教師なしマッピングによる言語横断テキスト分類

平尾 努　　岩田 具治　　永田 昌明
日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
{hirao.tsutomu,iwata.tomoharu,nagata.masaaki}@lab.ntt.co.jp

概 要

言語横断テキスト分類では，目的言語の文書をを原言語の文書が属する空間へと写像する必要がある．このためには，機械翻訳システムやコンパラブル・コーパス，対訳辞書などの言語資源から得た知識が必要であった．しかし，機械翻訳システムの構築，コーパス整備にかかるコストは多大である．本稿では，こうした資源を利用することなく，目的言語の文書を原言語の文書が属する空間へとマッピングする教師なしマッピングを用いて言語横断テキスト分類を行う．目的言語と原言語の文書をより低次の K 次元空間へと独立に写像した後，2つの空間の依存関係が最大になるように基底間の対応付けを行う．こうして目的言語の文書を原言語空間へマッピングすることで言語横断テキスト分類を可能にする．

1　はじめに

言語横断テキスト分類とは，原言語で記述されたカテゴリラベル付き文書で訓練した分類器を用いて，目的言語で記述されたカテゴリラベルなし文書を分類することである．例えば，カテゴリが付与された日本語の文書を用いて訓練した分類器を用いて英語の文書を分類することがこれに相当する．

この問題に対する最も単純な解決策は，機械翻訳システムを利用することであろう．上記の例では，英語文書を機械翻訳システムを用いて日本語に変換しておけば，日本語文書を用いて訓練した分類器でも問題なく英語文書を分類することできる．しかし，機械翻訳システムの構築には時間，金銭のコストが多大にかかるうえ，文書の翻訳にかかる時間も無視できない．また，機械翻訳システムを利用せずとも対訳辞書を利用し，単語単位で翻訳を行えば同様のことが可能であるが，どのような言語対，ドメインにおいても対訳辞書が利用可能なわけではない．

別の解決策として，原言語，目的言語の双方の文書データをそれらに共通する潜在トピック空間へと同時に写像し，同じ空間のデータとして取り扱うことが考えられる．CCA (Canonical Correlation Analysis) や LSA (Latent Semantic Analysis)，あるいは，トピックモデルなどを利用することで，潜在トピック空間へと文書データを写像することができるが，原言語と目的言語の間で文書単位での対応のとれたコーパス (コンパラブル・コーパス) や対訳辞書が必要となる．

このように，従来法では，対訳辞書，コンパラブル・コーパスなど人間の労力を必要とする言語資源や機械翻訳システムが必要であり，コストが高い．本稿では多大な労力を要する資源を必要としない言語横断テキスト分類手法を新たに提案する．提案手法では，原言語と目的言語との間の関係が与えられていないという前提のもと，原言語，目的言語の文書データをそれぞれ独立な K 次元空間へと写像し，それらの空間の依存関係が最大になるよう，2つの空間の間の対応をとる．そして，この低次元空間での対応関係を利用し，目的言語の文書を原言語文書の空間へとマッピングし，原言語で訓練した分類器を用いて目的言語の文書を分類する．なお，本稿では対応関係が与えられていない2つの空間に対し，一方に属するデータを他方の空間へとマッピングすることを教師なしマッピングと呼ぶ．

2　関連研究

原言語と目的言語との間に何らかの対応関係が与えられている場合，それらに共通する潜在トピック空間を仮定することができる．たとえば，文献 [Dumais 96] では，文書間の対応関係のついた 2 言語コーパスに対し LSA を適用することで，それらを同時に，より次元の低い空間へ写像する手法，文献 [Platt 10] では，同様の設定で，OPCA (Oriented Principle Component Analysis)，CCA を利用して写像する手法を提案している．LDA (Latent Dirichlet Allocation) [Blei 03] の拡張である MuTo (Mul-

tilingual Topic model) [Boyd-Graber 09]，JointLDA [Jagarlamudi 10]，PLSA (Probabilistic Latent Semantic Analysis) [Hofmann 99] の拡張である PCLSA (Probabilistic Cross Lingual Latent Semantic Analysis) [Zhang 10] などに代表される多言語対応のトピックモデルを利用することでも目的言語，原言語の文書データをそれらに共通する潜在トピック空間へ写像することができる．

しかし，これらの手法には文書単位，あるいは概念単位で対応付けのあるコンパラブル・コーパスや対訳辞書が大量に必要であり，多大なコストがかかる．さらに，マイナーな言語間ではコンパラブル・コーパスや対訳辞書を用意することがそもそも困難であるという問題もある．

一方，上述したように目的言語，原言語文書を 1 つの空間へと写像せず，それらを元の空間に残したまま文書間の対応関係を得ることでも言語横断テキスト分類は可能である．このようにコンパラブル・コーパスや対訳辞書を利用しない手法として，教師なしオブジェクトマッチング手法である KS (Kernelized Sorting)[Quadrianto 10]，MCCA (Matching CCA)[Haghighi 08]，LSOM (Least-Squares Object Matching)[Yamada 11] を利用することができる．だだし，教師なしオブジェクトマッチングは目的言語の文書数と原言語の文書数が同じであるという強い制約のもとでしか利用できないため，汎用性に欠ける．さらに，これらの手法では組合せ最適化問題の一種である線形割当問題を解く必要があり，それには $O(n^3)$(n はオブジェクトの数) の計算量が必要なため，オブジェクト数が多い場合には計算量が深刻な問題となる．

3 教師なしマッピング

先に述べた KS，MCCA，LSOM は一種の教師なしマッピング手法とみなすことができるが，目的言語，原言語の文書数が同じでなければ利用できないという問題がある．本稿ではこれを一般化し，文書数が異なる場合でも利用できる教師なしマッピング手法を提案する．

3.1 トピック抽出

いま，n 番目の原言語文書を多次元ベクトルデータ $\boldsymbol{s}_n = (\boldsymbol{s}_{ni})_{i=1}^{I}$ とし，そのクラスラベルをを y_n とする．分類器は $\{\boldsymbol{s}_n, y_n\}_{n=1}^{N}$ を訓練データとして学習する．同様に m 番目の目的言語文書も同様に多次元ベクトルデータ $\boldsymbol{t}_m = (\boldsymbol{t}_{mj})_{j=1}^{J}$ として表す．それぞれのデータを行列 $\boldsymbol{S} = (\boldsymbol{s}_n^\top)_{n=1}^{N}$，$\boldsymbol{T} = (\boldsymbol{t}_m^\top)_{m=1}^{M}$ として表す．すなわち，原言語文書集合は I 行 N 列，目的言語集合は J 行 J 列の行列として表される．

ここで，それぞれの行列を 2 つの低ランク行列の積として以下のように近似する．

$$\boldsymbol{S} \approx \boldsymbol{W_S}\boldsymbol{H_S} \tag{1}$$

$$\boldsymbol{T} \approx \boldsymbol{W_T}\boldsymbol{H_T} \tag{2}$$

$\boldsymbol{W_S}$ は I 行 K 列，$\boldsymbol{H_S}$ は K 行 N 列の行列であり，$\boldsymbol{W_T}$ は J 行 K 列，$\boldsymbol{H_T}$ は K 行 M 列の行列である．K は任意の整数であり，$K < \{I, J\}$ である．$\boldsymbol{H_S}$ の n 番目の列に対応するベクトル $\boldsymbol{h}_{\boldsymbol{s}_n}^\top$ は，I 次元の原言語文書ベクトル ($\boldsymbol{s}_n$) をより低次元の K 次元空間へと写象したデータとして解釈できる．また，$\boldsymbol{W_S}$ の k ($1 \leq k \leq K$) 番目の列に対応するベクトル $\boldsymbol{w}_{\boldsymbol{s}_k}^\top$ は，K 次元空間の基底 k に対する特徴の重みベクトル，すなわちトピック分布として解釈できる．同様に，$\boldsymbol{H_T}$ の m 番目の列に対応するベクトル $\boldsymbol{h}_{\boldsymbol{t}_m}^\top$ は，J 次元の目的言語文書ベクトル ($\boldsymbol{t}_m$) をより低次元の K 次元空間へと写象したデータとして解釈でき，$\boldsymbol{W_T}$ の k ($1 \leq k \leq K$) 番目の列に対応するベクトル $\boldsymbol{w}_{\boldsymbol{t}_k}^\top$ は，k 番目のトピックの分布として解釈できる．

3.2 トピック間の対応付け

$\boldsymbol{H_S}$ により，原言語文書集合を I 次元空間から K 次元空間へ，$\boldsymbol{H_T}$ により，目的言語文書集合を J 次元空間から原言語文書集合を写像した先と同じ大きさの K 次元空間へ写像することができる．こうして写像した空間はその大きさが K であるという共通点はあるが互いに独立な空間である．

ここで，原言語，目的言語から得たトピック，つまり，$\boldsymbol{W}_S$ の任意の列ベクトルと $\boldsymbol{W}_T$ の任意の列ベクトルとの間に対応関係があると仮定することはごく自然であろう．ただし，どちらのベクトルも異なる言語から得たトピック分布であるため直接類似度などを計算することはできない．よって，原言語，目的言語のそれぞれのドメイン内でのトピック間の類似性に依存性があると仮定する．すると，α を任意の正の定数として以下の式を考えれば良い．

$$||\boldsymbol{K_S} - \alpha\boldsymbol{K_T}||^2 = 0 \tag{3}$$

$\boldsymbol{K}$ は，$\boldsymbol{W}$ の列ベクトル同士の内積 (カーネル値) を要素とする対称行列，つまり，グラム行列である．こ

のような $\boldsymbol{W_T}$ が求まれば，原言語文書ををを写像した K 次元空間の各基底と目的言語文書を写像した K 次元空間の各基底が 1 対 1 の対応関係となる．すると，目的言語文書 $\boldsymbol{T}$ は，$\boldsymbol{W_S H_T}'$ によって I 次元の原言語文書と同じ空間にマッピングできるので，原言語文書を用いて訓練した分類器を用いて目的言語文書を分類できる．

最終的には，以下の目的関数を最小化するパラメータ集合 $(\boldsymbol{W_S}, \boldsymbol{H_S}, \boldsymbol{W_T}, \boldsymbol{H_T})$ を求めることで文書カテゴリの予測が可能となる．

$$\begin{aligned} E &= \|\boldsymbol{S} - \boldsymbol{W_S H_S}\|^2 \\ &+ \|\boldsymbol{T} - \boldsymbol{W_T H_T}\|^2 \\ &+ \beta||\boldsymbol{K_S} - \alpha \boldsymbol{K_T}||^2 \end{aligned} \quad (4)$$

パラメータ集合は，式 (4) に対し勾配法を適用することで求めることができるが，本稿では式 (4) の各項を個別に最小化する手法を採用した．具体的には第 1 項，2 項に対しては NMF (Non-negative Matrix Factorization) [Lee 00]，第 3 項には KS (Kernelized Sorting) を用いた．

4 評価実験

4.1 実験の設定

文献 [Prettenhofer 10] のアマゾンのレビューデータセットを用いて評価実験を行った．このデータセットは,「Book」,「DVD」,「Music」という 3 つのカテゴリそれぞれに対し，英語，フランス語，ドイツ語，日本語で記述されたレビュー文書から成る．原言語と目的言語の組み合わせは，以下の設定を試し，ラベルなし目的言語の文書を教師なしマッピングで原言語空間へとマッピングし，原言語で学習した分類器で先述した 3 つのカテゴリのいずれかに分類した．なお，分類器として k 最近傍法を採用した[1]．

- **原言語**:日本語，**目的言語**:英語，フランス語，ドイツ語
- **原言語**:ドイツ語，**目的言語**:英語，フランス語，日本語

目的言語の文書数は各言語とも 600，各カテゴリの文書数はそれぞれ均等に 200 とした．原言語の文書数は，目的言語と同じ場合の 600，それよりも少ない 300 の 2 通りを試した．目的言語の場合と同様各カテゴリに属する文書数は均等にした．原言語と目的言語の文書数が同じ場合 (600 文書の場合) には，教師なしオブジェクトマッチングを用いて言語横断テキスト分類が可能であるため，KS (Kernelized Sorting) と比較手法として採用した[2]．

提案手法は NMF を用いているが，NMF の結果は初期値に大きく依存する．そこで，初期値を変えて 100 回の試行を行い，その中で，目的関数 (式 (4)) が最も小さくなった時を評価結果に採用した．さらに，行列分解 (式 (1)，式 (2)) のパラメタ K は 60 とした．

4.2 実験結果と考察

表 1 に日本語を原言語とした場合の評価結果，表 2 にドイツ語を原言語とした場合の評価結果を示す．

表 1 より，KS の 3 カテゴリの平均正解率は 0.35 から 0.37 程度であり，ランダム (平均 0.33) よりもやや良い結果である．提案手法の場合，目的言語が日本語，訓練データ数が 600 の時に KS よりもやや劣る結果となったが，それ以外では KS よりも良い．また，訓練データ数は 600 よりも 300 の場合が良い結果であった．

カテゴリごとに提案手法と KS とを比較すると，その特性は大きく異なる．KS はどのカテゴリに対しても正解率の差が小さいことに対し，提案手法はカテゴリによってばらつきがある．原言語が日本語の場合，DVD カテゴリでの正解率が特に低いことに対し，Book，Music では 0.5 を超える正解率を得ており，KS との差も大きい．原言語がドイツ語の場合，DVD カテゴリの正解率が低いことは日本語を原言語とする場合と同様であるが， Book カテゴリの正解率も低い．しかしその一方，Music の正解率は非常に高く 0.8 を超える場合もある．

このように提案手法の正解率がカテゴリごとにばらつきがある理由は NMF を用いてトピック抽出を行ったことにある．たとえば，DVD カテゴリには映画 (物語) のレビューやライブ映像に関するレビューが混在しており，Book や Music カテゴリのレビューと似たレビューがある程度の数あろう．よって， NMF を用いてトピック抽出を行った際，DVD カテゴリに特有なトピックがうまく抽出されない可能性が高い．こうした問題を解決するためには，原言語側のトピック抽出をカテゴリ情報に基づき行うなどの改善が必要である．

[1] もちろん，他の分類器，サポート・ベクトルマシンやロジスティック回帰，パーセプトロンなどを用いても構わない．

[2] KS を用いて原言語と目的言語の文書間の 1 対 1 対応をとり，目的言語の文書に対して対応付けられた原言語文書のカテゴリをその予測とした．

表 1: 目的言語が日本語の場合の評価結果 (正解率)

	英語				フランス語				ドイツ語			
手法	Book	DVD	Music	平均	Book	DVD	Music	平均	Book	DVD	Music	平均
KS	.375	.370	.360	.368	.360	.370	.325	.352	.385	.300	.355	.347
提案手法 (600)	.325	.050	.660	.345	.320	.105	.535	.320	.355	.110	.525	.330
提案手法 (300)	.655	.030	.465	.383	.645	.055	.550	.417	.500	.030	.605	.378

表 2: 目的言語がドイツ語の場合の評価結果 (正解率)

	英語				フランス語				日本語			
手法	Book	DVD	Music	平均	Book	DVD	Music	平均	Book	DVD	Music	平均
KS	.380	.350	.310	.347	.300	.355	.315	.323	.355	.335	.375	.355
提案手法 (600)	.170	.225	.685	.360	.315	.210	.625	.383	.165	.390	.545	.367
提案手法 (300)	.255	.005	.840	.367	.205	.050	.815	.357	.190	.100	.845	.378

5 おわりに

本稿では，教師なしマッピング，つまり，原言語，目的言語の文書データをそれぞれ独立な K 次元潜在トピック空間へと写像した後，それらの空間の依存関係が最大になるように 2 つの空間の基底の対応を決定する手法を提案した．これにより，機械翻訳システムや対訳コーパスなどの資源を利用することなく言語横断テキスト分類が可能となる．アマゾンのレビューデータを対象として評価実験を行った結果，Kernelized Sorting と比較して良い結果を得た．

参考文献

[Blei 03] Blei, D. M., Ng, A., and Jordan, M.: Latent Dirichlet Allocation, Vol. 3, pp. 993–1022 (2003)

[Boyd-Graber 09] Boyd-Graber, J. and Blei, D. M.: Multilingual Topic Model for Unaligned Text, in *Proceedings of the UAI 2009*, pp. 75–82 (2009)

[Dumais 96] Dumais, S. T., Landauer, T. K., and Littman, M. L.: Automatic Cross-Linguistic Information Retrieval using Latent Semantic Indexing, in *Proceedings of the Workshop on Cross-Linguistic Information Retieval in SIGIR*, pp. 16–23 (1996)

[Haghighi 08] Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D.: Learning Bilingual Lexicons from Monolingual Corpora, in *Proceedings of ACL-08: HLT*, pp. 771–779 (2008)

[Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Analysis, in *Proceedings of the UAI 1999* (1999)

[Jagarlamudi 10] Jagarlamudi, J. and Daume III, H.: Extracting Multilingual Topics from Unaligned Corpora, in *ECIR* (2010)

[Lee 00] Lee, D. D. and Seung, H. S.: Algorithm for Non-negative Matrix Factorization, in *Advances in Neural Information Processing Systems* (2000)

[Platt 10] Platt, J. C., Toutanova, K., and Yih, W.-t.: Translingual Document Representation from Discriminative Projections, in *Proceedings of the EMNLP 2010*, pp. 251–261 (2010)

[Prettenhofer 10] Prettenhofer, P. and Stein, B.: Cross-Language Text Classification using Structural Correspondence Learning, in *Proceedings of the ACL* (2010)

[Quadrianto 10] Quadrianto, N., Smola, A. J., Song, L., and Tuytelaars, T.: Kernelized Sorting, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 32, No. 10, pp. 1809–1821 (2010)

[Yamada 11] Yamada, M. and Sugiyama, M.: Cross-Domain Object Matching with Model Selection, in *Proceedings of the UAI 2011* (2011)

[Zhang 10] Zhang, D., Mei, Q., and Zhai, C.: Cross-Lingual Latent Topic Extraction, in *Proceedings of the 48th ACL* (2010)