統計的意味論に基づく概念的類似度獲得手法の評価

久保田 豊久

近年では、分布仮説に基づいた単語の概念的類似度の自動抽出手法が盛んに研究されている。分布仮説に基づいた概念的な類似度の抽出手法として、word2vec や係り受け関係を用いた類似度計算手法が提案されている。この 2 種の手法は、文書集合を入力することで、概念的に類似した単語を出力することができるとされている。しかし、概念的な類似性は様々な観点から評価することができるものであり、分布仮説に基づいて抽出した単語類似度が、どのような特性をもつのかは自明ではない。人手で作成された意味辞書においては、単語の類義関係の集合の間の関係を記述した日本語 WordNet や、単語の語彙的な類似性に基づいて単語を分類している語彙分類表があり、その基準は明確に異なるとされている。またこれらの意味辞書と分布仮説に基づいて抽出した単語類似度との対応関係は明らかにされていない。

本研究では、word2vec と係り受け構造を考慮した類似度計算手法について、抽出される単語類似度と、人手で作成された既存の意味辞書との一致率を定量的に評価することにより、その特徴について新たな知見を得ることを目的とする。また、日本語の構文的知識を利用している係り受け構造を考慮した類似度計算手法に対して、word2vec との抽出結果の比較を行うことにより、日本語における word2vec の有用性を評価する。

実験では人手によって構築された既存の意味辞書との定量的比較において、両手法共にコーパスの影響を受けたとしても、約 1 割程度を網羅することができることを確認した。word2vec を用いた類似度計算手法では、サブサンプリングによってコーパスによる影響をより軽減し、日本人の分類感覚に近い語が上位として挙がることを、分類語彙表での一致度を通して確認できた。しかし、日本語 WordNet を用いた評価によって、兄弟語の抽出が難しいという特徴も確認した。係り受け構造を用いた類似度計算手法では、日本語 WordNet において word2vec と同等の同義語抽出性能を持ち、兄弟語を上手く抽出することができることを確認した。また、実験の評価対象では無いが、評価を行うまでの計算時間を考えると、日本語 WordNet の同義語のみの評価であれば、計算コストの小さな word2vec を用いた類似度計算手法の方がより有効であることを確認した。

今後の展望として、トピックモデルを用いた類似度計算手法といった異なる計算手法との比較を行う、コーパスの量を変化させる、別コーパスを学習させといった方法によって結果にどのような差異が現れるか、検証を行っていく。また今回は名詞のみの評価であったが、動詞や形容詞も追加することで精度の向上か可能であるかを検証し、辞書自動生成やその技術の一端としての利用を検討していく。

（指導教員　若林 啓）