潜在的意味推定に基づく文書分類と対訳語生成

江里口 瑛子　(指導教員：小林 一郎)

1　はじめに

情報通信網の発達により，E-mail やニュース記事などのテキスト情報が増加している．これらのテキスト情報を対象とした情報処理技術に，文書のカテゴリを自動推定する自動文書分類技術や，ある言語で書かれたテキストを別の言語に変換する機械翻訳技術などがある．テキスト処理を行う際，問題となるのが言葉の文脈処理である．一般に，言葉の意味やニュアンスは一意に定まることはなく，文脈に応じてそれらを適切に捉える必要があるからである．

この背景から，pLSI や LDA[1] に代表される潜在的意味推定を利用した文書処理技術が多く提案されている．一般に，テキスト情報は巨大でスパースなベクトルであることが多いが，潜在的意味推定では，このベクトルに対して次元削減を行い，ある話題のまとまり，すなわち潜在的トピックの分布として表現する．本研究は，文書に内在する潜在的トピックを捉える潜在的意味推定技術に着目し，自動文書分類技術と対訳語推定技術の改善を目指すものである．潜在的意味推定技術を利用した，グラフに基づく文書分類手法と対訳語生成手法を新たに提案し，先行研究との比較を行う．

2　潜在的意味推定を利用したグラフに基づくテキスト分類

グラフに基づく半教師あり学習法 (Graph-Based Semi-Supervised Learning; GBSSL) は，マルチラベル文書分類タスクにおいて，SVM などの代表的な教師あり学習法や半教師あり学習法と比較して，精度が高いことで知られている．GBSSL の精度は，学習時のグラフの構成法に依存することが報告されている．

ここでは，文書の潜在的トピックを加味した新たなグラフ構成法を提案し，Reuters-21578 コーパスを用いた文書分類実験により提案手法が有効であることを示す．

2.1　提案手法

以下のような，文書に内在する潜在的意味を加味したグラフ構成を行う．ここでは，n 個の文書を頂点とし，文書間の類似度を各頂点間を結ぶ辺の重みとして持つような，重み付き無向グラフを構成する．

式 (1) は，文書 S，T 間の類似度算出式である．$sim_{surf}(S,T)$ は，文書 S，T の表層的な情報に基づく類似度 (表層的類似度) を表し，その類似度指標には，文書 S，T の $tfidf$ ベクトル間の cos 類似度を用いる．$sim_{lat}(S,T)$ は，文書 S，T の潜在的意味情報に基づく類似度 (潜在的類似度) を表し，その類似度指標には，S，T それぞれの潜在的トピック分布 P，Q の L2 ノルム距離をシグモイド関数で $[0,1]$ 範囲に写像した値を用いる．潜在的トピック分布は LDA [1] により推定する．最後に，式 (1) 中の α は，これら表層的類似度と潜在的類似度に関する重みパラメータである．

$$
\begin{aligned}
sim(S,T) &\equiv (1-\alpha) * sim_{surf}(S,T) \\
&\quad +\alpha * sim_{lat}(S,T), \qquad (1)\\
sim_{surf}(S,T) &= \cos(tfidf(S), tfidf(T)), \qquad (2)\\
sim_{lat}(S,T) &= \frac{2}{1+\exp^{L2(P,Q)}}. \qquad (3)
\end{aligned}
$$

2.2　実験

2.2.1　実験設定

実験対象データには，Reuters-21578 (Reuters) [1] を用いる．Reuters のデータは，マルチラベルを有しており，その種類数は 10 である．対象データセットは以下の内訳で 15 組用意し，うち 5 組は最適パラメータ設定に使用した．各組のデータセットの内訳は，教師データとしてランダムに選出した 20 文書と，各組で共通のテストデータとしての 3,299 文書である．文書分類学習には，グラフに基づく半教師あり学習法の 1 つであるラベル伝播 [2] を用いた．ラベル伝播は 2 値分類学習器のため，文書のマルチラベルを予測するために one-versus-rest 法を用いた．手法の評価指標には PRBEP を使用した．PRBEP は，Precision 値と Recall 値が一致するときの値である．

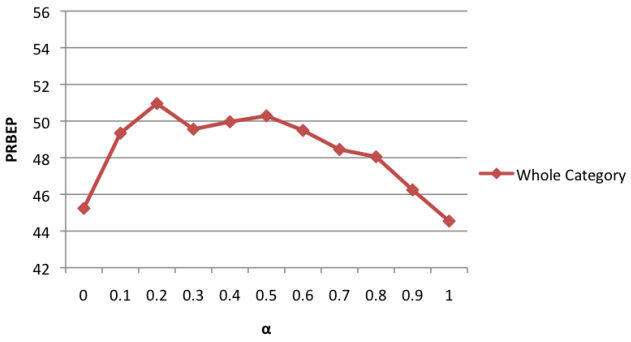

図 1: 全カテゴリ平均の PRBEP 値の変化

2.2.2　実験結果および考察

図 1 は，10 個のカテゴリに対する分類を行い，全カテゴリの平均値を求めたときの結果である．縦軸は PRBEP 値を，横軸はグラフ構成時に潜在的意味情報を考慮した割合 (α) を表す．

実験結果より，グラフ構成を行う際には，表層的な類似度と潜在的な類似度の両者を同時に用いた方が，分類精度が高くなることが分かった．また，今回実験に用いたコーパスでは，$\alpha = 0.2$ のときに精度は最大 (PRBEP= 51.0) となった．以上から，グラフ構成の際には，文書の表層的な情報に加えて，潜在的トピック情報を採用することで，グラフに基づく文書分類の精度が改善することが分かった．

[1] http://www.daviddlewis.com/resources/testcollections/

3　潜在的意味推定に基づく対訳語生成

対訳語生成とは．異なる言語間において翻訳関係にある単語の対応関係を求めるタスクである．代表的な手法に，分布仮説に基づく手法と単語アライメント手法がある．前者の手法の１つに，正準相関分析によるマッチング手法 [3] がある．この手法では，単語の素性として，文脈情報に加えて単語の綴り字情報を利用したマッチング手法を提案しており，特に，英語とスペイン語のような綴り字情報に相関性の高い言語対に対して有用性の高い手法となっている．しかしながら，日本語と英語のような異なる綴りを有する言語間において，Haghighi らの手法を適用することは難しい．そこで本研究では，複数言語間における潜在的意味推定を利用したマッチング手法を提案する．

3.1　提案手法

ここでは，MCCA で採用する単語の素性ベクトルとして，単語のスペリング情報ではなく，潜在的トピック分布を採用する．ここで扱う潜在的トピック分布は，LDA をマルチリンガルテキストを対象とするよう拡張した PLDA[4] により推定を行う．MCCA のモデルは以下に示す．

MCCA

$\mathbf{m}$ は一様分布で生成

各訳語対 $(i,j) \in \mathbf{m}$ に対して

(i,j) が対訳語ペアであるなら

$z_{i,j} \sim \mathcal{N}(0, I_d)$, [潜在空間]

$f_S(s_i) \sim \mathcal{N}(W_S z_{i,j}, \Psi_S)$, [s のベクトル空間]

$f_T(t_j) \sim \mathcal{N}(W_T z_{i,j}, \Psi_T)$. [t のベクトル空間]

言語 s の単語 i が対訳語に含まれない場合:

$f_S(s_i) \sim \mathcal{N}(0, \sigma^2 I_{d_S})$.

言語 t の単語 j が対訳語に含まれない場合:

$f_T(t_j) \sim \mathcal{N}(0, \sigma^2 I_{d_T})$.

3.2　予備実験

3.2.1　予備実験設定

ここでは，分布仮説に基づく手法として，文脈情報による手法 (Tf-Idf) と潜在的意味推定による手法 (LDA と PLDA) を，単語アライメント手法として，IBM-Model1，隠れマルコフモデルに基づいた HMM，潜在的意味推定による手法 (HM-BiTAM [5]) を行った．各手法における対訳語指標は，Tf-Idf では cos 類似度を，LDA と PLDA では Jensen-Shannon ダイバージェンスを利用した．単語アライメント手法では，各手法の学習の結果として求まる翻訳確率テーブルの値を利用した．潜在的意味推定におけるトピック数 k は $k \in \{5, 10, 20, 50, 100\}$ の範囲を動かす．

対象データには日本語と英語の１対１文対応コーパス (15,187 文対) を用いた．日本語の総単語数は 10,099 語，英語の総単語数は 11,132 語である．このうち，求める対訳語は日英共に名詞のみとする．本タスクにおける正解辞書には，日英・英日の EDR 電子辞書を用いる．また，手法の評価指標には Recall 値を用いる．

3.2.2　予備実験結果および考察

表 1 に，各手法で最大となったときの Recall 値をまとめる．但し，HM-BiTAM では計算が終わらなかったため，$k = 5, 20$ における結果の最大値が示されている．翻訳方向が日本語から英語 (日英)，英語から日本語 (英日) のいずれにおいても，全体を通して最も精度の高い手法は Tf-Idf となった．潜在的意味推定を利用した手法では，HM-BiTAM が最も精度が良くなった．これについては，HM-BiTAM が，潜在的意味推定に加え単語の翻訳確率も学習するモデルであるからだと考えられる．

表 1: Recall 値の比較

手法	日英	英日
Tf-Idf	**1**3.32	**1**8.78
LDA	0.19	0.06
PLDA	1.43	2.15
IBM-Model1	8.39	12.14
HMM Alignment	8.56	10.75
HM-BiTAM	7.00	**1**2.75

4　おわりに

本研究では，潜在的意味推定技術に基づいたテキスト分類手法と対訳語生成を提案した．前者では，潜在的トピック情報を反映したグラフ構成を提案し，マルチクラステキスト分類タスクにおけるラベル伝播法の精度を改善した．他方，後者では，潜在的トピック情報を利用した対訳語生成手法を提案した．対訳語生成タスクにおける代表的な手法および潜在的意味解析に基づく手法を，日英言語に適用し，各々の手法による精度を比較した．今後の課題として，今回予備実験で用いたコーパスに対して提案手法による同様の実験を行い，日英言語を対象とした潜在的意味推定に関する考察を行っていきたい．

参考文献

[1] D. M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol.3, pp. 993–1022, 2003.

[2] X. Zhu and Z. Ghahramani. Learning from Labeled and Unlabeled Data with Label Propagation. Technical report, Carnegie Mellon University, 2002.

[3] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein : Learning bilingual lexicons from monolingual corpora. In *Proc. of the ACL-08: HLT*, pp. 771–779, 2008.

[4] D. Mimno, H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Proc. of EMNLP*, pp. 880–889, 2009.

[5] B. Zhao and E. P. Xing. HM-BiTAM: Bilingual Topic Exploration, Word Alignment, and Translation. In *Proc. of NIPS*, pp. 1689–1696, 2007.