共起語グラフのクラスタリングによる単語の多義性抽出

鏑木 雄太†　　古宮 嘉那子‡　　小谷 善行‡
東京農工大学 工学部 情報工学科†
東京農工大学 工学研究院 先端情報科学部門‡
50006268014@st.tuat.ac.jp,{kkomiya, kotani}@.cc.tuat.ac.jp

1　はじめに

自然言語処理の分野では語義曖昧性解消という問題に対する研究が盛んに行われている。語義曖昧性解消とは、多義語が文中に出現したとき、その単語がどのような意味で使われたのかを推定するものであり、教師あり学習による方法と教師なし学習による方法がある。教師あり学習では教師データにない語義や辞書に定義されていない語義（新語義）は正しく語義を判定することができない。ウェブ上のテキストのように、教師データや辞書の更新よりも早い頻度で新語義が生まれるテキストに対して正しく語義を判定することは難しい。このような問題に対して、辞書にない語義を抽出する新語義発見（Word Sense Induction）に焦点を置いた研究が盛んに行われている。本研究では、単語の共起関係をグラフ構造にし、クラスタリングすることで単語の多義性を抽出するシステムを提案する。多義語は、語義ごと共起しやすい単語が異なると考えられる。グラフクラスタリングによって共起語のクラスタリングを行うことによって語義ごとに共起しやすい単語のクラスタを生成する。共起語のクラスタリング結果によって多義性の抽出、既知の語義を同定し新語義発見を行うことを目的とする。以下2章では関連研究について、3章では共起語から多義性を抽出する方法について、4章では多義性を抽出する具体的処理について、5章では実験の概要について、6章で実験結果に対しての評価を行う。

2　関連研究

グラフ構造を用いた自然言語処理の研究は多く行われている。グラフ構造を用いている代表的なものに概念辞書であるWordNet[2]がある。WordNetを辞書としたグラフベースの語義曖昧性解消に関する研究[3]も行われている。また、大規模なテキストコーパスから、名詞共起情報を用いて語義発見することをを目指した研究[5]も行われている。

3　共起語から多義性を抽出する方法

本研究は、多義語は、語義によって共起する単語が異なるという考えに基づいている。例えば、「ジャケット」という単語は、「上着の一種」と「レコード・本などを包む覆い」という語義を持っている。前者の語義では「着る」や「洋服」といった単語と共起することが考えられ、また後者の語義では「CD」や「本」といった単語と共起すると考えられる。また、前者の語義における共起語である「着る」と「洋服」は、お互いに共起しやすいと考えられる。共起語の共起関係を用いて、語義に対応する共起語集合を自動生成することで、単語の多義性を発見できると考えた。本研究では共起関係にグラフ構造を用い、共起語集合の生成にグラフクラスタリングを用いる。

4　共起語グラフのクラスタリングシステムの実現

まずコーパスから共起語グラフの生成を行う。共起語グラフの生成では、選択した共起語を基にグラフを生成する。

4.1　共起語グラフ生成の具体的処理

共起語グラフの生成を行うための共起語の選択とグラフ生成方法について述べる。

4.1.1　共起語の選択方法

本研究において共起とは、「同一の文中に出現すること」と定義する。任意の二単語が一文で共起したと

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

き、その二単語は一回共起したと数える。共起の対象とする単語は、名詞（形容動詞、サ変動詞を含む）、動詞、形容詞の各自立語とした。既存の研究 [5] では、語彙統語パターンを用いて並列関係にある名詞を対象としていたが、名詞に対して動詞や形容詞が語義クラスタリングや語義推定の手助けになると考えた。名詞、動詞、形容詞のみを抽出するために、事前に形態素解析を行ったコーパスを品詞情報を元にフィルタリングし、表記ゆれを考慮して対象単語の用言はすべて原形に変換した。

4.1.2 共起語グラフの生成方法

共起関係をコーパス全てにおいて調べ、その情報を基にグラフを生成する。以降この共起関係を表現したグラフのことを共起語グラフと定義する。共起語グラフにおいては、1 種類の単語は 1 つのノードによって表現され、共起関係はエッジによって表現される。ノードには単語の出現回数、エッジには両端の単語（ノード）の共起回数を保存する。例として、「太郎がりんごを食べた。」という文が現れたとき、図 1 のような共起語グラフを生成する。

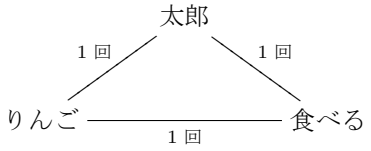

図 1: 共起語グラフの例

4.2 クラスタリングによる多義性抽出

次に、クラスタリングにより共起語グラフを基に多義性を抽出する

4.2.1 対象とする多義語を中心とした共起語による部分グラフの生成

多義性を抽出したい単語を一つ選び、その単語と直接共起した単語を全て列挙する。列挙した単語に対応するノードとノード間を結ぶエッジにより、グラフの一部分を抽出する。このグラフを以降では部分グラフと呼ぶ。この部分グラフには、ターゲット単語に対応するノードとそれに繋がっているエッジは含まない。結果、図 2 のようなグラフとなる。図 2 において実線は部分グラフに含めるエッジ、点線は共起語グラフに存在するが部分グラフに含めないエッジである。以降の処理はこの部分グラフを対象として行う。

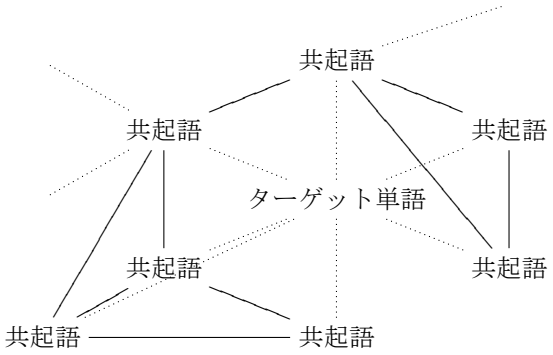

図 2: 部分グラフのイメージ

4.2.2 部分グラフエッジの重み計算方法

グラフクラスタリングを行うに当たってのグラフエッジの重みを設定する。重みには自己相互情報量を用いる。自己相互情報量 $I(x,y)$ は次の式で表わされる

$$I(x,y) = \log \frac{P(x,y)P(*,*)}{P(x,*)P(*,y)} \tag{1}$$

上記の式において、$P(x,y)$ は、グラフにおける単語 x,y の共起回数、$*$ はグラフに存在するすべての単語を意味する。本研究では、共起グラフ全体における自己相互情報量と部分グラフにおける自己相互情報量の両方を用いる。部分グラフにおける自己相互情報量 $I(x,y|part)$ は、

$$I(x,y|part) = \log \frac{P(x,y|part)P(*,*|part)}{P(x,*|part)P(*,y|part)} \tag{2}$$

となる。式 (2) における $part$ は、部分グラフに含まれることを指す。この式 (1) と式 (2) の積を部分グラフのエッジの重みとして用いる。ただし、いずれかの式の値が負となった場合は、重みを 0 とした。

4.2.3 グラフクラスタリングアルゴリズム

部分グラフに適用するグラフクラスタリング手法として。マルコフクラスタリングアルゴリズム [4] を用

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

いた。マルコフクラスタリングアルゴリズムは、グラフエッジの重みを遷移確率としてグラフ内をランダムウォークすることでクラスタリングを行う手法である。グラフの各ノードに自己ループを追加したものグラフにおける遷移確率行列 M に **inflation** と **expansion** を繰り返すことでクラスタリングを再現することができる。**inflation** と **expansion** の各定義式を次に示す。

expansion

$$M = M^2 \tag{3}$$

inflation

$$M = \Gamma_r(M) \tag{4}$$

$$\Gamma_r(M)_{pq} = (M_{pq})^r / \sum_{i=1}^{k} {M_{iq}}^r \tag{5}$$

式における"="は代入、M_{pq} は行列 M の要素 (p,q)、r は **inflation** パラメータ $(r > 1)$ である。行列遷移確率行列 M が収束するまで、**inflation** と **expansion** を繰り返す。収束した行列は、いくつかの小さなグラフの遷移確率行列が 1 つの大きな行列内に表現されている。現れた各小さなグラフに含まれているエッジ群（単語群）を 1 つのクラスタとする。以降、このクラスタを語義クラスタと呼ぶ。

5 実験

この章では、実際にコーパスを用いて単語の多義性抽出を行う実験を二つ行なう。言語資源は BCCWJ コーパス [1] の Yahoo!知恵袋コーパスを用いた。Yahoo!知恵袋コーパスは、1500 件の Yahoo!知恵袋の記事を形態素解析し岩波国語辞典 [6] の定義によって語義をタグ付けされたコーパスである。マルコフクラスタリングにおける **inflation** パラメータ r は 1.25 とした。

5.1 辞書を用いない多義性抽出実験の概要

コーパスから生成したグラフにクラスタリングを行い、辞書を用いずに多義性抽出を試みる。本実験では、コーパスのうち形態素情報のみを利用し、語義タグ情報は用いなかった。多義性抽出対象とした多義語は、コーパス内の文章に出現する単語群と wikipedia の曖昧性回避のページを参考に 16 種類の名詞を選んだ。

5.2 辞書を用いた多義性抽出実験の概要

辞書を用いない多義性抽出では、既知の語義であるかどうかを判断することが難しい。そこで、語義クラスタと語義を同定し、新語義を推定する手がかりとして、辞書の定義文と用例文を用いる。辞書として岩波国語辞典第五版を用いた。辞書の定義文と用例文の共起関係を共起語グラフに追加することで、辞書内単語が語義クラスタにどれだけ含まれているかを指標として語義の推定をすることができる。辞書内単語が含まれていない語義クラスタは、新語義を示す語義クラスタであると仮定した。

6 評価

本章では 5 章で示した二つの実験の結果得られた語義クラスタを示し、それに対する評価と考察を行う。

6.1 辞書を用いない多義性抽出結果

辞書を用いない多義性抽出実験の結果得られた語義クラスタのうち、「ソース」「ジャケット」「マーチ」のクラスタリング結果を表 1 に示す。

「ソース」については、「情報源」と「調味料」の各語義に対応する二つのクラスタが生成されている。「ジャケット」については、「洋風の上着」と「レコード・本などの覆い」の各語義に対応するクラスタが生成されている。二番目のクラスタを見ると、前述の 2 語義に関係する単語が混ざったクラスタが生成され、過分割が発生している。「マーチ」については、「音楽のジャンル」の他、岩波国語辞典には定義されていない「自動車の車種」「大学群の略称」という語義クラスタが生成されたが、一番目と三番目のクラスタには語義に相応しくない単語が多く含まれていた。性能評価のために生成した語義クラスタに正解と不正解のラベルを付与する。語義クラスタのうち、岩波国語辞典で定義されている語義に相当するクラスタには、正解ラベルを人手で付与する。ただし、一つの語義に対応する正解ラベルは一つのクラスタにのみ付与し、残りのクラスタは不正解とした。16 個の単語に対して実験を行った結果、適合率、再現率、F 値の各平均値は表 2 のようになった。

6.2 辞書を用いた多義性抽出結果

辞書を用いた多義性抽出実験の結果得られた語義クラスタを表 3 に示す。

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 「ソース」「ジャケット」「マーチ」の多義性抽出結果

ソース

ID	クラスタ内の単語	推定語義
1	アスペクト, コーディング, コンパイル, マルチメディア, 比, 比率, ボリューム, マイク, 録音, ...	情報源
2	ヤング, ナポリ, 少な目, ミート, 蒸し焼き, ポーク, コンデンスミルク, のっける, ソテー, 松屋, ...	調味料

ジャケット

ID	クラスタ内の単語	推定語義
1	硬質, ジン, 冠, プラスチック, まわす, 番目, きく, 薄い, 歯, レ	覆い
2	陰干し, 鼻血, 通称, 手洗い, 脱水, すすぐ, 柔軟, 真っ白, 中学生, 材, アルバム, ダウン, 印刷, どの,...	不明
3	ベロア, キルティング, 濃いめ, コーデュロイ, 夏服, キタ, がま, 射光, キャミソール, 春物, 脱げる, ...	洋服の種類

マーチ

ID	クラスタ内の単語	推定語義
1	シトロン, イエロー, 旧型, いっこ, ニッサン, 半音, リック, トルコ, 廃盤, 日産, メタ, 目線, レンタカー,...	自動車
2	ラブソング, 怖気, さんぽ, オルゴール, ごねる, 半音, ミッキーマウス, トルコ, サントラ, 音源, ...	音楽
3	立教, 法政, 青山, 脱走, 栃木, 国立, 茨城, 群馬, 勢力, 吹奏楽, 勝負, 偏差, 年度, 埼玉, 千葉, 明治, 中央...	大学群の略称

表 2: 辞書を用いない多義性抽出の実験結果

適合率	再現率	F 値
0.36	0.51	0.40

表 3: 「ソース」の多義性抽出結果

ID	クラスタ内の単語	推定語義
1	アスペクト, コーディング, コンパイル, アクセサリ, 比, 比率, ボリューム, マイク, 録音, ...	
2	**西洋**, **料理**, **調味**, **汁**, **ホワイト**, ヤング, 少な目, ナポリ, ミート, 蒸し焼き, ポーク, コンデンスミルク, のっける, ソテー, 松屋, ...	調味料
3	**出どころ**, **源泉**, **ニュース**, 広島	情報源

「ソース」の定義文と用例文に含まれている単語は**太字**で示した。語義に対応するクラスタは、その語義の定義文と用例文の単語が最も含まれているクラスタとした。結果、辞書を用いない多義性抽出実験結果と異なる語義推定結果となった。

7 おわりに

本研究では、コーパスと辞書定義文と用例文を基にした共起語グラフをクラスタリングすることで単語の多義性抽出を提案した。実験の結果、いくつかの単語から辞書にない新語義を発見することができた。本研究手法では、単語によってクラスタ分けにばらつきが発生していた。今後は、提案手法の改良を行い、辞書をより有効活用し共起語から新語義を発見しやすくすることを目指す。

謝辞

データを提供していただいた東京工業大学 奥村研究室に深く感謝する。

参考文献

[1] Kikuo Maekawa. Balanced corpus of contemporary written japanese. In *Proceedings of the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102, 2008.

[2] G.A. Miller. WordNet: a lexical database for English. *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41, 1995.

[3] R. Navigli and M. Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.

[4] Stijn van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.

[5] 田淵史郎, 鍛治伸裕, 吉永直樹. 大規模コーパスからの語義のマイニング. 日本データベース学会論文誌, Vol. 8, No. 1, pp. 77–82, 2009-06.

[6] 西尾実, 岩淵悦太郎, 水谷静夫. 岩波国語辞典 第五版. 岩波書店, 1994.

Copyright(C) 2011 The Association for Natural Language Processing.
All Rights Reserved.