特集論文 「人工知能学会設立 30 周年記念論文特集」

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

Translation of Word Vectors by Exploiting Translatable Context Pairs

石渡 祥之佑 Shonosuke Ishiwatari
東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, the University of Tokyo
ishiwatari@tkl.iis.u-tokyo.ac.jp

鍛治 伸裕 Nobuhiro Kaji
ヤフー株式会社 *1
Yahoo Japan Corporation
nkaji@yahoo-corp.jp

吉永 直樹 Naoki Yoshinaga
東京大学 生産技術研究所 *2
Institute of Industrial Science, the University of Tokyo
ynaga@tkl.iis.u-tokyo.ac.jp

豊田 正史 Masashi Toyoda
東京大学 生産技術研究所
Institute of Industrial Science, the University of Tokyo
toyoda@tkl.iis.u-tokyo.ac.jp

喜連川 優 Masaru Kitsuregawa
国立情報学研究所，東京大学 生産技術研究所
National Institute of Informatics / Institute of Industrial Science, the University of Tokyo
kitsure@tkl.iis.u-tokyo.ac.jp

keywords: word representation, cross-lingual projection, bilingual dictionary induction, distributional semantics

Summary

While vector-based representations of word meanings (word vectors) have been widely used in a variety of natural language processing applications, they are not meant for capturing the similarity between words in different languages. This prevents using word vectors in multilingual-applications such as cross-lingual information retrieval and machine translation. To solve this problem, we propose a method that learns a cross-lingual projection of word representations from one language into another. Our method utilizes translatable context pairs obtained from a bilingual dictionary and surface similarity as bonus terms of the objective function. In the experiments, we evaluated the effectiveness of the proposed method in four languages, Japanese, Chinese, English and Spanish. Experiments shows that our method outperformed existing methods without any additional supervisions.

1. は じ め に

単語の意味を計算処理可能な形式で表現することは，自然言語テキストを対象としたアプリケーションの高度化に必要となる要素技術である．現在，単語の意味は頻度に基づくベクトルとして表現する方法が一般的である．こうした単語の意味ベクトル (以下，単語ベクトルと呼ぶ) は,「単語の意味は，その単語と共に出現している単語群によって特徴づけられる」という分布仮説 [Firth 57, Harris 54] に基づいている．単語の意味をベクトルで表現することにより，単語間の意味的な類似性の計算を，ベクトル間の類似度計算 (コサイン類似度など) に帰着させることが可能になる．このような単語のベクトル表現は類義語検出 [Baroni 14] や言い換え検出 [Erk 08]，対話分析 [Kalchbrenner 13] など多くの自然言語処理アプリケーションで広く活用されている．

しかし，従来の意味表現には，異なる言語の単語間の類似性判定，すなわち単語の対訳関係の判定に用いることが難しいという問題点が存在する．これは，異なる言語のコーパスから上記のようなベクトル表現を学習した場合，共通の共起語がほとんど出現しないためである．たとえば，日本語のコーパスから学習された “猫” のベクトルと，英語のコーパスから学習された *“cat”* のベクトルの類似度は正しく計算することができず，“猫” と “*cat*” が同義であることを計算により判定することができない．このため，共起する単語に基づく意味表現は言語横断検索や機械翻訳など，多言語にまたがる自然言語処理アプリケーションにおいて利用することができない．

このような問題意識から，近年では，単語のベクトル表現を言語横断的に活用するための手法の研究が進められている [Dinu 14, Fung 98, Mikolov 13b]．特に Mikolov

*1 本研究は著者が東京大学 生産技術研究所に所属していたときに行われた．
*2 本研究は著者が国立研究開発法人 情報通信研究機構主任研究員を兼務していたときに行われた．

らは，少量の対訳辞書を学習データとして用いて，異なる言語の単語ベクトル間の線形変換，m 行 n 列の翻訳行列 (m, n はそれぞれ目的言語のベクトルの次元数と，原言語のベクトルの次元数) を学習する手法を提案した．

本研究では，単語ベクトルの翻訳において，翻訳行列を教師あり学習によって得るという Mikolov らの枠組みで，(1) Fung らが用いた辞書に基づく共起語間の対応関係や，(2) 表層の類似性に基づく共起語間の対応関係といった事前知識を同時に活用する新しい手法を提案する．特に本研究が対象としているのは，カウントベースの単語ベクトル [Baroni 14] である．これは各次元が他の単語との共起頻度に対応しているベクトルのことであり，意味が似た単語のベクトル間の類似度が高くなることが知られている [Erk 12, Turney 10]．翻訳行列の学習に用いる対訳辞書には，こうしたカウントベースのベクトルの次元間に存在する対応関係に関する情報が部分的に含まれている [Fung 98] (例: 学習データに “犬” - “*dog*” という単語対が含まれていた場合，日本語のベクトルの “犬” の次元と，英語のベクトルの “*dog*” の次元は対応していると考えられる)．また，単語の表層の類似度もベクトル間の写像を学習する際の有用な手がかりとなりうる (例: 英語のベクトルの “*cocktail*” の次元と，スペイン語のベクトルの “*cóctel*” の次元はその綴りが類似していることから，対応していると考えられる)．こうした知識は明らかにカウントベースのベクトルの翻訳において有用であるが，教師あり学習の枠組み内で活用する手法は未だ提案されていない．そこで本研究では，ベクトル間の翻訳行列を学習する際，上述した共起語間の対応関係に関する事前知識を目的関数の報酬項として利用することにより，より精度の高い翻訳を実現させる．

本稿では，提案手法の有効性を検証するため，日本語，英語，中国語，スペイン語の単語ベクトルについて，既存研究にならって翻訳による訳語選択を行うというタスクを設定し，Wikipedia から学習した単語ベクトルを用いてその精度を評価した．ニューラルネットワークを用いて学習される単語ベクトルを用いた手法を含め，3 つの既存手法との比較実験を行った結果，提案手法が最も正確にベクトルを翻訳できるという結果を得た．

以降，2 章では関連研究を紹介し，3 章ではある言語の単語のベクトルを他の言語のベクトルに翻訳する手法を述べる．続く 4 章ではベクトル翻訳の精度を評価するための実験について述べ，最後に 5 章で本研究をまとめる．

2. 関 連 研 究

カウントベースの単語ベクトルの翻訳に関する研究は，機械翻訳における対訳辞書の自動抽出を目的として始められた [Fung 98]．この研究で提案されている単語ベクトルの翻訳手法では，既存の対訳辞書を用いて対応する次元間の値を直接移行することでベクトルを（部分的に）変換する．この手法では，翻訳に用いる対訳辞書のカバレッジが低いと，翻訳元言語の単語ベクトルの多くの次元が翻訳先言語のベクトルの次元に対応させることができないという問題が存在する．これに対し，1 章で述べた研究 [Mikolov 13b] はベクトル間の対応をベクトル間の線形変換（翻訳行列）を求める問題として定式化し，教師あり学習によってベクトル間の対応関係を間接的に求めることで，この問題に対処している．Mikolov らの研究ではカウントベースの単語ベクトルは用いられておらず，ニューラルネットワーク言語モデルにより学習するベクトル [Bengio 03, Mikolov 13a] が利用されている．

単語ベクトルの翻訳とは異なるアプローチであるが，異言語に共通のベクトル表現を学習する手法 [Chandar A P 14, Faruqui 14, Gouws 15, Hermann 14, Klementiev 12, Xiao 14] も研究されている．特に Hermann らは「複数の言語から単語の表現を学習することで，言語非依存なもの (すなわち，意味) を精度良く捉えられる」と主張し，翻訳の必要ない言語非依存なベクトル表現を提案している．しかし，こうした単語の表現を学習するためには，莫大な対訳コーパスが必要となるため，対訳コーパスの存在しない言語対には適用できない．また，対訳が学習データに存在しない新語や，語の新用法にも適用できないという短所も存在する．

また，統計的機械翻訳における対訳辞書の自動拡張を目的とし，上記の手法とは異なるアプローチを採用した研究 [Marton 09, Razmara 13, Saluja 14] も多く行われている．中でも Razmara らと Saluja らの手法で用いられるグラフ伝搬アルゴリズムは計算量が大きく，翻訳に先んじて辞書を生成する必要がある．これに対し，提案手法による翻訳はベクトルと行列の乗算が 1 回と単語ベクトル間の類似度計算のみであり，即時的に訳語を与えることが可能である点において応用上優れている [Ishiwatari 16]．また，Marton らは，未知語の類義語を分布類似度等の尺度により探索し，得られた類義語の訳語を未知語の訳とする手法を提案した．この手法は，未知語に類義語 (あるいは言い換え) が存在するという仮定に基づくため，全く新しい概念を持つ新語の翻訳を行うことは難しい．

本研究では，以上で述べた異言語に共通のベクトル表現を学習する手法や対訳辞書拡張の手法の欠点をふまえ，Fung らや Mikolov らと同じく，単語ベクトルを翻訳するアプローチを採用する．このアプローチは，(1) 小規模な単語の対訳辞書のみを学習データとして利用する点と，(2) 学習データに存在しない語に対しても訳語を与えられる点で，上述した他のアプローチよりも適用範囲が広く，優れている．本研究の提案手法は，Mikolov らの教師あり学習を用いたベクトル翻訳という枠組みに，Fung らが用いた辞書の情報を学習の手がかりとして取り込んだ点で，これらの短所を補い，長所を組み合わせた手法であると言える．提案手法とこれらの先行研究の手法との詳細な比較は 4 章で述べる．

3. 提 案 手 法

1 章で述べたように，一般的に単語ベクトルは同じ言語のテキストコーパスから作られるため，異なる言語に属する 2 単語のベクトルをそのまま比較することはできない．本章ではこの問題に対処するため，我々が提案する，ある言語における単語のベクトルを他の言語におけるベクトルに翻訳する手法について述べる．

3·1 ベクトル表現の翻訳

まずはじめに，提案手法の基礎となる Mikolov らの手法 [Mikolov 13b] について説明する．Mikolov らは行列変換に基づくベクトル表現の翻訳手法を提案している．これは，ある言語の単語 $\boldsymbol{x}$*3に翻訳行列 $\boldsymbol{W}$ を乗じることによって，その訳語のベクトル表現の近似を得るというものである．翻訳行列 $\boldsymbol{W}$ は，対訳語ペアの集合 $\{\boldsymbol{x_i}, \boldsymbol{z_i}\}_{i=1}^{n}$ を用いて，以下の最適化問題を解くことによって得られる．

$$\boldsymbol{W}^{\star} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{z}_i\|^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|^2 \quad (1)$$

ここで 2 つめの項は L_2 正則化項である．この正則化項は，Mikolov らの元論文では用いられていない．しかし，正則化項には過学習を防ぐ効果があり，一般的にモデル学習の精度を高めるために有効である．予備実験においてもその効果が確認されたため，本論文では上記の学習方法をベースラインとして議論を進める．

3·2 翻訳可能な文脈語ペア

本節では，提案手法の核となるアイデアについて述べる．カウントベースの単語ベクトルを用いる場合，前節で述べた翻訳行列の学習に用いる原言語のベクトルと目的言語のベクトルの次元は，互いに異なるコーパスにおける共起単語，すなわち文脈語に対応している．しかしながら，この 2 言語のベクトルの次元間には，対訳や翻字，発音の類似など，何らかの意味的な対応関係が存在する可能性がある．

例として，スペイン語と英語の単語ベクトルをそれぞれ考える．スペイン語のベクトルの各次元は “*amigo*” や “*comer*”，“*cóctel*” といった，スペイン語の共起語と対応している．一方で，英語のベクトルの各次元は “*eat*” や “*friend*”，“*small*”，“*cocktail*” といった，英語の共起語と対応している．ここで，2 つの共起語 “*amigo*” と “*friend*” は互いの対訳になっている．こうした言語をまたいだ文脈語間，すなわち次元間の対応関係は，式 (1) の翻訳行列 $\boldsymbol{W}$ を学習する際にも活用しうるものである．

本研究では以下 2 つの方法により，上記で述べたような 2 カ国語のベクトルの次元間の対応関係を自動的に見つける．1 つめは，翻訳行列を学習するための学習データを転用する方法である．この学習データは単語の対訳辞書であるため，次元間の対訳関係を見つけるのにそのまま用いることができる．以降，この方法で見つけられた文脈語ペアの集合は $\mathcal{D}_{train}$ と記す．ただし，ある言語における単語 (例: スペイン語の “*amigo*”) が，他の言語においては複数の訳語を持つ (例: 英語の “*friend*”, “*fan*”, “*supporter*”) 場合は少なくないため，$\mathcal{D}_{train}$ 内の文脈語ペアは 1:1 に対応するとは限らない点に留意する．

次元間の対応関係を自動的に見つける 2 つめの方法は，共起語間の表層の近さを用いる方法である．言語はその進化の過程で，他の言語から語や概念を借用することがある．こうした現象によって他言語から取り入れられた言葉は，もとの言語における言葉と似通った (あるいは完全に同じ) 綴りを持つことが多い (例: スペイン語の “*cóctel*” と英語の “*cocktail*”)．本研究ではこの点に着目し，下記の距離関数で単語の表層の近さを測った．

$$\text{DIST}(r, s) = \frac{Levenshtein(r, s)}{\min(len(r), len(s))} \quad (2)$$

ただし，ここで $Levenshtein(r, s)$ は 2 単語の編集距離を表し，$len(r)$ は単語 r の文字数を表す．また，式 (2) の分母は編集距離の正規化のために導入している．本研究では，適当な閾値 *4を設定し，閾値よりも表層の距離が小さい単語のペア (r, s) を全て翻訳可能な文脈語ペアとした．以降，上記の方法で見つけられた文脈語ペアの集合は $\mathcal{D}_{sim}$ と記す．

3·3 文脈語ペアの目的関数への導入

前節では，翻訳可能な文脈語ペアの集合 $\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ を見つける方法についてそれぞれ述べた．本節では得られた $\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ を，翻訳行列を学習する際の手がかりとして用いる方法について述べる．

前節と同様に，スペイン語から英語へベクトルの翻訳を行うことを考える．このとき，直観的にはスペイン語のベクトルの “*amigo*” の次元は英語のベクトルの “*friend*” の次元と，スペイン語のベクトルの “*cóctel*” の次元は英語のベクトルの “*cocktail*” の次元と強い相関を持つはずである．この直観に従い，仮にスペイン語のベクトルの j 次元目が “*amigo*” との共起頻度を表し，英語のベクトルの k 次元目が “*friend*” という語との共起頻度を表すとすると，翻訳行列 $\boldsymbol{W}$ の k 行 j 列目の値が他の成分と比較して相対的に大きくなるよう補正することで，より正確なベクトルの翻訳が得られるはずである．同様に，スペイン語のベクトルの l 次元目が “*cóctel*”，英語のベクトルの m 次元目が “*cocktail*” とそれぞれ対応している場合，$\boldsymbol{W}$ の m 行 l 列目の値が相対的に大きくなるようにするべきである．

このように対訳関係にある次元のペアにかかる重みを強くするため，式 (1) に報酬項を加えた下記の最適化問

*3 正確には単語ベクトルであるが，文脈から明らかな場合には，単語とそのベクトル表現のことは区別しないものとする．

*4 今回は 0.5 に固定した．

題を解くことによって，翻訳行列を学習することを提案する．

$$\boldsymbol{W}^{\star} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{z}_i\|^2 + \frac{\lambda}{2}\|\boldsymbol{W}\|^2 - \beta_{train} \sum_{(j,k)\in\mathcal{D}_{train}} w_{kj} - \beta_{sim} \sum_{(l,m)\in\mathcal{D}_{sim}} w_{ml}$$

ここで，上記の式における第 3 項と第 4 項は，$\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ に含まれている文脈語ペアに関して，$\boldsymbol{W}$ の対応する成分の値を大きくするように働く報酬項である．この 2 項の係数 β_{train} と β_{sim} はハイパーパラメータであり，$\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ の最適な重みを開発データにより求める．本稿で行った実験では，予備実験の結果から $\beta_{train} = \beta_{sim} = 5.0$ とした．

3·4 確率的勾配降下法による学習

前節で提案した最適化問題を，確率的勾配降下法 [Bottou 04] の一種である Pegasos [Shalev-Shwartz 11] で解く．τ 番目の学習例 $(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau)$ が与えられたとき，翻訳行列 $\boldsymbol{W}$ を次式に基づいて更新する．

$$\boldsymbol{W} \leftarrow \boldsymbol{W} - \eta_\tau \nabla E_\tau(\boldsymbol{W})$$

ここで η_τ は学習率であり，Pegasos では λ をハイパーパラメータとして $\eta_\tau = \frac{1}{\lambda\tau}$ で表現される．また，$\nabla E_\tau(\boldsymbol{W})$ は τ 番目の学習例 $(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau)$ に基づいて求められた勾配であり，次式で表すことができる．

$$2(\boldsymbol{W}\boldsymbol{x}_\tau - \boldsymbol{z}_\tau)\boldsymbol{x}_\tau^{\mathrm{T}} - \beta_{train}\boldsymbol{A} - \beta_{sim}\boldsymbol{B} + \lambda\boldsymbol{W}$$

ここで $\boldsymbol{A}$ と $\boldsymbol{B}$ は 3·3 節で導入した 2 つの報酬項にそれぞれ対応している．$\boldsymbol{A}$ は $(j,k) \in \mathcal{D}_{train}$ のとき $a_{kj} = 1$ となり，それ以外の場合は $a_{kj} = 0$ となる行列である．$\boldsymbol{B}$ も同様に，$(l,m) \in \mathcal{D}_{sim}$ のとき $b_{ml} = 1$ となり，それ以外の場合は $b_{ml} = 0$ となる行列である．

4. 実　　験

我々は，日本語 (Ja)，中国語 (Zh)，英語 (En)，スペイン語 (Es) の 4 カ国語の単語ベクトルを用い，それぞれの間の翻訳を通して提案手法の評価を行った．それぞれの言語の組み合わせにおけるベクトル翻訳の実験を行い，学習の目的関数に組み込まれた 2 種類の文脈語ペアがそれぞれ与える影響を評価した．

4·1 データセット

日本語，中国語，英語，スペイン語の 4 カ国語の単語ベクトルを作成するため，下記の手順で各言語の Wikipedia ダンプデータ [*5] を加工した．まず，wp2txt[*6] を用いて XML タグの除去を行う．得られた 4 カ国語のテキストデータのうち，文中の単語が空白等で句切られない言語である日本語と中国語は形態素解析ソフトウェア MeCab[*7] と Stanford Word Segmenter[*8] をそれぞれ用いて単語単位に分割した．続いて，語形変化が存在する英語，スペイン語，日本語のデータについてはそれぞれ Stanford POS tagger[*9]，Pattern[*10]，MeCab を用いて見出し語化し，最終的なテキストコーパスとした．

次に，上記のテキストコーパスを用いてカウントベースの単語ベクトルを作成する．この際，前後 5 単語以内に出現している語を共起語とみなすが，接続詞や助詞といった機能語は共起語から除外する．こうして作られた単語ベクトルは非常に高次元でスパースであるため，全言語について頻出の共起語上位 10,000 語のみを文脈語として残し，10,000 次元のベクトルを得た．また，単語の出現頻度の差異を吸収するため，全てのベクトルの各次元の値を正の相互情報量 (PPMI; Positive Pointwise Mutual Information) [Church 90] に変換した後，正規化を行ってスケールを統一した．

最後に，Open Multilingual Wordnet[*11] を用いて，学習データと評価データとして用いる対訳辞書を作成した．本研究では，提案手法を応用して対訳辞書の拡張を行う状況を想定し，頻出語であるほど既に辞書に含まれている可能性は高いことをふまえて，得られた対訳辞書を学習データと評価データに分割した．具体的には，原言語のコーパス内での出現頻度が高い順に全ての単語を並べ，上位 10,000 位までの単語を学習データとし，続く 10,001 位から 11,000 位までの単語を開発データ，11,001 位から 12,000 位までの単語を評価データとした．このとき，多義語が存在した場合はそれぞれを独立した対訳語ペアとして学習データに含める．そのため，表 2 に示すように，学習データ内の原言語の語彙数は全て 10,000 語で固定されているのに対し，目的言語の語彙数と学習データの数は 10,000 件よりも大きくなっている．また，原言語と目的言語の単語が 1:1 に対応していないため，$\mathcal{D}_{train}$ の件数も 10,000 件を超える場合が存在する．表 2 には言語対 (Ja → Es)，(Es → Ja)，(Zh → Es)，(Zh → Es) が含まれていないが，これは Open Multilingual Wordnet に含まれる対訳データが少なく，12,000 位までの単語について対訳辞書を作成できなかったためである．そのため，この 4 種の言語対を除いた計 8 つの言語対について，上記の対訳辞書を作成した．

---

*5 http://dumps.wikimedia.org/から入手．バージョンはそれぞれ Ja: 2014/11/04，Zh: 2014/12/04，En: 2014/10/08，Es: 2015/02/07 である．
*6 https://github.com/yohasebe/wp2txt/
*7 http://taku910.github.io/mecab/
*8 http://nlp.stanford.edu/software/segmenter.shtml
*9 http://nlp.stanford.edu/software/tagger.shtml
*10 http://www.clips.ua.ac.be/pages/pattern
*11 http://compling.hss.ntu.edu.sg/omw/

4·2 比 較 手 法

提案手法の有用性を評価するため，以下に述べる 3 種類の比較手法を実装した．

ベースライン 式 (1) に基づき，翻訳行列を学習する．このとき，提案手法と同様にカウントベースの単語ベクトルを翻訳に用いる．ベースラインと提案手法を比較することで，提案手法で用いる文脈語ペアの集合 $\mathcal{D}_{train}$ と $\mathcal{D}_{sim}$ が単語ベクトルの翻訳精度に与える影響を明らかにする．

CBOW 式 (1) に基づき，翻訳行列を学習する．このとき，提案手法とは異なる，ニューラルネットワークによって学習された Continuous bag-of-words モデル (CBOW) の単語ベクトルを翻訳に用いる [Mikolov 13b]．この手法と前項のベースラインを比較することで，ベクトル表現手法の差異が翻訳精度に与える影響を明らかにする．ただし，4·1 節で述べたとおり，提案手法及び前項のベースラインで用いたカウントベースのベクトルは出現頻度上位 10,000 位までの単語のみ共起語として考慮している．一方で，CBOW モデルのベクトルはコーパス内に出現した全ての単語を共起語として考慮している．本研究では，word2vec[*12] を用いて 4 カ国語の単語ベクトルを得た [*13]．ベクトルの最適な次元数については，原言語のベクトルの次元数が目的言語のベクトルの次元数の 2〜4 倍になるように設定した際の翻訳精度が最も良いという報告 [Mikolov 13b] がある．本研究ではこれをふまえ，原言語のベクトルを m 次元 ($m = 100, 200, 300$) に，目的言語のベクトルを n 次元 ($n = 2m, 3m, 4m$) に設定し，開発データを用いて最適な m と n の組み合わせを探索した．

Direct Mapping 学習された翻訳行列による翻訳が有用であることを確認するため，翻訳行列を用いずに単語ベクトルを翻訳する手法を実装した．この手法では，提案手法と同様にカウントベースの単語ベクトルを翻訳に用いる．3·2 節で述べた学習データから得られる $\mathcal{D}_{train}$ を用いて，原言語のベクトルの次元を，目的言語のベクトルの対応する次元に直接変換する [Fung 98]．この際，原言語の次元と目的言語の次元が 1:1 に対応しない場合も存在する．その場合，各訳語に対して目的言語内での出現頻度に基づき，Reciprocal Rank (順位の逆数) による重み付け和 [Prochasson 09] を用いた．

4·3 評 価 手 法

3·1 節で述べたように，学習した翻訳行列がどれほど正確に言語間の翻訳を行うことができるかを，ある単語のベクトル $\boldsymbol{x}$ を翻訳した結果 $\boldsymbol{W}\boldsymbol{x}$ が，どれほど対訳語のベクトル $\boldsymbol{z}$ に近くなるかで評価する [Mikolov 13b]．具体的には，原言語の各単語ベクトルについて，下記 3 つの手順で評価を行った．(1) 原言語から目的言語へベクトルの翻訳を行う，(2) 得られた翻訳後のベクトルを，目的言語の全ての単語ベクトルと比較し，コサイン類似度が高いもの上位 n 語 ($n = 1, 5$) を選択する．(3) n 語の中に，正解の訳語が含まれているかどうかを評価する．

4·4 実 験 結 果

§1 訳語選択タスクにおける翻訳精度

表 1 に各言語対における単語ベクトルの翻訳精度を示す．**提案手法**は**ベースライン**と比較して，ほぼ全ての言語対において翻訳精度が大きく向上していることがわかる．**CBOW** はカウントベースの単語ベクトルのかわりに CBOW モデルのベクトルを用いることで，**ベースライン**よりも良い精度を達成しているが，**提案手法**と比較すると僅かな改善に留まっている．

提案手法 (sim 無し) は文脈語ペアを学習データのみから探し，文脈語間の表層に関する情報は利用しない．すなわち，$\beta_{sim} = 0$ と設定することで，文脈語ペアの集合 $\mathcal{D}_{sim}$ の影響力を除外する．この手法は特に (Ja, En)，(En, Es) の各組み合わせについて **Direct Mapping** よりも大きく優れており，翻訳行列を学習する手法の有用性を示している．ただし，(Zh → Ja) と (Zh → En) については **Direct Mapping** の方が**提案手法 (sim 無し)** よりも高精度となっている．また，逆方向の (Ja → Zh) と (En → Zh) についても，**Direct Mapping** から**提案手法 (sim 無し)** への改善は僅かである．この現象に関して，表 2 を参照すると，(Ja, Zh) や (Zh, En) の各組み合わせと比較して，(Ja, En) や (En, Es) の各組み合わせは $\mathcal{D}_{train}$ の 1.6 倍から 2 倍程度多いことが読み取れる．すなわち，特に $\mathcal{D}_{train}$ として活用できる対訳辞書が十分存在している状況であれば，辞書から得られる情報 $\mathcal{D}_{train}$ が翻訳精度の向上に大きく貢献するといえる．

また，全ての言語対について，**提案手法**の結果は**提案手法 (sim 無し)** を上回っており，その差は特に (Ja, Zh)，(En, Es) といった言語対において顕著である．日本語と中国語，英語とスペイン語はそれぞれ文字体系が近く，語彙の交換が頻繁に存在する．こうした表層的に類似している言語対においては，表層の類似度から得られる情報 $\mathcal{D}_{sim}$ が翻訳精度の向上に大きく貢献すると考えられる．

なお，言語対をまたいで翻訳精度を比較することは，単語ベクトルを学習する際の Wikipedia コーパスのサイズが異なるため困難である点に注意されたい．今回は各言語で頻度順位 11,001 位から 12,000 位の単語を評価データとしたが，これらの単語はコーパス中での出現頻度はバラつきがある．高頻度の単語ほど，単語ベクトルが密となり翻訳するための情報が多く得られるが，一方で低頻度の単語と比べて意味的曖昧性が多くなり，必ずしも低頻度の単語に比べて翻訳が容易であるとは限らない．

∗12 https://code.google.com/p/word2vec/

∗13 “*a*” や “*the*” といった頻出語の影響を抑制するため，サブサンプリングの閾値パラメータを 1e-3 に設定した．

表 1 実験結果: 単語ベクトルの翻訳精度

言語対	ベースライン		CBOW		Direct Mapping		提案手法 (sim 無し)		提案手法	
	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5
Ja → Zh	0.6%	1.6%	5.4%	13.8%	9.3%	22.2%	11.1%	26.2%	**15.5%**	**34.0%**
Zh → Ja	0.3%	1.2%	2.9%	11.3%	11.6%	26.9%	7.8%	21.6%	**13.1%**	**27.9%**
Ja → En	0.2%	1.0%	6.5%	19.1%	22.3%	37.4%	32.3%	51.0%	**32.5%**	**51.9%**
En → Ja	0.3%	1.1%	4.9%	13.3%	5.4%	13.9%	18.5%	36.4%	**19.3%**	**37.1%**
Zh → En	0.2%	0.8%	3.4%	11.8%	**23.3%**	40.6%	22.3%	40.4%	23.1%	**42.0%**
En → Zh	0.2%	1.1%	5.1%	13.7%	4.5%	11.8%	9.1%	22.1%	**9.5%**	**23.0%**
En → Es	0.2%	1.0%	7.1%	18.9%	11.9%	26.1%	28.7%	45.7%	**31.3%**	**49.6%**
Es → En	0.0%	0.6%	7.5%	22.0%	45.7%	61.1%	46.6%	62.4%	**54.7%**	**67.6%**

表 2 語彙，学習データ，$\mathcal{D}_{train}$ 及び $\mathcal{D}_{sim}$ の件数

	語彙 (原言語)	語彙 (目的言語)	学習データ	$\mathcal{D}_{train}$	$\mathcal{D}_{sim}$
(Ja → Zh)	10,000	10,641	42,037	9,552	3,189
(Zh → Ja)	10,000	20,356	69,619	9,552	3,189
(Ja → En)	10,000	15,060	50,300	18,296	2,234
(En → Ja)	10,000	28,275	84,451	18,296	2,234
(Zh → En)	10,000	15,784	41,144	9,292	3,551
(En → Zh)	10,000	14,770	38,854	9,292	3,551
(En → Es)	10,000	10,247	34,034	15,567	12,764
(Es → En)	10,000	19,917	48,125	15,567	12,764

表 3 $\mathcal{D}_{sim}$ の有無による正解例の数の変化
(全言語対において，評価データは 1,000 例存在する)

	不正解 → 正解	正解 → 不正解
(Ja → Zh)	53	9
(Zh → Ja)	60	7
(Ja → En)	9	7
(En → Ja)	11	3
(Zh → En)	14	6
(En → Zh)	8	4
(En → Es)	52	26
(Es → En)	108	27

文脈語の表層の類似度が翻訳行列に与える影響を分析するため，表 3 に $\mathcal{D}_{sim}$ の有無による正解例の数の変化を示す．(Ja, Zh)，(En, Es) といった言語対においては $\mathcal{D}_{sim}$ が翻訳精度の向上に大きく貢献することは前述したが，この 2 種類の言語対における $\mathcal{D}_{sim}$ の役割は少し異なっていることが表 3 から読み取れる．すなわち，(Ja, Zh) 間の翻訳では，不正解→正解となる例が増える一方で正解→不正解の例はあまり増加しておらず，$\mathcal{D}_{sim}$ は確実に正しい方向に，慎重にバイアスをかけているといえる．しかし，(En, Es) 間の翻訳では不正解→正解の増加とともに正解→不正解も大きく増加していることから，(En, Es) 間の翻訳行列を学習する場合，$\mathcal{D}_{sim}$ は小さい間違いは無視して，よりアグレッシブにバイアスをかける働きをしていると考えられる．

§2 学習データの規模による影響

図 1 は，各言語対について，学習データの大きさを変化させたときの翻訳精度 (Precision@1) の変化をプロットしたものである．ただし，学習データは翻訳行列の学習だけではなく，**Direct Mapping** における次元の変換と，提案手法の報酬項にも用いられている．図から，**Direct Mapping** と**提案手法 (sim 無し)** を比較すると，後者は学習データが少ないときには有効でないことがわかる．これは後者が学習する翻訳行列のパラメータ数と比較して，学習データが少ないことから起こる過学習の影響であると推測される．しかし，**Direct Mapping** では学習データを増加させても精度向上に寄与しにくいことから，十分学習データが存在している場面では翻訳行列の学習を行った方が良いと言える．

また，表層が似ている言語対 (Ja, Zh) と (En, Es) の各組み合わせについては，特に学習データが多いときに**提案手法**の精度が**提案手法 (sim 無し)** を大きく上回っていることがわかる．このことから，文脈語の類似度による手がかり $\mathcal{D}_{sim}$ が翻訳行列の学習を正しい方向に誘導する役割を果たしているといえる．

§3 各翻訳手法の出力例

表 4，表 5，表 6 に (Zh → Ja)，(En → Ja)，(Es → En) の出力結果を一部示す．(太字は正解の翻訳語を表す) 3 つの言語対全てにおいて，**ベースライン**は設問に依存せず，どれも似通った翻訳候補を出力している．この現象は Hubness 問題 [Lazaridou 15] として報告されており，線形回帰を写像関数として用いることによって，説明変数が目的変数よりも原点に近い位置に写像される傾向があることが原因である [Shigeto 15] ことが知られている．**CBOW** の出力を**ベースライン**と比較すると，正解の関連語が多く存在していることがわかる．たとえば，(En → Ja) における sorceress (魔法使い/魔女) の訳語候補には「化け物」や「生霊」が，また xenon (キセノン) の訳語候補には「微粒子」が存在している．こうした例から，CBOW モデルのベクトルを用いることにより，正解の関

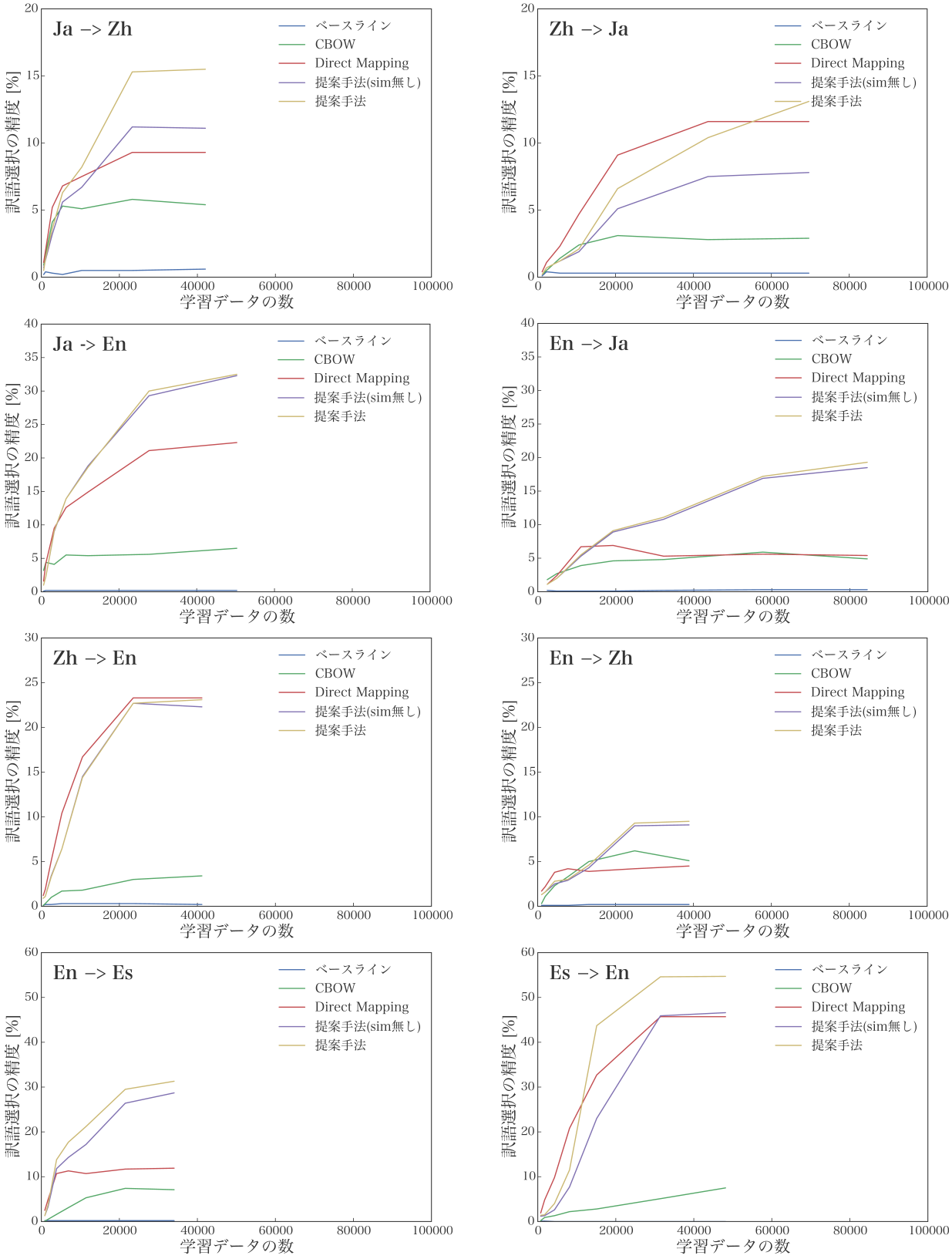

図 1 学習データの規模が翻訳精度に与える影響

連語が出力されやすくはなるものの，正解は出力できていないことがわかる．

Direct Mapping と **提案手法 (sim 無し)** の出力は多くが共通している．これら 2 つの手法は，いずれも学習データに含まれる対訳例をベクトルの翻訳に用いている．前者はベクトルの次元の変換に直接対訳例を用いるのに対し，後者は $\mathcal{D}_{train}$ として学習の目的関数の報酬項として取り込まれている．両手法で翻訳に利用できる手がかりは完全に等しいため，出力結果や翻訳精度が類似していると考えられる．

表 4 (Zh → Ja) の翻訳例

	ベースライン	CBOW	Direct Mapping	提案手法 (sim 無し)	提案手法
校験位 → パリティ/パリティビット/パリティー					
1 位	違う	考慮	プリミティブ	縫い目	**パリティビット**
2 位	動く	相	クライアント	言葉	**パリティ**
3 位	持つ	把握	結び目	**パリティ**	縫い目
4 位	周囲	規準	用語	正しい	クライアント
5 位	十分	正しい	ディレクトリ	一見	言葉
焼瓶 → フラスコ					
1 位	周囲	卵殻	空気	**フラスコ**	**フラスコ**
2 位	軽い	微粒子	**フラスコ**	空気	空気
3 位	動く	**フラスコ**	溶解	磨る	活栓
4 位	小さい	小片	滴	寒天	磨る
5 位	持つ	薄片	寒天	天井	寒天
小農 → 小作農					
1 位	周囲	変質	**小作農**	困窮	困窮
2 位	見る	溜め込む	困窮	把握	保護
3 位	現れる	無駄	配慮	好ましい	把握
4 位	かなり	不潔	把握	配慮	**小作農**
5 位	動く	腐敗	作物	保護	配分

表 5 (En → Ja) の翻訳例

	ベースライン	CBOW	Direct Mapping	提案手法 (sim 無し)	提案手法
sorceress → 魔法使い/魔女					
1 位	思う	化け物	魔物	魔物	**魔法使い**
2 位	恐ろしい	生霊	恐ろしい	**魔法使い**	魔物
3 位	捨てる	狂気	**魔法使い**	呪い	**魔女**
4 位	怒る	暴君	思う	**魔女**	呪い
5 位	邪魔	人殺し	呪い	怪物	エルフ
xenon → キセノン					
1 位	逆	微粒子	気体	放射	**キセノン**
2 位	実際	蒸散	放射	**キセノン**	放射
3 位	ある程度	変化	粒子	気体	気体
4 位	弱い	吸い込む	特性	粒子	粒子
5 位	小さい	縮む	小さい	重力	重力
abduct → 連れ去る					
1 位	思う	追い払う	逃げる	**連れ去る**	襲う
2 位	捨てる	騙す	逃げ出す	襲う	**連れ去る**
3 位	恐ろしい	庇う	襲う	殺害	殺害
4 位	逃げる	責める	思う	殺し	殺し
5 位	怒る	見捨てる	恐ろしい	逃げ出す	逃げ出す

表 6 (Es → En) の翻訳例

	ベースライン	CBOW	Direct Mapping	提案手法 (sim 無し)	提案手法
clericalismo → clericalism					
1 位	call	attitude	struggle	attitude	**clericalism**
2 位	describe	banality	attitude	negativity	attitude
3 位	intend	self-consciousness	turn	struggle	struggle
4 位	make	fatalistic	**clericalism**	**clericalism**	negativity
5 位	ignore	egoism	espouse	fatalistic	chauvinism
papio → baboon					
1 位	call	crab	elephant	cow	**baboon**
2 位	turn	dwarf	antelope	ichthyosaur	crocodile
3 位	make	elephant	cow	parcel	dwarf
4 位	describe	crocodile	parcel	elephant	elephant
5 位	intend	hairy	bovid	crocodile	obscure
yambo → iamb					
1 位	call	fairy	call	**iamb**	interrogative
2 位	turn	pluck	turn	caesura	stanza
3 位	describe	stick	stanza	interrogative	**iamb**
4 位	make	dark	set	gesture	caesura
5 位	intend	croak	**iamb**	stanza	gesture

提案手法では，**提案手法 (sim 無し)** で用いる対訳例の情報に加え，さらに文脈語の表層の類似度を手がかりとして活用している．この手がかりを $\mathcal{D}_{sim}$ として目的関数に加えることで，**提案手法**はより正確な翻訳を出力できている．たとえば，(Zh → Ja) における校験位 (パリティビット) の訳語候補を比較すると，**提案手法 (sim 無し)** では翻訳候補の 3 位に正解「パリティ」が出現しているが，**提案手法**では 1 位，2 位がともに正解 (「パリティビット」「パリティ」) となっている．(Zn → Ja) における小農（小作農）や (Es → En) の yambo (iamb) の翻訳のように，提案手法で正解を出力できなくなった例も存在するが，多くの正解は翻訳候補の上位に出力されていることがわかる．

5. お わ り に

本稿では，単語ベクトルを他の言語のベクトルへ翻訳する際に事前知識を取り込む新しい手法を提案した．我々は (1) 学習データと，(2) 表層の類似度という 2 種類の手がかりから得られる文脈語ペアの集合を考慮可能な目的関数を設計した．提案手法は，Mikolov らが提案した教師あり学習を用いたベクトル翻訳という連続的な数値最適化の枠組み [Mikolov 13b] に，対訳辞書や表層の類似性で定義される単語間の離散的な対応関係 [Fung 98] を学習の手がかりとして取り込むことで，各手法の短所を補い，長所を組み合わせた手法と言える．実験から，我々の手法が日本語，中国語，英語，スペイン語における単語ベクトルの翻訳精度を大きく向上させることを示した．

本稿で用いたカウントベースの単語ベクトルは，同一言語内の意味の類似性判定タスクに用いる場合，ニューラルネットワークベースの単語ベクトルと比べて精度が劣るため [Baroni 14]，単語ベクトルの翻訳が適切に行えたとしても，類義性判定の失敗により適切な訳語が選べないことがありうる．ただ，Levy らによる最新の報告によれば，カウントベースのベクトルに特異値分解などを施すことで，その質をニューラルネットワークベースの単語ベクトルと同程度に向上させることが可能である [Levy 15]．そこで，今後の課題としては，目的言語の単語ベクトルと本研究で翻訳した単語ベクトルに，それぞれ Levy らの手法を適用することでその質を改善し，得られた単語ベクトルの間の類義性判定を行うことで訳語選択の精度を改善することが挙げられる．

謝　　辞

本研究の実験を行うにあたり，Facebook AI Research の Tomas Mikolov 氏より大変有益な助言を頂きました．

本論文は，Proceedings of the 19th Conference on Computational Natural Language Learning において発表した内容をもとに，実験と議論を追加し，再構成したものです [Ishiwatari 15]．

本研究の一部は JSPS 科研費 25280111，16H02905，16K16109 の助成を受けたものです．

♢ 参　考　文　献 ♢

[Baroni 14] Baroni, M., Dinu, G., and Kruszewski, G.: Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 238–247 (2014)

[Bengio 03] Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C.: A neural probabilistic language model, *Journal of Machine Learning Research*, Vol. 3, pp. 1137–1155 (2003)

[Bottou 04] Bottou, L.: Stochastic learning, in *Advanced Lectures on Machine Learning*, pp. 146–168, Springer (2004)

[Chandar A P 14] Chandar A P, S., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., and Saha, A.: An autoencoder approach to learning bilingual word representations, in *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 1853–1861 (2014)

[Church 90] Church, K. W. and Hanks, P.: Word association norms, mutual information, and lexicography, *Computational Linguistics*, Vol. 16, No. 1, pp. 22–29 (1990)

[Dinu 14] Dinu, G. and Baroni, M.: How to make words with vectors: Phrase generation in distributional semantics, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 624–633 (2014)

[Erk 08] Erk, K. and Padó, S.: A structured vector space model for word meaning in context, in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 897–906 (2008)

[Erk 12] Erk, K.: Vector space models of word meaning and phrase meaning: A survey, *Language and Linguistics Compass*, Vol. 6, No. 10, pp. 635–653 (2012)

[Faruqui 14] Faruqui, M. and Dyer, C.: Improving vector space word representations using multilingual correlation, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 462–471 (2014)

[Firth 57] Firth, J. R.: A synopsis of linguistic theory, *Studies in Linguistic Analysis*, pp. 1–32 (1957)

[Fung 98] Fung, P.: A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora, in *Proceedings of the Third Conference of the Association for Machine Translation in the Americas (AMTA)*, pp. 1–17 (1998)

[Gouws 15] Gouws, S. and Søgaard, A.: Simple task-specific bilingual word embeddings, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technology (NAACL-HLT)*, pp. 1386–1390 (2015)

[Harris 54] Harris, Z. S.: Distributional structure, *Word*, Vol. 10, No. 2-3, pp. 146–162 (1954)

[Hermann 14] Hermann, K. M. and Blunsom, P.: Multilingual models for compositional distributed semantics, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 58–68 (2014)

[Ishiwatari 15] 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, in *Proceedings of the 19th Conference on Computational Natural Language Learning (CoNLL)*, pp. 300–304 (2015)

[Ishiwatari 16] Ishiwatari, S., Yoshinaga, N., Toyoda, M., and Kitsuregawa, M.: Instant translation model adaptation by translating unseen words in continuous vector space, in *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)* (2016)

[Kalchbrenner 13] Kalchbrenner, N. and Blunsom, P.: Recurrent convolutional neural networks for discourse compositionality, in *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, pp. 119–126 (2013)

[Klementiev 12] Klementiev, A., Irvine, A., Callison-Burch, C., and Yarowsky, D.: Toward statistical machine translation without parallel corpora, in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 130–140 (2012)

[Lazaridou 15] Lazaridou, A., Dinu, G., and Baroni, M.: Hubness and pollution: Delving into cross-space mapping for zero-shot learning, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 270–280 (2015)

[Levy 15] Levy, O., Goldberg, Y., and Dagan, I.: Improving distributional similarity with lessons learned from word embeddings, *Transactions of the Association for Computational Linguistics*, Vol. 3, pp. 211–225 (2015)

[Marton 09] Marton, Y., Callison-Burch, C., and Resnik, P.: Improved statistical machine translation using monolingually-derived paraphrases, in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 381–390 (2009)

[Mikolov 13a] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, in *Proceedings of Workshop at International Conference on Learning Representations (ICLR)* (2013)

[Mikolov 13b] Mikolov, T., Le, Q. V., and Sutskever, I.: Exploiting similarities among languages for machine translation, *arXiv preprint* (2013)

[Prochasson 09] Prochasson, E., Morin, E., and Kageura, K.: Anchor points for bilingual lexicon extraction from small comparable corpora, in *Proceedings of the 12th Machine Translation Summit (MT SUMMIT XII)*, pp. 284–291 (2009)

[Razmara 13] Razmara, M., Siahbani, M., Haffari, R., and Sarkar, A.: Graph propagation for paraphrasing out-of-vocabulary words in statistical machine translation, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1105–1115 (2013)

[Saluja 14] Saluja, A., Hassan, H., Toutanova, K., and Quirk, C.: Graph-based semi-supervised learning of translation models from monolingual data, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 676–686 (2014)

[Shalev-Shwartz 11] Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A.: Pegasos: Primal estimated sub-gradient solver for SVM, *Mathematical programming*, Vol. 127, No. 1, pp. 3–30 (2011)

[Shigeto 15] Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., and Matsumoto, Y.: Ridge regression, hubness, and zero-shot learning, in *Machine Learning and Knowledge Discovery in Databases*, pp. 135–151, Springer (2015)

[Turney 10] Turney, P. D., Pantel, P., et al.: From frequency to meaning: Vector space models of semantics, *Journal of Artificial Intelligence Research (JAIR)*, Vol. 37, No. 1, pp. 141–188 (2010)

[Xiao 14] Xiao, M. and Guo, Y.: Distributed word representation learning for cross-lingual dependency parsing, in *Proceedings of the 18th Computational Natural Language Learning (CoNLL)*, pp. 119–129 (2014)

〔担当委員：栗原聡，山川宏，矢入健久〕

2015 年 12 月 10 日　受理

著　者　紹　介

石渡　祥之佑

2014 年 東京大学工学部電子情報工学科卒業．2016 年 同大学院情報理工学系研究科修士課程修了．現在，同研究科博士課程在学中．自然言語処理の研究に興味を持つ．

鍛治　伸裕

2005 年 東京大学大学院情報理工学系研究科博士課程修了．博士 (情報理工学)．東京大学生産技術研究所特任准教授，情報通信研究機構主任研究員などを経て，2015 年よりヤフー株式会社 Yahoo! JAPAN 研究所上席研究員．CGM テキストの解析などの自然言語処理の研究に興味を持つ．

吉永　直樹（正会員）

2000 年 東京大学理学部情報科学科卒業．2002 年 同大学院理学系研究科修士課程修了．2005 年 同大学院情報理工学系研究科博士課程修了．博士 (情報理工学)．2002 年より 2008 年まで日本学術振興会特別研究員 (DC1, PD)．2008 年 東京大学生産技術研究所特任研究員，特任助教，特任准教授，情報通信研究機構主任研究員などを経て，現在，東京大学生産技術研究所准教授．自然言語処理・計算言語学の研究に従事．

豊田　正史

1994 年 東京工業大学理学部情報科学科卒．1996 年 同大学院情報理工学研究科修士課程修了．1999 年 同研究科博士後期課程修了．博士 (理学)．同年，科学技術振興事業団計算科学技術研究員．2001 年 東京大学・生産技術研究所学術研究支援員，同研究所産学官連携研究員，同研究所特任助教授，助教授を経て　現在，同研究所准教授．ウェブマイニング，情報可視化に興味を持つ．

喜連川　優

1983 年東京大学大学院工学系研究科情報工学専攻博士課程修了，工学博士．東京大学生産技術研究所教授，2013 年 4 月より国立情報学研究所所長．データベース工学の研究に従事．情報処理学会功績賞，ACM SIGMOD E.F Codd Innovations Award 受賞．紫綬褒章．21 世紀発明賞, C&C 賞, ACM，IEEE，電子情報通信学会 ならびに情報処理学会フェロー．