ミドルソフトタグのトピック素性を利用した語義曖昧性解消

國井 慎也　　新納 浩幸　　佐々木 稔
茨城大学 工学部情報工学科
09t4026l@hcs.ibaraki.ac.jp, {shinnou, msasaki}@mx.ibaraki.ac.jp

1　はじめに

本論文ではトピックモデルを語義曖昧性解消 (Word Sense Disambiguation, WSD) に利用する手法を提案する．

WSD は文中の多義語である単語に対して，その単語がどの語義で使われているかを識別するタスクである．このタスクは自然言語処理の中心課題であり，従来より多数の手法が提案されているが，近年は教師付き学習を用いる手法が中心である．そこではまず少量の対象単語の用例を用意し，その用例中の対象単語に語義を付与する．次に用例中の対象単語の周辺の文脈を素性リストで表し，先に付与した語義とのペアを作る．このペアの集合が訓練データとなり，この訓練データから様々な機械学習手法を利用して語義の分類器を学習することで WSD を解決する．

このアプローチで問題となるのは，素性リストのスパース性 (Feature Sparseness) である．例えば素性として周辺の単語表記だけを利用する場合，訓練データから得られる素性の集合は，取り得る素性全体に対してスパースとなる．そのためテストデータでは訓練データには全く出現しない単語のみで素性リストが構成される場合も多く，その場合，正しく語義識別ができない可能性が高い．

スパース性を解消するためにシソーラスを利用することが一般に行われているが，トピックモデルを利用することも１つの方法である．トピックモデルとは文書 d の生起に K 個の潜在的なトピック z_i を導入した確率モデルである．

$$p(d) = \sum_{i=1}^{K} p(z_i)p(d|z_i)$$

トピックモデルである Latent Dirichlet Allocation (LDA) [1] を用いた場合，単語 w に対して $p(w|z_i)$ が得られることから，ある文脈 s を単語の集合 $\{w_1, w_2, \cdots, w_m\}$ で表し，$p(s|z_i)$ を以下で近似することで WSD の対象単語の用例に対してトピック z_i への関連度が得られる．

$$p(s|z_i) \approx \prod_{j=1}^{K} p(w_j|z_i)$$

これを利用して素性のスパース性に対処できる可能性がある．ここではトピックへの関連度によって作られた素性ベクトルを**トピック素性**と呼ぶことにし，トピック素性以外の従来使われていた WSD の素性ベクトルを**基本素性**と呼ぶことにする．

トピック素性を WSD に直接利用する方法は，基本素性にトピック素性を結合させた素性ベクトルを作成し，その素性ベクトルに対して学習手法を適用することである．ここでトピック素性の表現方法として，**ハードタグ**と**ソフトタグ**がある．ハードタグとはトピック素性のなかで最も関連度の高いトピックの次元のみを 1 にし，他は 0 としたベクトルである．またソフトタグとは各トピック z_i との関連度 $p(s|z_i)$ を第 i 次元の値としたベクトルである．Cai [4] は WSD においてハードタグよりもソフトタグの方が効果があることを示した．ただしトピック素性はシソーラスの情報であることを考えると，関連度の低いトピックの値は小さい値ではなく 0 である方が好ましいはずである．そこで本論文ではハードタグとソフトタグの中間のタイプである**ミドルソフトタグ**を提案する．ミドルソフトタグではトピック z_i との関連度 $p(s|z_i)$ がある閾値 θ よりも小さい場合に，第 i 次元の値を 0 にし，そうでなければ $p(s|z_i)$ とするものである．

実験では SemEval-2 の日本語 WSD タスク [8] のデータを利用した．このタスクは 50 単語の曖昧な各単語に対して，50 用例の訓練データと 50 用例のテストデータが用意されている．実験の結果，基本素性のみを用いた分類器による平均正解率は 76.76%，基本素性にハードタグのトピック素性を結合した素性を用いた分類器による平均正解率は 76.88%，基本素性にソフトタグのトピック素性を結合した素性を用いた分類器による平均正解率は 76.96%，そして基本素性に提案手法であるミドルソフトタグのトピック素性を結合した素性を用いた分類器による平均正解率は 75.96% となった．トピック素性を利用する効果は確認できたが，ミドルソフトタグの効果は確認できなかった．

2 トピックモデル

文書には潜在的にトピックが存在すると考え，そのトピックを用いることで文書の生成過程を確率的に表現したモデルがトピックモデルである．

古典的なトピックモデルとしては Probabilistic Latent Semantic Indexing (PLSI) が存在する [5]．ただし PLSI は一種のソフトクラスタリングであり，対象とする文書群の各文書に対するトピックへの帰属度は得られるが，文書群以外の文書（未知文書）に対してのトピックへの帰属度は得られないという問題がある．そこで提案されたトピックモデルが LDA である [1]．LDA ではトピックを語彙上の分布として，また文書をトピック上の分布としてモデル化する．このため未知文書に対してもそのトピックを推定することができる．

LDA では M 個の文書からなる文書群の生成を以下のように行う．文書のトピック分布 θ がディリクレ分布 $Dir(\alpha)$ からサンプリングされる．次に文書の語数 N 個になるまで以下の (a) と (b) を繰り返す．(a) トピック z_n が多項分布 $Mult(\theta)$ からサンプリングされる．(b) 単語 w_n が確率 $p(w_n|z_n, \beta)$ からサンプリングされる（図 1）．LDA は変分ベイズ法やギブズサンプリングを利用することで，α と β を求める．トピックの数を K，語彙数を V とすると，α は K 次元ベクトル，β は $K \times V$ の行列となる．そして β の i 行 j 列の要素が $p(w_j|z_i)$ となる．

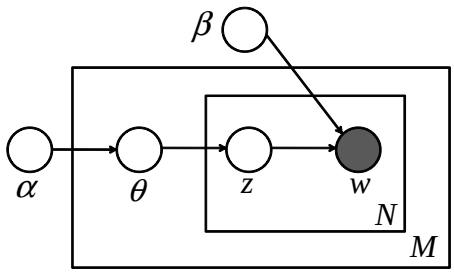

図 1: LDA のグラフィカルモデル

3 トピックモデルを利用した WSD

WSD にトピックモデルを利用する場合，間接的な利用と直接的な利用に分けられる．

間接的な利用とは WSD で必要とするリソースをトピックモデルで補強するようなタイプである．Cai は WSD に対してベイジアンネットワークを利用しているが，そのベイジアンネットワークに LDA から求めたトピック素性を組み込むことでもとのベイジアンネットワークを改良している [4]．Boyd-Graber は追加の潜在変数として WordNet の語義を LDA に組み込んだ．これによって WordNet から同義語グループ（synset）を探す処理の中に，トピックモデルを利用している [3]．Li はコーパスから語義の事前分布が得られる場合，そうでない場合，そしてコーパスの言い換えのリソースが不足していた場合の 3 つの状況に合わせた WSD のための確率モデルを構築する手法を提案している．この構築のためにトピックモデルを利用している [6]．

直接的な利用とはトピックモデルから得られるトピック素性を，直接 WSD に利用するタイプである．本研究もこのタイプに属する．Boyd-Graber は LDA によって単語と単語の周辺分布を求め，周辺分布の類似性から単語の語義を推定した [2]．ただし，これは教師なしの枠組みであるため，本論文での基本素性を WSD に利用しておらず，通常の教師あり学習により得られる分類器をトピックモデルを利用して改善する形ではない．前述した Cai の研究 [4] では，提案手法との比較手法として基本素性にトピック素性を結合した素性を利用した手法も実装している．そこではトピック素性を最大の関連度をもつトピックだけ 1 にし，残りを 0 にしたハードタグと，全てのトピックについてその関連度を使ったソフトタグを試しており，ソフトタグの方が優れていることを報告している．

4 提案手法

対象単語 w の用例 s の基本素性を $\vec{b}$，トピック素性を $\vec{t}$ で表す．

$$\vec{b} = (b_1, b_2, \cdots, b_N)$$

$$\vec{t} = (t_1, t_2, \cdots, t_K)$$

$\vec{b}$ のみを使った識別が通常の WSD となる．また基本素性にトピック素性を結合した素性 $append(\vec{b}, \vec{t})$ による識別がトピックモデルを直接的に利用した WSD である．

$$append(\vec{b}, \vec{t}) = (b_1, b_2, \cdots, b_N, t_1, t_2, \cdots, t_K)$$

ここで t_i はトピック z_i との関連度であり，LDA を利用して得られる $p(s|z_i)$ に対応する．

t_i の値として $p(s|z_i)$ を直接用いるのがソフトタグである．そして $\hat{i} = \arg\max_i p(s|z_i)$ のとき

$$t_i = \begin{cases} 1 & i = \hat{i} \\ 0 & i \neq \hat{i} \end{cases}$$

としたものがハードタグである．

LDA におけるトピックとは単語のクラスターであるため，トピックとは概念に相当する．そう考えた場合，単語は通常複数の意味があるので，1 つの概念に対応させるハードタグは現実的ではない．しかし単語をすべての概念と何らかの関連があると考えるのも妥当ではない．そこで本論文ではハードタグとソフトタ

グの中間にあたるミドルソフトタグを提案する．ミドルソフトタグとは t_i の値として $p(s|z_i)$ がある閾値 θ よりも小さい場合に 0 にし，そうでなければ $p(s|z_i)$ とするものである．閾値の設定が問題であるが，ここではトピック数 K の逆数に設定した．つまり $\theta = 1/K$ とした．

5 基本素性とトピック素性

5.1 基本素性の作成

本論文で利用した素性は以下の 8 種類である．(e0) w の表記，(e1) w の品詞，(e2) w_{-1} の表記，(e3) w_{-1} の品詞，(e4) w_1 の表記，(e5) w_1 の品詞，(e6) w の前後 3 単語までの自立語の表記，(e7) e6 の分類語彙表の番号の 4 桁と 5 桁．これらが基本素性となる．なお対象単語の直前の単語を w_{-1}，直後の単語を w_1 としている．

例えば以下は WSD の対象単語が 16 単語目の“経済”である文の形態素解析結果である．

```
<sentence>
<mor pos="名詞-固有名詞-組織名" rd="デンツー">電通</mor>
<mor pos="補助記号-読点" rd="、">、</mor>
<mor pos="名詞-固有名詞-組織名" rd="ハクホー">博報</mor>
<mor pos="接尾辞-名詞的-一般" rd="ドー">堂</mor>
<mor pos="助詞-格助詞" rd="オ">を</mor>
<mor pos="名詞-普通名詞-副詞可能" rd="ハジメ">はじめ</mor>
<mor pos="名詞-普通名詞-一般" rd="ジョーイ">上位</mor>
<mor pos="名詞-数詞" rd="ゴ">五</mor>
<mor pos="接尾辞-名詞的-助数詞" rd="シャ">社</mor>
<mor pos="助詞-副助詞" rd="クライ">くらい</mor>
<mor pos="助動詞" rd="ナラ" bfm="ダ">なら</mor>
<mor pos="名詞-普通名詞-一般" rd="エイチピー">ＨＰ</mor>
<mor pos="助詞-格助詞" rd="オ">を</mor>
<mor pos="動詞-一般" rd="ツクル" bfm="ツクル" >作る</mor>
<mor pos="形状詞-一般" rd="ジンテキ">人的</mor>
<mor pos="名詞-普通名詞-一般" rd="ケーザイ" sense="X">経済</mor>
<mor pos="接尾辞-形状詞的" rd="テキ">的</mor>
<mor pos="名詞-普通名詞-一般" rd="ヨユー">余裕</mor>
<mor pos="助詞-係助詞" rd="モ">も</mor>
<mor pos="動詞-非自立可能" rd="アル" bfm="アル" >ある</mor>
<mor pos="助動詞" rd="デショー" bfm="デス">でしょう</mor>
<mor pos="助詞-接続助詞" rd="ガ">が</mor>
<mor pos="補助記号-読点" rd="、">、</mor>
<mor pos="名詞-普通名詞-一般" rd="チューショー">中小</mor>
<mor pos="助詞-格助詞" rd="ノ">の</mor>
<mor pos="名詞-普通名詞-サ変可能" rd="ダイリ">代理</mor>
<mor pos="接尾辞-名詞的-一般" rd="テン">店</mor>
<mor pos="助詞-格助詞" rd="デ">で</mor>
<mor pos="助詞-係助詞" rd="ワ">は</mor>
<mor pos="連体詞" rd="ソンナ">そんな</mor>
<mor pos="名詞-普通名詞-一般" rd="ヨユー">余裕</mor>
<mor pos="助詞-係助詞" rd="ワ">は</mor>
<mor pos="動詞-非自立可能" rd="アリ" bfm="アル" >あり</mor>
<mor pos="助動詞" rd="マセ" bfm="マス">ませ</mor>
<mor pos="助動詞" rd="ン" bfm="ヌ">ん</mor>
<mor pos="補助記号-句点" rd="。">。</mor>
</sentence>
```

図 1. 対象単語「経済」の例文

ここから以下の基本リストが作成される．これをベクトル表現したものが基本素性となる．

e0=経済，e1=名詞-普通名詞-一般，e2=人的，
e3=形状詞，e4=的，e5=接尾辞，e6=人的，
e6=作る，e6=ＨＰ，e6=余裕，e6=ある，
e6=中小，e7=2386，e7=1197，e7=11972

5.2 トピック素性の作成

本論文ではトピック数 K を 100 とした．またコーパスとしては現代日本語書き言葉均衡コーパス（BCCWJ コーパス [7]）の一部を用いた[1]．次に，各文書から名詞を取り出し，名詞をインデックスとした文書索引語行列を作成した．ここから LDA ツール[2]を用いて，トピック $z_i (i = 1 \sim K)$ の下で名詞である単語 w の分布 $p(w|z_i)$ を求めた．

今，対象単語の用例 s 中に含まれる対象単語以外の名詞のリストを $\{w_1, w_2, \cdots, w_m\}$ とする．s のトピック素性は K 次元のベクトルとなり，第 i 次元の値はトピック z_i への関連度に対応する．関連度は基本的には LDA から求まる以下の $p(s|z_i)$ を利用する．

$$p(s|z_i) = \frac{1}{Z} \prod_{j=1}^{m} p(w_j|z_i)$$

ここで Z はトピック素性の大きさを 1 にする正規化定数である．

$$Z = \sum_{i=1}^{K} \prod_{j=1}^{m} p(w_j|z_i)$$

6 実験

データセットとして SemEval-2 の日本語 WSD タスクのデータを使用する．そこでは WSD の対象単語が 50 単語設定されている．各々の単語に対して 50 件のテストデータと約 50 件の訓練データが存在する．単語に対する 50 件のテストデータに対する正解率の平均をとったものを手法の正解率する．

また本論文では素性からの分類器の学習に SVM を利用した．SVM のツールとしては LIBSVM[3] を用いた．またカーネルは線形カーネルを利用した．

まず基本素性のみを用いた場合，正解率は 76.76% であった．次に基本素性にハードタグのトピック素性を結合した素性の場合，正解率は 76.88% であった．また基本素性にソフトタグのトピック素性を結合した素性の場合，正解率は 76.96% であった．最後に基本素性にミドルソフトタグのトピック素性を結合した素性の場合（提案手法），正解率は 75.96% であった．

この実験によりトピック素性を直接利用することで WSD の精度を改善できることが示された．ただし提案したミドルソフトタグの効果は確認できなかった．

7 考察

本論文ではハードタグとソフトタグの中間にあたるミドルソフトタグを提案した．一般にソフトタグが最も情報量が多いために，理論的には，これを使うのが最良であるが，現実的には値の低い部分は誤差と見なせるものであり，逆に悪影響があると考えた．またトピックが概念に相当するものだとすれば，関連度の低

[1] 具体的には SemEval-2 日本語 WSD タスクで配布されたもの 1,421 文書である．

[2] http://chasen.org/~daiti-m/dist/lda/

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

いトピックには 0 を与えた方が適切だと思える．これらの考えが妥当であれば，ミドルソフトタグが現実的には最も効果が高いはずである．ただし実験では効果は全くなかった．

いくつかの問題が考えられる．1 つの問題は閾値 θ の設定法である．ここではトピック数 K の逆数を閾値に設定している（$\theta = 1/K$）．この設定であれば少なくとも 1 つ以上の次元の値が 0 より大きな値になるので，ミドルソフトタグが自動で生成されるという利点がある．しかし閾値としてトピック数の逆数を使う明確な理由はなく，何らかの考えから閾値を設定すべきであった．今後，閾値を変化させた実験も行い，適切な閾値の設定法を考察する予定である．

またミドルソフトタグに割り当てる数値も問題であった可能性がある．ミドルソフトタグの場合，次元の値は閾値より小さいものを 0 にし，閾値以上のものはそのままの値を使っているので，ミドルソフトタグのトピック素性の大きさは 1 以下である．一方，ハードタグもソフトタグもトピック素性の大きさは 1 であり，この点が悪影響を及ぼした可能性もある．

またトピック数の問題もある．トピックモデルを利用した WSD では利用するトピック数が精度に大きく影響する．一般にトピック数は大きい方がよいと考えられるが，一概に大きければ良いというものでもない．参考として，トピック数 K を 50 にした実験も行った．この場合，ソフトタグを利用した正解率は 77.04%，ハードタグを利用した正解率は 77.28%，そしてミドルソフトタグを利用した正解率は 75.92% となった．全体的にトピック数が 100 のときよりもかなり良い値が出ている．しかもハードタグによるトピック素性を利用した場合の正解率は 77.28%であり，これは SemEval-2 の日本語 WSD タスクではかなりの高い値である[4]．またトピック数 50 ではハードタグの方がソフトタグよりもかなり正解率が高く，Cai [4] の実験とは逆の結果が得られている．つまりトピック数は有効なタグのタイプにも影響していると考えられる．

最後にトピック素性を直接利用するアプローチとしては，単純ではあるが，基本素性にトピック素性を結合するアプローチが有効であることを述べておきたい．トピック素性を直接利用するアプローチとしては，トピック素性を分離し，トピック素性のみからの識別の結果と基本素性のみからの識別の結果をアンサンブルするアプローチも考えられる [9]．実験結果は割愛するが，トピック素性のみからの識別の結果 (68.92%) が最大頻度を出力する識別の結果 (68.96%) よりも悪く，どのようにアンサンブルさせても良い正解率は得られなかった．この点からトピック素性を直接利用するのは，単純ではあるが，基本素性にトピック素性を結合するというアプローチが有効である．その場合，本論文で提案したミドルソフトタグにより更なる精度向上が可能だと考えている．

[4]SemEval-2 の日本語 WSD タスクの参加システム中最高の正解率は RALI-2 の 76.36% であった．

8 おわりに

本論文では WSD にトピックモデルを利用する手法を提案した．トピックモデルを直接 WSD に利用するには，基本素性にトピック素性を結合させる方式が，単純ではあるが，効果が高い．ただしその際にトピック素性をハードタグで表現するか，ソフトタグで表現するかのオプションがある．ここではそれらの中間に位置するミドルソフトタグを提案した．SemEval-2 の日本語 WSD タスクを用いた実験では，提案手法の効果は確認できなかった．ただし閾値，割り当てる数値およびトピック数が精度に関連しており，適切な設定が行えなかったためだと考えている．今後はこれらの適切な設定を見つけ，ミドルソフトタグの効果を示したい．

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Machine Learning Reseach*, Vol. 3, pp. 993–1022, 2003.

[2] Jordan Boyd-Graber and David Blei. Putop: Turning Predominant Senses into a Topic Model for Word Sense Disambiguation. In *SemEval-2007*.

[3] Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A Topic Model for Word Sense Disambiguation. In *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1024–1033, 2007.

[4] Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. Improving Word Sense Disambiguation using Topic Features. In *the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 1015–1023, 2007.

[5] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

[6] Linlin Li, Benjamin Roth, and Caroline Sporleder. Topic Models for Word Sense Disambiguation and Token-based Idiom Detection. In *ACL-2010*, pp. 1138–1147.

[7] Kikuo Maekawa. Balanced Corpus of Contemporary Written Japanese. In *the 6th Workshop on Asian Language Resources (ALR)*, pp. 101–102, 2008.

[8] Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. SemEval-2010 Task: Japanese WSD. In *The 5th International Workshop on Semantic Evaluation*, pp. 69–74, 2010.

[9] Ming wei Chang, Wen tau Yih, and Christopher Meek. Partitioned logistic regression for spam filtering. In *the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '08*, pp. 97–105, 2008.