Vol.2017-NL-233 No.14
2017/10/25

単語の表層類似性を用いた多言語単語分散表現の教師なし学習手法

佐久間 仁[1,a)]　吉永 直樹[2,b)]

概要：異なる言語の単語を同一の意味空間に写像した多言語単語分散表現は，英語を始めとする言語資源の豊かな言語において学習した高精度の解析モデルを言語資源の乏しい言語の解析に転用するのに役に立つため注目を集めている．しかしながら，既存の多言語単語分散表現の学習手法の多くは対訳辞書や対訳コーパスなどの対訳資源を手がかりとして利用するため，適用可能な言語が制限される問題がある．そこで本研究では，異言語間の単語について，借用語や翻字，さらには語源を同じくする語などで表層に共通性が見られることを手がかりとして，単一言語コーパスのみから多言語単語分散表現を学習する方法を模索する．具体的には，出現文脈に加えて単語自身を構成する部分文字列を考慮した分散表現獲得手法を利用して，1）個々の言語ごとに独立に学習した単語分散表現を，語の表層類似性を手がかりとして学習した直交行列で写像することで，多言語分散表現を獲得する手法と 2）単一言語コーパスを連結して得られる複数言語コーパスから多言語単語分散表現を同時に学習する手法を提案する．実験ではこれらの手法の有用性を検証するために，同一言語内で意味的に近い語が意味空間でも近くなるとこを確認する言語内評価と，言語を越えて意味的に近い語が意味空間でも近くなることを確認する言語間評価を行った．

1. はじめに

現在，多くの自然言語処理タスクは教師あり学習に基づく統計的なモデルを用いて解かれるようになっているが，英語などの一部の言語を除く多くの言語資源（学習データ）に乏しい言語では高い解析精度を達成することが難しい．この問題に対して，豊富な言語資源を持つ言語上で学習した高精度の解析モデルを言語資源の少ない言語の解析に転用する手段として多言語単語分散表現が注目を集めている [1]．多言語単語分散表現とは，異なる言語の単語を同じ意味空間に写像したもので，意味の近い単語間の意味空間上での距離が言語を越えて近くなるような単語の分散表現である．

多言語単語分散表現については，これまで様々な学習手法が提案されているが，これらの手法の多くは個々の言語の単一言語コーパスだけでなく，大規模な対訳辞書や対訳コーパスなどの対訳資源を必要とする [1], [2], [3], [4]．このような大規模な（教師あり）対訳資源への依存は，豊富な言語資源が利用できる言語のモデルを言語資源の乏しい言語に転用する上では大きな障害となる．そこで近年の研究では，より少ない対訳資源から多言語単語分散表現を獲得する試みが行われている [5], [6]．

本研究では単語の表層類似性に着目することで，対訳資源に全く依存しない多言語単語分散表現の教師なし学習手法を提案する．異なる言語間であっても言語学的に近い言語間や文化的，地理的に交流の多い言語間では借用語や翻字，同じ語源をもち，意味的にも近い語が多く存在する．これらの単語が文字列として完全に一致することは稀であるが，単語を構成する部分文字列においては共通性が見られることが多い．このような部分文字列の共通性に着目することで，対訳資源に依存しない 2 種類の多言語単語分散表現の学習手法を提案する．一つはそれぞれの言語の単語分散表現を単語の部分文字列を手がかりに一つの意味空間に写像する手法（3.1 節）であり，もう一つは部分文字列を手がかりに 2 言語の単語分散表現を同時に学習する手法（3.2 節）である．

本論文は以下のように構成される．2 節では単一言語における単語分散表現について述べ，本研究で用いた Skip-gram と Subword Information Skip-gram (SI-Skip-gram) を導入する．3 節では，多言語単語分散表現の既存研究を紹介し，本研究の位置づけを明確にする．次に，4 節で単語の部分文字列を使用した 2 つの多言語単語分散表現の獲得手法を提案する．これらの手法の有効性を検証するために，5 節

[1] 東京大学大学院 情報理工学系研究科
[2] 東京大学 生産技術研究所
[a)] jsakuma@tkl.iis.u-tokyo.ac.jp
[b)] ynaga@iis.u-tokyo.ac.jp

で同一言語内での単語分散表現の評価と，言語間での評価を行いその結果を報告する．最後に 6 節で本研究をまとめると共に，今後の課題と研究の方向性について述べる．

2. 事前知識: 単語分散表現

単語を計算機で扱う際に，アトミックな要素として定義する，すなわち単語固有の次元のみ 1 の値を持つような離散値ベクトルを用いた場合，複数の単語間の意味的な関係を柔軟に捉えることは難しい．この問題に対し，意味的に近い単語に対して意味空間上の距離の近い連続値ベクトルを与える単語分散表現に関する研究が，活発に行われてきた．多くの研究では「単語の意味は共起する語から類推できる」という分布仮説 [7] に基づいて単語分散表現を構成しており，特に Mikolov らが提案したニューラルネットに基づく Skip-gram と Continuous Bag of Words (CBOW)[8] は多くのタスクで用いられる単語分散表現獲得手法となっている．これらの手法は，単純な 2 層のニューラルネットワークにより各単語の分散表現を学習する手法であり，ラベルのついていない言語コーパスのみから教師なしで高速に学習できるという特性をもつ．ここでは，本研究で使用した Skip-gram について述べ，その後 Skip-gram で単語の表層の情報を考慮した手法で，本研究でも利用した Subword Information Skip-gram（SI-Skip-gram）[9] について説明する．

Skip-gram では，コーパス内における分散表現を学習する単語（目的単語）からその前後 K 単語（文脈単語）を予測する 2 層のニューラルネットワークを用いて単語分散表現の学習を行う．ここで文脈単語の集合を文脈集合といい，それらの単語を文脈単語と呼ぶことにする．各単語 w に対して，2 つのベクトル u_w，v_w が学習される．ここで $w_1, w_2, \cdots, w_N$ を入力の生コーパスとすると，Skip-gram は目的関数

$$L = \sum_{i=1}^{N} \sum_{k \in \{-K, \cdots, -1, 1, \cdots, K\}} -\log p(w_{i+k}|w_i)$$

を最大化するように学習する．ただし，任意の単語 c と w に対して $p(c|w)$ は softmax 関数を使用して以下のように計算される．

$$p(c|w) = \frac{\exp(v_w \cdot u_c)}{\sum_{c' \in V} \exp(v_w \cdot u_{c'})}$$

ここで，学習された u_w を単語 w の分散表現として使う．ただし，この softmax 関数は計算コストが大きいので Negative Sampling [10] を用いて計算コストを下げることが多く，本研究でも Negative Sampling を用いる．

Skip-gram では，文字列が一致する単語を同一の単語として扱い，単語自身の表層から得られる情報は利用しない．したがって，例えばドイツ語のような屈折語においては単語が構成的に生成され出現回数が少なくなりがちであるため，効率よく学習が行われない [9]．

そこで，Bojanowski らは単語の出現文脈に加えて単語の部分文字列を利用して単語分散表現を学習する手法である Subword Information Skip-gram (SI-Skip-gram) を提案した [9]．この手法では単語に加えてその語を構成する部分文字列（文字 n-gram）に対しても分散表現の学習を行う．単語の分散表現は単語自身の分散表現と単語を構成する部分文字列の分散表現の線形和として表現されるため，$\mathcal{G}_w$ を単語 w を含んだ w に現れる部分文字列の集合とし，z_g を部分文字列の分散表現，v_c を単語 c の分散表現とすれば，単語 w，c の類似度 $s(w, c)$ は

$$s(w, c) = \sum_{g \in \mathcal{G}_w} z_g^T v_c$$

と計算できる．また，この手法は単語の表層に着目することで未知語に対しても，その語を構成する部分文字列の分散表現の和として分散表現が計算可能であることが特徴である．

3. 関連研究

本節では多言語単語分散表現の関連研究を述べ，本研究の背景を明確にする．多言語分散表現を得る手法としては，単言語コーパスから独立に単語分散表現を学習したのち，対訳辞書などを手がかりに共通の意味空間に写像する線形変換を学習する手法 (3.1 節) と，対訳コーパスなどを手がかりに直接多言語分散表現を学習する手法（3.2 節）が存在する．以下で，これら 2 つのアプローチに関する既存研究を順に説明する．

3.1 （非）線形変換に基づく多言語単語分散表現の獲得

Mikolov らは，異なる 2 つの言語の単語分散表現が意味空間上の単語の配置に同型写像があることを仮定し，一方の言語の単語分散表現を対訳関係にある他方の言語の単語分散表現に写像する線形変換を学習することで言語間に共通する単語分散表現を得る手法を提案した [2]．以下，2 つの言語の単語分散表現の集合を X と Y で示す．ここで，X の i 行目は対訳辞書を用いて Y の i 行目の訳語となるように並び替えられているとする．これら 2 つの行列を用いて，変換行列 W は目的関数

$$L = ||XW - Y||_F^2$$

が小さくなるように勾配法を使って学習する．

Faruqui らは，Canonical Correlation Analysis (CCA) を使い， それぞれの言語の単語分散表現を線形変換して同一の空間に写像した [3]．これを拡張し，Lu らは Deep Canonical Correlation Analysis (Deep CCA) を使い非線形写像を学習している [11]．Xing らは，各単語の分散表現のノルムが 1 になるように事前に正規化かつ，変換行列 W

が直交行列であると制約を加えることで，変換前の単語間の距離と変換後の単語間の距離が不変になるように工夫した [4]．Artetxe は [3], [4], [11] をまとめ，さらに理論的な考察を与えた [1]．この手法では，ノルムの正規化と直交性の制約に加え分散表現の各次元の平均が 0 になるように正規化し，変換行列 W を学習した．さらにこの条件において，変換行列 W に特異値分解を使用した解析的な解が存在することを示した．

石渡らは，本手法と同様に表層の類似性を用いて 2 言語間の単語分散表現の間の線形変換を学習している [12], [13]．彼らは共起語が次元に対応した分布ベクトルを採用しており，表層の類似する次元の対応付けが強くなるよう最適化の目的関数に報酬項を追加している．さらにこの研究では，比較的小規模な対訳辞書に対しても実験を行い，高い精度を達成している．我々の手法との相違は，我々は訓練データの自動生成に表層の類似性を用いている点である．

Artetxe らは，変換行列の学習と辞書の生成を繰り返し行うことで非常に小さな対訳辞書から多言語単語分散表現を学習する方法を提案している [5]．この実験では 25 個と小規模な対訳辞書に加えて数などの自明な対訳関係を用いて変換行列を学習している．

Smith らは，表層が完全に一致する語のみを対訳辞書とし多言語単語分散表現を教師なし学習している [6]．この手法では，softmax に工夫を加えた inverted softmax 関数を用いて学習を行った．

3.2 多言語単語分散表現の同時学習

一方，各言語の単語分散表現を同時に学習することで多言語単語分散表現を直接獲得する手法も提案されてきた．Cao らは，任意の 2 つの言語に対してその単語分散表現の分布が近くなるように目的関数を変更することで対訳資源なしで単語分散表現を同時に学習した [14]．Ammar らは，多言語単語分散表現を各言語に対して同時に学習する 2 種類の手法を提案した [15]．1 つ目は，単語分散表現を学習する言語対について対訳辞書内で同じ意味をもつ単語に対して同じ単語 ID を与え各言語の単一言語コーパスを連結して複数言語コーパスを作り，その上で Skip-gram や CBOW などの既存手法を適用することで多言語単語分散表現を獲得する手法で，2 つ目は入力文の単語に対してその訳文に含まれる単語も文脈単語として用いるように Skip-gram を拡張し，多言語単語分散表現を獲得する手法である．

本研究では，2 つの単語の部分文字列に着目することで対訳資源に一切依存しない多言語単語分散表現の学習手法を提案する．1 つ目の手法は，[5], [6] をさらに発展させた補助的な線形変換を用いた手法と考えることができる．2 つ目の手法は，2 言語の単語分散表現を同時に学習する手法 [15], [16] を対訳辞書を用いないよう拡張したものである．

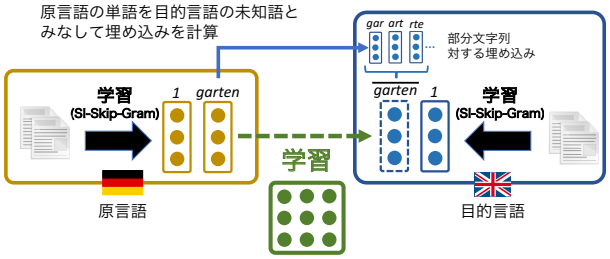

図 1　事後写像手法の概要．

4. 提案手法

既存の多言語単語分散表現の学習手法の多くは大規模な対訳辞書を手がかりにして言語間の写像を学習するものが多い．しかし，大規模な対訳資源は言語対によっては入手が非常に困難である．また，特定のドメインに特化した多言語単語分散表現を獲得したい場合，対訳資源の入手は言語資源の豊富な言語においても困難になる．

そこで本研究では，語の表層の類似性に着目することで対訳資源に一切依存せず多言語単語分散表現を学習することを目指す．ドイツ語や英語など同一語族の言語対においては，同じ語源を持つ単語が多く存在する．さらに日本語と中国語など地理的，文化的に近接する言語対においては，借用語や翻字のように一方の言語の単語が他方の言語に輸入され使用されることがある．これらの単語対は意味が同一，あるいは近いことが多く，文字列上完全に一致するとは限らないが，その部分文字列には共通箇所が見られることが多い．そこで本研究では，単語の部分文字列を手がかりとした以下の 2 つの手法を提案する．

4.1 線形変換に基づく多言語単語分散表現の学習

この手法（図 1）では各言語で独立に学習した単語分散表現を線形変換により一つの意味空間へ写像することで多言語単語分散表現を得る．以降，線形写像を学習する際，入力となる単語分散表現の言語を原言語，出力となる単語分散表現の言語を目的言語と呼ぶ．

この手法では，まず原言語と目的言語に対して SI-Skip-gram を利用して独立に単語分散表現を学習する．こうして得られた単語分散表現をそれぞれ X（原言語），Y（目的言語）とする．この時，SI-Skip-gram は未知語に対してもその部分文字列から単語分散表現を与えることができる．そこで原言語の単語 w とその分散表現 x_w が与えれた時，この w は目的言語での未知語とみなせるため，その単語分散表現 $\bar{x}_w$ を与えると，この単語分散表現 $\bar{x}_w$ はその部分文字列のみから計算されるため目的言語の意味空間に属すると考えることができる．こうして，原言語の各単語に対して原言語の単語分散表現 X と目的言語の単語分散表現 $\bar{X}$ が得られる．そこで，原言語の単語分散表現 X が $\bar{X}$ に近くなるように線形変換 W を学習する．

$$\hat{W} = \arg\min_{W} ||XW - \bar{X}||_F^2$$

これにより，部分文字列に共通性が見られない単語に対しても多言語単語分散表現を与えることが可能となる．目的関数の最適化には 3.1 節で紹介した先行研究 [1] を用いる．

このとき，原言語の全ての単語の分散表現 x_w と計算した $\bar{x}_w$ を用いて W の学習を行うことは可能であるが，原言語の単語の多くが目的言語の単語と部分文字列の一致など表層的な関係がないことを考えると，これは必ずしも最適であるとは限らない．そこで，言語を越えて表層的な関係を持つと期待できる借用語や翻字，語源を同じにする単語のみを用いて学習するために，変換行列の学習に使用する単語のフィルタリングを行う．まず，先行研究 [6] に倣い目的言語と原言語で文字列が完全に一致する単語は学習に使用する．今回，文字列長が長い単語は翻字であったり語源が同じである可能性が高いこと，出現頻度の少ない単語は正確な単語分散表現が与えられていない場合があることを考慮し，文字列長が十分長くかつ出現頻度の高い単語のみを変換行列 W の学習に用いる．文字列の最低長と最低出現頻度は開発データを用いたチューニングによって決定する．以降，この手法を事後写像手法と呼ぶ．

4.2 同時学習

この手法では，2 つの言語の単語分散表現を同時に学習する．先行研究 [6], [15] を参考に，文字列が一致する単語を辞書としてそれらの単語に同一の単語 ID を与えたコーパスで多言語単語分散表現を学習する．このためには，複数言語のコーパスを連結した複数言語コーパスに対して任意の単語分散表現の獲得手法を適用するだけで十分である．

そこで，分散表現の学習手法としてここでも表層文字列を考慮した SI-Skip-gram を用いることで，この問題を間接的に回避する．4 節で述べた通り異なる言語間において文字列が一致する単語は多くないが，その部分文字列には共通性が見られる．そこで，SI-Skip-gram を適用することで単語の部分文字列の共通性を踏まえた単語分散表現の学習が可能になり，より高精度の分散表現が学習できると期待できる．以降，この手法を同時学習手法と呼ぶ．

5. 実験

実験では，前節で提案した手法の有効性を検証し，単語の表層の類似性を考慮することが多言語単語分散表現にどのような影響を与えるかを確認した．

個々の手法の有効性を検証するために，それぞれ以下の手法と比較した．

- 事後写像手法に対しては，表層の類似性を考慮することの効果を検証するために文字列が完全に一致するもののみを辞書として用いる先行研究 [6] で SI-Skip-gram を用いた手法と比較した．またフィルタリングが与える影響を検証するために，全ての単語を変換写像の学習に用いた手法とも比較した．
- 同時学習手法に対しては，SI-Skip-gram の代わりに Skip-gram を用いて学習したものと比較した．これは，[15] を教師なしとした場合と同じであり，単語の部分文字列を用いない多言語単語分散表現が学習できる．

データセット	言語	# of sentences	# of tokens
Europarl	En	765,378	61,417,235
Europarl	It	749,153	60,801,119
Europarl	De	719,010	54,853,240
Europarl	Fi	707,864	40,453,294
Twitter	En	1,000,000	16,183,882
Twitter	It	1,000,000	12,069,667
Twitter	De	1,000,000	11,765,326
Twitter	Fi	1,000,000	6,864,736

表 1 データセットの概要.

学習には公開されている Europarl[*1]のデータと Twitter のデータを使用した．トークン化には Europarl に付随する tokenizer.perl を使用し XML コードは削除した．Twitter のデータは 2016 年 8 月 1 日から 29 日までの 29 日間のデータから各言語に対して 1,000,000 ツイートをランダムに選び，Europarl と同様にトークン化した．さらに@から始まるユーザー ID は全て特別なトークンに置換し，URL は削除した．表 1 に得られたデータセットの概要を示す．

本研究の手法は言語学的に近い語族に属する言語対や，文化的・地理的に近い言語対において効果を発揮することが期待される．そこでそのような言語対として En-It, En-De を使用した．さらに，これらの手法が言語学的に離れた言語においてどの程度効果を発揮するかを調べるために， En-Fi を使用した．

単語分散表現の学習には Skip-gram[*2], SI-Skip-gram[*3]共に公開されている C 言語の実装を使用した．また分散表現の次元数は 256 とし，それ以外のパラメータは全てデフォルトの値を使用した．また，Skip-gram，SI-Skip-gram 共に Negative Sampling を用いて学習を行った．獲得した多言語単語分散表現の評価では，先行研究に倣い同一言語内の単語分散表現の精度の評価として言語内評価（5.1 節）と，言語横断的な精度の評価として言語間評価（5.2 節）を行った．

5.1 言語内評価

単語分散表現の同一言語内の精度の評価として言語内評価を行った結果を報告する．言語内評価は英語とドイツ語の単語分散表現について行い，英語に対しては WordSim-

[*1] http://www.statmt.org/europarl/
[*2] https://github.com/svn2github/word2vec
[*3] https://github.com/facebookresearch/SI-Skip-gram

			単一言語		同時学習	
データセット	言語対	言語	**Skip-gram**	**SI-Skip-gram**	ベースライン	提案手法
Europarl	-	En	45	**49**	-	-
Europarl	-	De	68	**74**	-	-
Europarl	En-De	En	-	-	**43**	41
Europarl	En-De	De	-	-	36	**58**
Europarl	En-It	En	-	-	**39**	34
Europarl	En-Fi	En	-	-	**40**	38
Twitter	-	En	51	**56**	-	-
Twitter	-	De	41	**54**	-	-
Twitter	En-De	En	-	-	**54**	44
Twitter	En-De	De	-	-	11	**38**
Twitter	En-It	En	-	-	39	**47**
Twitter	En-Fi	En	-	-	46	**47**

表 2 WordSim-353 と GUR350 を用いた言語内評価の結果．評価尺度は人手による類似度との相関．事後写像手法の言語内評価は，単語分散表現の不変性により単一言語と同値．

		事後写像手法			同時学習	
データセット	言語対	文字列一致	＋フィルタ	全単語	ベースライン	提案手法
Europarl	En-De	64.21	**65.23**	47.21	19.12	**34.49**
Europarl	En-It	70.71	**71.30**	62.71	12.77	**20.15**
Europarl	En-Fi	**56.13**	55.16	13.88	16.57	**24.39**
Twitter	En-De	36.84	**37.80**	30.14	16.29	**29.63**
Twitter	En-It	**40.74**	40.12	35.80	12.53	**23.51**
Twitter	En-Fi	42.38	**43.05**	31.13	**26.06**	25.00

表 3 対訳辞書構築を用いた言語間評価の結果．評価尺度は訳語の予測精度．

353[*4]を，ドイツ語に対しては GUR353[*5]を用いた．これらは各言語で抽出して 2 つの単語ペアのそれぞれに人手で類似度を付与したものであり，単語間のベクトルのコサイン類似度と人手による類似度の相関により評価した．

結果を表 2 に示す．単一言語において学習された単語分散表現では，全ての場合において SI-Skip-gram で学習した単語分散表現が Skip-gram で学習したものを上回った．またデータセット間で比較をすると，En-De 言語対上で学習されたドイツ語単語分散表現を除いて Twitter で学習した単語分散表現が Europarl で学習したものを上回った．これは Europarl は小さなドメインに限られたデータであり，幅広い単語対に対して評価を行う WordSim-353 では高い精度を発揮できなかったことが原因として考えられる．なお，事後写像手法では変換行列が直交行列であるという制約により単語間のコサイン類似度が変換後も不変であることを保証するため，原言語，目的言語共に，その言語内評価は単一言語上で SI-Skip-gram を使用して学習した結果と同じである．

同時学習により学習した単語分散表現の言語内評価は，データセットと言語対によって大きく異なった．Europarl データセットにおいては，多くの言語対において表層の情報を含めないベースラインが高い精度を発揮している中で En-De 言語対の多言語単語分散表現のドイツ語単語分散表現においては表層情報を含む提案手法が高い精度を発揮した．Twitter データセットにおいては，En-De 言語対で学習した英語の単語分散表現においてはベースラインが上回り，それ以外の場合は提案手法が上回った．

また，これらの結果を単一言語において学習した単語分散表現と比較すると同時学習では言語内の精度の低下が見られる．言語内での 2 単語間のコサイン類似度の不変性を保証する事後写像手法では，このような言語内の精度低下は起こらない．同時学習の手法においていかに言語内の類似度の精度低下を避けるかは，今後の課題である．

5.2 言語間評価

次に，言語横断的な精度を評価するために対訳辞書構築をによる言語間評価を行なった．対訳辞書構築は，一つの言語の単語から他の言語の対応する訳語を予測するタスクであり，評価尺度は予測の精度である．結果を表 3 に示す．

まず事後写像手法においては，Europarl の En-Fi 言語対と Twitter の En-It 言語対を除いて，文字列が完全に一

[*4] http://www.cs.technion.ac.il/~gabr/resources/data/wordsim353/
[*5] https://www.ukp.tu-darmstadt.de/data/semantic-relatedness/german-relatedness-datasets/

致する単語のみを用いた場合と比較して提案手法が上回った．Europarl の En-Fi 言語対で，提案手法が高精度を発揮できない原因として，英語とフィンランド語は他の言語対に比べて言語的に遠く表層の共通性が少ないことがあげられる．Europarl と Twitter を比較すれば，全ての言語対において Europarl で学習した単語分散表現が Twitter で学習したものを大きく上回った．これは Europarl が対訳コーパスであり，言語間でデータのドメインに大きな差がないことに起因すると考えられる．

また，全ての単語を変換行列の学習に用いた場合とフィルタリングを行う提案手法を比較すると全ての言語対において提案手法が大きく上回っており，フィルタリングの重要性が確認できた．これは単語の多くは言語を横断した対応関係を持っておらず，その表層の情報は多言語単語分散表現の獲得に寄与しないことが原因だと思われる．今回の研究では，単語のフィルタリングは単語の出現回数と単語の文字列長という 2 つのシンプルな指標によるものであったが，今後の課題としてこの手法をさらに工夫していくことがあげられる．

次に，同時学習手法に着目する．言語間の評価においては Twitter の En-Fi 言語対を除いて，SI-Skip-gram を用いることで単語の表層情報を活用する提案手法がベースラインを大きく上回る結果となった．Twitter の En-Fi 言語対においてベースラインが提案手法より高精度を達成した理由として，1）英語とフィンランド語にはその表層に大きな共通点が見られないこと，2）Twitter 特有の表現（ユーザーを示す@やリツイートを示す RT など）が多く，これらが手がかりとなりベースラインが高精度を発揮したことがあげられる．

事後写像手法と同時学習手法を比べると事後写像手法が大きく上回る結果となった．これは各言語の単語分散表現の不変性が保証される事後写像手法に対して，同時学習手法では言語内の単語分散表現の精度を担保することが難しく，これが言語内評価だけでなく言語間の評価にも影響を与えていることが原因と考えられる．

6. おわりに

本研究では， 対訳辞書に依存しない多言語単語分散表現の教師なし学習手法を提案した．具体的に，異なる言語間であっても共通する語源を持つ単語や，借用語，翻字が存在していることに着目し，単語の部分文字列を手がかりにする 2 種類の多言語単語分散表現手法を提案し，同一言語内の単語類似性判定と，対訳辞書構築タスクを通してその評価を行なった．この結果，多くの場合において提案手法が単語の部分文字列を用いない手法を上回る結果を残した．しかし同時に，今後取り組まなければならない課題も明らかになった．

言語横断的な表層の共通性に着目するこれらの提案手法は適用する言語対に大きく影響を受ける．今回は主にヨーロッパ系の言語に対しこれらの手法の有用性を確認したが，En-Ja のように文字体系が大きく異なるような言語学的な差異の大きな言語対や，文化的，地理的な交流の少ない言語対に対して有用であるかはわかっていない．

個々の手法については，事後写像手法は言語内評価，言語間評価共に高い精度を達成した．しかし，5.2 節で述べたように変換の学習に使用する単語の選び方などさらなる工夫によってより良い多言語単語分散表現が獲得できる可能性がある．

同時学習の手法では，単語の表層情報を活用することで多言語単語分散表現の精度向上に成功したが，5.2 節で述べたように各言語で独立して学習した単語分散表現と言語内評価を比較すれば多くの場合大きく劣る結果となった．同時学習の枠組みの中でより高精度な多言語単語分散表現の獲得を目指すことは今後の課題である．

謝辞　本研究は JSPS 科研費 16K16109 の助成を受けたものである。

参考文献

[1] Artetxe, M., Labaka, G. and Agirre, E.: Learning principled bilingual mappings of word embeddings while preserving monolingual invariance, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2289–2294 (2016).

[2] Mikolov, T., Le, Q. V. and Sutskever, I.: Exploiting Similarities among Languages for Machine Translation (2013).

[3] Faruqui, M. and Dyer, C.: Improving Vector Space Word Representations Using Multilingual Correlation, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 462–471 (2014).

[4] Xing, C., Wang, D., Liu, C. and Lin, Y.: Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation, *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 1006–1011 (2015).

[5] Artetxe, M., Labaka, G. and Agirre, E.: Learning bilingual word embeddings with (almost) no bilingual data, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 451–462 (2017).

[6] Smith, S. L., Turban, D. H. P., Hamblin, S. and Hammerla, N. Y.: Offline Bilingual Word Vectors, Orthogonal Transformations and The Inverted Softmax, *Proceedings of 5th International Conference on Learning Representations (ICLR)* (2017).

[7] Harris, Z. S.: Distributional Structure, *WORD*, Vol. 10, No. 2-3, pp. 146–162 (1954).

[8] Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *Proceedings of 1st International Conference on Learning Representations (ICLR)* (2013).

[9] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T.: Enriching Word Vectors with Subword Information, *Transactions of the Association of Computational Lin-*

guistics (TACL), pp. 135–146 (2016).

[10] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems 26 (NIPS 2013)* (Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. and Weinberger, K. Q., eds.), pp. 3111–3119 (2013).

[11] Lu, A., Wang, W., Bansal, M., Gimpel, K. and Livescu, K.: Deep Multilingual Correlation for Improved Word Embeddings, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 250–256 (2015).

[12] 石渡祥之佑，鍜治伸裕，吉永直樹，豊田正史，喜連川優：文脈語間の対訳関係を用いた単語の意味ベクトルの翻訳，人工知能学会論文誌，Vol. 31, No. 6, pp. AI30–A_1–10 (2016).

[13] Ishiwatari, S., Kaji, N., Yoshinaga, N., Toyoda, M. and Kitsuregawa, M.: Accurate Cross-lingual Projection between Count-based Word Vectors by Exploiting Translatable Context Pairs, *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL)*, pp. 300–304 (2015).

[14] Cao, H., Zhao, T., ZHANG, S. and Meng, Y.: A Distribution-based Model to Learn Bilingual Word Embeddings, *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pp. 1818–1827 (2016).

[15] Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C. and Smith, N. A.: Massively Multilingual Word Embeddings, *arXiv preprint arXiv: 1602.01925* (2016).

[16] Duong, L., Kanayama, H., Ma, T., Bird, S. and Cohn, T.: Multilingual Training of Crosslingual Word Embeddings, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 894–904 (2017).