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２６日現在

研究種目：若手研究(B)
研究期間：2007～2009
課題番号：19700140
研究課題名（和文）　単語間の関係表現を利用した多言語横断質問応答システムの研究

研究課題名（英文）　Multi Lingual Question Answering System Using Word Relationship Expressions

研究代表者
土屋　雅稔（TSUCHIYA MASATOSHI）
豊橋技術科学大学・工学部・助教
研究者番号：70378256

研究成果の概要（和文）：多言語横断質問応答システムを実現するには，複数の言語を対象として単言語質問応答システムを用意する必要がある．従来の質問応答システムにおいては，様々な箇所で人手規則に基づく方法を採用していたため，複数の単言語質問応答システムを用意することは構築コスト的にかなり困難だった．そこで，本研究では，質問応答システムに機械学習的アプローチを適用することにより，従来手法に比べて低コストでありながら，ほぼ同等の性能を有する質問応答システムを作成する方法を提案した．加えて，質問応答システムの性能を改善するために 2 つの研究を実施した．第 1 に，キーワード翻訳の精度を改善するために，既存の小規模対訳辞書を自動的に拡充し，大規模対訳辞書を構築する方法を提案した．第 2 に，回答選択の性能を改善するために，キーワード間の機能的関係を解析する方法について検討した．

研究成果の概要（英文）：In order to realize a multilingual question answering system, several monolingual question systems will be required as its components. However, rule based approach is too expensive to prepare several monolingual question systems. Therefore, machine learning approach is employed to prepare them in this study, instead of rule based approach. We showed that machine learning approach is effective to prepare question answering systems in cheaper costs than rule based approaches with competitive performance. In order to improve their performances, two researches were conducted: large scale bilingual dictionary to improve keyword translation performance, and analysis of functional relationship between keywords to improve answer selection performance.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,100,000	0	1,100,000
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,000,000	300,000	1,300,000
年度			
年度			
総　計	3,200,000	630,000	3,830,000

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：自然言語処理

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

近年の計算機の普及と大規模ストレージの低価格化にともない，個人の利用できる情報は爆発的に増加しつつある．しかし，一人の人間が参照し，利用できる情報には限りがあるため，大量の情報利用を支援する技術の開発が喫緊の課題となっている．そのような技術としては，大量の文書中からユーザの情報要求に適合する文書を検索・提示する**情報検索**や，大量の文書中からユーザの質問に対する回答を発見・提示する**質問応答**などが研究されている．

また，ネットワークの発達に伴って，母国以外の言語で記述された情報に接する機会も増え続けているため，複数の言語を横断した情報利用支援技術も必要とされている．例えば，日本語を母国語とするユーザの情報要求に対して，適合する回答を英語の文書集合から発見・提示するような**日英言語横断質問応答**などが研究されている．質問文を直接機械翻訳した結果を用いて質問応答を行うには，機械翻訳の精度が問題になることが多いため，多くのシステムでは，質問文から抽出されたキーワードを翻訳して得られた翻訳キーワード集合を用いて，英語文書集合から回答を含む文書を検索するという手法をとっている．

しかし，このような言語横断質問応答システムには２つの欠点がある．第１に，適当な回答が文書集合中に含まれていない場合に対する対応が十分ではない．第２に，質問文を含まれる単語の集合として処理しており，単語間の関係に関する配慮が少ないため，複雑な質問には対応できない．第１の問題に対応するには，複数の言語で記述された文書集合を対象として動作する多言語横断質問応答システムが必要である．第２の問題に対応するには，質問文中に含まれる単語間の機能的関係の解析手法の研究が必要である．特に，日本語においては，よく似た機能的関係を表現する場合であっても，多数の表現が用いられ得るため，そのような多数の表現を取り扱う方法が必要である．

２．研究の目的

先に述べた問題を解決するため，本研究では，ある言語で記述された自然な質問文を入力として取り，単語間の関係を考慮して質問文を翻訳し，複数言語で記述されている文書集合を参照して回答を提示する**多言語横断質問応答システム**を研究する．

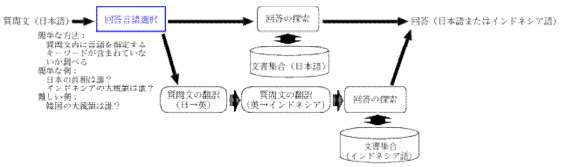
３．研究の方法

多言語横断質問応答システムとは，ある１つの言語で記述された自然な質問文を入力とし，複数の言語で記述されている文書集合を参照として回答を提示するシステムである．例として，以下のような質問文を考える．

- 日本の総理大臣は誰?
- アメリカの大統領は誰?
- ワシントン州の知事は誰?

これらの質問に対して，日本語の文書集合(例えば，新聞)だけに基づいて回答するシステムを考える．このシステムは，日本語の新聞でも大きく扱われているような事象については正しく回答することができると期待されるが，日本語の新聞では殆ど扱われていないような事象については正しく回答することは期待できない．この問題を解決するには，日本語の文書集合だけでなく，英語の文書集合をも同時に考慮することができるような多言語横断質問応答システムが有効と考えられる．

そこで本研究では，複数の単言語質問応答システムを組み合わせることにより，多言語横断質問応答システムを構築する方法を研究する．模式図を以下に示す．



このようなシステムを実現するためには，複数の単言語質問応答システムを用意する必要がある．従来の質問応答システムは，様々な人手作成の規則に基づいているため構築コストが高価であり，複数の単言語質問応答システムを用意することは現実的ではない．そこで，本研究では，従来は人手規則によって実現されていた部分に機械学習手法を適用することによって，複数の単言語質問応答システムを用意する．

実際に機械学習手法を適用した質問応答システムを作成したところ，従来の人手規則に基づくシステムと比較して，ほぼ同等の性能を得た．これは，従来の人手規則に基づく方法では，様々な経験的知識が組み込まれているのに対して，機械学習手法に基づくシステムでは，かなり表層的な知識しか組み込まれていないことが原因だと考えられる．そこで本研究では，２つの方向での検討を行う．第１に，キーワードの翻訳精度とカバー率を高めるため，既存の小規模対訳辞書を自動的に拡充して，大規模対訳辞書を自動的に作成する方法を研究する．小規模対訳辞書に既登録の語を手がかりとして，未登録語の文脈的類似度を比較し，適切な訳語の獲得を行う．第２に，キーワード間の機能的関係性を明らかにして，より深い言語的知識を組み込む方法を研究する．特に，日本語においては，よく似た機能的関係を表現する場合であっても，多数の表現が用いられ得る．そのため，複数の類似した機能表現を１つのグループにまとめて取り扱うことにより，多数の表現を

取り扱う方法を研究する．

４．研究成果

多言語の質問応答システムを実現するには，複数の単言語を対象とする質問応答システムを組み合わせる必要がある．従来の質問応答システムの多くは，人手によって作成された規則などを多く用いているが，そのような方法では，複数の質問応答システムを効率よく作成することはできない．そのため，申請者らは，機械学習手法を用いて質問応答システムを作る方法を検討した．提案システムは，以下の 4 段階の処理からなる．

① 質問文解析：入力された質問文を解析し，回答タイプと質問キーワードを明らかにするモジュール．例えば，「アメリカの大統領は誰ですか」という質問文が入力された場合，回答タイプは「人名」，質問キーワードは「アメリカ」「大統領」の 2 語である．
② 文書検索：回答を含む可能性がある文書集合から，質問キーワードを含む文書を検索する．
③ 回答候補抽出：検索された文書を対象として固有表現抽出を行い，回答タイプと一致する固有表現を列挙する．
④ 回答選択：複数の回答候補に対して順位付けを行い，実際に回答である回答候補を選択する．

これらの処理の内，①質問文解析，③回答候補抽出，④回答選択の 3 つの処理を，機械学習手法として SVM を用いて実現した．機械学習手法を適用するには，適切なラベルが付与された訓練データが必要である．そこで，申請者らは，インドネシア語と日本語の 2 言語対訳質問文コーパスを構築し，このコーパスに対して質問文タイプを付与して，①質問文解析の訓練データとした．③回答候補抽出には，IREX 実行委員会によって作成された固有表現ラベルつきコーパスを用いた．④回答選択のためには，NTCIR QA トラックによって作成された回答データに対して，適切な回答候補位置のラベル付与を行って訓練データとした．このように，従来の質問応答システムにおいて人手規則を必要としていた箇所を全て機械学習手法により実現することにより，様々な言語を対象として容易に質問応答システムを作成できるようにした．実際に，提案手法により，インドネシア語に対する質問応答システム，および，インドネシア語-英語を対象とする言語横断質問応答システムが作れることを示した．ただし，提案手法によって作成したインドネシア語-日本語を対象とする言語横断質問応答システムは，先行研究と比べて性能が良くなかった．

言語横断の質問応答システムを作成するには，その言語対に対する大規模対訳辞書が不可欠である．しかし，世界中にはマイナーな言語も多く存在するので，そのような大規模対訳辞書は常に利用可能であるとは限らない．そこで，小規模な対訳辞書を大規模に自動的に拡充する手法を提案した．本手法は，対象となる言語対（A→B）について，容易に入手できる言語 A の単言語コーパス，言語 B の単言語コーパスおよび小規模な A→B 対訳辞書のみを用いる．言語 A の単言語コーパス上で獲得された文脈ベクトルを，小規模対訳辞書を用いて翻訳し，言語 B の単言語コーパス上での文脈ベクトルと比較することにより適切な訳語対を獲得する．実験により，3,000 語程度の小規模対訳辞書があれば，自動拡充が可能であること，および提案手法によって作成された辞書が，言語横断情報検索タスクにおいて人手によって作成された大規模対訳辞書と同等に有効であることを示した．

質問応答システムを実現するには，そのサブタスクとして固有表現抽出が重要である．固有表現抽出には，教師有り機械学習が有効であることが既に知られている．しかし，現実世界では常に新規な固有表現が生まれ続けているので，常に豊富な学習データが利用できることは期待できない．申請者らは，毎日新聞記事 95 年，96 年，98 年および 2005 年の記事の一部を対照として人手による固有表現ラベル付与作業を行い，このデータに基づいて固有表現出現傾向の経年的な変化の調査した．ある年の新聞に出現した固有表現の 70%から 80%が翌年にも出現すること，言い換えれば，20%から 30%は 1 年前の新聞には出現していない未知固有表現であることなどを明らかにした．この結果より，そのような未知固有表現に対しても有効な手法が必要であることは明らかである．そこで，少量の固有表現ラベル付き学習データと大量の固有表現ラベル無しデータを併用した半教師有り機械学習手法の適用を検討した．固有表現解析の入力文には，少量の固有表現ラベル付きデータにとっての頻出語と非頻出語（未知語を含む）が出現する．この内，非頻出語に対して，大量の固有表現ラベル無しデータに基づいて文脈ベクトルを求めて，その文脈ベクトルの観点で最も類似した頻出語を素性として追加する．このような処理を行うことにより，少量の固有表現ラベル付きデータだけを用いていては，正確に抽出できないような場合でも抽出できるようになることが期待される．実験では，少量の固有表現ラベル付きデータにとって非頻出であるような固有表現の抽出精度を 55.7%から 65.4%に改善できた．

単語間の機能的関係を明らかにするための研究を実施した．日本語には，「について」などのように複数の単語がひとかたまりと

なって非構成的かつ機能的な意味を持つようになり，文の構造に対して重要な意味を持つ表現が多数存在する．しかも，これらの表現は「について」「についての」「に関して」「に関しての」「に関する」のように，基本となる表現から派生する表現が多数に上るという特徴がある．これらを効率良く扱うために，基本となる表現（約 1,000）と派生表現に分けて処理する方法を研究した．

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３件）

① 土屋雅稔，脇田敏行，Ayu Purwarianti，中川聖一，中間言語を用いたインドネシア語-日本語対訳辞書の拡充，自然言語処理，査読有，Vol.15，No.5，pp.23-43，2008.
② Ayu Purwarianti，Masatoshi Tsuchiya，Seiichi Nakagawa，A Machine Learning Approach for an Indonesian-English Cross Language Question Answering System，IEICE Transaction on Information and Systems，査読有，Vol.E90-D，No.11，pp.1841-1852，2007.
③ Ayu Purwarianti，Masatoshi Tsuchiya，Seiichi Nakagawa，Indonesian-Japanese Transitive Translation using English for CLIR，Journal of Natural Language Processing，査読有，Vol.14，No.2，pp.95-123，2007.

〔学会発表〕（計８件）

① 長坂泰治，宇津呂武仁，松吉俊，土屋雅稔，階層的機能表現辞書に基づく日本語機能表現の分析と検出，言語処理学会第16回年次大会，pp.970-973，2010.
② 長坂泰治，坂本明子，宇津呂武仁，森下洋平，松吉俊，土屋雅稔，階層的機能表現辞書に基づく新聞記事中の機能表現の調査・分析，NLP若手の会，2009.
③ Masatoshi Tsuchiya，Shoko Endo，Seiichi Nakagawa，Analysis and Robust Extraction of Changing Named Entities，2009 Named Entities Workshop，pp.161-167，2009.
④ 遠藤翔子，土屋雅稔，中川聖一，固有表現の経年変化と頑健な抽出，言語処理学会第15回年次大会，2009.
⑤ Masatoshi Tsuchiya，Shinya Hida，Seiichi Nakagawa，Robust Extraction of Named Entity Including Unfamiliar Word，ACL2008，pp.125-128，2008.
⑥ 土屋雅稔，肥田新也，中川聖一，非頻出語に対して頑健な日本語固有表現の抽出，情報処理学会研究報告，Vol. 2008-NL-46，pp.1-6，2008.
⑦ Ayu Purwarianti，Masatoshi Tsuchiya，Seiichi Nakagawa，A transitive translation for Indonesian-Japanese CLQA，情報処理学会研究報告，Vol.2007-NL-182，pp.93-100，2007.
⑧ Masatoshi Tsuchiya，Ayu Purwarianti，Toshiyuki Wakita，Seiichi Nakagawa，Expanding Indonesian-Japanese Small Translation Dictionary Using a Pivot Language，ACL2007，pp.197-200，2007.

６．研究組織
(1)研究代表者

土屋雅稔（TSUCHIYA MASATOSHI）
豊橋技術科学大学・工学部・助教
研究者番号：70378256