様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１５　日現在

機関番号：　９４３０５
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３０１２６

研究課題名（和文）ハッシュ法を統合した多様で大規模な言語データの解析技術に関する研究

研究課題名（英文）Large-scale text data analysis using hashing techniques

研究代表者
林　克彦（Hayashi, Katsuhiko）

日本電信電話株式会社ＮＴＴコミュニケーション科学基礎研究所・協創情報研究部・研究員

研究者番号：５０７２５７９４

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：多様なドメインの文、文書の内容理解を効率的に行うことは自然言語処理の重要な課題である。ここではハッシュ法や行列分解に基づく手法によりこれらの課題解決に取り組んだ。まず、新聞記事の談話構造を高速、かつ、高精度に解析する手法を考案し、文書要約での有効性を示した。また、音声対話/新聞データを対象とした高速かつ高精度な単語省略補完システムを考案し、崩れた文の整形が行えるシステムを開発した。さらに、単語シソーラスを使った単語の類似度検索を高速、かつ、高精度に行う手法を考案した。これらは文、文書の高度な内容理解を行うもので、今後も自然言語処理の様々なアプリケーションへの応用が期待される。

研究成果の概要（英文）：I investigated hashing and matrix factorization techniques to efficiently analyze large-scale text data in various domains. First, I proposed a fast and accurate parsing algorithm for discourse tree structure analysis of English newswire texts. I also presented a text summarization method using discourse trees, and achieved an improvement in text summarization accuracy. Second, I proposed a method to automatically detect and insert missing elements in English and Japanese speech/newswire texts. Finally, I proposed a knowledge (word thesaurus) embedding method for fast word similarity computation. In future, I will apply these methods to such more advanced NLP applications as machine translation and question answering.

研究分野：　自然言語処理

キーワード：　談話構造解析　省略補完　行列分解　ハッシュ法

１．研究開始当初の背景

(1) 自然言語処理における主要な応用タスクである機械翻訳、文書要約、質問応答を高精度化するには、入力となる文や文書の構造を解析し明らかにすること(省略補完、構文解析、談話解析)が重要である。これらの基盤言語解析技術では多様なドメインのテキストデータを高精度に解析する必要があるため、統計的識別学習モデルによりモデル化するのが一般的である。しかし、識別モデルでは特徴量が増大すると、メモリコストや解析コストもそれに比例して増大するという問題がある。そのため、大規模かつ長文を含むような新聞記事、音声対話、ブログ記事を解析の対象にする場合、その解析処理速度の向上がきわめて重要な課題となっていた。

(2) 単語分散ベクトル表現は単語に N 次元の特徴量を与える。これは基盤言語解析技術の特徴量、単語の類似度検索、質問応答などに利用される。このような分散表現はテキストデータ、あるいは、単語シソーラスデータから獲得されるが、データが大規模であるため、効率的な類似度計算手法及びその計算に基づく分散表現学習アルゴリズムの開発が重要な課題となっていた。

２．研究の目的

(1) 省略補完、構文解析、談話解析などの自然言語処理基盤技術における処理速度を向上させることで、機械翻訳や文書要約などの応用タスクの実用化を促進させることが本研究の目的である。また、Web 上のブログや新聞記事データは日々増えており、このようなテラバイト規模のテキストデータを効率的に解析し、構造化することは情報検索の文脈でも非常に重要となる。

(2) 単語シソーラスデータは一般に知識グラフと呼ばれるものの一種である。知識グラフは普通の辞書データではなく、単語間の関係を詳細に記述した辞書データベースであり、質問応答システムの根幹を成すデータである。質問応答では辞書データベースへの問い合わせをするが、信頼度スコアの高い解答を高速に検索することがもとめられる。そのため、質問応答システムにおいて、データベースに登録されている要素間の類似度を高速に計算する手法は不可欠である。

３．研究の方法

(1) 統計的識別モデルでは特徴量を表す非固定長のキーとその ID を対応付けたハッシュ表を持つ。そのため、特徴量が増大すると、メモリ使用量が増大し、また、非固定長のキーに基づく表へのアクセスに大きな時間を消費する。この課題を解決するため、特徴量ハッシング法を導入し、全ての特徴量を 0～N の有限の ID へハッシュ関数で対応付けることを考えた。この方法では特徴量の衝突が起こり、モデル精度の低下を招く恐れもあるが、上記したハッシュ表を排除できるため、メモリ使用量と ID 検索速度は大幅に向上する。

(2) 知識グラフから単語分散表現を獲得するための新たな行列分解法を考えた。知識グラフにおけるある関係に着目すると、辞書に登録されている単語と単語がその関係を持つかどうかを隣接行列で表現することができる。この隣接行列は一般に非対称行列となるため、実正規行列に対する新たな固有値分解法を考案した。その結果、単語と関係はそれぞれ複素ベクトルを持ち、そのベクトルの次元数に対して線形時間で類似度計算を行うことができるようになる。

４．研究成果

(1) 特徴量ハッシング法に基づく識別学習モデルを使った「談話解析」、「省略解析/構文解析」システムを開発し、従来法よりも解析処理速度が優れる事を実験的に示した。前者は下記学会発表の[1,3]に対応し、後者は[2,4]に対応する。解析速度の向上については、「省略解析/構文解析」システムに対する結果を下図 1 に示した。

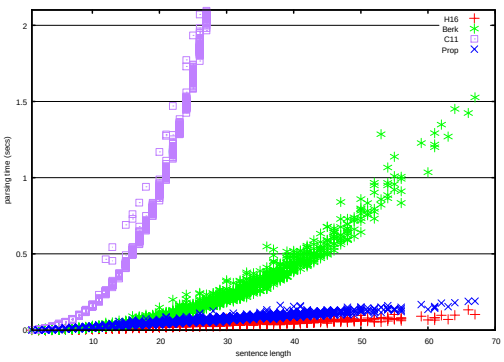

図 1 省略解析の処理速度比較

青線及び赤線が提案法の結果であり、従来法に比べて大幅な解析速度の向上を達成していることがわかる。また、これらの成果に対して問い合わせがあった研究者に、学術利用のみに限定して配布を行った。また、下記 URL には特徴量ハッシング法を組み込んだ依存構造解析システムを公開している。依存構造解析とは構文解析の一種である。

http://cl.naist.jp/~katsuhiko-h/
[1,3]の成果においても同様の解析速度向上を達成しており、また、文書要約に応用した際の精度も従来技術を使った結果と遜色がなかった。このことからここで開発した技術は自然言語処理全般に大きな恩恵をもたらすものと考えられる。

(2) 単語の分散表現学習に関する結果を下図2 に示す。ここでは単語類似度計算の処理速度を示している。

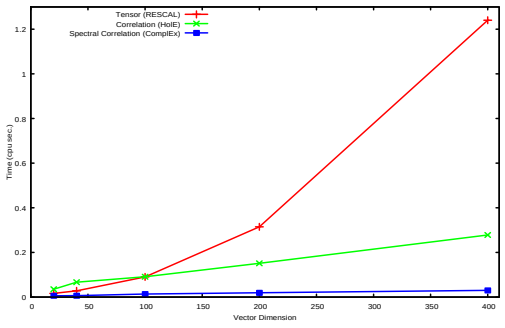

図 2 単語類似度計算の処理速度比較

青線が提案法の類似度計算速度であり、ベクトルの次元数(横軸)が高くなっても従来法より高速であることがわかる。また、検索精度を検証するため、ある単語と特定の関係にある単語をどれだけ正確に検索できるかを検証した。その結果を下図 3 に示す。

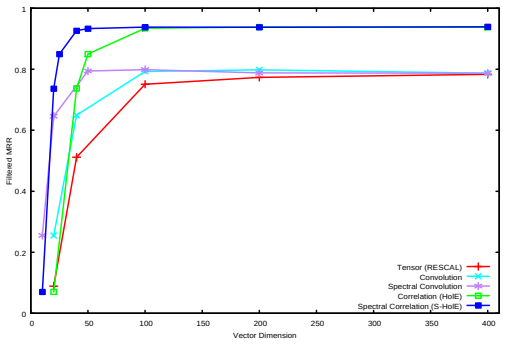

図 3 単語類似度検索精度の比較

ここでは青色の線が提案法の実験結果を示しており、ベクトルの次元数(横軸)が大きな場合、従来法よりも検索精度が高いことがわかる。このような高速かつ高精度な類似度計算手法は確率的関係データベースなどより汎用的な枠組みと統合されることが期待される。

5．主な発表論文等

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 5 件）

(1) 林克彦，平尾努，吉田康久，永田昌明，修辞構造木から自動変換した談話依存構造木の性質について，言語処理学会第 21 回年次大会, pp.369–372，査読無し, 2015 年 3 月.

(2) Katsuhiko Hayashi, Masaaki Nagata, Empty element recovery by spinal parser operations, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp.95–100，査読有り，August，2016.

(3) Katsuhiko Hayashi, Tsutomu Hirao, Masaaki Nagata, Empirical comparison of dependency conversions for RST discourse trees, Proceedings of the SIGDIAL 2016 conference, pp.128–136, 査読有り，September，2016.

(4) Katsuhiko Hayashi, Masaaki Nagata, K-best Iterative Viterbi Parsing, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp.305–310，査読有り，April，2017.

(5) Katsuhiko Hayashi, Masashi Shimbo, On the Equivalence of Holographic and Complex Embeddings for Link Prediction, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 査読有り，July-August, 2017.（発表予定）

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 1 件）

名称：単語学習装置、単語学習方法及び単語学習プログラム
発明者：林克彦、新保仁、永田昌明
権利者：同上
種類：特許
番号：2017039543
出願年月日：2017 年 3 月 2 日
国内外の別： 国内

○取得状況（計 0 件）

〔その他〕
ホームページ等
http://www.kecl.ntt.co.jp/icl/lirg/members/hayashi/index-j.html

6．研究組織

(1)研究代表者
林　克彦（Hayashi, Katsuhiko）

日本電信電話株式会社 NTT コミュニケーション科学基礎研究所・協創情報研究部・研究員
研究者番号：50725794