単語分散表現のタスク横断写像に基づく高精度多言語モデル

佐久間 仁
東京大学大学院 情報理工学系研究科
jsakuma@tkl.iis.u-tokyo.ac.jp

吉永 直樹
東京大学 生産技術研究所
ynaga@iis.u-tokyo.ac.jp

1 はじめに

英語など言語資源の豊富な言語では，膨大な注釈付きデータを用いた深層学習（表現学習）により，多くの自然言語処理タスクで大幅な精度向上が報告されている．しかし実世界に存在する数千の言語に対して各自然言語処理タスクの大規模注釈付きデータを用意するのは非現実的である．その結果，自然言語処理タスクにおける精度の言語間格差が大きく広がっている．

この問題を解決するために，言語資源の豊富な言語（原言語）で膨大な訓練データを用いて獲得したモデルを，言語資源に乏しい（特に，対象タスクに付いて注釈付きデータが存在しない）言語（目的言語）に対して適用する多言語モデルが提案されている．これらのモデルの多くは，多言語単語分散表現という言語に依存しない単語分散表現を注釈なしデータから事前学習し，深層学習モデルの単語埋め込み層に用いることで，言語間での語彙の差を吸収する．ここで，単語埋め込み層はモデル学習時に固定され，与えられたタスクに対して最適化されないため，深層学習の真価を発揮することができないという課題がある [1, 2].

そこで本研究では，事前学習した多言語単語分散表現から，原言語の訓練データを用いて学習した深層学習モデルの単語埋め込み層（タスク特化単語分散表現）へのタスク横断写像を獲得することで高精度な多言語モデルを実現する．具体的には，事前学習した多言語単語分散表現において近い表現を持つ単語はタスク特化単語分散表現においても近い表現を持つと仮定し，局所的な単語間の関係性を保存するように写像を学習する手法（locally linear mapping）を提案する (§ 3)．得られた写像を用いて目的言語のタスク特化単語分散表現を獲得すれば，原言語で学習した高精度の深層学習モデルをそのまま用いることが可能となる．

実験では文書分類と感情分析に本手法を適用し，英語を原言語として複数の目的言語で評価した (§ 4)．その結果，全言語・タスクにおいて性能向上を確認した．

2 関連研究

言語資源の乏しい言語を対象として，機械翻訳を活用し言語資源豊かな言語の資源を活用する手法が広く試みられている [3, 4, 5]．しかし，高精度の機械翻訳モデルの学習には大規模な対訳コーパスが必要であるため，このアプローチは適用状況が大きく制限される．

この問題に対し，深層学習モデルの単語埋め込み層を多言語単語分散表現に固定する手法が多言語モデルとして近年盛んに研究されている [6, 7, 8, 9, 10]．4 節でも実験的に確認するように，これらの手法ではタスクに特化した単語分散表現を用いることができず，深層学習（表現学習）の利点を十分に活かすことができない．文字レベルの分散表現を用いることで，多言語単語分散表現に依存せず多言語モデルを獲得する手法 [11, 12] も提案されているが，これらの手法は異なる文字体系を持つ言語対間では適用することができない．

Gouws らは品詞解析タスクに特化した対訳辞書を構築することで，タスクに特化した多言語単語分散表現を獲得した [13]．この手法では，特定の自然言語処理タスクに特化した対訳辞書を構築した（例えば品詞解析であれば同一品詞の単語が対訳関係にあるとする）．しかし，多くのタスクにおいては対訳関係を決定するための基準が明確でない上，辞書を構築する人的コストも必要となる．これに対し本研究では，原言語で推定したタスク特化単語分散表現からタスク横断写像を学習することで対訳辞書に依存せずタスク特化多言語単語分散表現を獲得した．

3 タスク特化多言語モデル

本節では，単語埋め込み層を含めた全パラメタが対象タスクに対して最適化された多言語深層学習モデルの構築手法を提案する．提案手法では，より広範な言語・タスクに足して手法を適用できるよう，原言語にのみラベル付きデータが存在すると仮定する．

以下に，タスク特化多言語モデルの獲得手法を示す．

Step 1: 多言語単語分散表現の獲得 まず既存手法 [14] を用いて，原言語と目的言語の単語を同一意味空間に埋め込んだ多言語単語分散表現 X^{gen}, Y^{gen} を獲得する．以降，ここで獲得した単語分散表現を汎用多言語単語分散表現と呼ぶ．

Step 2: 原言語に対する深層学習モデルの学習 次に，原言語の注釈付きデータを用いて深層学習モデルを学習する．この結果得られるモデルの単語埋め込み層 X^{spec} は原言語のタスクに特化した単語分散表現となっている．

Step 3: 単語分散表現のタスク横断写像の学習 Step 1 で得られた汎用的な多言語単語分散表現 X^{gen} から Step 2 で得られたタスク特化単語分散表現 X^{spec} へのタスク横断写像を学習する (後述)．得られた写像を用いて，Y^{gen} を Step 2 で得られた深層学習モデルのための単語埋め込み Y^{spec} に変換する．

Step 4: 目的言語に対する深層学習モデルの構築 Step 2 で獲得した深層学習モデルの単語埋め込み層 X^{spec} を Step 3 で獲得した目的言語のタスクに特化した単語分散表現 Y^{spec} に置き換えることで目的言語に適用可能な深層学習モデルを得る．

Locally Linear Mapping に基づく単語分散表現のタスク横断写像　注釈なしデータから事前学習した原言語と目的言語の汎用多言語単語分散表現 X^{gen}, Y^{gen} と原言語のタスク特化単語分散表現 X^{spec} から，目的言語のタスク特化単語分散表現 Y^{spec} を計算する．本手法は，locally linear embeddings [15] から着想を得たもので，学習時に考慮したタスクに依らず，単語分散表現の間の局所的な構造に共通性があることを仮定している．言い換えると，あるタスクに対する単語分散表現で十分に表現が近い単語は別のタスクの単語分散表現でも表現が近くなることを仮定する．

まず，汎用多言語単語分散表現に注目し，目的言語の各単語 i に対して原言語の近傍単語 k 個を cosine 類似度を用いて得る．次に，Y_i^{gen} を得られた近傍単語の汎用単語分散表現の重み付け平均で近似する．

$$\alpha_{ij} = \arg\min_{\alpha_{ij}} \left| Y_i^{\text{gen}} - \sum_{j \in N_i} \alpha_{ij} X_j^{\text{gen}} \right|, \quad \sum_j \alpha_{ij} = 1$$

ここで，N_i は目的言語の単語 i の近傍となる単語集合であり，α_{ij} は X_j^{gen} に対する重みである．この最適化問題は解析解

$$\alpha_{ij} = \frac{\sum_k C_{ijk}^{-1}}{\sum_j \sum_k C_{ijk}^{-1}}$$

を持つ．ただし，

$$C_{ijk} = \left(Y_i^{\text{gen}} - X_j^{\text{gen}}\right) \cdot \left(Y_i^{\text{gen}} - X_k^{\text{gen}}\right)$$

である．

得られた α_{ij} を用いて，前述の局所性仮定のもと，目的言語のタスク特化単語分散表現を計算する．

$$Y_i^{\text{spec}} = \sum_{j \in N_i} \alpha_{ij} X_j^{\text{spec}}.$$

得られた Y^{spec} は原言語のタスクに特化した単語分散表現 X^{spec} と同一の意味空間にあり，局所的な構造は写像の前後で保存されている．この手法のハイパーパラメタは近傍の単語数 k のみであり，簡単な計算により最適解が得られるという利点を待つ．

ハイパーパラメタ k の調整　本手法が想定する状況ではハイパーパラメタ k の調整のための開発データが存在しない．そこで様々な k に対し原言語においてタスク横断写像を適用してタスクに特化した単語分散表現表現を獲得し，その単語分散表現を用いた深層学習モデルを原言語の開発データにおいて評価し，最も高い性能を発揮した k を採用する．実験では，この手法を用いた結果と目的言語の小さな開発データ（100 サンプル）を用いて k を調節した結果を比較する．

4 実験

既存研究 [6, 3] に倣い，文書分類タスクと感情分析タスクで提案手法の評価を行った．全ての実験において英語を原言語とし，他の言語を目的言語とした．

4.1 設定

本節では実験設定について説明を行う．まず，評価タスクのデータセットを説明し，その後，実験に用いた汎用多言語分散表現と深層学習モデルについて述べる．なお，目的言語の開発データは，3 節末尾で述べた提案手法のハイパーパラメタ k を最適化する手法でのみで用いた．

言語	例数	平均単語数
English (en)	673,768	237.0
Spanish (es)	14,997	159.0
German (de)	86,550	195.8
Danish (da)	8,366	172.2
French (fr)	71,292	256.9
Italian (it)	21,594	137.5
Dutch (nl)	1,690	229.0
Potuguiss (pt)	6,263	249.0
Swidish (sv)	10,383	162.3

表 1: 文書分類タスクに用いたデータセット．

文書分類タスク　文書分類タスクのデータセットは既存研究 [6] に倣い RCV1/RCV2 コーパス [16] を用いた．このデータセットでは，各文書に対し “Corporate/Industrial”, “Economics”, “Government/Social”, “Markets” の 4 カテゴリのいずれかが付与されている．原言語（英語）のデータセットは，評価データと開発データとして 10,000 サンプルづつをランダムに選び残りを訓練データとした．目的言語のデータセットは，100 サンプルをランダムに開発データとし，残りを評価データとした．データセットの詳細を表 1 に示す．

評判分析タスク　評判分析の原言語（英語）の注釈付きデータには，Yelp Review dataset[1] を用いた．このデータセットではレストランのレビューに対して 1 から 5 の評価が書き手により与えられている．既存研究 [2] に倣い，評価が 1，2 のレビューを “negative”，評価が 4，5 のサンプルを “positive” とし，評価が 3 のレビューは除外した．既存研究 [3] に倣い，ラベル間のサンプル数を揃えるために “positive” ラベルに対してダウンサンプリングを行った．その後，評価データと開発データとして 100,000 サンプルずつをランダムに選び，残りを訓練データとした．目的言語のデータとして，ABSA dataset [17] を用いた．このデータは各言語でのレストランのレビューと，その各文に対して感情極性が与えられている．各言語に対して，100 文を開発データとして用いて残りを評価データとした．データセットの詳細を表 2 に示す．

全データセットに対し，NLTK[2] ツールを用いて単語分割を行った上で，単語は小文字化した．

汎用多言語単語分散表現の学習　次に，既存手法 [14] を用いて汎用多言語単語分散表現を獲得した（手法の実装には著者の実装[3] を用いた）．この手法は二言語において学習した単語分散表現を，直交変換により回転し多言語単語分散表現を得る．各言語の単語分散表現としては Subword-information skip-gram を用いて Wikipedia コーパスから得られた学習済み単語分散表現[4] を用いた．

コーパス	言語	例数	平均単語数
Yelp Review	English (en)	4,406,965	133.0
ABSA	English	1,513	14.0
	Spanish (es)	1,411	15.1
	Dutch (nl)	1,148	14.1
	Turkish (tr)	878	9.7

表 2: 評判分析タスクに用いたデータセット．

モデル　本手法は，単語埋め込み層を持つ任意の深層学習モデルに対して適用可能である．本研究ではシンプルな bag-of-embeddings モデルを対象タスクを解く深層学習モデルとして用いて提案手法の評価を行った．このモデルは，入力中の単語の単語分散表現の平均に対して一層の順伝播型ネットワークを適用する．

多言語モデルにおいて，提案手法で獲得するタスク特化単語分散表現の効果を調べるため，以下の手法を比較する．

汎用固定 単語埋め込み層を事前学習した汎用多言語単語分散表現に固定した深層学習モデル [6].

タスク特化 学習時に単語埋め込み層を含めて深層学習モデルを最適化し，タスク横断写像により目的言語のタスクに適用可能とする提案手法（§ 3）．3 節で述べた 2 つのハイパーパラメタ調整手法を比較する．

汎用固定＋ FFNN 上記の深層学習モデルの単語埋め込み層の直後に embedding-wise の 2 層順伝播型ニューラルネットワークを追加し，単語埋め込み層を事前学習した汎用多言語単語分散表現に固定した深層学習モデル．追加した順伝播型ニューラルネットワークは汎用多言語単語分散表現からタスクに特化した単語分散表現への写像を学習すると期待される．

各モデルは 3 回学習し，分類精度の平均を示す．

[1] https://www.yelp.com/dataset
[2] https://www.nltk.org/api/nltk.tokenize.html
[3] https://github.com/artetxem/vecmap
[4] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md

手法（単語埋め込み層）	en-es	en-de	en-da	en-fr	en-it	en-nl	en-pt	en-sv
汎用固定	0.376	0.767	0.635	0.665	0.542	0.662	0.477	0.803
汎用固定＋ **FFNN**	0.669	0.697	0.571	0.778	0.569	0.744	0.424	0.466
タスク特化（原言語で k を調整）	0.666	0.753	0.697	**0.854**	0.569	0.799	0.557	0.812
タスク特化（目的言語で k を調整）	**0.724**	**0.788**	**0.718**	0.840	**0.617**	**0.823**	**0.588**	**0.820**

表 3: 文書分類タスクにおける分類精度.

手法（単語埋め込み層）	en-es	en-nl	en-tr
汎用固定	0.802	0.736	0.695
汎用固定＋ **FFNN**	0.773	0.705	0.679
タスク特化（原言語で k を調整）	0.825	0.759	**0.712**
タスク特化（目的言語で k を調整）	**0.826**	**0.763**	0.709

表 4: 感情分析タスクの分類精度.

4.2 結果

文書分類タスクと感情分析タスクでの評価結果をそれぞれ表 3 と表 4 に示す．全言語対において，提案手法（**タスク特化**）がベースライン（**汎用固定**）を上回った．これにより，我々が提案するタスク横断写像で獲得したタスク特化単語分散表現の有効性が確認できた．また，ハイパーパラメタ k の調整は，目的言語の小さな開発データを用いた場合の方が精度は良かったが，原言語を用いても十分な精度向上が得られている．

次に，提案手法（**タスク特化**）が比較手法（**汎用固定＋ FFNN**）の精度を上回ったことから，本手法は多言語モデルにおいてより複雑な深層学習モデルを用いるよりも効果的であることが分かった．いくつかの言語において，比較手法（**汎用固定＋ FFNN**）はベースライン（**汎用固定**）よりも低い分類精度となっているが，これはニューラルネットがより多層になることによりモデルが単語分散表現のノイズにより敏感になったためと考えられる．

5 おわりに

本研究では，目的言語の教師データや言語横断的な資源に依存せずに完全にタスクに特化した多言語モデルを獲得する新たな手法を提案した．この手法では，タスク横断写像を用いてタスクに特化した多言語単語分散表現を獲得する．実験を通して提案したタスク横断写像が正しくタスクに特化した単語分散表現を獲得していること，また，この手法により最終的なモデルの分類精度が大きく向上することが確認できた．

謝辞 本研究の一部は、情報通信研究機構の委託研究の成果です。

参考文献

[1] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014.

[2] B. Shi, Z. Fu, L. Bing, and W. Lam. Learning domain-sensitive and sentiment-aware word embeddings. In *ACL*, 2018.

[3] X. Wan. Co-training for cross-lingual sentiment classification. In *ACL*, 2009.

[4] P. Prettenhofer and B. Stein. Cross-Language Text Classification Using Structural Correspondence Learning. In *ACL*, 2010.

[5] K. Xu and X. Wan. Towards a universal sentiment classifier in multiple languages. In *EMNLP*, 2017.

[6] L. Duong, H. Kanayama, T. Ma, S. Bird, and T. Cohn. Multilingual training of crosslingual word embeddings. In *EACL*, 2017.

[7] N. Pappas and A. Popescu-Belis. Multilingual hierarchical attention networks for document classification. In *IJCNLP*, 2017.

[8] S. Upadhyay, M. Faruqui, G. Tur, D. Hakkani-Tur, and L. Heck. (almost) zero-shot cross-lingual spoken language understanding. In *IEEE ICASSP*, 2018.

[9] X. Chen, Y. Sun, and Athiwaratkun B. Adversarial deep avaraging networks for cross-lingual sentiment classification. In *EMNLP*, 2017.

[10] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *TACL*, 2018.

[11] J. Kim, Y. Kim, R. Sarikaya, and E. Fosler-Lussier. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *EMNLP*, 2017.

[12] Z. Yang, R. Salakhutdinov, and W. W. Cohen. Transfer learning for sequence tagging with hierarchical recurrent networks. 2017.

[13] S. Gouws and A. Søgaard. Simple task-specific bilingual word embeddings. In *NAACL*, 2015.

[14] M. Artetxe, G. Labaka, and E. Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *ACL*, 2018.

[15] S. T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Vol. 290, No. 5500, 2000.

[16] D. D. Lewis, Y.Yang, T.G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *JMLR*, Vol. 5, No. Apr, 2004.

[17] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. AL-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, V. Hoste, M. Apidianaki, X. Tannier, N. Loukachevitch, E. Kotelnikov, N. Bel, S.M. Jiménez-Zafra, and G. Eryiğit. Semeval-2016 task 5: Aspect based sentiment analysis. In *SemEval*, 2016.