辞書の定義文に関する用例の分布を用いた未知語義の検出

システム情報工学研究科 コンピュータサイエンス専攻
博士前期課程 1 年 201220629 大西健志
指導教員 山本幹雄・乾孝司

2012 年 10 月 25 日

1 はじめに

自然言語処理における重要な課題のひとつとして, 語義曖昧性解消（WSD：Word Sense Disambiguation）がある. 語義曖昧性解消とは, 単語がある文脈で出現した時に, その単語がどのような語義で使われているかを判定することである. ここで, 語義とは「単語の意味」のことであり, 辞書などを引いた時のそれぞれの単語の説明に相当する.

例えば, 「壺」という単語を辞書で調べると,

つぼ【壺】（岩波国語辞典 第六版 [西尾他, 2000]）

1. 口がせまく胴がふくらんだ形の容器.
2. つぼ (1) に似た形のもの.
3. ここと見込んだ所.

とあり, 3 つの語義を持っていることが分かる. 語義曖昧性解消は, 例えば次のような文があったとき,

大切な壺を割ってしまった.

この文の「壺」は 1 の意味で使われていることを判定する. この技術は, 機械翻訳や情報検索などに有用である.

しかし, 語義曖昧性解消には辞書に定義されていない語義は判定できないという大きな問題がある. 例えば, 次のような文の語義曖昧性解消を考える.

これだから壺の住人は...

ここでの「壺」は「２ちゃんねる（電子掲示板群の一つ）の隠語」として使われている. しかし, そのような語義は上の辞書の「壺」の定義には含まれないため, 語義曖昧性解消を行えば, 誤って 1〜3 のどれかの語義が付与されてしまうことになる. この問題に対応するために, 辞書は全ての語義を網羅し, また新しい語義に対応するために頻繁に更新を行う必要があるが, それは現実的ではない.

そこで, 文中で出現した単語が既存の辞書に定義されていない語義（未知語義）で使われている, ということを判定することを考える. これは未知語義検出, または新語義検出（New Sense Detection）と呼ばれる. 未知語義であることが分かれば, 語義曖昧性解消の高精度化や, さらに辞書編纂の助けにもなる. 本研究では, 高精度な未知語義検出を行う手法を検討する.

本稿では, まず未知語義発見に関する研究を 2 節で紹介する. 3 節ではその中の一つである [田中, 2009] の方法に基づいて予備実験を行い, その手法を改善するための手法を 4 節で述べる. 最後に今後の計画などについて述べる.

2 関連研究

2.1 語義に注目したその他の課題設定

2.1.1 語義識別

上述した語義曖昧性解消の問題に対応するために, 語義識別（Word Sense Discrimination）がある. 辞書の語義定義を用い, どの語義で使われているかを判定する語義曖昧性解消に対し, 語義識別では辞書による語義定義は用いず, ある用例と別の用例の中で, 単語が違う語義で使われているかどうかを判定する. 語義識別では単語がどのような語義で使われているかは知ることはできないが, 辞書を用いないことにより, 辞書に定義されていない未知語義も区別できる. 例えば, 次のような 3 つの「壺」の用例中の「壺」の語義識別を考える.

a. 大切な壺を割ってしまった.
b. 昨日, 大きな壺を買った.
c. これだから壺の住人は...

語義識別では, この例では a と b の「壺」は同じ語義で使われており, c の「壺」はまた別の語義で使われていることが分かる.

2.1.2 語義推定

語義推定（Word Sense Induction）とは, ある単語の全ての語義を見つけるタスクである [Brody and Lapata, 2009]. 語義識別と同様, 語義推定も辞書の語義定義は用いない. また, 語義識別と同じようなプロセスが考えられるが, 目的が異なる.

2.2 語義推定による未知語義検出

[田中, 2009] では, クラスタリングによって語義推定を行い, 結果のクラスタと辞書の語義定義とを対応付けすることによって未知語義を発見する手法を提案している.

田中は語義推定のために, [九岡, 2008] により提案された, 用例をクラスタリングすることにより語義を弁別（語義識別）する手法を利用している. 久岡は, 単語の語義を知るにはその単語の周りの文脈が手がかりになるという考えから, 用例中の語義曖昧性を解消したい語（対象語）の周りの文脈を特徴ベクトルで表し, 用例のクラスタリングをしている. ここで, 特徴ベクトルの作り方として, 大きく分けて次の 4 種類を考えた.

隣接ベクトル
対象語の周り 1, 2 単語を素性とする.

文脈ベクトル
対象語の周り s 単語の自立語を素性とする.

連想ベクトル
対象語の周り s 単語の自立語と共起する語を素性とする.

トピックベクトル
対象語のトピックを推定し, 素性とする.

以上のように特徴ベクトルを作成しクラスタリングを行った後, 用例のクラスタと辞書の語義との対応付けを行う. まず, クラスタを特徴ベクトルで表し, また辞書の語釈文からも特徴ベクトルを作成する. この 2 つの特徴ベクトルのコサイン類似度が最も大きくなるものを対応する語義とする. ここで, 類似度の最大値が小さいものについてはどの語義とも類似していない, つまり未知語義であると考えられるが, 実際には閾値を設けてこれを判定するのは難しい. そこで, 語義を類似度の降順に並べ, 類似度の差が大きいところに境界を設け, その境界以下の類似度の語義を未知語義としている.

2.3 その他の方法

[菊田, 2006] では, 辞書に定義されている語義に,「未知語義」を表す語義を加え, 語義曖昧性解消を行っている. 菊田は, Naive Bayes 分類器を用いこれを行い, そのモデルパラメータは EM アルゴリズムを用いて計算している.

[新納・佐々木, 2012] では, データマイニングで用いられる外れ値検出手法を元にして, 未知語義を検出するための教師付き外れ値検出手法を提案している.

[Lau et al., 2012] では, 単語のどの語義が使われやすいかは文書のトピックによるとして, 語義をトピックとして扱い, トピックモデルを用いて未知語義の検出を行っている. トピックモデルとは文書の生成モデルの一つで, これを用いてどんなトピックが出現しやすいか, どのようなトピックの時にどんな単語が出現しやすいかを学習することができる. トピックモデルの一つである LDA（Latent Dirichlet Allocation）は, トピックの数（語義数に相当する）を予め決めなければならないため, Lau らはトピック数を自動的に決定する HDP（Hierarchical Dirichlet Processes）を用いている. Lau らは古いコーパスと新しいコーパスの 2 つについて, HDP を用いトピック（語義）の出現確率を計算し, 古いコーパスである語義が出現しにくく, 新しいコーパスでその語義が出現しやすいならば, その語義は時代が変わるにつれてその単語に付与された新語義であるとしている.

2.4 Semeval-2

Semeval-2 は, 2010 年に開催された言葉の意味を解析・評価するワークショップである. 同様のワークショップは Senseval や Semeval-1 などあるが, Semeval-2 には Japanese WSD というタスクが含まれる. Japanese WSD は, 日本語の語義曖昧性解消のタスクであるが, 未知語義の判定も含まれているということが特徴である. [田中, 2009] や [菊田, 2006] で提案された方法もこのワークショップに参加している. このタスクへの参加者には, 人手で語義がアノテートされたデータが配布される. そのデータセットは, 正解データとして本研究に有用であるといえる.

3 予備実験

3.1 実験項目

現在, ベースラインとなる関連研究の実装を行っている. 具体的には, [田中, 2009] の方法の, 語義推定を

するモジュールを作成中である．田中は [九岡, 2008] で提案された 4 種類の特徴ベクトルを用いているが，今回はそのうちの隣接ベクトルを用い，k-means 法を使って用例をクラスタリングした結果を示す．今回は数量的な結果の評価は行っていない．

ここでは，毎日新聞の 1991 年から 1993 年のコーパスを用いて実験した結果を示す．まず，コーパスから対象語の用例を抜き出し，次にそれらを隣接ベクトルで表し，クラスタリングを適用する．今回の対象語には「キリン」を選んだ．「キリン」には，動物としての意味と，ある飲料会社としての意味があると想定した．結果は，各クラスタごとに，各用例の対象語の周りの 1, 2 単語を示す．各クラスタに含まれるこれらの語を見ることによって，各クラスタがどのような語義を表すかを知ることができる．同じような文脈で使われる語が集まっていたら，そのクラスタはその文脈での対象語の語義を表しているといえる．また，クラスタ数は 10 とした．

3.2 結果

クラスタリングの結果を表 1 に示す．「キリン」の用例は，毎日新聞のコーパスから 358 個作成された．なお，表中の「HEAD」は行頭を表す記号である．

表 1: 隣接ベクトルを用いた「キリン」の用例のクラスタリング結果

クラスタ	各用例の対象語の前後 2 単語
1	そもそも, プラザ, 大阪, シマウマ, …
2	頭, 代表, カップ, ミズノ, サントリー, …
3	説明, 商品, ラガー, サッポロ, 」, …
4	投入, 記録, HEAD, 最高, カップ, …
5	ラガー, ドライ, 頭, アサヒ, 大阪, …
6	する, ドラフト, ドライ, マンモス, …
7	他社, シマウマ, 絶滅, ラガー, …
…	…

表 1 を見ると，動物のキリンに関係のありそうな「シマウマ」，「頭」，「絶滅」などの単語，飲料会社のキリンに関係のありそうな「サントリー」，「アサヒ」，「ラガー」などの単語が，各クラスタに混在している．またその他の語義として，建物名のキリンプラザ大阪に関係のありそうな「プラザ」「大阪」といった語や，サッカーの大会であるキリンカップに関係のありそうな「カップ」「代表」といった語も混在している．

ここでは，複数の語義に関係の有りそうな語がひとつのクラスタにまとめられることは好ましくないため，この結果は良好ではない．これは，隣接ベクトルは比較的局所的な情報を反映するため文脈を表しきれなかったこと，自立語以外も考えることにより助詞や記号など関係のないものも素性に含まれること，語の出現場所や順番も考慮することにより同じ語でも違う素性として表されることなどが原因ではないかと考えられる．

次に，田中の方法の問題点を挙げる．一つ目は，辞書定義文の情報が特徴空間において，1 つの特徴ベクトルでしか表されないことである．これにより，辞書定義文の細かい特徴が表されず，本当はその語義に割り当てられるべき用例が，割り当てられないかもしれない．二つ目は，用例のクラスタリングの結果が悪いことである．クラスタリング結果が悪ければ，クラスタと辞書定義文との対応付けもされにくくなる．

4 提案手法

ここでは，3 節で示した一つ目の問題点に対応するために，現在考えているアイデアを示す．本手法では，辞書の語義定義文中の語はその語義に関係している語であるという仮定から，辞書の語義定義文中の自立語の用例を集め，それを特徴空間に配置し，対象語の用例の語義が近いかどうかを見る．例えば，前述した「壺」の語義 1 を考える．語義 1 の定義文には，口，形，容器などの自立語が含まれる．これらの語を含む用例をコーパスから抜き出し，それぞれ特徴ベクトルで表し特徴空間に配置したものが，図 1 の上段と左下の図である．次に，特徴空間上に配置した特徴ベクトルが多く集まっているところを見つける．これが，その語義の分布となる．図 1 の右下の図では，特徴ベクトルが多く集まっているところを赤色で示した．

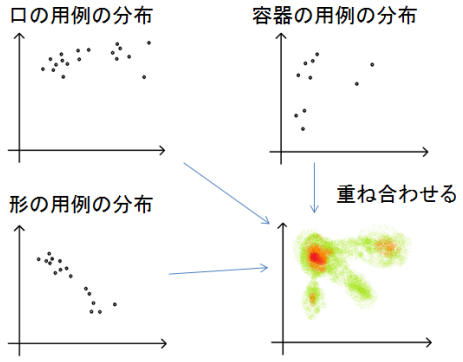

図 1: コーパスから抜き出した各用例を特徴空間上にプロットし，その用例が多く集まっているところを見つける．

次に,「壺」の用例を特徴空間に配置した時にこの領域に配置されれば, それは語義 1 で使われているかもしれないという事が分かる. これは, 語義の定義文中に現れる語はその語義と関係があるだろうという仮定に基づいている. このようにして, ある程度の広さを持たせて語義定義文を特徴空間に表すことで, 田中の手法の問題点が改善されると考える.

5 今後の予定

まず, 上述した関連研究や, その他の関連研究について調査・理解・実装を行う. 特に, 3.2 節で述べたように隣接ベクトルを用いたクラスタリング結果が悪かったので, その他の文脈ベクトルなどを実装するなどし, その原因を調べたい. また, Semeval-2 のデータセットが手に入ったので, そちらのデータでも実行してみたい.

そして, 提案手法を実装する. 現在, 4 節の提案手法はアイデアであるので, これが実際に実装できるのか, どのように定式化するのかを調べなければならない. またこれでうまく語義が割り当てられるのか, 未知語義が検出できるのかなども調べる必要がある.

また, 語義推定の際のクラスタ数はどうするのかなど, 検討していきたい.

参考文献

Samuel Brody and Mirella Lapata, 2009. "Bayesian Word Sense Induction," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 103–111.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin, 2012. "Word Sense Induction for Novel Sense Detection," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 591–601.

菊田篤史, 2006 年.「未定義語義を含む語義曖昧性解消」, 修士論文, 北陸先端科学技術大学院大学情報科学研究科.

九岡佑介, 2008 年.「コーパスからの単語の意味の発見」, 修士論文, 北陸先端科学技術大学院大学情報科学研究科.

新納浩幸・佐々木稔, 2012 年.「外れ値検出手法を利用した新語義の検出」,『言語処理学会 第 18 回年次大会発表論文集』, pp.1304–1307.

西尾実・岩淵悦太郎・水谷静夫, 2000 年.『岩波国語辞典』, 岩波書店, 第 6 版.

田中博貴, 2009 年.「用例のクラスタリングに基づく単語の新語義の発見」, 修士論文, 北陸先端科学技術大学院大学情報科学研究科.