単語の多義性を考慮した対訳単語対の抽出

中本　裕大[1]，　田村 晃裕[2]，　二宮　崇[2]
[1] 愛媛大学 工学部 情報工学科，[2] 愛媛大学 大学院理工学研究科 電子情報工学専攻
{nakamoto@ai., tamura@, ninomiya@}cs.ehime-u.ac.jp

1　はじめに

近年，自然言語処理において基本的かつ重要なタスクの一つである単語のベクトル表現として，ニューラルネットワークを活用した単語分散表現が盛んに研究されている [1, 2]．当初は，単言語コーパスからの単語分散表現の獲得が中心であったが，近年では，多言語コーパスを対象として，多言語の単語を共通のベクトル空間で表現する多言語単語分散表現の研究も盛んに行われている [3]．多言語の単語を共通のベクトル空間上に表現できれば，ベクトル間類似度を算出することにより，異なる言語の単語間の類似度を求めることが可能となる．そのため，高品質な多言語単語分散表現が，機械翻訳や言語横断文書分類，対訳辞書獲得などの様々な多言語処理で求められている．本研究では，多言語単語分散表現の品質を改善させることで，対訳単語対抽出の精度改善を試みる．

従来の多言語単語分散表現の獲得手法の代表的な手法として，Translation Matrix を用いる手法がある [4]．この手法は，まず各言語のコーパスから Word2Vec[1, 2] により単語分散表現を獲得し，その後，教師データである対訳単語対を用いて，言語依存の単語ベクトル空間を同一空間に線形写像することで，多言語で共通の単語ベクトル空間を獲得する（詳細は 2.2 節参照）．この従来手法を含め，多くの多言語単語分散表現獲得手法は，異なる言語間で単語ベクトル空間の線形性を仮定している．しかしながら，通常，言語が異なれば単語の意味の粒度や単語の多義性は異なるため，単語ベクトル空間の線形性は必ずしも成り立たない．例えば，英単語 “bank” は日本語の “銀行” という意味に加えて “土手” の意味も持つが，日本語の単語 “銀行” は “土手” の意味を持たない．そこで本稿では，各言語で単語の語義曖昧性を解消した上で言語依存の単語ベクトル空間を構築することで，異なる言語間で単語ベクトル空間の線形性を高めた上で多言語単語分散表現を獲得する手法を提案する（3.1 節）．そして，語義曖昧性を考慮した多言語単語分散表現に基づき対訳単語対を抽出する手法を提案する（3.2 節）．日本語と英語の Wikipedia を用いた対訳単語対抽出の評価実験において，従来の Translation Matrix に基づく手法と比較することで，提案手法の有効性を検証する．

2　関連研究

本節では，2.1 節で単言語の分散表現獲得手法の代表的手法である Word2Vec について説明し，2.2 節で本研究のベースラインである Translation Matrix に基づく手法を説明する．

2.1　Word2Vec

Word2Vec[1, 2] は，2 層のニューラルネットワークを用いて，前後の文脈単語の情報から単語のベクトル表現を学習するモデルである．大量のテキストデータを学習に用いることで，従来の Bag-of-Words による単語ベクトル表現よりも質の良いベクトル表現を獲得可能であることが報告されている．Word2Vec のモデルには，CBOW（Continuous Bag-Of-Words）と Skip-Gram の 2 種類が存在する．

CBOW は，単語列 $w_1, w_2, ..., w_T$ が与えられたとき，文脈単語 $\{w_{t-S}, ..., w_{t-1}, w_{t+1}, ..., w_{t+S}\}$ の one-hot ベクトル表現から得られる文脈ベクトルに基づき単語 w_t を推定するニューラルネットワークを学習する．ここで，S は文脈窓のサイズである．

一方，Skip-Gram は，単語 w_t からその周辺単語 $\{w_{t-S}, ..., w_{t-1}, w_{t+1}, ..., w_{t+S}\}$ のベクトル表現を推定するニューラルネットワークを学習する．

Word2Vec で獲得した分散表現には大きく 2 つの特徴がある．1 つ目は，ベクトル同士の “意味” の足し引きが可能な点である．“$king''$ − “man'' + “$woman''$ のような単語ベクトルの意味の足し引きを実行すると，“queen” に近いベクトルを得られることが知られている．2 つ目は，異なる言語間において意味の似ているベ

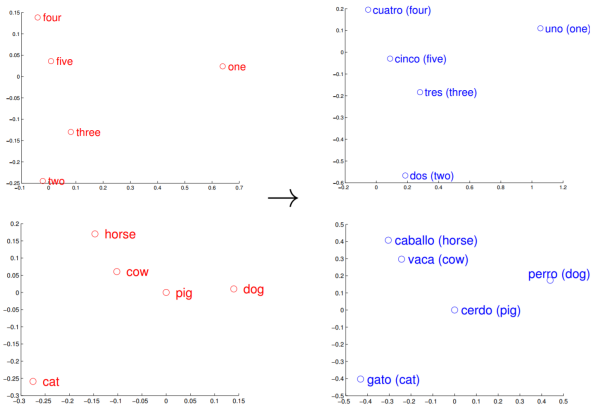

図 1: 英語とスペイン語のベクトル配置 [4]

クトルは似通った配置になることである．Word2Vec の学習は前後の単語関係によって行うため，文構造が似ている言語間の単語ベクトル空間では単語間の関係が類似する．

2.2 Translation Matrix に基づく手法 [4]

2.1 節で述べた通り，Word2Vec では，単語は前後の文脈単語により得られる文脈ベクトルで意味付けられる．したがって，文構造の似通った言語間では単語間の関係性も似通るため，それぞれの言語における単語のベクトル配置に線形な関係がみられる．図 1 に文献 [4] で報告されている実際の英語とスペイン語のベクトル空間を示す．図 1 より, 英単語 “horse” とその対訳となるスペイン語の単語 “caballo” は似通った位置関係にあり，その他の単語も同様であることから，単語間に線形な関係が見受けられることが分かる．Translation Matrix に基づく手法 [4] は，この線形な関係に着目し，Word2Vec により得られた言語 $\{L_f, L_e\}$ のベクトル空間 V_f, V_e を，少量の対訳単語対に基づき，もう一方のベクトル空間上へと線形写像することで L_f と L_e で共有するベクトル空間を獲得する．具体的には，対訳関係にある $x_i \in L_f$ と $z_i \in L_e$ $(i = 1, \cdots, n)$ に対して，式 (1) で表される目的関数を最小化する写像関数 W を学習する：

$$\min_W \sum_{i=1}^{n} ||WV_f(x_i) - V_e(z_i)||^2. \quad (1)$$

ここで，$V_f(x_i)$ と $V_e(z_i)$ は，それぞれ，Word2Vec により得られた，x_i と z_i に対する d_1, d_2 次元のベクトルであり，写像関数 W は $d_2 \times d_1$ の行列である．

Word2Vec により得られるベクトルは，言語が異なっても単語間の幾何学的位置関係が似通っているため，少量の対訳単語対により学習された W を用いてベクトル空間全体を写像することにより，写像後のベクトル空間上においては，教師データ以外の対訳関係にある単語対も近く配置される．したがって，写像後のベクトル空間上でのベクトル間類似度が高い単語対を抽出することで対訳単語対を抽出することができる．実験では，ベクトル間類似度としてコサイン類似度を用いた．具体的には，言語 L_f の単語 x と言語 L_e の単語 z 間の類似度を次の通りに算出した：

$$Cos(WV_f(x), V_e(z)). \quad (2)$$

3 提案手法

3.1 多言語単語分散表現の獲得

従来の多言語単語分散表現の獲得手法は単語の多義性を考慮しないため，複数の意味を持つ単語も一つのベクトルで表現される．したがって，複数の語義が混在するベクトルに基づき言語間の対応関係が学習される．しかし，通常，単語の多義性は言語間で異なるため，言語間で多義性の異なる単語に対しては，写像できない語義が存在してしまう．例えば，図 2 の左図においては，英単語 “bank” は “銀行” には対応付くが，“川岸” には対応付いていない．

そこで提案手法では，各言語で予め語義曖昧性を解消し，語義毎に Word2Vec により単語ベクトルを獲得する．そして，語義毎に規定された単語ベクトル空間に基づき Translation Matrix を学習する．こうすることで，異なる言語間での線形性を高めた後で言語共通のベクトル空間を構築することができるため，従来手法よりも適切な写像関数を学習でき，得られる多言語単語分散表現の品質も向上すると思われる．例えば，図 2 においては，予め，英単語 “bank” を “銀行” の意味を持つ “bank$_1$” と “川岸” の意味を持つ “bank$_2$” に分けて英語の単語ベクトル空間を構築しておくことで，写像後のベクトル空間では “bank” が “川岸” と “銀行” の両単語に近い位置に配置することができる．各言語での単語の語義曖昧性解消には文脈ベクトルを k-means でクラスタリングする手法などの様々な手法を用いる事ができるが，各単語の語義数は予め分かっていない場合が通常であるため，本研究では，文脈ベクトルをノンパラメトリックな手法である DP-means[5] でクラスタリングすることで実現した．ただし，各データとクラスタの重心との類似度は，ユー

Algorithm 1 多言語単語分散表現獲得アルゴリズム

Input: $D = \{D_f,\ D_e\}, S, \lambda$
1. for $l = f, e$ do
2. $V_l \leftarrow$word2vec(D_l)
3. $C_l \leftarrow$dp_means(V_l, D_l, λ)
4. $D'_l \leftarrow$relabel(D_l, C_l)
5. $V'_l \leftarrow$word2vec(D'_l)
6. end for
7. $W \leftarrow$ translation_matrix$(V'_f,\ V'_e,\ T)$
Output: W, V'_f, V'_e

クリッド距離ではなくコサイン類似度を用いた．したがって，単語の文脈ベクトルと既存の全クラスタとのコサイン類似度が λ 未満の場合は，新しい意味クラスタが作成される．

提案の多言語単語分散表現獲得アルゴリズムを Algorithm 1 に示す．提案手法は，多言語テキストコーパス D_l $(l = f, e)$，文脈窓のサイズ S，DP-means のパラメータ λ を入力として受け付け，まず，Word2Vec により，コーパス中の全単語をベクトル化した単語ベクトル集合 V_l を生成する（ステップ 2）．そして，DP-means により各出現位置の単語をクラスタリングし，クラスタリング結果 C_l を獲得する（ステップ 3）．その後，各出現位置の単語に対して，属するクラスタをその出現位置での単語の意味とみなし，単語を「単語_意味」のラベルに変更する（ステップ 4）．例えば，ある出現場所の単語 “bank” のクラスタが “2” であった場合，その位置の単語 “bank” を “bank_2” に変更する．この変更後のコーパス D'_l に対して，Word2Vec を適用することで，単語の語義毎に規定された単語ベクトル空間 V'_l を獲得することができる（ステップ 5）．最後に，語義毎に規定された単語ベクトル空間 V'_l $(l = f, e)$ と教師データである対訳単語対 T を用いて，V'_f を V'_e に写像する行列 W を式 (1) に基づき学習する（ステップ 7）．教師データ T 中の各単語の語義は判定できないため，提案手法では T の中で，DP_means により語義が 1 つと判定された単語同士の対訳単語対のみを用いる．

3.2 対訳単語対の抽出

提案手法においても，従来手法と同様，写像後のベクトル空間上でベクトル間類似度が高い単語対を対訳対として抽出する．類似度の尺度としてはコサイン類似度を用いる．ただし，提案手法のベクトル空間は語義毎に規定されているため，言語 L_f の単語 x と言語 L_e の単語 z の類似度は式 (3) の通り算出する：

$$max_{c_f \in C_f(x), c_e \in C_e(z)} Cos(WV'_f(x_{c_f}), V'_e(z_{c_e})). \quad (3)$$

式 (3) において，$C_f(x)$ と $C_e(z)$ はそれぞれ x と z のクラスタを表し，x_{c_f} と z_{c_e} はそれぞれ意味 c_f の単語 x，意味 c_e の単語 z を表す．

3.1 節で述べた通り，提案手法では写像行列 W の学習時に，教師データ T 中の一部の対訳単語対のみを用いる．一方で従来手法は，教師データ T 全体を用いることができる．したがって，提案手法と従来手法で捉えられる対訳関係が異なる可能性がある．そこで，従来手法による単語間類似度 (式 (2)) と提案手法による単語間類似度 (式 (3)) の和を単語間類似度とし，その類似度が高い対訳対を抽出する手法も提案手法とする．以降では，式 (3) による対訳単語対抽出手法を「提案手法 1」，式 (2) と (3) の和による対訳単語対抽出手法を「提案手法 2」と記す．

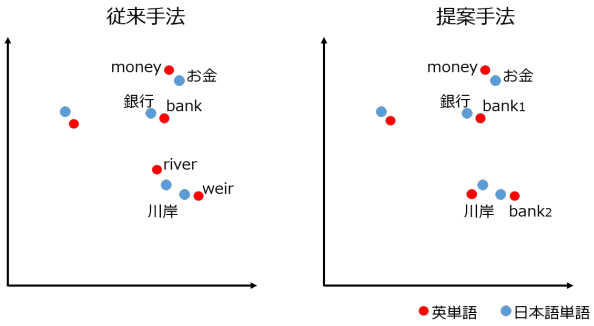

図 2: マッピングの例

4 実験

4.1 実験設定

本節では，Wikipedia の日本語と英語の単言語コーパス[1]からの日英対訳単語対抽出実験を通じて提案手法の有効性を検証する．日本語の Wikipedia データの単語分割は Kytea[6] [2]を使用した．また，データの前処理として，重複文の削除，特殊文字の削除，数値の特殊文字 ⟨NUM⟩ への置換，アルファベットの小文字への統一を行った．また，従来手法，提案手法における写像行列 W の教師データ及びテストデータに用いる対訳単語対は，Wiktionary[3]から作成した．具体的に

[1] https://sites.google.com/site/rmyeid/projects/polyglot#TOC-Download-Wikipedia-Text-Dumps
[2] http://www.phontron.com/kytea/
[3] http://hlt.sztaki.hu/resources/

表 1: 実験データ

	単語延べ数	語彙数
英語コーパス	89,949,091	175,485
日本語コーパス	107,237,345	142,746

	対訳単語対の数
従来手法の学習データ	4,300
提案手法の学習データ	3,382

表 2: 対訳単語対抽出性能 (%)

	1 位正解率	10 位正解率
従来手法 [4]	6.4	19.4
提案手法 1	6.7	16.2
提案手法 2	7.3	19.8

は，Wiktionary 中の対訳単語対の内，日英 Wikipedia コーパスに含まれる単語からなる単語対の中で頻度が高い 4,300 対を教師データとして用い，続く 1,000 対をテストデータとして用いた．ここで，提案手法における，多言語単語分散表現の写像行列 W の学習時には，4,300 対の中で，クラスタリングの結果，1 つの意味しか持たなかった単語同士の対のみを用いることを再確認しておく．また，DP_means における λ の値は，英語では 0.42，日本語では 0.38 とした．表 1 に実験に使用したデータの統計値を記す．

4.2 結果

従来手法と提案手法の対訳対抽出性能を表 2 にまとめる．評価指標は，1 位正解率と 10 位正解率を用いた．

表 2 より，1 位正解率は提案手法 1，2 ともに従来手法を上回ったことが確認できる．この結果より，多義性を考慮することで対訳対抽出性能を改善できることが分かる．一方で，10 位正解率に関しては，提案手法 1 は従来手法より劣っている．これは，写像関数 W を学習する際，提案手法では従来手法で使った学習データより小規模なデータしか学習データとして使えなかったことが原因と考えられる．今後は，全ての学習データを使って写像関数 W を学習する手法に改良していきたい．

5 おわりに

本論文では，単語の多義性を考慮した多言語単語分散表現の獲得手法を提案した．そして，この提案手法により獲得した分散表現に基づき対訳単語対を抽出する手法を提案し，日英対訳単語対抽出実験を通じて，提案手法の有効性を示した．今後は，提案手法を日英以外の言語対にも適用し，提案手法の有効性を検証する予定である．また，特定非営利活動法人言語資源協会が保持する講談社和英辞典データに適用して提案手法の性能を確かめる予定である．

謝辞

本研究は，特定非営利活動法人言語資源協会からの受託研究「和英辞典コーパスの活用と拡張に関する研究」の中で行ったものである．ここに謝意を表する．

参考文献

[1] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*, 2013.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pp. 3111–3119, 2013.

[3] Shyam Upadhyay, Manaal Faruqui, Chris Dyer, and Dan Roth. Cross-lingual models of word embeddings: An empirical comparison. In *ACL*, pp. 1661–1670, 2016.

[4] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *CoRR abs/1309.4168*, 2013.

[5] Brian Kulis and Michael I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *CoRR abs/1111.0352*, 2013.

[6] Graham Neubig, Yosuke Nakata, and Shinsuke Mori. Pointwise prediction for robust, adaptable japanese morphological analysis. In *ACL*, pp. 529–533, 2011.