潜在的トピックに基づく対訳語生成への取り組み

江里口 瑛子　　小林 一郎
お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻
{g0920506, koba}@is.ocha.ac.jp

1 はじめに

対訳語生成手法の代表的な手法に，分布仮説に基づく手法と単語アライメントに基づく手法がある．後者の手法は，1 対 1 の対訳文コーパスから単語アライメントモデルの学習が行われるが，対訳文コーパスの収集過程では，一般に，翻訳家による対訳文生成が行われるため，莫大な時間と費用を要するという問題点がある．他方，前者の手法ではコーパスに対する制約はないものの，高精度を報告する既存の対訳語生成手法は，単語の綴り字情報に強く依存した手法である．日本語と英語のように異質な綴り字情報をもつ言語データ，および言語構造の系統が遠い言語データに対して有効な手法を検討するためには，各単語の用法や意味をコンピュータで処理可能なデータ表現で記述する必要がある．本論文では，複数の異なる言語に対応した潜在的意味解析手法に着目し，綴り字情報に代わる情報として複数言語で記述されたテキストに内在する潜在的意味を採用する対訳語生成手法を提案する．

2 関連研究

対訳語生成手法には，単語の文脈情報を利用した手法 [7] や Latent Dirichlet Allocation (LDA) に代表されるような潜在的意味解析を利用した手法 [8] がある．また，LDA を拡張させたモデルに，複数の言語で書かれた文書を対象とした多言語トピックモデル (Polylingual Latent Dirichlet Allocation; PLDA)[2] がある．このモデルでは，コンパラブルコーパスを学習することで，異なる言語で書かれた文書から言語横断情報としての潜在的トピックを抽出する．Zhu ら [4] は，PLDA によって得られた潜在的トピック情報の比較方法を提案し，英語と中国語のコンパラブルコーパスに適用している．

他方，Haghighi らは正準相関分析によるマッチング手法 [1] の提案を行っている．Haghighi らは，単語の素性ベクトルとして文脈情報と綴り字情報を統合したものを用いており，これらに対して正準相関分析によるマッチング (MCCA) 推定を行って，訳語候補の共起確率を計算した．この結果，言語構造の関係が近しいとされる英語とスペイン語のコーパスや，英語とフランス語のコーパスに関して，彼らは，高い精度で対訳語生成に成功した．しかし，英語と中国語のコーパスなど全く異質な言語同士では同様の精度は得られていない．これは，綴り字情報が単語の素性ベクトルとして適当ではなかったからことが理由として考えられる．これに対して，林ら [3] は日英コーパスを対象に，特定の単語に対してヒューリスティック値を設け，最大エントロピーモデルを用いて，素性ベクトルの重み付けに改良を加えたが，十全な結果は得られず，一部の単語ペア推定に対する改善に留まっている．

3 潜在的トピックに基づく手法の比較

本章では，潜在的意味推定に基づく対訳語生成手法の検討を行うため，予備実験として，分布仮説に基づく手法並びに単語アライメントに基づく手法の性能比較を行う．それぞれの手法のうち，顕在的情報のみに基づく手法，及び，既に提案されている潜在的意味解析手法に基づく手法を用いて，一対一文対訳コーパスにおける対訳語生成タスクの実験を行う．

3.1 分布仮説に基づく手法

「ある言語で共起する語があれば，翻訳後の言語でもそれらの翻訳語は共起する.」という分布仮説 [12] に基づき，着目している単語それぞれを，その単語の周辺に出現する単語の情報からなるベクトル表現 (周辺文脈ベクトル) で表す．この手法では，各単語はある言語共有空間 (周辺文脈ベクトル空間) に写像され，この空間における距離が近いもの同士を翻訳単語とする．本研究では，文脈ベクトルとして，Tf-Idf ベクトル，LDA によって推定された単語の潜在的トピック

分布，PLDA によって推定された単語の潜在的トピック分布，を採用する．

Tf-Idf では，単語の出現頻度 (Tf) 情報並びに文書の逆頻度 (Idf) 情報が考慮される．一方，LDA では，言語毎に対象文書群から単語の潜在的トピックが推定され，各言語で推定された潜在的トピック分布に対して制約は設けられていない．他方，PLDA では，対象文書群間では潜在的トピック分布が共有されるという仮定の下で，単語の潜在的トピックは推定される．

3.2 単語アライメントに基づく手法

単語アライメントに基づく手法として，本研究では，IBM Model 1[9]，HMM に基づく単語アライメント手法 [10]，HM-BiTAM[11] を用いる．IBM Model 1 は，単語の共起頻度情報に基づく単語アライメント手法である．HMM に基づく単語アライメント手法は，この IBM Model 1 で一様と仮定されていた単語アライメント確率を 1 次マルコフ過程に基づいて拡張したモデルとして提案されている．そして，HM-BiTAM は，更に，この HMM に基づく単語アライメント手法に対してトピックモデルの概念を導入した，潜在的トピックに基づく単語アライメント手法である．本予備実験では，HM-BiTAM による結果を潜在的意味推定に基づく単語アライメント手法の結果とする．

4 実験

4.1 実験仕様

データセットは，Business Letter から日英パラレルコーパスをそれぞれ 15,187 文用いた．実験前処理として，英語の文に対しては，オープンソースの統計的機械翻訳システムである Moses[1] の tokenizer を用いて単語と記号を切り分ける処理並びに，単語の直後に続く句点や読点の間に空白を挿入した．また，日本語の文に対しては，通常単語の区切り位置が文からでは不明虜なため，MeCab を用いて分かち書き処理を行い，単語間に空白の挿入を行った．

分布仮説に基づいた手法に関する実験設定を以下に記す，潜在的意味解析を利用する手法 (LDA, PLDA) では，潜在的トピック数パラメータ K は $K \in \{5, 10, 20, 50, 100\}$ の範囲を動かした．潜在変数を推定するアルゴリズムには，Collapsed Gibbs Sampler (CGS) を用い，その反復回数はいずれも 200 回とした．それぞれのモデルにおける Dir 分布のハイパーパラメータは，$\alpha = 0.1$，$\beta = 0.1$ とした．実験では，各手法に基づく周辺文脈素性の類似度あるいは距離を算出し，高い類似度を有する単語同士，あるいは近い距離にある単語同士を翻訳語とみなした．具体的には，Tf-Idf に基づく手法では cos 類似度を，LDA 並びに PLDA 手法では，潜在的トピック分布の分布間距離として，Jensen Shannon 距離を用いる．

単語アライメントに基づく手法に関する実験設定を以下に記す．HMM に基づくアライメント手法において NULL に割り当たる確率 p_0 は 0.01 に設定し，単語のジャンプ幅の初期値は 7 とした．IBM-Model 1, HMM に基づく単語アライメント手法では共に，EM アルゴリズムの反復回数は 5 回とする．翻訳単語候補は，学習後に得られる翻訳確率テーブルを参照し，翻訳確率が高い単語同士とする．

上記の手法によって得られた全訳語候補のうち，名詞単語に限定して評価を行う．英語の形態素解析には，TreeTagger を用い，日本語の形態素解析には，MeCab を用い，全ての単語に対して品詞を付与した．そして，日英辞書並びに英日辞書の正解辞書として，株式会社 日本電子化辞書研究所 (Japanese Electronic Dictionary Research Institute, Ltd.: EDR) が作成した EDR 電子化辞書をそれぞれ用いた．手法の評価指標には Recall 値を用いる．

4.2 実験結果

潜在的トピック数 $K \in \{5, 10, 20, 50, 100\}$ の範囲において，LDA，PLDA，HM-BiTAM が最大値の精度をとったのは，それぞれ $K = 20, 50, 20$ の時である．また，全ての潜在的トピック数にわたって，各言語において潜在的トピック推定を個別に行う LDA の結果よりも，複数の言語において同時に潜在的トピック推定を行う PLDA の結果の方が，精度は大幅に改善されている．そして更に，複数の言語において同時に潜在的トピック推定を行い，翻訳モデルの概念を導入している HM-BiTAM の結果は，$K = [5, 20]$ までの範囲に限定されるものの，LDA，PLDA の結果を大きく上回っている．

表 1 は，全比較手法における Recall 値をまとめたものである．表 1 には，最も精度の高かったときの値を示す．精度が最大値となったときのトピック数 K は，LDA，PLDA，HM-BiTAM (日英)，HM-BiTAM (英日) において，それぞれ $K = 20, 50, 20, 20$ の時である．

[1] http://www.statmt.org/moses/

表 1: 各手法における Recall 値.

分布仮説に基づく手法	日英	英日
Tf-Idf	**13**.32	**18**.79
LDA	0.19	0.06
PLDA	1.41	2.15
単語アライメント手法	日英	英日
IBM-Model1	8.39	12.14
HMM	8.56	10.75
HM-BiTAM	7.00	12.75

5 考察

表 1 から，英日，日英の両結果において，潜在的意味推定を用いた手法は，言語毎に潜在的意味推定を行う LDA よりも，複数言語で同時に潜在的意味推定を行う PLDA や HM-BiTAM の方が精度が改善しているのは，後者の手法の方が，言語を横断して同一の潜在的意味を推定することができているためだと考えられる．また，複数言語を対象とした潜在的意味推定手法のうち，HM-BiTAM による結果が PLDA の結果を遥かに上回っているのは，複数言語を対象とした潜在的意味推定に加えて，今回実験に利用したコーパスは一対一文対応の翻訳コーパスであり，HM-BiTAM モデルに最も適したコーパスである．そして，LDA 並びに PLDA は，文書 (複数の文からなる文集合) を対象としてモデル化されている．LDA 並びに PLDA に基づく手法による精度が HM-BiTAM の精度よりも低くなったのは，文レベルにおける潜在的トピック推定を十分に行うことができなかったためだと考えられる．

また，本実験では，日英，英日共に，潜在的意味推定を利用しない Tf-Idf による手法が最も精度が高かった．まず，分布仮説に基づく手法間においては，上記に挙げた理由と同様に，今回使用したコーパスにおいては，LDA 並びに PLDA では，潜在的意味推定を十分に行うことができなかったことが理由として考えられる．各単語アライメント手法においては，日英においては，IBM-Model1 の結果よりも HMM の結果が上回っているが，英日においては，逆転している．HM-BiTAM では，英日においては，IBM-Model 1，HMM の精度は上回ったものの，日英においては，これら 2 つの手法と比べて精度は下回った．日本語から英語へ，あるいは英語から日本語へ翻訳する方向によって，それぞれの学習モデルの結果は異なり，いずれのモデルにも一長一短がある結果となった．日本語と英語の言語構造が大きく異なることはよく知られている．HMM や HM-BiTAM では，単語アライメントの際にアライメントのジャンプをモデルに取り入れており，語順を考慮した推定が理論的には可能となっている．しかしながら，いずれのモデルにおいても，大きなジャンプになるほどより小さな確率値が与えられるため，語順の大きく異なる日英言語間に適用した際に，その特徴を捉えきれていないことが示唆される．

6 提案手法: 潜在的トピックに基づく対訳語推定

本研究では，林ら [3] と同様に日英コーパスを対象に，MCCA の抱える素性ベクトルの綴り字情報の問題点に対して，トピックモデルを用いてコーパスから推定される潜在的トピック情報を新たな素性として採用する手法の提案を行う．本研究における潜在的トピック推定手法には，多言語テキストを対象とした潜在的意味解析として有用であることが報告されている PLDA を用いる．

6.1 MCCA モデル

MCCA(Matching Canonical Correlation Analysis) は，多変量解析手法の 1 つである正準相関分析を利用した，対訳語推定手法である [1]．単語の素性ベクトルとして，その単語の文脈情報と綴り字情報を統合したものを採用し，正準相関分析と割当問題を反復して解くことで対象にしている複数言語のパラレルな単語ペア (対訳語) をそれぞれ求める．

$\mathbf{s} = (s_1, s_2, \cdots, s_{n_S})$ は翻訳元言語 (ソース言語) の単語集合を，$\mathbf{t} = (t_1, t_2, \cdots, t_{n_T})$ は翻訳先言語 (ターゲット言語) の単語集合を表し，$(i, j) \in \mathbf{m}$ は単語 s_i, t_j が対応関係にある (対訳語である) ことを表している．

～MCCA における対訳語生成モデル～

$\mathbf{m}$ 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:

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

6.2 EM アルゴリズムによるパラメータ推定

式 (1) は，MCCA モデルにおける対数尤度関数である．EM アルゴリズムを用いて，モデルのパラメータ $\theta = (W_S, W_T, \Psi_S, \Psi_T)$ の最尤推定を行う．θ は，各言語の素性ベクトルの生成過程として定めた，多変量正規分布のパラメータである．

MCCA 手法では，対訳語推定問題は，2 部グラフの重み付き最大マッチングとして扱われる．まず，E-step では，モデルパラメータ θ 下で異なる言語間の単語に与えられた重み $w(i,j)$ が最大となるときの，単語のマッチング関係 $\mathbf{m} \in \mathcal{M}$ を求める．次に，M-step では，E-step で重み付き最大となったマッチング $\mathbf{m}$ 下で正準相関分析 (Canonical Correlation Analysis; CCA) を行い，多変量正規分布のパラメータ θ の更新を行う．

$$Likelihood(\theta) = \log p(\boldsymbol{s}, \boldsymbol{t}; \theta) = \log \sum_{m} p(\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{t}; \theta). \quad (1)$$

7 実験

7.1 実験仕様

対象データセットとして，Business Letter(日英パラレルコーパス) からそれぞれ 15,187 文を用いた．MCCA で取り扱う単語データには，対象コーパスで出現頻度の高かった 100 語の名詞を用いた．各単語の潜在的トピック推定は PLDA の CGS によって行い，その反復回数は 200 回とした．潜在的トピック数のパラメータ K は $K \in \{5, 10, 20, 50, 100\}$ の範囲を動かし，Dir 分布のハイパーパラメータは，$\alpha = 0.1$，$\beta = 0.1$ とした．EDR 辞書を正解辞書とし，Recall 値による評価を行う．

7.2 実験結果および考察

表 2 は，提案手法による実験結果である．日英，英日それぞれにおいて，最適なトピック数は K=10, 20 となった．

得られた対訳語，及び正解データを詳しく見てみると，適切な対訳語が生成されていても，正解辞書に記載されていない場合や，$n = 100$ 件の中に対訳語が含まれていない場合が見受けられた．精度を改善するためには，MCCA で扱うデータ数を増やし，対象コーパスに合致した正解データを用いる必要がある．

表 2: 提案手法による Recall 値.

	日英	英日
提案手法	**1**.0	**2**.0

8 おわりに

MCCA の抱える素性ベクトルの綴り字情報の問題点を改善するため，単語の素性ベクトルとして，PLDA によって推定される潜在的トピック分布を新たに採用した，潜在的トピックに基づく対訳語推定手法の提案を行った．日英パラレルコーパスから対訳語を生成する実験では，頻度情報に基づく Tf-Idf の結果が最も良くなったものの，対象コーパスに対して適切なモデルを作り，潜在的トピック推定手法を導入することが精度向上に繋がることも確認した．今後，提案手法による精度を更に改善するため，パラレルでない日英コーパスデータを用いた実験や，MCCA で扱う単語数を更に増やした場合の変化に関する考察を行って行きたい．

参考文献

[1] A. Haghighi et al.. Learning bilingual lexicons from monolingual corpora. In *Proc. of the ACL, pp. 771–779, 2008.*

[2] D. Mimno et al.. Polylingual topic models. In *Proc. of EMNLP, Vol. 2, pp. 880–889, 2009.*

[3] 林 克彦ら.MCCA モデルの日英辞書構築への適用. 言語処理学会第 *16* 回年次大会,*pp. 982–985, 2010.*

[4] Z. Zhu et al.. Building comparable corpora base on bilingual LDA model. In *Proc. of the ACL, pp. 278-282, 2013.*

[5] I. Vulić et al.. Identifying words translations from comparable corpora using latent topic models. In *Proc. of the ACL: Human Language Technologies, Vol. 2, pp. 479–484, 2011.*

[6] X. Ni et al.. Mining multilingual topics from Wikipedia. In *Proc. of Web Search and Data Mining, pp. 375–384, 2011.*

[7] P. Fung and L. Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of COLING and ACL, pp. 414–420, 1998.*

[8] J. Preiss. Identifying comparable corpora using LDA. In *Proc. of the NACACL, pp. 558–562, 2012.*

[9] P. F. Brown et al.. The mathematics of statistical machine translation: parameter estimation. *Journal of Computational Linguistics – Special issue on using large corpora: II, Vol. 19, pp. 263–311, 1994.*

[10] S. Vogel et al.. HMM-based word alignment in statistical translation. In *Proc. of Computational linguistics*, Vol. 2, pp. 836–841, 2006.

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

[12] Zellig Harris. Distributional structure. Word, 10:146-162, 1954.