報源となる文書群からその答そのものを見つけ出す質問応答などがある。

情報検索や質問応答のような情報アクセ

ス手法においては、回答候補取得の理論的な基礎として、利用者が与えた情報要求と、回答候補となる文書や文書断片との間の関連性に注目したモデルが、今まで検討されてきている。また、情報源となる文書群における事物の記述の正否については、利用者の判断に任せるものとし、処理の上では正しいものとして扱われてきた。すなわち、従来研究においては、回答候補群について、利用者がどのようにしてそれを検証し、自分なりに納得のいく回答を選ぶかという観点での議論がなされていなかった。しかし、Web 文書に記されている情報は玉石混交であるので、情報アクセスシステムが利用者にとって有益なツールとなり得るためには、利用者の判断を積極的に手助けすることが必要である。

２．研究の目的

本研究では、情報アクセスシステムが提示する回答候補の採否を利用者が最終的に判断をしなければいけないことを当然の前提とし、利用者の納得に寄与する情報アクセス技術について検討を行う。具体的には、個々の回答候補の観点および回答候補間の関係の観点から、利用者の納得への寄与の検討をし、利用者からの納得に関する対話的フィードバックの検討を行う。

３．研究の方法

(1) 理論的な枠組みに関する検討として、情報アクセスに纏わる言語モデルの利用の検討を行った。利用者の納得の観点から言語モデルを構築することを見据えて、情報アクセスシステムの代表例である質問応答システムを題材とし、Q&A コミュニティサービスの事例から得られる質問と回答の仕方に関する言語モデルを利用し、求解精度を向上する手法を検討した。

(2) 個々の回答候補の観点からの、利用者の納得への寄与の検討の一部として、 時空間位置に注目した情報アクセス法の検討を行った。

(3) 個々の回答候補の観点からの、利用者の納得への寄与の検討の一部として、利用者が様々な情報抽出部品を組み合わせて利用できる情報抽出型検索エンジンを提案した。

(4) 回答候補間の関係の観点からの、利用者の納得への寄与の検討として、我々が「調停要約」として提案した、情報統合・提示の枠組みについて検討を行い、その精度向上手法を提案した。特に、利用者から疑問に思う点について対話的にフィードバックを受け、それに基づき、出力結果の詳細化を行うことを検討した。

４．研究成果

(1) 質問の型によらない non-factoid 型質問応答システムに対する確率的言語モデル、トピックモデルの導入

Q&A コミュニティサービスの事例から得られる質問と回答の仕方に関する情報を non-factoid 型質問応答に活用し、質問の型に依存しない有力な手法を従来提案していた。これに対して、入力質問と類似する質問事例を取得する過程、ならびに、取得された質問事例と対をなす回答事例に類似する解候補を見つける過程に、確率的言語モデルを導入することにより、精度が向上することを示した(学会発表⑦等)。

さらに、質問と解候補の間の内容関連度を確率的トピックモデルにより精密にモデル化することにより、さらに精度が向上することを示した(学会発表⑧等)。

提案手法の概略を図 1 に示す。

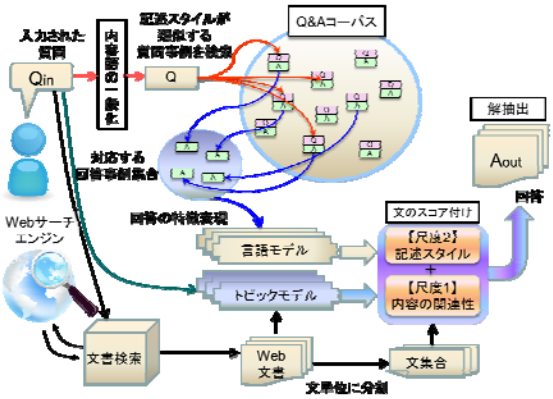


図 1 質問の型によらない non-factoid 型質問応答システムにおける、確率的言語モデル、トピックモデルの導入

導入した確率言語モデルは、表層表現とその品詞情報の両者を考慮する 2-gram モデルである。入力された質問と記述スタイルが類似する質問事例を収集するにあたっては、質問の生起確率が最大となるような言語モデルを生成する Q&A コーパスの部分集合を得ることが理想であるが、計算量が膨大となってしまう。そのため、クラスタリングと山登り探索を組み合せた近似手法を検討した。評価型ワークショップ NTCIR-6 の QAC formal run テストセットを用いて評価した結果、言語モデルの導入前後で 0.338 から 0.438 に上昇した。なお、MRR は最上位正解順位の逆数の、質問にわたる平均であり、1 に近いほど良好な精度である。

確率的トピックモデルとしては、LDA (Latent Dirichlet Allocation)を用い、検索された Web 文書集合から潜在的意味モデルを抽出する。いくつかの予備実験の結果、得

られた複数のトピックのうち、確率が最大のものを一つ選択し、トピックモデルとすること、ならびに、質問に含まれるキーワードについては、確率値を 1 に固定して重要視することが有効であることが分かった。このトピックモデルを導入することにより、MRR がさらに上昇し、0.521 となった。

(2) 時空間位置に注目した情報アクセス法の検討

利用者の納得を支えるもののうち基本となるものは、利用者が理解可能である(と仮定できる)情報である。特に固有表現は、ある事物を指し示す名前であるので、テキストと実世界をつなぐという意味で重要である。ここでは、特に時空間位置を特定する表現(日時表現、地域を表す表現)に注目し、「～は、いつ、どこで起こったか」といった時空間位置に焦点を当てた情報要求に対する情報アクセスについて検討した。特に、質問応答エンジンのスコアを利用すると、適切な文書を得られることを示した(学会発表⑧)。

図 2 に手法の概要を示す。

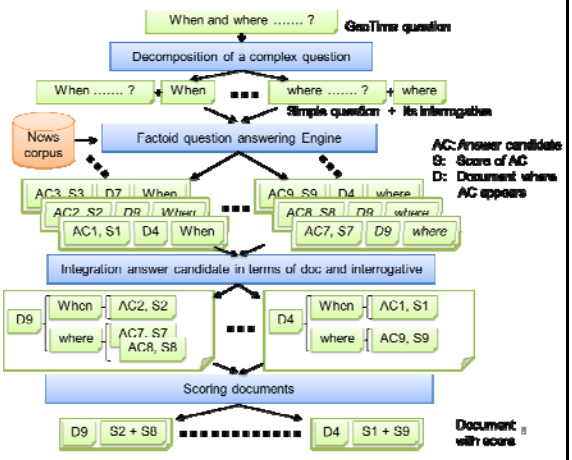


図 2 時空間位置に注目した情報検索手法

(3) Web 文書を対象とする情報抽出型検索エンジンのプラットホームの提案

情報アクセスシステムの多くが情報検索と情報抽出の組み合わせで構成されるが、細部が利用者から見えず、回答候補への利用者の納得に繋がらない。これを問題意識として、利用者が様々な情報抽出部品を組み合わせて利用できる情報抽出型検索エンジンを提案した。具体的には、抽出部品の開発と利用を通じて、利用者と開発者が協働し、コミュニティを醸成することにより、抽出部品の特徴を理解しつつそれらを組み合わせて利用可能な基盤を構築した(学会発表⑤)。

図 3 に提案手法の概要を、図 4 にシステムの利用例を示す。

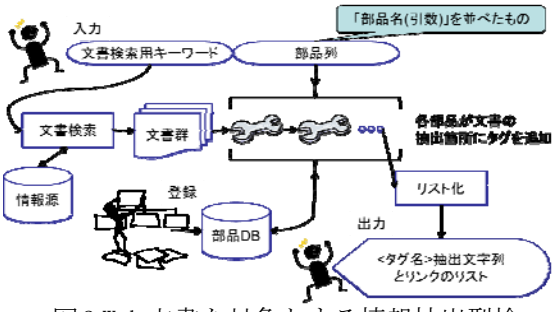
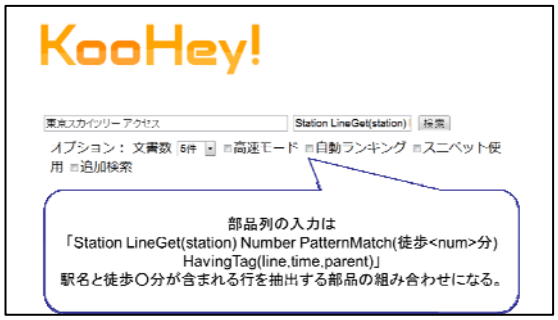


図 3 Web 文書を対象とする情報抽出型検索エンジン



(a) 入力画面

(b) 出力その 1

(c) 出力その 2（その 1 の該当箇所を Click した後）

図 4 利用例

(4) 「調停要約」の高度化に関する検討

回答候補間の相対的な位置づけを利用者に提示する際に情報信憑性の判断と納得に

寄与するために、我々が従来「調停要約」として提案した、情報統合・提示の枠組みについて検討を行った。調停要約生成手法は、複数文書要約の一つであり、要約対象文書群から「まとめ文章」を取り出すことにより要約する手法に属する。すなわち，互いに対立しているようにみえる二言明が両立可能となるような状況を簡潔にまとめている文章を要約対象文書群から探し出し，該当する文章を１つのパッセージとして抽出して提示する。図５に調停要約生成の例を示す。

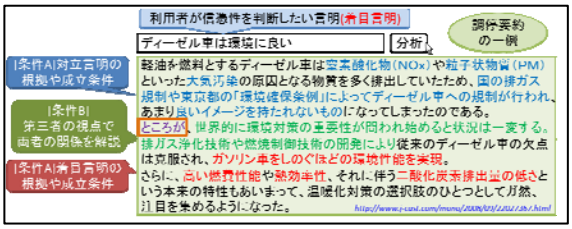


図５ 調停要約生成の例

本研究においては、次の成果を得た。

① 手がかり表現の利用と重要度計算手法の検討

調停要約文章に現れる対比構造を認識するために、逆接や条件等を表す手掛かり表現に注目した。特に、表現の有無による単純なフィルタリングと比較して、文章(パッセージ)の重要度計算に手掛かり表現の有無を反映させる手法が有効であることを示した(学会発表③⑥，雑誌論文②)。

② 利用者からの対話的フィードバックが存在する環境下での調停要約生成手法の検討

現実的な状況として、情報検索の出力を出発点とした次のような対話的なフィードバック環境を想定する。まず、利用者は、提示された文書集合(の要約であるスニペット集合)を読み，互いに矛盾しているように見えるために信憑性が疑わしく思える ２ 文をマウス操作等によりマーキングする。提案手法はこれら２文の記述を手掛かりにして、直接調停要約文章の発見を行う。図６に生成例を、図５に提案手法の概略をそれぞれ示す。

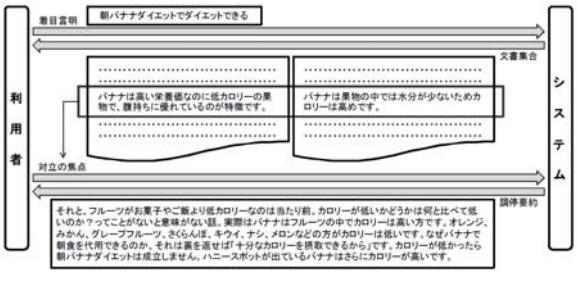


図 ６ 対話的フィードバックが存在する環境下での調停要約の生成例

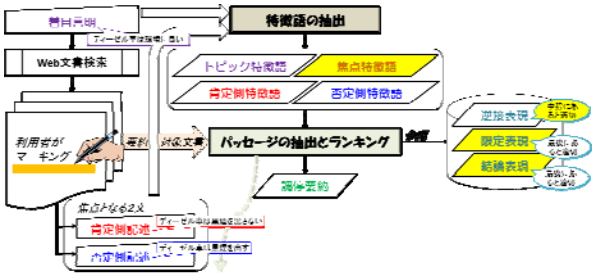


図７ 対話的フィードバックが存在する環境下での調停要約生成手法

評価実験の結果、明確な対比構造を持つ疑似的な対比文を入力とした場合に比べて、利用者がマークした２文を焦点としたほうが精度が良いことが確認された。具体的には、システム出力の上位 10 件の適合率が 0.050 から 0.231 に向上した。この理由として、抽出される特徴語が増加したことによる焦点の明確化、および、焦点に関連するパッセージの絞り込みが容易になったことが考えられる(学会発表④, 雑誌論文①)。

③ 語彙的連鎖を用いた調停要約生成手法の提案

上記①、②で用いている手法では、抽出する要約文章の範囲について、意味的なまとまりに基づいて判断しているわけではないため、意味的なまとまりがない文群が１ つのパッセージとして切り出されてしまい、結果として、調停要約として不適切な文を含むことがあった。そのために、意味的なまとまりを判断する一手法である語彙的連鎖に基づく語彙的結束性を判断し、調停要約生成に役立てる手法を検討した(学会発表①)。

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２件）

① 渋木英潔，永井隆広，中野正寛，石下円香，松本拓也，森辰則．情報信憑性判断支援のための対話型調停要約生成手法．自然言語処理 (2013)(採録決定)，査読有

② 中野正寛，渋木英潔，宮崎林太郎，石下円香，金子浩一，永井隆広，森辰則．情報信憑性判断支援のための直接調停要約生成手法．電子情報通信学会論文誌，Vol. J94-D, No. 11, pp. 1919-1930 (2011)，査読有
http://search.ieice.org/bin/summary.php?id=j94-d_11_1919&category=D&year=2011&lang=J

〔学会発表〕（計８件）

① 朱丹，渋木英潔，森辰則．語彙的連鎖を

用いた調停要約生成手法の提案. 言語処理学会第 19 回年次大会, P5-15, 2013 年 3 月 15 日, 名古屋

② Kyosuke Yoshida, Taro Ueda, Madoka Ishioroshi, Hideyuki Shibuki, and Tatsunori Mori. Introduction of a Probabilistic Language Model to Non-Factoid Question-Answering Using Example Q&A Pairs. In Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26), 2012 年 11 月 9 日, Bali (Indonesia)

③ 永井隆広, 石下円香, 中野正寛, 松本拓也, 渋木英潔, 森辰則. 調停要約生成の精度向上に向けた手がかり表現の利用と重要度計算手法の検討. 言語処理学会第 18 回年次大会発表論文集, F4-3, 2012 年 3 月 16 日, 広島

④ 渋木英潔, 永井隆広, 中野正寛, 石下円香, 松本拓也, 森辰則. 情報信憑性判断支援におけるユーザが焦点を明確化した状況下での調停要約生成手法. 言語処理学会第 18 回年次大会発表論文集, P2-27, 2012 年 3 月 15 日, 広島

⑤ 菅原晃平, 森辰則. Web 文書を対象とする情報抽出型検索エンジンのプラットホーム. 情報アクセスシンポジウム 2011, 2011 年 9 月 14 日, 東京

⑥ 渋木英潔, 中野正寛, 石下円香, 永井隆広, 森辰則. 調停要約生成手法の改善と調停要約コーパスを用いた評価. 第10回情報科学技術フォーラム(FIT 2011)講演論文集, RE-003, 2011 年 9 月 7 日, 函館

⑦ 吉田恭輔, 上田太郎, 石下円香, 森辰則. 質問・回答事例を利用した non-factoid 型質問応答に対する確率的言語モデルの導入. 自然言語処理研究会報告 2011-NL-201, 情報処理学会, 2011 年 5 月 17 日, 東京

⑧ Tatsunori Mori. A Method for GeoTime Information Retrieval based on Question Decomposition and Question Answering. In Proceedings of the Eighth NTCIR Workshop Meeting, pp.167-172, 2010 年 6 月 17 日, 東京

６．研究組織

(1)研究代表者

森　辰則 (MORI TATSUNORI)

横浜国立大学・大学院環境情報研究院・教授

研究者番号: 70212264

(2)研究分担者

(3)連携研究者