用例文拡張辞書を利用したトピックモデルに基づく新語義検出

神宮理織[1]　佐々木稔[2]　古宮嘉那子[2]　新納浩幸[2]

[1]茨城大学大学院理工学研究科情報工学専攻

[2]茨城大学工学部情報工学科

{16nm713g,minoru.sasaki.01,kanako.komiya.nlp,hiroyuki.shinnou.0828}@vc.ibaraki,ac,jp

1. はじめに

近年、インターネットや電子文書の普及により、様々な表現技法や言い回し方が存在するようになった。新たな単語や言葉（新語）が増加してきており、また既存の言葉にも新たな意味、用法（新語義）が存在してきている。そこで重要となるのが辞書の改善、作成であり、その新語や新語義を見つけ出す必要がある。

本稿では、日本語の用例文集合と辞書定義文を用いたトピックモデルに基づく新語義検出を行う。対象単語の用例文集合から得られたトピック行列に対し、辞書の語義に対応するトピックが存在するのかを表す評価値の計算により、新語義を含むトピックの推定を行う。

既存研究では英語の用例文で実験が行われていたため、予備実験として日本語の用例文集合に対する新語義検出実験を行った。その結果、日本語辞書では新語義検出の効果があまり出ていないことが分かった。辞書定義文だけでは情報が少ないことが原因だと考えられるので、各語義の語釈文に用例文を追加して新語義推定実験を試みる。語義とトピックの対応関係を示す評価値の計算によって、新語義を含む用例文の検出可能性について検証するとともに、用例文を追加した語義定義文を用いた場合の新語義検出精度を検証する。

2. トピックモデルに基づく新語義検出手法

本稿では、対象単語に対して新語義が含まれる用例文集合を実験データとして使用する。用例文を単語毎に分け、その頻度を要素とする行列データを作成する。この行列に対し、Hierarchical Dirichlet Process（HDP）を用いてトピックモデルを作成する。得られたトピックと岩波国語辞典の定義文との類似性を求めるために、Jensen-Shannon ダイバージェンス（JS 距離）による単語類似度の計算手法を用いて各トピックに語義を割り当てる。トピックに対応する語義が存在しなければ、辞書未記載の語義を使用した用例文が存在すると考えられる。

2.1 トピックモデルの作成

実験データとして、Semeval2010 日本語 WSD タスクの用例文を使用する。新語義を含む 6 単語を対象単語とし、各単語について与えられた 50 件の訓練データと 50 件のテストデータをまとめた用例文集合を用いる。この用例文集合に対して、MeCab を用いて各用例文から名詞、動詞、形容詞(以下単語とする)を抽出する。各単語に対して頻度を求めて、各単語の頻度を要素とするベクトルを計算し、行列データ作成する（図 1）。

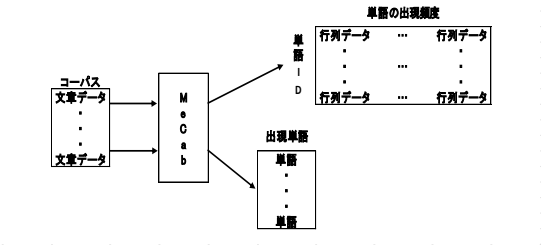

図 1：行列データと登録単語の作成

この行列に対して HDP を実行し、トピックモデルを作成する（図 2）。

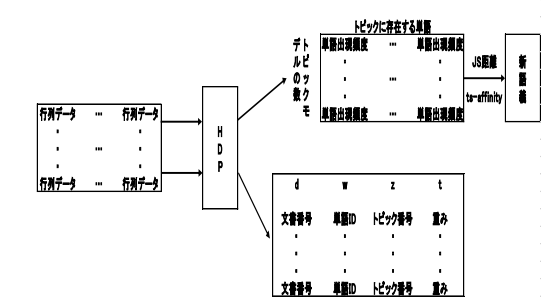

図 2：トピックモデルの作成

2.2　トピックへの語義の割り当て

HDP によって得られた各トピックに対し、岩波国語辞典の語義を割り当てる。トピックと辞書定義文の類似度は JS 距離を用いて求める。類似度 sim は、語義s_iの定義文に含まれる単語集合 S、トピックt_jで出現する単語集合 Tとすると、以下のように定義される。

$$\mathrm{sim}(s_i, t_j) = 1 - \mathrm{JS}(S||T)$$

類似度 sim の値から、トピックと語義の類似度を推定することができる。この値が大きいほど、トピックと語義が類似していると推測することができる。一方でその値が小さいほど、そのトピックと語義が類似せず、意味合いが異なることが推測される。

2.3　新語義判定のための評価関数

2.2 節で計算した類似度を利用して、辞書にない新語義が含まれているのかどうかを推定する。その新語義判定の計算方法は、以下のように定義される。

$$\text{ts-affinity}(t_j) = \frac{\sum_i^S \mathrm{sim}(s_i, t_j)}{\sum_l^T \sum_k^S \mathrm{sim}(s_k, t_l)}$$

この値を用いて新語義が含まれているかどうかを推定する。この値が大きいほど、トピックt_jにはふさわしい語義が存在すると推測することができる。一方で、この値が小さいほど、トピックt_jにはふさわしい語義がないということになり、そのトピックに含まれる用例文は新しい意味で使用されているということが推測される。本稿ではその小さい値に注目し、得られたトピックに新語義を使用した用例文が含まれているかどうかを検証する。

3　用例を追加した辞書定義文を用いた新語義検出

本節では、岩波国語辞典の辞書定義文に用例文を追加し、語義情報を拡張する。各語義定義文に、訓練データを形態素解析した単語（名詞、動詞、形容詞）を追加する。拡張された辞書定義文を用いて、トピックとの類似度 sim を計算する。類似度 sim から評価値 ts-affinity を計算し、トピックに新語義が存在するかどうか判定する。

4　実験

4.1　使用データ

本実験では、Semeval2010 日本語 WSD タスクの用例文集合を使用する。本稿で

は、新語義が含まれる 6 単語(「あげる」、「意味」、「手」、「始める」、「前」、「求める」)を対象として実験を行った。各対象単語について、訓練データとテストデータとして与えられた 100 用例を使用して、トピックモデルを求めた。

5　実験結果

2.1 節で作成したトピックモデルを用いて語義との類似度を求め、新語義判定実験を行った。その結果の一部を表 1 に示す。

表 1：対象単語「あげる」の各トピックに対する評価値

トピック 0　=　0.056583
トピック 1　=　0.056980
トピック 2　=　0.058187
トピック 3　=　0.059152
トピック 4　=　0.059434
トピック 5　=　0.058534
・・・

上の表 1 の結果から、トピック 0 と 1 の評価値が比較的小さかったので、これらのトピックと新語義用例を見比べ、適切なトピックが選ばれたかどうか調査した。今回、新語義が含まれている文書は、それぞれトピック 1 に属していたので、新語義評価の結果とマッチしている。同様に新語義が含まれる 6 単語において、岩波国語辞典の辞書定義文に用例文を追加する前と後で、適切なトピックが選出できたかどうかの比較結果を以下に示す（表 2)。

表 2：新語義用例数における新語義評価の結果とトピック番号の合致数

対象単語	追加前	追加後
あげる	0/2	0/2
意味	0/1	0/1
手	0/3	0/3
始める	2/2	0/2
前	0/7	7/7
求める	0/1	0/1

6　考察

表 2 の結果から、辞書に用例文を追加する前と後で精度が向上していることがわかったが、その向上は大幅なものではなかった。その原因としては、もともと評価関数の計算手法が日本語ではなく英語に対しての計算手法だからだと推測されるため、評価関数の計算手法を日本語でも当てはまるように見直す必要がある。岩波国語辞典には語義定義文に名詞単語が少ないことから、適切に語義とトピックとの類似性を計算できていないと考えられる。また、ただ単に岩波国語辞典に訓練データの単語を後ろに付け加えるのではなく、対象単語の前後数単語を追加することで有効な語義の特徴が得られるのではないかと考えられる。

もし仮に、評価関数の値が突出して小さいトピックが得られたとすれば、新語義を含む用例文が一つのトピックにほとんど集中しているため、そのトピックに高い確率で新語義が含まれている可能性があると推測できる。だが、トピックによっては用例文の数が異なるため、新語義を見つけにくい場合がある。以下にそのトピックに含まれる用例文の数と、新語義を含む用例文の数の内訳を示す（表 3)。

表 3：それぞれの用例文の数の内訳

対象単語 （トピック番号）	新語義を含む用例文の数	そのトピックに含まれる 用例文の数	確率
あげる（1）	2	17	2/17
意味（8）	1	4	1/4
手（1）	3	5	3/5
始める（0）	2	46	1/23
前（1）	7	9	7/9
求める（9）	1	2	1/2

実験結果では、対象単語「始める」だけが新語義評価でのトピック番号と、実際の新語義が含まれている用例文のトピック番号が合致していた。しかし、仮に評価関数の計算結果が正しいとすれば、対象単語「あげる」の場合、そのトピックに含まれる用例文の数が 17 個であるから、残りの 83 個には新語義が含まれていないことが分かり、17 個の用例文を調べるだけで新語義を含む用例文を検出することが可能である。他の対象単語も同様に、比較的高い確率で新語義を含む用例文を抽出することができるため、評価関数の計算結果が正しいものであれば新語義を含む用例文を比較的高い確率で抽出できると考えられる。

7　おわりに

本論文では、日本語の用例文集合と辞書の語釈文を用いたトピックモデルに基づく新語義検出を行った。実験の結果、辞書に用例文を追加したほうが、新語義を含む用例文を見つける精度が向上していることがわかった。また評価関数の計算結果が正しいものだとすれば、比較的高い確率で新語義を含む用例文を抽出できることも分かった。正しい結果が得られない原因としては、辞書定義文とは関係のない単語の追加が考えられる。したがって、更なる検出性能の向上を目指すには、ただ単に岩波国語辞典に訓練データの単語を後ろに付け加えて実験を行うのではなく、前後数単語のように単語を厳選する必要があると考えられる。また今回は六つの単語でしか実験を行っていないため、さらに実験対象とする単語数を増やし、考察を深める予定である。

参考文献

[1]　本間康允, 渋木英潔, 森辰則, “利用者の状況に応じた用語委開設抽出システムの提案とその実現に向けた検討”, 人工知能学会インタラクティブ情報アクセスと可視化マイニング第 6 回研究会(2014).

[2]　Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, Timothy Baldwin, “Learning Word Sense Distributions, Detecting Unattested Senses and Identifying Novel Senses Using Topic Models”, In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), pp. 259-270 (2014).