様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　１０　月　１４　日現在

機関番号：　１３５０１
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2015
課題番号：　２５３３０２５５

研究課題名（和文）時系列文書を対象とした語義に関する局所・大域的特徴量の抽出と続報記事判定への適用

研究課題名（英文）Local and Global Feature Extraction for Senses and its application to Topic Tracking

研究代表者

福本　文代（FUKUMOTO, Fumiyo）

山梨大学・総合研究部・教授

研究者番号：６０２６２６４８

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：本研究は，長期間に渡る時系列文書データを対象とした検索に有効な語彙的意味処理技術の開発を目的とする．具体的には，(1) 分野語義辞書を開発し，(2) 時系列モデルに基づき語義の局所・大域特徴量を抽出することにより，意味に基づく時系列データ処理を実施した．またこれらを用いることで，訓練データと作成時期が異なるテストデータを高精度で分類することが可能となることを示した．

研究成果の概要（英文）：This study proposed a method for lexical semantic extraction which is effective for topic tracking and text categorization that training data may derive from a difference time period from the test data. We present a method that minimizes the impact of temporal effects by using term smoothing and transfer learning techniques. The results showed that integrating term smoothing and transfer learning improves overall performance of topic tracking and text categorization, especially it is effective when the creation time period of the test data differs greatly from the training data.

研究分野：　自然言語処理

キーワード：　分野語義辞書　転移学習　素性選択　文書分類　続報記事抽出

１．研究開始当初の背景

近年，WWWの爆発的な普及を背景に，特定の話題に関するバースト性，すなわちある話題がどの時期に集中的に報道されているかを分析したり，ユーザが指定した出来事に関する一連の内容を検索･提示する続報記事抽出に関する研究が盛んに行われている．これらの研究は，確率モデルや機械学習に基づく手法が主流である[Allan’03, Larley’04, Leskovec’09, He’10, Tang’13]．しかしそれらの多くは意味を排除した枠組みとなっているため，精度面で課題が残る．

語の意味を考慮した初期の研究としては，YangやAllan の研究がある[Yang’94, Allan’98]．彼らは,辞書情報を利用した単語の類推や構文解析などの言語処理を積極的に利用することで続報記事の抽出を試みた．しかし，いずれの場合にも大幅な精度の向上がみられなかったことから，言語処理の利用は必ずしも有効ではないと結論づけている．以来，テキストマイニング，あるいは情報検索分野においても依然，語の表層に基づき，新たな確率･統計手法や機械学習法を適用した研究が主流となっている．

2000 年以降，文書はいくつかの潜在的なトピックで表現されていると仮定し，時系列データをトピック集合で集約する潜在的意味解析(LDA, Blei’03) が脚光を浴びるようになった．その後，LDAを用いて話題のバースト性を解析する研究が国内外で多くなされている[Alsumait’09, Caballero’13]．BleiらはLDA を拡張し，文書集合中の時系列情報を考慮したDynamic Topic Models(DTM) を提案した
[Blei’06]．高橋らは,DTMとKleinberg の確率モデル[Kleinberg’02] を組み合わせたバースト解析手法を提案している[高橋’12]．しかし，いずれの手法も各トピック集合の要素であるキーワードは単語の表層情報であり，各トピック，及びキーワードの意味的解釈は人手により行っているため，時系列データから抽出された単語が必ずしもその時期に注目された話題となっていない場合や，時間の経過とともに内容が刻々と変遷していくような場合にはバーストが正しく検出できないという問題が残る．

２．研究の目的

本研究は，長期間データ系列を対象としたバースト解析や続報記事を高精度で抽出するためには，意味を中心に据えた自然言語処理技術が必要不可欠であるという主張のもとに，長期間に及ぶ時系列データを対象とした検索に有効な語彙的意味処理技術を開発することを目的とする．具体的には，(1) 分野語義辞書を開発し，(2) 時系列モデルに基づき語義の局所･大域特徴量を抽出することで，時系列データに対する意味処理を実施する．またこれらが意味処理において有効な知識ベース言語モデルであることを検証するため，これらを用いることにより訓練データと作成時期が異なるテストデータを高精度で抽出･提示できることを示す．さらに，時系列データを対象とした意味処理において適切な意味の粒度，及び異なる言語体系(日本語と英語) における分野語義の類似･相違点についても明らかにする．

３．研究の方法

本研究は3つの課題から成る．第１の課題は，分野語義辞書の開発である．分野語義辞書は，既存の辞書の語義にその語義が頻繁に使用される分野名を自動的に付与した辞書であり，話題と背景を同定するために用いる．研究代表者は，文書の話題語，及び背景語の語義は，その文書が属する分野に依存して決まることに注目した．例えば，“テキサスの洪水” に関する一連の報道記事が“災害” という分野に属していることがわかっているとする．この報道記事中，“Relief help has been granted” の“relief” はこの文書の話題語と捉えることができる．多義解消の結果，Relief help は救済/救助の意味であると判

定されたとする．一方，WordNet に記載されているrelief の語義中，上記の語義が災害の分野で使用されることが明記されていれば，relief help は災害の分野において話題語であると判断することができると考えた．そこで，既存の辞書と大量の分野ラベル付きテキストコーパスを利用することで，辞書に記載されている語義に分野を付与する手法を提案した．

第2 の課題は，局所･大域特徴量の抽出である．TDTコーパスの調査から，背景語は長期間重要であり，その出来事が勃発した時期は密集して出現するが，時間の経過と共に疎となる場合が多いこと，話題語は，短期間にのみ頻出することが明かになった．例えば95年に起きた神戸の地震に関する一連の記事において，勃発当時”earthquake” は頻出しているが，半年後の被災者の医療問題に関する記事では，2 回のみである．一方，医療問題の記事において話題語である”medical” は神戸地震に関する一連の記事の中で，この記事にのみ頻出している．さらに同一分野に属する2 種の出来事に関する記事系列から抽出した話題語(背景語) の分布は類似傾向にある．そこで，続報記事に関する訓練データを用い，話題語(背景語) が出現した時期以降の時間差と重要性の度合いを推定することで，出来事の推移が把握できるのではないかと考えた．具体的には，転移学習の一つであるTrAdaBoost を用い，重要性の度合いを学習する手法を提案した．

第3 の課題は，文書分類である．課題2 で得られた学習法を用い，時系列記事中の各記事を高精度で分類する手法を提案した．

４．研究成果

第１の課題では，先ず，分野ラベル付きコーパスとしてReuters’96，及び毎日新聞96，97 年を用い，各分野ごとに名詞単語を抽出した．次に辞書としてWordNet，EDR を用い，各分野ごとに，名詞の各語義をノード，語義同士の類似度をエッジとするグラフを作成し，固有値計算を用いることで，語義のスコアリングを行うことで，各分野の主要語義を求めた．この成果については現在，論文を執筆中である．

第2，及び第3の課題である訓練データと作成時期が異なるテストデータの分類については，局所・大域的特長量（素性）を抽出した後，これらを用い，TrAdaBoostを適用することで分類器を作成，テストデータを分類する手法を提案した．TrAdaBoostは転移学習の一つであり，訓練データとは異なる分野のテスト事例を分類するために考案された手法である．本研究はこれを時系列データに適用することにより分類を行った．

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 F. Fukumoto and Y. Suzuki, “Smoothing Temporal Difference for Text Categorization”, Lecture Notes in Computer Science, Vol. 9460, pp. 203-214, 2015, 査読有.

② F. Fukumoto and Y. Suzuki, “Exploiting Guest Preferences with Aspect-based Sentiment Analysis”, Communications in Computer and Information Science, vol. 553, pp. 34-49, 2015, 査読有.

③ F. Fukumoto and Y. Suzuki, “Identifying Event and Subject of Continuous News Streams for Multi-Document summarization”, New Advances in Human Language Technologies, Z. Vetulani and H. Uszkoreit (eds.), Springer, 2014, 査読有.

〔学会発表〕（計　5　件）

① F, Fukumoto and Y. Suzuki, “Short Text Categorization by Smoothing Word Distribution”, Proc. of the 7th Language and Technology Conference, pp. 95- 99, 27th Nov, 2015, Poznan Poland, 査読有.

② F. Fukumoto and Y. Suzuki,

“Temporal-based Feature Selection and Transfer Learning for Text Categorization”, Proc. of the 7th International Conference on Knowledge Discovery and Information Retrieval, pp. 17-26, 5th Nov, 2015, Lisbon Portugal, 査読有.

③ F. Fukumoto and Y. Suzuki, Learning Timeline Difference for Text Categorization, Proc. of the Conference on Empirical Methods in Natural Language Processing, pp. 799-804, 19th Sept, 2015, Lisbon, Portugal, 査読有.

④ Y. Suzuki and F. Fukumoto, “Detection of Topic and its Extrinsic Evaluation through Multi-Document Summarization”, Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 241-246, 27th Oct, 2014, San Francisco, USA, 査読有.

⑤ F. Fukumoto, Y. Suzuki, and A. Takasu, “Timeline Adaptation for Text Classification”, Proc. of the 22nd ACM Conference on Information and Knowledge Management, pp. 1517-1520, 23rd, Jun, 2013, Baltimore, USA, 査読有.

〔その他〕
ホームページ等
cl.cs.yamanashi.ac.jp

６．研究組織
(1)研究代表者
福本　文代　(FUKUMOTO Fumiyo)
山梨大学・大学院総合研究部・教授
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
鈴木 智弘 (　SUZUKI TOMOHIRO　)
山梨大学・大学院総合研究部・准教授
研究者番号：　70235977