様式 C-19

科学研究費補助金研究成果報告書

平成２３年　４　月　８　日現在

機関番号：１３５０１
研究種目：基盤研究（C）
研究期間：　２００８　～　２０１０
課題番号：２０５００１２８
研究課題名（和文）　語義に基づく日英関連文書の抽出と続報記事判定への適用

研究課題名（英文）　Retrieving Bilingual Documents based on Word Sense and its Application to Topic Tracking

研究代表者
福本　文代（FUKUMOTO FUMIYO）
山梨大学・　大学院医学工学総合研究部・教授
研究者番号：６０２６２６４８

研究成果の概要（和文）：インターネットの普及により，膨大かつ多様な情報がネットワーク上に溢れている．このような状況において，ユーザが指定した出来事に対し，その発生から後の経過を示す一連の内容を提示する技術（続報記事抽出）は，情報活用としての知的アクセス基盤を提供するだけでなく，予測発見の技術としても期待されている．本研究では，意味を考慮することでユーザが指定した出来事に対する一連の記事を高精度で抽出・提示することを目的とする．具体的には，日英報道記事に焦点をあて，各記事に対し，(1) 話題，及び話題の推移を示す名詞と動詞に注目し,それら品詞単語の多義を解消した後,同義クラスに置き換える，(2) 照応解析を行った後，不要な文を削除し，各記事を話題を示す文のみで示す，(3) 文で表現された日英の各記事を比較することで日英関連記事を抽出する，(4) (3) の結果を用いてユーザが指定した出来事に関する一連の記事を高精度で抽出し，提示する手法を提案した．

研究成果の概要（英文）：With the exponential growth of information on the Internet, it is becoming increasingly difficult to find and organize relevant materials. Topic tracking is a research to attack the problem. One of the major problems in the tracking task is how to make a clear distinction between a topic and an event. A wide range of statistical and machine learning techniques have been applied to topic tracking. However, one encounters quite a large number of referring expressions and ambiguous word senses. We proposed a topic tracking approach based on semantic analysis, especially we focused on word sense disambiguation, overt pronoun resolutions and retrieving relevant documents by using cross-language category hierarchies.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	２，２００，０００	６６０，０００	２，８６０，０００
２００９年度	９００，０００	２７０，０００	１，１７０，０００
２０１０年度	５００，０００	１５０，０００	６５０，０００
年度			
年度			
総　計	３，６００，０００	１，０８０，０００	４，６８０，０００

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：多義解消，クラスタリング，照応解析，続報記事抽出，多言語コーパス，教師なし学習

１．研究開始当初の背景　　インターネットの普及により，膨大かつ多

様な情報がネットワーク上に溢れている．このような状況において，ユーザが指定した出来事に対し，その発生から後の経過を示す一連の内容を提示する技術（続報記事抽出）は，情報活用としての知的アクセス基盤を提供するだけでなく，過去の事例から将来起こりうる問題を予測し事前に対処するための知識発見の技術として，産業界における様々な分野での利用が期待できる．続報記事の抽出に関する研究は，統計や機械学習を用いて情報を示すタグが付与された少数から成る事例の特徴抽出を行う手法が主流となっている．しかし抽出対象となるデータは，新聞やニュース記事など，異種の文書が混在していること，また時系列データであるため，時間の経過とともに話題が刻々と変化することから，少数事例のみを用いた学習法では制度面で限界があり，多様なコンテンツを扱う現実世界において十分対処可能な枠組みとはいえない．この問題を解決するためには，各事例の話題を正確に捉える必要があり，意味を中心に据えた言語処理技術が必須となる．

２．研究の目的

本研究では，英文報道記事に対し，意味を考慮することでユーザが指定した出来事に関する一連の記事を高精度で抽出・提示することを目的とする．具体的には，日英報道記事に焦点をあて，各記事に対し，(1) 話題，及び話題の推移を示す名詞と動詞に注目し，それら品詞単語の多義を解消した後，同義クラスに置き換える，(2) 照応解析を行った後，不要な文を削除し，各記事を話題を示す文のみで示す，(3) 文で表現された日英の各記事を比較することで日英関連記事を抽出する，(4) (3) の結果を用いてユーザが指定した出来事に関する一連の記事を高精度で抽出し，提示する．

３．研究の方法

（１）多義解消

本研究では，分類のための一手法として，Reichardt らにより提案されたグラフベースの教師なしクラスタリング手法(RB アルゴリズム)を用いる．RB アルゴリズムは磁性体分類のためのソフトクラスタリング手法であり，エネルギーが最小になるように磁性体を分類する．我々はこの手法を動詞の意味分類に適用した．一般に,意味的に類似した動詞は同じ格構造を持つことから,動詞を格構造パターンを次元とするベクトルで表現し，ベクトル同士の分布間類似度に基づき，クラスタリングを適用する手法が多く用いられている．しかし,コーパスなどから抽出した動詞は，多数の格構造パターンをもつことから，結果的に高次元空間での分類となるため，クラスタリングの精度に悪影響を与える場合が多い．我々はこの問題に対処するため，リンク解析を利用することで，動詞の意味分類に必要となる格構造パターンのみを抽出し，これを用いて動詞の意味分類を行う手法を提案した．

（２）代名詞の照応解析

申請者は，これまでタイトルと固有表現の中で，機関，人名，固有物名が話題と強く関係していることを明らかにした．本研究で対象とするＴＤＴは，ラジオ放送を含む報道記事であるため，タイトル情報は利用できない．そこでこれら固有表現に加え，新たに照応解析に注目した．照応解析では，代名詞の先行詞を正確に同定するため，代名詞を含む文から得られる代名詞の取り得る意味素性と各候補となる先行詞の意味素性との関連の度合いを学習することで，意味の粒度を考慮した同定手法を提案した．

（３）日英関連記事の自動抽出

本研究では，文書分類手法をもちいることで分野間の類似性を推定し，階層構造を統合する手法を提案した．さらに統合された階層構造に位置する英語と日本語の記事に対して類似度尺度を用いることで互いに関連する日英文書を抽出する手法を提案した．本研究における関連文書抽出の精度は，階層構造の統合精度に依存する．申請者は，文書を分類するための手法として機械学習 SVMs を用いた．また，類似性を推定するためにカイ 2 乗を用いた．さらに，類似分野対として得られた分野の各組に対して，アプリオリアルゴリズムを適用することで類似分野対の精度向上をねらった．

（４）続報記事抽出

続報記事抽出では，対象となる記事集合に対してこれまで提案した（１）多義解消，（２）代名詞の照応解析，（３）日英関連記事の自動抽出を行った結果を適用した結果に対して続報記事の抽出を行った．記事抽出では，Allan らが提案した Adaptation 手法をそのまま適用した．

４．研究成果

（１）多義解消

実験では 1991 年から 2007 年までの 17 年間の毎日新聞を利用した．全ての記事に対して構文解析 Cabocha を用いて構文解析を行い，その結果から計算機用日本語基本動詞辞書 IPAL に記載されている 2,042 語とその格構造パターンを全て抽出した．実験では 2,042 語のうち，出現頻度の高い上位 500 語の動詞を抽出した．IPAL の語彙数は 1,081 語

と少ないため，500 語に対する格構造パターンの格要素の意味素性は，EDR で定義されている意味素性を用いた．その結果，総計 29,690 の格構造パターンが得られた．リンク解析を用い，これらのパターンの中から有用な格パターンを抽出した．

評価用のデータは IPAL 辞書を用いて作成した．IPAL 辞書は，2,042 語の動詞を総計 2,814 のクラスに分類している．本研究では，2,814 クラスを 2 つのセット，すなわち RB アルゴリズムで用いられているパラメータを推定するための訓練データと意味分類のためのテストデータに分けた． 動詞分類の評価尺度として，再現率，適合率，F 尺度を用いた．また比較手法として，ソフトクラスタリング手法の中の代表的な手法の一つである EM (Expectation Maximaization)を用いた．EM における分布確率は，動詞の格構造パターンを用いて求めた．また確率の初期値は,k-means を用いて求めた．モデル学習の繰り返し数は 50 回とした． 実験の結果，本手法はＦ値 0.51 であり，関連研究である EM(0.40)よりも優れた精度が得られることが確認できた．

（２）代名詞の照応解析

実験では，ＴＤＴ３ English コーパスを用いた．ＴＤＴ３コーパスは ABC ニュースなど 8 種類の報道機関から成り，1998 年の 10 月から 12 月までの 34,600 記事で構成されている．本研究では，これら 8 種類の報道機関から各 10 記事を無作為に抽出し，代名詞の照応解析を行った．これらの記事中，代名詞は 1,082 種類存在した．平均精度は 62.8%であった．精度は，各報道機関により異なる．もっとも高い精度が得られた報道記事は，NYT であり 75.6%であった．一方，もっとも精度が低いものは，CNN であり 46.2%であった．実験の結果から，ニュース記事よりも TV やラジオニュースの方が精度は低かった．理由として，後者は，インタビュー記事で構成されているものが多いため，cataphora や formal subject が多く含まれていることが考えられる．また書き言葉ではないため、構文解析結果の誤りがあるため，格情報を誤って抽出したことも原因として考えられる．

（３）日英関連記事の自動抽出

実験では Reuters’96 と UDC コード階層を用いた．Reuters’96 は 121 種類のカテゴリィからなり，UDC は 9,951 種類のカテゴリィからなる．本研究ではこれらのうち，Reuters’96 は 102, UDC は 4,739 からなるカテゴリィを用いて実験を行った．実験では，本手法である階層構造を用いた分野対の抽出精度の効果を検証するため，階層構造を利用しない手法との比較を行った．実験の結果，本手法は,0.482 のＦ値が得られたのに対し，階層構造を用いない手法では 0.422 であったことから階層構造の有効性が確認できた．関連文書抽出の実験では，本手法の有効性を検証するため，階層構造を用いずに，日英文書の類似度により関連文書を抽出する手法との比較を行った．実験の結果，本手法はＦ値が 0.689 であったのに対し，baseline では 0.363 の精度であったことから，本手法の有効性を確認することができた．

（４）続報記事抽出

ＴＤＴ３コーパスを用い、多義解消を行った結果と行わずに Adaptation 手法を適用した結果との比較を行った．多義解消を行った結果，Macro-averaged F-score 0.530 に対して多義解消を適用しない結果は 0.480 であった．

照応解析を用いた実験では，訓練データが 4 記事という少数データに対して，照応解析を適用した結果を用いて，続報記事を抽出した結果，Macro-averaged F-score が 0.59 であった． 一方，照応解析を用いずに続報記事を抽出した結果は 0.480 であった．

日英関連文書を抽出した結果については，関連する日本語文書を機械翻訳を用いて英語文書に翻訳した結果を続報記事抽出の対象となる訓練文書に追加したものと追加しないものとの比較を行った．実験の結果，訓練データが 4 記事に対して，前者は Macro-averaged F-score が 0.595 であるのに対して後者は 0.480 であったことから，本手法の有効性が確認できた．さらに関連研究との比較として，UMass の relevance model と比較した結果,本手法の方が優れた DET curve が得られることを確認した．実験の結果から，日英関連文書抽出が続報記事抽出にもっともよい効果を与えることが明らかになった．一方，多義解消の精度差は 6%と低く，多義解消単独では，続報記事抽出に効果があるとは言えないことも確認できた．

５．主な発表論文等

〔雑誌論文〕（計 13 件）

① F. Fukumoto and Y. Suzuki and K. Yamashita, “Polysemous Verb Classification using Subcategorization Acquisition and Graph-based Clustering”, Human Language Technologies, Lecture Notes in Computer Science (*To Appear*), 2011, 査読有.

② F. Fukumoto and Y. Suzuki, “Identifying Domain-Specific Senses and its Application to Text Classification”,

Proc. of 2nd International Conference on Knowledge Engineering and Ontology Development, pp. 59-64, 2010, 査読有.

③ F. Fukumoto, A. Sakai and Y. Suzuki, “Eliminating Redundancy by Spectral Relaxation for Multi-Documents Summarization”, Proc. of 2010 Workshop on Graph-based Methods for Natural Language Processing, pp. 98-102, 2010, 査読有.
④ F. fukumoto and K. Yamashita and Y. Suzuki, “Classifying Polysemies using a Graph-based Clustering”, Proc. of the 4th Language and Technology Conference, pp. 210-214, 2009, 査読有.
⑤ F. Fukumoto and Y. Suzuki, “Effect of Overt Pronoun Resolution in Topic Tracking”, Proc, of the 4th Language and Technology Conference, pp. 350-354, 2009, 査読有.
⑥ N.A.H.B.Zahri and F. Fukumoto, “Example-assignment to WordNet Thesaurus based on Distributional Similarity of Words and its Application to Word Sense Disambiguation”, Proc. of 11th Conference of the Pacific Association for Computational Linguistics, pp. 188-193, 2009, 査読有.
⑦ F. Fukumoto and Y. Suzuki, “Using Graph-based Indexing to Identify Subject-Shift in Topic Tracking”, Human Language Technologies, Lecture Notes in Computer Science, Vol. 5603, pp. 392-404, 2009, 査読有.
⑧ F. Fukumoto and N. A. H. B. Zahri and Y. Suzuki, “Example-assignment to WordNet Thesaurus based on Distributional Similarity of Words”, Proc. of Workshop on Matching and Meaning, pp. 1-5, 2009, 査読有.
⑨ 市岡健一, 福本文代, “Web から取得した共起頻度と音象徴によるオノマトペの自動分類”, 電子情報通信学会論文誌, J92-D, No. 3, pp. 428-438, 2009, 査読有.
⑩ F. Fukumoto and Y. Suzuki, “The Effect of Combining Similarity Measures for Onomatopieia Clustering”, Proc. of Empirical Method in Asian Natural Language Processing, pp. 37-51, 2008, 査読有.
⑪ F. Fukumoto and K. Ichioka, “Graph-based Clustering for Semantic Classification of Onomatopoetic Words”, Proc. of the 3rd Textgraphs Workshop on Graph-based Algorithms in Natural Language Processing, pp. 33-40, 2008, 査読有.
⑫ F. Fukumoto and Y. Suzuki, “Retrieving Bilingual Verb-Noun Collocations by Integrating Cross-Language Category Hierarchies”, Proc. of the 22nd International Conference on Computational Linguistics, pp. 233-240, 2008, 査読有.
⑬ F. Fukumoto and Y. Suzuki, “Integrating Cross-Language Hierarchies and its Application to Retrieving Relevant Documents”, ACM Trans. of Asian Language Information Processing, 7(3), pp. 1-22, 2008, 査読有.

６．研究組織

(1)研究代表者

福本　文代　　(FUKUMOTO FUMIYO)

研究者番号：６０２６２６４８

(2)研究分担者

なし

研究者番号：

(3)連携研究者

なし

研究者番号：