多言語トピックモデルによるパラレルコーパス生成

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

機械翻訳とは，1 つの言語を他の言語へ機械的に変換する作業のことである．この機械翻訳には，大きく分けて 2 種類の手法があり，1 つは規則ベース機械翻訳手法であり，もう 1 つは統計的機械翻訳手法である．両手法に共通する問題点としては，機械翻訳が扱う対象の自然言語には曖昧性や例外が多分に含まれているということがある．前者の手法は，言語間の翻訳規則を恣意的に定める．しかし，全ての翻訳規則を網羅的に記述することが難しいという欠点がある．これに対して，後者の手法は，翻訳規則を統計的・確率的に定める．これによって，規則を網羅的に記述することが可能となり，後者の手法には，自然言語の曖昧性や例外に対応できるという利点がある．

この後者の手法は，Noisy channel モデル [1] によって，更に翻訳モデルと言語モデルに大別され，これら 2 つのモデルは対訳コーパス (パラレルコーパス) を用いて自動学習される．しかし，複数の言語でパラレルに書かれた文書は希少である．一般的に，翻訳家によるパラレルコーパスの生成が極めて高コストであるからである．他方，Web 上においては，Wikipedia やニュース記事などに見られるような，同一内容に関してそれぞれの言語で書かれた文書 (コンパラブルコーパス) は多く存在する．今日，これらコンパラブルコーパスを利用したパラレルコーパスの自動生成に対する関心が高まっている．

本研究は，データに内在する潜在的トピック，並びに，データに基づいて仮定した潜在空間に注目し，それらを用いたパラレルコーパスの自動生成の手法を提案するものである．多言語トピックモデルの手法を用いて，複数の言語で書かれた文書から潜在的トピックを推定し，得られた言語横断情報に対して正準相関分析によるマッチング (MCCA; Matching Canonical Correlation Analysis) 推定を行い，パラレルコーパスを自動生成することを目的とする．

2　関連研究

コンパラブルコーパスを利用したパラレルコーパス生成手法には，単語の文脈情報を利用した手法 [2, 3]，翻訳用辞書と単語の出現頻度回数を利用した手法 [4]，そして Latent Semantic Indexing (LSI) [5] や Latent Dirichlet Allocation (LDA) [6] などの潜在的意味解析を利用した手法 [7, 8, 9] がある．

多言語文書を対象に LDA を拡張させたモデルとして，多言語トピックモデル [10, 11, 12] が提案されている．コンパラブルコーパスを提案モデルで学習することにより，文書に内在すると仮定した潜在的トピックに基づく言語横断情報の抽出を行うことができる．Vulić ら [13] は，多言語トピックモデルに基づく潜在的トピックの観点から単語の類似度測定を行う手法を提案し，英語とイタリア語のコンパラブルコーパスに適応した．また，Zhu ら [14] は，多言語トピックモデルによって得られた言語横断情報の比較方法を提案し，英語と中国語からなるコンパラブルコーパスに適応した．この他，多言語文書分類 [15, 16] などの多言語文書処理タスクにおいても多言語トピックモデルは利用されている．

他方，Haghighi らは正準相関分析によるマッチング手法 [17] を提案している．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，高い精度のパラレルコーパス生成に成功した．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高い精度は得られなかった．その理由としては，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だと考えられている．これに対して，林ら [18] 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 正準相関分析による対訳語推定

MCCA(Matching Canonical Correlation Analysis) とは，単一言語で書かれた文書集合 (単言語コーパス) から対訳語を抽出するために提案された確率的手法である [17]．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
$\mathbf{m}$ 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:
　$f_T(t_j) \sim \mathcal{N}(0, \sigma^2 I_{d_T})$.

3.1 パラメータ推定

対数尤度関数 (式 (1)) を最尤推定することによってパラメータ θ の推定を行う．ここで，$\theta = (W_S, W_T, \Psi_S, \Psi_T)$ は各言語の素性ベクトルの多変量正規分布モデルのパラメータである．パラメータ θ の推定には，EM アルゴリズムを用いる．

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

E-step では，現在のモデルパラメータから重み付き最大となる単語の関係 $\mathbf{m} \in \mathcal{M}$ を求める．M-step では，E-step で得られた $\mathbf{m}$ の下で正準相関分析を行い，各多変量正規分布モデルパラメータの更新を行う．

3.2 M-step

M-step では，正準相関分析を用いて最適パラメータ θ の推定を行う．与えられた単語の対応関係 $\mathbf{m}$ に対して対数尤度関数を最大にするパラメータを求めるため，式 (1) は式 (2) に置き換えることができる．

$$\max_{\theta} \sum_{(i,j) \in \mathbf{m}} \log p(s_i, t_j; \theta). \quad (2)$$

式 (2) によって新たに示された最尤推定問題は，正準相関分析によって解くことができる．言語の特徴ベクトルをそれぞれ射影し，射影先の各特徴ベクトルを比較した際，相関が最大となるように固有値ベクトル U_S, U_T を固有値問題として求めることで，パラメータ θ は式 (3–6) より求まる．

$$W_S = C_{SS} U_S P^{\frac{1}{2}}, \quad (3)$$
$$W_T = C_{TT} U_T P^{\frac{1}{2}}, \quad (4)$$
$$\Psi_S = C_{SS} - W_S W_S^{\mathrm{T}}, \quad (5)$$
$$\Psi_T = C_{TT} - W_T W_T^{\mathrm{T}}, \quad (6)$$
$$C_{SS} = \frac{1}{|\mathbf{m}|} \sum_{(i,j) \in \mathbf{m}} f_S(s_i) f_S(s_i)^{\mathrm{T}}. \quad (7)$$

ここで，P は $d \times d$ の正準相関係数行列を表す．C_{TT} は，C_{SS} と同様に共分散行列の計算で求めることができる．

3.3 E-step

E-step では，単語間の重み付き最大マッチング $\mathbf{m} \in \mathcal{M}$ を求める．M-step で求めた θ と式 (8) を用いることで，ソース言語の単語とターゲット言語の単語の対応関係情報を求めることができる．

$$\mathbf{m} = \operatorname*{argmax}_{\mathbf{m}'} \log p(\mathbf{m}', \mathbf{s}, \mathbf{t}; \theta). \quad (8)$$

ただし，計算量を抑えるために，式 (8) をそのまま解くのではなく，単語のマッチング最大化問題 (式 (9)) に置き換えて解く．ここで，式 (10) は，ソース言語の単語 i とターゲット言語の単語 j 間のマッチング辺の重み (対訳語となる確率) を表す．

$$\log p(\mathbf{m}, \mathbf{s}, \mathbf{t}; \theta) = \sum_{(i,j) \in \mathbf{m}} w_{i,j} + C, \quad (9)$$
$$w_{i,j} = \log p(s_i, t_j; \theta) - \log p(s_i; \theta) - \log p(t_j; \theta). \quad (10)$$

4 提案手法: 潜在的トピックによるパラレルコーパス生成

本研究では，林ら [18] 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を改善する手法の提案を行う．具体的には，多言語トピックモデルによって得た言語横断情報 (単語のもつ潜在的トピック分布 ϕ^l) を単語の素性ベクトルに採用し，MCCA 推定を行い，パラレルコーパスを生成し，精度の確定を行う．

4.1 多言語トピックモデル

PLDA(Polylingual Latent Dirichlet Allocation) [10] とは，複数言語で書かれた文書を文書組とみなし，この文書組を同時に分析するため，トピックモデルの枠組みに基づいて提案された手法である．我々は，PLDA を日英コーパスに対して用いる．

パラレルでない多言語文書を対象にした処理手法では，「同一内容に関して書かれた文書であれば同じ意味の単語が同じ頻度で出てきやすい」という仮定の下で，単語の共起情報や文脈情報などに着目した研究がなされてきた [2, 3, 4]．Mimno ら [10] は，この仮定を「同一内容に関して複数言語で書かれた文書組であれば，各文書組内に含まれる話題 (潜在的トピック) の比率 (θ) は等しい」という仮定の下，多言語トピックモデルを提案した．

図 1 は PLDA のグラフィカルモデルを表す．背景が白色の変数は潜在変数を表し，背景が灰色の変数は観測変数を表す．各言語 $l = 1, \cdots, L$ に対して，言語毎のトピック分布集合 $\Phi^1, \cdots, \Phi^L$ ($\Phi^l = \{\phi^l_1, \cdots, \phi^l_K\}$) が存在する．

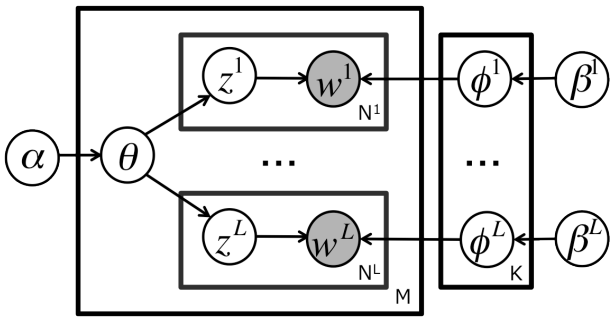

図 1: PLDA のグラフィカルモデル

PLDA の生成過程は以下の通りである．$\boldsymbol{w} = (\boldsymbol{w}^1, \cdots, \boldsymbol{w}^L)$ は L 種類全ての言語の文書集合を表す．ここで，$Dir(\cdot)$ はディリクレ分布を表し，$Dir(\cdot, \cdot)$ はディリクレ過程，$\boldsymbol{m}$ は base measure，w^l_n は言語 l の n 番目の単語，z^l_n は言語 l の n 番目の単語の潜在的トピック，ϕ^l_k は言語 l のトピック k の単語分布，そして θ_k はトピック k の文書分布を表す．ただし，本研究で用いる多言語トピックモデルは，$L = 2$ のときの PLDA とする．

1. 言語 l の各トピック $k = 1, \cdots, K$ について:

$$\phi^l \sim Dir(\beta^l). \tag{11}$$

2. 言語 l の各文書 $d^l = 1, \cdots, M$ について:

$$\boldsymbol{\theta} \sim Dir(\boldsymbol{\theta}, \alpha \boldsymbol{m}). \tag{12}$$

(a) 言語 l の各単語 $w^l_n = 1, \cdots, N^l$ について:

$$\boldsymbol{z}^l \sim P(\boldsymbol{z}^l | \boldsymbol{\theta}), \tag{13}$$
$$\boldsymbol{w}^l \sim P(\boldsymbol{w}^l | \boldsymbol{z}^l, \Phi^l). \tag{14}$$

5 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，MCCA 推定の際に，多言語トピックモデルを用いて得た言語横断情報を単語の素性ベクトルに採用し，対訳語推定を行う手法の提案を行った．

今後，日英コーパスデータを用いて，PLDA による多言語文書への潜在的トピック情報の推定と，MCCA による訳語対マッチングを行い，提案手法の検証を行う．

参考文献

[1] P. F. Brown, V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: parameter estimation. *Journal of Computational Linguistics – Special issue on using large corpora: II, Vol. 19, Issue 2, pp. 263–311, 1993.*

[2] R. Rapp. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting on the ACL, pp. 320–322, 1995.*

[3] P. Fung and L. Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th COLING and 36th ACL, pp. 414–420, 1998.*

[4] T. Vu, A. T. Aw, and M. Zhang. Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th Conference of EACL, pp. 843–851, 2009.*

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science, Vol. 41, pp. 391–407, 1990.*

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

[7] M. Littman, S. T. Dumais, and T.K. Landauer. Automatic cross-language information retrieval using latent semantic indexing. *Cross-Language Information Retrieval, Chap. 5, pp. 51–62, 1998.*

[8] Y. Tam, I. Lane, and T. Schultz. Bilingual-LSA based LM adaption for spoken language translation. In *Proceedings of the 45th Annual Meeting of the ACL, pp. 520–527, 2007.*

[9] J. Preiss. Identifying comparable corpora using LDA. In *Proceedings of the 2012 Conference of the NAACL: Human Language Technologies, pp. 558–562, 2012.*

[10] D. Mimno. H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on EMNLP, Vol. 2, pp. 880–889, 2009.*

[11] X. Ni, J. Sun, J. Hu, and Z. Chen. Mining multilingual topics from Wikipedia. In *Proceedings of the 18th International Conference on WWW, pp. 1155–1156, 2009.*

[12] W. De Smet and M. Moens. Cross-language linking of news stories on the Web using interlingual topic modeling. In *Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining, pp. 57–64, 2009.*

[13] I. Vulić, W. De Smet, and M. Moens. Identifying words translations from comparable corpora using latent topic models. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies, Vol. 2, pp. 479–484, 2011.*

[14] Z. Zhu, M. Li, L. Chen, and Z. Yang. Building comparable corpora based on bilingual LDA model. In *Proceedings of the 51st Annual Meeting of the ACL, Vol.2, pp. 278-282, 2013.*

[15] W. De Smet, J. Tang, and M. Moens. Knowledge transfer across multilingual corpora via latent topics. In *Proceedings of the 15th PAKDD, pp. 549–560, 2011.*

[16] X. Ni, J. Sun, J. Hu, and Z. Chen. Cross lingual text classification by mining multilingual topics from Wikipedia. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 375–384, 2011.*

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

[18] 林 克彦, 福西 孝章, 西田 昌史, 山本 誠一. MCCA モデルの日英辞書構築への適用. 言語処理学会第 *16* 回年次大会発表論文集, *pp. 982–985, 2010.*