科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１１　日現在

機関番号：　１４６０３
研究種目：　研究活動スタート支援
研究期間：　2012～2013
課題番号：　２４８０００４１

研究課題名（和文）言語処理のための頑健な語彙表現の機械学習

研究課題名（英文）Learning Robust Word Representations for Natural Language Processing

研究代表者
Ｄｕｈ　Ｋｅｖｉｎ（Duh, Kevin）

奈良先端科学技術大学院大学・情報科学研究科・助教

研究者番号：８０６３７３２２

交付決定額（研究期間全体）：（直接経費）　　2,100,000円、（間接経費）　　630,000円

研究成果の概要（和文）：言語というものは際限なく新しい表現が作れる。しかし、現在の自然言語処理技術は、単語の意味をモデル化していないため、新しい表現に対して正確に取り扱いできない。本研究の目的は、単語の意味をモデル化し、様々なテキストを頑健に扱う自然言語処理技術を目指す。具体的に、単語をそれぞれ連続ベクトルとしてモデル化し、ディープラーニング（深層学習）による最適化の枠組みを開発する。また、大規模テキストデータから学習した単語ベクトルを構文解析や機械翻訳システムに取り込んで、精度を向上させることを実現した。

研究成果の概要（英文）：Language is a highly productive phenomenon; new words and expressions are constantly being invented. Current Natural Language Processing techniques have difficulty handling new words, so their performance on real-world text suffers. To address this, we develop robust models of word semantics, focusing on Deep Learning methods for learning vector word representations. Furthermore, we show improvements in parsing and translation performance using systems that incorporate these word representations.

研究分野：　情報学

科研費の分科・細目：　人間情報学・知能情報学

キーワード：　自然言語処理　機械学習

１．研究開始当初の背景

言語というものは際限なく新しい表現が作れる。言語学では、言語の生産性、創造性と呼ぶ。例えば、近年ソーシャルメディアで「なう」という新しい言葉が一般的になって、様々な用例がある：

「会議なう」
「当選確実なう」
「ラーメンなう」

人間なら、未知語でも用例から意味の推定ができる。この場合、「なう」は「今」の意味を示す：

「今、会議中です」
「今、当選確実になりました」
「今、ラーメンを食べている」

しかし、計算機はこういった未知語は理解できない。現在の自然言語処理技術は、単語の意味をモデル化していないため、未知語に対して正確に取り扱いできない。一方、近年ソーシャルメディアのような一般大衆による情報発信が盛んになり、未知語の問題は増えつつある。自然言語処理技術の実用化のため、頑健な語彙表現モデルは重要な課題になる。

２．研究の目的

本研究の目的は、上で述べたように単語の意味をモデル化し、未知語に含んだ様々なテキストを頑健に扱う自然言語処理技術を開発する。具体的に、単語をそれぞれ連続ベクトルとしてモデル化し、ディープラーニング（深層学習）による最適化の枠組みを研究する。

連続ベクトルの利点は、単語の意味を分割し、単語と単語の関連性を計算できるようになる。図１に示すように、「なう」と「今」のベクトルが近ければ、意味は似ていることを推測できる。如何に大規模テキストデータからベクトルを得られ、そして如何に自然言語処理システムに取り込むのは本研究の課題です。

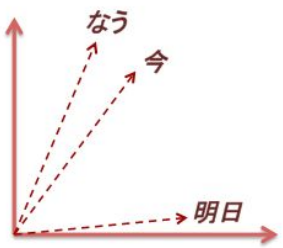

図 １：語彙表現をベクトルとしてモデル化

目標として以下の２つを設定する：

（１） ディープラーニング及びニューラルネットワークによる語彙表現の学習方法を開発する。特に、言語学の知識を生かして、新しいモデルを提案する。

（２） 大規模テキストデータから学習した語彙表現を様々な自然言語処理システムに取り込む。また、未知語を含むテキストで評価する。

３．研究の方法

本研究は、語彙表現の学習方法と使い方、それぞれ並列で開発を行った。

（１）ディープラーニングによる語彙表現の学習方法の開発：これまでの先行研究(Turian2010,Mikolov2011,Collobert2011)はディープラーニングやニューラルネットワークを利用して、大規模テキストデータから単語のベクトルが学習できることを示した。

しかし、先行研究のモデルは、テキストを系列データとして扱うため、言語の本質を考慮していない。そこで、我々のアプローチは従来のモデルの上に、如何に階層的な構成情報や外部の言語知識資源を取り込む方法を考えた。

（２）連続ベクトルの語彙表現を取り込める自然言語処理システムの開発と評価：自然言語処理分野で最も重要なタスクは構文解析と機械翻訳です。我々は、これらのタスクを専念して、連続ベクトルを取り込めるシステムを構築し、未知語を含むテキストデータで評価した。

構文解析の場合は、ブログ記事など最も未知語が多いウェブテキストで実験した。現在、新聞記事の解析精度（係り受け）は９０％に超えるが、ブログ記事の解析精度は８０％以下です。精度が下がる原因は、おそらく未知語と新しい表現の多さだと思われる。この問題を解決するために連続ベクトルを導入する。

機械翻訳の場合は、ニューラルネットワークのモデルを使って、対訳データを集めるタスクで実験した。機械翻訳システム構築する時は、人手で翻訳した対訳データが必要だか、コストを下げるために、様々の対訳データから、本来のタスクに類似しているデータサップセットを選択する。これによって、よりいい翻訳システムを構築することが可能です。このタスクも未知語の問題が発生するので、ベクトルモデルの活用を検討した。

４．研究成果

（１）<u>ディープラーニングによる語彙表現の学習方法の開発</u>：言語学の知識をディープラーニングに取り込んで、二つの新たな語彙表現の学習手法を提案した。一つ目は、WordNetのような知識ベースと大規模テキストを統合する多目的学習手法である。

二つ目は、単語から文までの意味構成性を考慮した語彙表現の再学習手法である。図２のように、英語の「run」は２つ意味を持ち、「run marathon」の場合は「走る」の意味だが、「run company」の場合は「経営」の意味になる。この例を見れば分かるように、複数意味持つ単語は一つのベクトルでモデル化するのは限界がある。我々は、動詞と目的語の関連性を考慮し、動的に正しいベクトルを計算するモデルを提案した。

これらは、言語知識を取り組む込むことにより、従来よりも適切な語彙表現を得ることができたため、先行研究の手法を改善したことを確認した。

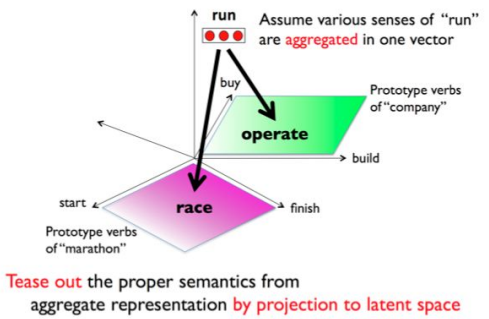

図 2：意味構成性を考慮したモデル

（２）<u>連続ベクトルの語彙表現を取り込める自然言語処理システムの開発と評価</u>：構文解析の実験では、様々なモデルとウェブテキストの組み合わせを試した。連続ベクトルを導入するシステムは、一般的に（係り受け）解析精度が上がることが見られた。導入の仕方により精度の上がりは異なるが、連続ベクトルは頑健な語彙表現を確認した。

また、機械翻訳システムに適用し、精度を向上させることを実現した。図３に示すように、ベクトルモデルで集めたデータは従来法より精度高い翻訳システムを構築できる。

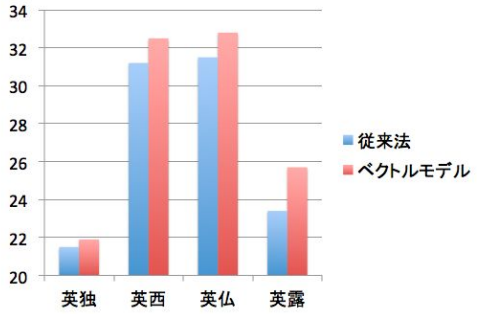

図 3：翻訳精度(BLEU score)

５．主な発表論文等

〔雑誌論文〕（計　　件）

〔学会発表〕（計　３　件）

(1) M. Tsubaki, <u>K. Duh</u>, M. Shimbo, Y. Matsumoto, Modeling and Learning Semantic Co-Compositionality through Prototype Projection and Neural Networks, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), October 18, 2013, Seattle, USA.

(2) <u>K. Duh</u>, G. Neubig, K. Sudoh, H. Tsukada, Adaptation Data Selection using Neural Language Models: Experiments in Machine Translation, Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), August 6, 2013, Sofia, Bulgaria

(3) S. Hisamoto, <u>K. Duh</u>, Y. Matsumoto, An Empirical Investigation of Word Representations for Parsing the Web, 言語処理学会第１９回年次大会発表論文集, March 14, 2013, 名古屋大学

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
Duh Kevin　(Duh, Kevin)
奈良先端科学技術大学院大学・情報科学研究科・助教
研究者番号：80637322

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
(　　)

研究者番号：