D-021

係り受け関係に基づく多義語の語義曖昧性解消法
A Word Sense Disambiguation Method based on Dependency Relations

宮崎 義隆 † 塩井 隆円 † 波多野 賢治 †
Yoshitaka Miyazaki Takamitsu Shioi Kenji Hatano

1. はじめに

現在，情報技術の発達により，ブログや EC サイトのレビューなどのサービスにおいて個人が手軽に情報コンテンツを Web 上に発信できるようになっている．しかし，Web 上に大量のコンテンツが氾濫しているため，ユーザが目的のコンテンツを容易に探し出すことが困難になってきている．

ユーザのコンテンツ探索を支援する手法の一つに，情報コンテンツに付与されたタグやテキスト中の単語に対して出現頻度から重み付けを行い，それを基に重要語を抽出し，フォントや色にその重みを反映させ，視覚的に表示するタグクラウドを用いた視覚化がある．近年では，より効果的なタグクラウドの作成を目的とした研究も存在する [1][2]．しかし，既存のタグクラウドでは抽出した重要語をそれぞれ別々に表示するため，一語で二つ以上の語義を持つ多義語が重要語として抽出された場合，ユーザはその語義を特定することができず，コンテンツ理解の妨げになるという問題がある．

そこで本稿では，単語間の関係性を表す係り受け関係を利用し，多義語の語義曖昧性を解消し，ユーザに多義語の語義の想起を促すことが出来るような単語対の抽出手法を提案する．また単語間の関係性の一つである共起関係と提案手法において，それぞれのカバー率を算出し，比較を行い，提案手法の有用性を確認する．

2. 関連研究

松崎ら [1] は EC サイトを利用するユーザが商品の特徴を素早く捉えることが出来るようなタグクラウドを生成するため，係り受け解析器，日本語概念辞書を用いて同義語を統合し，機械的に EC サイトの商品紹介ページやレビュー文書から表記揺れのない重要語を用いたタグクラウドの生成を行っている．このような同義語の統合や単語の類似性の判別は，文書分類の精度の向上に必要な技術であり，昔から研究が行われている．例えば稲子ら [3] は各語に対し，共起頻度を元に重み付けを行い，共起単語から成る共起単語ベクトルを生成し，それらが類似する単語が類似性の高い単語であるとしている．

文献 [1] では，日本語 WordNet[4] を利用することで語の意味の統一を行った．具体的にはまず，対象となる文を形態素に分割し，連体語・接続語・独立語以外の自立語を重要語候補として抽出している．しかし，同じ商品のレビューは複数で書かれていることが多いため，同じ内容のレビューでも異なる表記で書かれることが多い．この問題の解消のため，日本語 WordNet[4] を用いて同義語のグルーピングを行った．そして抽出した重要語候補が，作成したグループに属するか否かを判断し，各概念にグルーピングすることで同義語の統合を行い，表記の揺れを解消した．このように作成したそれぞれのグループの中の単語の重要度を tf-idf 法によってそれぞれ算出，比較し，最も値の大きい単語をそのグループを表す重要語であると定義している．そのような重要語群を用いて tf-idf 値が大きい程フォントサイズが大きくなるようなタグクラウドを構築している．

3. 提案手法

従来手法では，多義語が重要語として抽出され，表示された場合，ユーザはその語義を特定することが出来ない．タグクラウド上でユーザに多義語の語義の特定を促すためには，単語間の関係性を考慮し，元となるテキストから多義語と共にその語義の想起を促すことが出来るような単語を対で抽出し，表示する必要がある．

そのため本稿では，多義語の語義をユーザに想起させるために単語間の関係を示す係り受け関係を考慮した単語対の抽出手法の提案を行う．

提案する処理手順の概要は以下の通りである．

1. 文書を形態素単位に分割
2. 各形態素の品詞，係り受け関係の把握
3. 係り受け関係にある自立語を対で抽出
4. 抽出した単語対の重み付け

まず，対象の文書を形態素解析システム JUMAN[5] を用いて形態素単位に分割する．これは後述する構文解析システムで係り受け関係を把握する際に形態素単位での入力が必要であるからである．

JUMAN から得られた出力結果を日本語構文・格解析システム KNP[6] を用いて構文解析を行い，品詞の判別と係り受け関係の把握を行う．KNP は日本語構文解析，格解析，照応解析を行い，KNP 独自で用いられる単位である基本句および文節間の係り受け関係，格関係，照応関係を出力するシステムである．それらの関係は Web 上に存在する様々なテキストを収集し，作成したコーパスから構築した大規模格フレームに基づく確率的モデルにより決定している．

この出力を利用して係り受け関係にある自立語の対を抽出する．なお自立語は文献 [7] を参照し，名詞・動詞・形容詞・副詞・感動詞・接続詞と定義した．自立語を抽出候補としたのは，自立語はその語のみで意味が理解できるとされているからである [7]．

抽出した係り受け関係にある自立語の対の係り先の単語の出現頻度の集計を係り元の単語ごとに行い，集計した出現頻度が高い単語から順に重みを付ける．係り元の単語に対して重み付けの結果の上位 10 件をそれぞれ提示することで，ユーザに係り元の単語の語義の想起を促す．

†同志社大学，Doshisha University

表 1: 品詞別の集計結果

名詞	動詞	形容詞	副詞	感動詞	接続詞
247,797	7,044	3,855	3,674	125	91

4. 評価実験

本来評価実験として，提示した単語対によってユーザが多義語の語義を想起出来るか否かを評価をする必要があるが，文脈によって多義語の語義は変化するため，まずは提案手法によって想起することが出来る多義語の語義数が多義語自身が持つ語義の内，どれくらいをカバーしているのかを検証する必要がある．そのため，想起することが出来る多義語の語義数を用いることで提案手法を評価した．評価指標には式 (1) で定義されるカバー率を用いた．これは各多義語が持つ語義の内，想起できた語義の割合である．

$$\text{カバー率} = \frac{\text{想起できた語義数}}{\text{多義語が持つ語義数}} \tag{1}$$

提案手法の有用性を示すため，語義曖昧性解消に一般的に使用されている共起関係を利用し，共起頻度により重み付けを行った単語対でも同様の評価を行い，提案手法とカバー率の比較を行った．

実験データは 1999 年〜2003 年の 5 年分の毎日新聞の記事を用いた．新聞データに提案手法を適用し，係り受け関係にある単語対の係り元となる単語を品詞別に集計した結果が表 1 である．係り受け関係にある単語対の総数は 262,586 語であり，そのうち全体の約 94%である 247,797 語が名詞であった．そのため本稿では係り元が名詞である単語対 247,797 語に着目した．

本稿では多義語の定義を日本語の意味辞書である日本語 WordNet を参照し，二つ以上の語義を持つ単語であるとし，この名詞群から多義語を抽出したところ，その数は 18,866 語であった．この多義語を母集団とし，統計的推論に基づき母集団から母集団を表すような必要標本数を式 (2) から算出し，単純無作為抽出によって評価対象となる多義語を抽出した [8].

$$n \geq \frac{N}{(\frac{a}{k})^2(\frac{N-1}{P(1-P)})+1} \tag{2}$$

a は収めたい誤差の範囲を表す変数であり，一般的に使用されている 0.05 を用いる．その結果，信頼率は 0.95 となり，対応する信頼度係数 k の値は 1.96 となった．母集団の比率 P は予測が困難であるため最も安全な標本の大きさが決められるとされている 0.5 を用いる．今回の場合，母集団 N は 18,866 であるので，必要な標本数 n は式 (2) より約 377 語となった．そのためそれを上回る 400 語をランダムに抽出し評価の対象とした．

被験者である 20 代から 50 代の男女 10 名に重みが高い上位 10 語の単語を提示した際に想起出来る係り元の多義語の語義数を検証し，それを元にカバー率を算出した．表 2 に実験の結果を示す．

表 2 より，単語対でユーザに多義語の語義を想起させるためには，共起頻度を利用するより提案手法を用いた方が有用である可能性があると判明した．これは共起関係が文中で共に用いられやすいという語の関係であるのに対し，係り受け関係が意味の上で結びついた関係であるため，その結果カバー率が高くなったと考えられる．

表 2: 実験結果

	係り受け	共起
カバー率 (%)	0.50	0.33

5. おわりに

本稿ではタグクラウド上に表示される多義語の語義の想起をユーザに促すことが出来るような単語対を文中の係り受け関係を利用し抽出する手法を提案した．共起関係との比較実験を行った結果，提案手法の方がカバー率が高くなり，多義語の想起に有用である可能性が高いことを示した．

今後の課題として，抽出した多義語の元の文書から多義語の語義の正解データを作成し，その正解を提案手法の単語対から想起させることが出来るかを検証する実験を行うことが挙げられる．

謝辞

本研究の一部は JSPS 科研費 26280115，15H02701 の助成を受けたものである．

参考文献

[1] 松崎, 波多野. EC サイトにおける購買行動促進のための重要語抽出とタグクラウド生成. 情報処理学会研究報告, Vol. 2014-IFAT-114, No. 10, pp. 1–6, 2012.

[2] A. Zubiaga, A. P. Garcia-Plaza, V. Fresno, and R. Martinez. Content-based Clustering for Tag Cloud Visualization. *Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in*, pp. 316–319, 2009.

[3] 稲子, 笠原, 松澤. 複合語内単語共起による名詞の類似性判別. 情報処理学会論文誌, Vol. 41, No. 8, pp. 2291–2298, 2000.

[4] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki. Development of the Japanese WordNet. *Proceedings of the Sixth International Conference on Language Resources and Evaluation ELRA*, pp. 2420–2423, 2008.

[5] 黒橋･河原研究室. 日本語形態素解析システム JUMAN version 7.0 使用説明書, 2012. `http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN`.

[6] 黒橋・河原研究室. 日本語構文解析システム KNP version 4.0 使用説明書, 2012. `http://nlp.ist.i.kyoto-u.ac.jp/?KNP`.

[7] 小池. 現代日本語探究法. 朝倉書店, 2001.

[8] 林, 佐藤, 青山, 林. 統計学の基本. 朝倉書店, 2000.