トピックモデルを用いた語義曖昧性解消

西野太樹
茨城大学工学部情報工学科

新納浩幸
茨城大学工学部情報工学科

佐々木稔
茨城大学工学部情報工学科

1　はじめに

本論文では語義曖昧性解消（WSD: Word Sense Disambiguation）にトピックモデルを利用することを提案する．

WSDは自然言語処理の根幹技術のひとつであり，従来より多数の手法が提案されてきているが，現在は教師あり学習手法を用いるのが一般的である．そこではWSDの対象単語wを含む文sから，素性ベクトルfを作り，それをインスタンスとして学習や識別を行う．ここではwを含む文sだけではなく，文sを含む文書dが利用できるという設定のもとでwの語義曖昧性解消を行う．つまり訓練データやテストデータは対象単語wを含む文書セットとなる．この文書セットからトピックモデルを構築する．トピックモデルとは，文書のトピックをK個設定し，各トピック$z_i(i = 1 \sim K)$から文書dが生成される確率$P(d|z_i)$定めた確率モデルである．文書セットからトピックモデルが構築できれば，文書dがトピックz_iに属する確率$P(z_i|d)$が求まるので，文書dに対して以下のベクトルが構築できる．

$$(P(z_1|d), P(z_2|d), \cdots, P(z_K|d))$$

このベクトルをここでは トピックベクトル と呼ぶことにする．

本論文では通常作成される素性ベクトルfに上記のトピックベクトルをアペンドした形で，素性ベクトルを拡張する．この拡張された素性ベクトルを利用して，通常のWSDを行う．

実験ではSemEval-2のJapanese WSDタスク[6]により，トピックモデルを利用した効果を示す．また考察において，過去のトピックベクトルを利用したWSDの研究との比較も行う．

2　素性ベクトルの作成

現在，WSDの標準手法としては，教師あり学習である．ここではSVMを利用する．SVMを利用するには，WSDの対象単語wが与えられたときに，wに対する素性ベクトルを作成しなくてはならない．素性ベクトルは素性リストの各素性値をベクトルの各次元に対応させることで作成できる．

本論文で利用した素性は以下の8種類である．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．

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の周辺自立語の表記

e7 e6の分類語彙表の番号の4桁と5桁

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

ここからは以下の素性リストが作成される．

```
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
```

3 トピックベクトルの作成

本実験ではトピックモデルの構築に pLSI[4] を用いる．

pLSI では，アスペクトモデルと呼ばれる統計的なモデルを用いる．このアスペクトモデルは潜在的なクラス変数 $z \in Z = \{z_1, ..., z_K\}$ のための潜在変数モデルである．文書 d と単語 w の同時確率モデルは以下で定義される．

$$P(d,w) = P(d)P(w|d) \quad (1)$$

$$P(w|d) = \sum_{z \in Z} P(w|z)P(z|d) \quad (2)$$

以上から，ベイズの定理を用いて等価モデルを導出すると以下のようになる．

$$P(d,w) = \sum_{z \in Z} P(z)P(d|z)P(w|z) \quad (3)$$

モデルの最尤推定には EM アルゴリズムを用いる．EM アルゴリズムでは以下の E ステップと M ステップを交互に行う．E ステップでは，潜在変数 z の事後確率を算出する．計算には以下の方程式を用いる．これらは既出の式より得ることができる．

$$P(z|d,w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z' \in Z} P(z')P(d|z')P(w|z')} \quad (4)$$

M ステップでは，E ステップの事後確率よりパラメータを更新する．同様に得ることができる以下の式により，計算を行う．

$$P(w|z) \propto \sum_{d \in D} n(d,w)P(z|d,w) \quad (5)$$

$$P(d|z) \propto \sum_{w \in W} n(d,w)P(z|d,w) \quad (6)$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d,w)P(z|d,w) \quad (7)$$

pLSI を行うには前処理として，文書セットの各文書を形態素解析し，索引語文書行列を作成する．索引語文書行列は以下のように，各行が文書 d を表し，各列は単語 ID と文書内での単語の出現回数のペアの羅列となる．

```
0:1 1:1
2:1 3:2 4:1
1:2 4:1 5:1
```

トピックの数 K には対象単語 w の語義の数を設定する．

ここでは w 含む文 s 以外に文 s を含む文書 d も与えられるので，トピックモデル構築後は，w に対して，以下のトピックベクトルを作成することができる．

$$(P(z_1|d), P(z_2|d), \cdots, P(z_K|d))$$

4 実験

本手法の効果を確認するために，素性ベクトルをそのまま用いる場合と，素性ベクトルにトピックベクトルを追加した場合との各々の WSD の精度を比較する．

データセットには SemEval-2 の Japanese WSD タスクを使用する．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．更に各データを含む文書も提供されているので，本実験の設定も満たされている．

pLSI の学習はテストデータと訓練データの各データに対する文書，合計約 100 文書が対象である．この学習を各単語について行う．また pLSI の学習には，工藤拓氏のツール[1]を使用する．実行時のオプションで，対象単語の語義数をクラス数（トピック）として指定した．出力されたファイル群より，確率 $P(d|z_i)$ をモデル化した pzd ファイルをトピックベクトルとして用いる．

SVM による識別には，LIBSVM[2]を使用する．用いたカーネルは線形カーネルである．

実験結果を表 1 に示す．素性ベクトルだけを用いた場合と比べ，トピックベクトルを加えた場合は (+0.36%) の改善となった．また，トピックベクトルだけを用いた場合は素性ベクトルだけを用いた場合に比べ，(-9.0%) と精度が大幅に下がった．最後に，トピックベクトルの重み付けを変化させたところ，2 倍した場合は (-0.64%) と減少し，0.5 倍とした場合は (+1.12%) 上昇した．

[1] http://chasen.org/ taku/software/plsi/
[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm/

5　考察

トピックベクトルの重みについて考察する．今回の実験ではトピックベクトルの重みを大きくするほど全体の平均の精度は下がった．しかし，個別の単語について見ていくと，重みを大きくしていくと精度が下がるものと，精度が上がるものがある．つまり，単語によってトピックベクトルの有効性の度合いが異なるということである．有効性の度合いはトピックベクトルへの重みに対応するので，どのように適切な重みを求めていくかは今後の課題である．

次にトピックモデルを用いた語義曖昧性解消の関連研究について概説する．Cai らは，トピックモデルの一つである LDA(Latent Dirichlet allocation)[3] によって，ラベルなしコーパスからトピック特性を抽出した，教師ありシステムの手法を提案している [2]．Boyd-Graber は，追加の潜在変数として WordNet の語義を LDA に組み込んだ手法を提案している [1]．これは WordNet-WALK と呼ばれる，WordNet の同義語グループ（synset）を辿る確率的プロセスの考え方と，トピック特性とを組み合わせて利用する．Li らは，LDA を基に，コーパスから語義の事前分布が得られる場合と，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モデルを構築する手法を提案している [5]．第 1 のモデルではトピックと文書による語義の条件付き確率を最大化する．第 2 のモデルでは，第 1 モデルの条件の余弦値から語義の条件付き確率を間接的に最大化する．第 3 のモデルでは，語義の言い換えを構成する単語を使って語義の条件付き確率を最大化する．

上記 3 つの研究と本研究の大きな違いは，本研究では対象単語に対して文書が与えられるという仮定を置いている点である．そのためにトピックモデルを単語毎に学習している．単語毎にトピックモデルを学習する利点としては，トピック数をいくつに設定するかという問題を避けることができる点である．

また上記 3 つの研究はトピックモデルを学習するために LDA を用いている．本研究ではトピックモデルを利用する効果を示すことが目的であるので，簡易な pLSI を用いたが，当然 LDA を用いることも可能である．LDA を用いた場合は，更に精度が高まると予想している．

6　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．ただし対象単語 w を含む文 s 以外に，s を含む文書 d が利用できることを前提としている．この前提のもとで対象単語に対する文書群からトピックモデルを学習し，インスタンスのトピックベクトルを作成する．このトピックベクトルを通常の素性ベクトルに追加する形で WSD を行う．実験では SemEval-2 の Japanese WSD タスクのデータを用いて，提案手法の有効性を示した．トピックベクトルへの適切な重み付けが今後の課題である．

参考文献

[1] J. Boyd-Graber, D. Blei, and X. 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.

[2] J. Cai, W. S. Lee, and Y. W. 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.

[3] M. I. Jordan. D. M. Blei, A. Y. Ng. Latent dirichlet allocation. *Machine Learning Reseach*, Vol. 3, pp. 993–1022, 2003.

[4] 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.

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

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

表 1: 各手法による正解率

対象単語	素性ベクトルのみ	トピックベクトル追加 (重み 0.5)	トピックベクトル追加 (重み 1.0)	トピックベクトル追加 (重み 2.0)	トピックベクトルのみ
相手	84.00	84.00	84.00	84.00	82.00
あう	90.00	90.00	92.00	92.00	66.00
あげる	36.00	36.00	36.00	38.00	12.00
与える	76.00	78.00	76.00	74.00	58.00
生きる	94.00	94.00	94.00	94.00	94.00
意味	52.00	56.00	56.00	60.00	42.00
入れる	70.00	70.00	70.00	68.00	72.00
大きい	94.00	94.00	94.00	94.00	94.00
教える	42.00	42.00	44.00	44.00	18.00
可能	60.00	60.00	60.00	58.00	56.00
考える	98.00	98.00	98.00	98.00	98.00
関係	98.00	98.00	98.00	96.00	78.00
技術	84.00	84.00	84.00	84.00	84.00
経済	98.00	98.00	98.00	98.00	98.00
現場	76.00	78.00	76.00	76.00	78.00
子供	68.00	68.00	66.00	66.00	42.00
時間	86.00	86.00	86.00	88.00	88.00
市場	66.00	66.00	60.00	48.00	32.00
社会	86.00	86.00	86.00	86.00	86.00
情報	82.00	82.00	82.00	82.00	84.00
すすめる	82.00	86.00	86.00	86.00	32.00
する	60.00	62.00	62.00	54.00	38.00
高い	86.00	86.00	86.00	86.00	86.00
出す	44.00	44.00	42.00	40.00	28.00
立つ	54.00	54.00	54.00	52.00	52.00
強い	90.00	90.00	90.00	90.00	92.00
手	78.00	78.00	78.00	78.00	78.00
出る	60.00	58.00	58.00	58.00	58.00
電話	80.00	80.00	80.00	80.00	84.00
とる	38.00	40.00	40.00	30.00	28.00
のる	84.00	84.00	80.00	74.00	48.00
場合	86.00	88.00	88.00	88.00	86.00
入る	58.00	60.00	62.00	62.00	46.00
はじめ	96.00	96.00	96.00	94.00	66.00
はじめる	82.00	82.00	82.00	82.00	78.00
場所	96.00	96.00	96.00	96.00	96.00
早い	70.00	72.00	72.00	72.00	52.00
一	92.00	92.00	92.00	92.00	92.00
開く	92.00	92.00	92.00	90.00	90.00
文化	96.00	96.00	96.00	96.00	98.00
他	100.0	100.0	100.0	100.0	100.0
前	76.00	74.00	74.00	74.00	62.00
見える	76.00	78.00	78.00	74.00	40.00
認める	72.00	72.00	78.00	76.00	76.00
見る	80.00	80.00	80.00	80.00	80.00
持つ	66.00	68.00	68.00	68.00	68.00
求める	78.00	78.00	78.00	78.00	78.00
もの	88.00	88.00	88.00	88.00	88.00
やる	94.00	94.00	94.00	94.00	94.00
良い	38.00	40.00	40.00	40.00	24.00
平均	76.64	77.12	77.00	76.00	68.00