対訳抽出におけるハブの影響

重藤 優太郎　　新保 仁　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{yutaro-s, shimbo, matsu}@is.naist.jp

1　はじめに

1.1　背景

統計的機械翻訳では，パラレルコーパスから翻訳知識を獲得する．しかし，パラレルコーパスの作成はコストの高い作業である．また，現存するパラレルコーパスは限られた言語対や分野の文書を対象にしたものがほとんどである．これらの理由から，コンパラブルコーパスを用いた自動での対訳抽出に関する研究が盛んである [1, 4, 5, 7, 9, 11, 13, 14].

対訳抽出は *distributional hypothesis* [6] を二言語に拡張した仮説に基づいており，対訳対が類似した文脈で用いられ，類似したトピックを持つことを期待している．この仮説より，対訳抽出は二言語の単語を共通の素性空間で表現し，二言語間の単語対の類似度を計算する．その後，類似度が最も高い単語対を対訳対として抽出する．

二言語の単語を共通の素性空間で表現するために，シード（既知の対訳対）を用いる手法 [1, 4, 11, 13] やトピックモデル [5, 9] を用いる手法が提案されている．

英仏のような同族言語間の対訳抽出にスペリングの類似度が大きく寄与することが報告されている [5, 7]. しかし，スペリングの類似度は英日や英中などの非同族言語間の対訳抽出には有効でないことが考えられる．

1.2　研究目的

本論文ではコーパスと少量のシードのみが存在すると仮定し，スペリング類似度を用いず，少数のシードのみを用いた対訳抽出を行う．対象とする言語は非同族言語である日英とする．

また，本論文は対訳抽出におけるハブ [10] の影響に注目する．ハブとはデータセット中の多数のオブジェクトの近傍に存在しているオブジェクトを指しており，近年，次元の呪いの一種として注目を集めている．

1.3　本論文の貢献

本論文の貢献を以下に示す．

- 対訳抽出におけるハブの存在を示した．実験の結果，コサイン類似度を用いた手法とラベル伝搬法を用いた手法 [13] のどちらにもハブが発生していることを確認した．また，ハブの発生が対訳抽出の精度に悪影響を与えていることを確認した．
- 中心化が対訳抽出の精度を改善させることを示した．これは中心化によってハブの発生が抑制されていることが理由である．中心化はハブの発生を抑制する効果があることが発見されており [12], 本研究は中心化が対訳抽出の精度の改善に繋がることを示した初めての論文である．また，中心化を用いた対訳抽出は先行研究の精度を上回ることを確認した．

2　共通の素性空間

対訳抽出は単語対の類似度を計算するために，全ての単語を共通の素性空間で表現する必要がある．本論文ではシード対訳対を用いて共通の素性空間を構築する．

シード対訳対を用いて共通の素性空間を作る方法は単語の直接共起を用いる方法 [4, 11] と単語の分布類似度を用いる方法 [1, 7, 13] が提案されている．共起ベクトルを用いた場合，シード単語と共起しなかった単語は零ベクトルとなる．一方で，分布類似度ベクトルはシード単語との類似度をベクトルの要素とするので密ベクトルを得ることが期待できる．従って，シード単語と共起しなかった単語でも，類似度を得る可能性がある．これより，本論文では単語の分布類似度を素性ベクトルとして採用する．

分布類似度ベクトルの素性空間は，(単言語内の) 単語とシード単語の分布類似度で表現される．従って，シード単語が共通の素性空間の基底となり，単語の素性ベクトルの次元数はシード単語数となる．

まず, n 組のシード対訳対 $\mathcal{X} = \{(s^{(i)}, t^{(i)}) \mid i = 1, \ldots, n\}$ が与えられているとする．$s^{(i)}$ は原言語のシード単語であり, $t^{(i)}$ は目的言語のシード単語を表す[*1]．原言語の任

[*1] 上付き文字が単語のインデックスであり，下付き文字はベクトルの要素を表す．

意の単語 s は分布類似度ベクトル $\mathbf{s} = [s_1, \ldots, s_n]^{\mathrm{T}} \in \mathbb{R}^n$ で表現され，分布類似度ベクトルの j 番目の要素 s_j は s と $s^{(j)}$ の類似度である．

本論文ではベクトルの要素 s_j はコサイン類似度で計算する．つまり，単語 s の共起ベクトル $\mathbf{c}$，単語 $s^{(j)}$ の共起ベクトル $\mathbf{c}^{(j)}$ が与えられた場合，s_j は次のように定義される．

$$s_j = \cos(\mathbf{c}, \mathbf{c}^{(j)}) = \frac{\left\langle \mathbf{c}, \mathbf{c}^{(j)} \right\rangle}{\|\mathbf{c}\| \cdot \|\mathbf{c}^{(j)}\|}.$$

同様に，目的言語でも分布類似度ベクトルを構築する．

得られた単語 s の分布類似度ベクトル $\mathbf{s}$ と目的言語の対訳候補の類似度を計算し，順位付けを行う．

3 ハブの影響を考慮した対訳抽出

3.1 高次元空間におけるハブの影響

近年，次元の呪いの一種としてハブの存在が報告されている [10]．ハブは，高次元空間における近傍法で多数のオブジェクトの近傍に出現するオブジェクトのことを指し，ハブの出現は近傍法の精度低下の一因となっている．

Radovanovic ら [10] は種々のデータセットにおいてハブが発生していることを示し，次元数が 20 の素性空間においてもハブが発生する可能性があることを報告した．また，類似度尺度に内積（もしくはコサイン類似度）を用いた場合のハブが発生する理由も報告されている [12].

自然言語処理で扱う素性ベクトルは高次元であることが一般的であり，文書のクラスタリングにおいてハブの発生によって精度が低下していることが報告されている [12]．対訳抽出は高次元空間でのランキングタスクに定式化されるので，ハブの発生により精度の低下が生じていると考えられる．

本論文では対訳抽出における，ハブの発生を抑制する手法として中心化を行う．

3.2 中心化

中心化 [2, 3, 8] はデータセットのセントロイドを素性空間の原点に移動させるものである．データセット $\{\mathbf{x}^i \mid i = 1, \ldots, n\}$ が与えられている場合，データセットのセントロイドは

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}^i$$

で求まる．セントロイドを原点に移動させることは，各オブジェクト $\mathbf{x}^i$ を

$$\mathbf{x}^i_{cent} = \mathbf{x}^i - \bar{\mathbf{x}}$$

に置き換えることで実現される．中心化を行ったデータセット $\{\mathbf{x}^i_{cent}\}$ のセントロイドは $\mathbf{0}$ となる．

データのバイアスを無くすことを目的とした中心化は古くから提案されていたが，内積（もしくはコサイン類似度）を類似度尺度として用いた場合，中心化はデータ中のハブの発生を抑制する効果があることが解明された [12]．中心化は全てのオブジェクトからセントロイドを減算することに相当するので，オブジェクト間の距離が変化することはなく，内積（もしくはコサイン類似度）の値のみが変化する．

4 実験設定

本実験では先行研究に従って対訳抽出をランキングタスクとして定式化する．すなわち，与えられた原言語の単語に対して，目的言語の対訳候補に順位付けを行う．正解である対訳単語にできるだけ高い順位を付けることが目標となる．本実験では対訳抽出を行う前に，コンパラブルコーパスとシード対訳対が与えられているとし，原言語の単語と目的言語の対訳候補は共通の素性空間で表現されているとする．

4.1 データセット

評価のために，2 種類の英日コンパラブルコーパスを用いる．

- MED-PNE: 英語のコーパスとして MEDLINE[*2] の 2006 年の概要 139404 文と日本語のコーパスとして 1985 年から 2006 年までの PNE[*3]の記事 512504 文を用いる．
- WIKI: 内部リンクのある英語版 Wikipedia 5000 記事，334886 文と日本語版 Wikipedia 5000 記事，162138 文を用いる．

本実験にはこれらのデータセットに品詞付与を行い，機能語を削除したものから bag-of-words を構築する．品詞付与は MEDLINE に GENIA tagger[*4] を用い，英語版 Wikipedia には hunpos[*5]を用いた．日本語コーパスは MeCab[*6] を用い，名詞が連続した場合には 1 単語の複合名詞として取り扱った．原言語の単語と目的言語の対訳候補には 10 回以上出現する名詞を用いた．

シード対訳対と評価用の対訳対は同じ辞書を用いる．MED-PNE にはライフサイエンス辞書[*7] を用い，WIKI

[*2] http://www.ncbi.nlm.nih.gov/pubmed
[*3] http://lifesciencedb.jp/pne/
[*4] http://www.nactem.ac.uk/GENIA/tagger/
[*5] https://code.google.com/p/hunpos/
[*6] https://code.google.com/p/mecab/
[*7] http://lsd.pharm.kyoto-u.ac.jp/en/index.html

seed size	method	MAP	top 1	top 5	top 10	top 20	top 30	N_{10} skewness
20% (*243* 対)	cos	18.2	12.1	24.0	29.1	35.8	39.5	5.00
	lp	14.1	9.0	18.2	23.7	29.6	33.7	9.23
	cos+centering	**25.7**	**18.5**	**32.4**	**39.1**	**47.4**	**52.5**	**2.42**
40% (*486* 対)	cos	19.9	13.9	25.6	30.5	36.6	41.6	5.14
	lp	24.0	16.8	30.6	36.1	43.8	47.6	6.40
	cos+centering	**31.2**	**22.9**	**39.2**	**46.7**	**55.5**	**61.1**	**2.72**
60% (*727* 対)	cos	21.5	15.3	26.6	32.2	38.7	44.1	5.51
	lp	30.3	22.7	39.0	44.8	51.2	55.5	3.33
	cos+centering	**35.8**	**27.3**	**44.2**	**51.4**	**58.6**	**65.1**	**2.66**

(a) Medline / PNE + ライフサイエンス辞書

seed size	method	MAP	top 1	top 5	top 10	top 20	top 30	N_{10} skewness
20% (*420* 対)	cos	2.7	1.2	2.9	5.2	7.4	8.9	11.46
	lp	2.2	1.0	3.2	4.1	5.6	6.5	12.32
	cos+centering	**5.6**	**2.8**	**7.5**	**10.4**	**14.1**	**17.2**	**3.24**
40% (*840* 対)	cos	3.1	1.5	4.2	5.7	7.5	9.6	11.46
	lp	4.7	2.5	6.8	8.1	10.2	11.3	13.76
	cos+centering	**7.0**	**3.9**	**11.2**	**14.8**	**19.1**	**22.3**	**3.67**
60% (*1262* 対)	cos	3.2	1.7	4.0	5.4	7.6	9.5	11.90
	lp	5.6	2.9	7.5	10.1	13.4	16.1	14.25
	cos+centering	**9.4**	**5.3**	**13.2**	**17.6**	**21.5**	**24.2**	**3.57**

(b) Wikipedia en/ja + EDR 日英対訳辞書

表 1　実験結果: Mean-averaged precision と top k 精度 ($k = 1, 5, 10, 20, 30$).

には EDR 日英対訳辞書[*8]を用いた．

MEDLINE には名詞が 65477 語，PNE には 415819 語出現しており，その内ライフサイエンス辞書に掲載されていたものは MEDLINE は 2633 語，PNE は 2579 語だった．この内の 1213 語が対訳対であった．英語版 Wikipedia には名詞が 334886 語，日本語版 Wikipedia には 162138 語あり，その内 EDR 日英対訳辞書に掲載されていたものは英語版が 6916 語，日本語版は 5474 語だった．この内の 2012 語が対訳対であった．

4.2 素性ベクトル

2 節で述べた通り，シード単語を基底とした分布類似度ベクトルを求めるために共起ベクトル $\mathbf{c}$ を定義する必要がある．本実験の共起ベクトル $\mathbf{c}$ は左右 4 単語の bag-of-words で構成される．この際，左文脈と右文脈は区別した．ベクトルの標準化に自己相互情報量の正の値のみを用いた．

4.3 比較手法

本実験では以下の手法の評価を行った．

- *cos*: 素性ベクトルのコサイン類似度（ベースライン）．
- *lp*: Tamura ら [13] が提案したラベル伝搬法を用いた対訳抽出手法．枝狩りで保存するエッジ数 $m \in \{50, 100, 200, 300\}$，ラベル伝搬の繰り返し回数 $t \in \{1, 5, 10\}$ は開発セットを用いて決定した．
- *cos+centering*: 中心化を行ったベクトルを用いたコサイン類似度．Suzuki ら [12] による分析は，（順位付けの対象となるオブジェクトではなく）順位付けの基準となるオブジェクトの分布平均（データ集合のセントロイドで近似できる）を原点に移動させることによって，ハブが削減されることを示している．本実験では原言語の単語に対して，目的言語の対訳候補に順位付けを行うため，彼らの分析に従い，セントロイドを原言語の開発セットから計算した．

4.4 評価

4.4.1 評価尺度

本実験では対訳抽出をランキングタスクとして定式化した．よって，ランキングタスクの評価で用いられる Mean Average Precision (MAP) をメインの評価に用いる．参考のため $k \in \{1, 5, 10, 20, 30\}$ ベスト精度も報

[*8] http://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html

告する．

また，対訳抽出の精度とハブの相関を調査するため N_{10} 分布の歪度（N_{10} Skewness）の評価も行う．N_{10} 分布は対訳候補が上位 10 位内に何回出現したかを表現した分布であり，その歪度は手法がどれだけハブの影響を受けているかを示す指標として用いられる [10, 12]．N_{10} 分布の歪度は次式によって計算される．

$$(N_{10}\text{ skewness}) = \frac{\sum_{i=1}^{n}(x_i - \mu)^3 / n}{\sigma^3}.$$

この n は対訳候補の数であり，x_i は i 番目の対訳候補が上位 10 位内に出現した回数である．μ は N_{10} 分布の平均，σ は標準偏差である．N_{10} 分布の歪度が高い場合，上位 10 位内に頻出するハブ対訳候補が多く発生していることを意味する．

4.4.2 データ分割

本実験では既知の対訳対をシード対訳対，開発データ，評価データに分割する．開発データは *lp* のパラメータ選択と *cos+centering* のセントロイドの計算に用いる．既知の対訳対の 60%（ME-PNE: 727 対，WIKI: 1262 対）をシード対訳対，残りの各 20%（MED-PNE: 243 対，WIKI: 420 対）を開発データと評価データとした．

シード対訳対の数と対訳抽出の精度の相関を調査するため，シード対訳対の数を 20% と 40%（MED-PNE: 486 対，WIKI: 840 対）に減らした場合の評価も行う．

本実験ではランダムサンプリングにより異なるデータ分割を 4 セット作り，各々における評価の平均値を報告する．

5 実験結果と考察

5.1 対訳抽出の精度とハブの影響

実験結果を表 1 に示す．各評価指標で最も精度の良かったものを太字で示している．

表 1 より，中心化を行ったベクトルを用いたコサイン類似度（*cos+centering*）が最も良い精度を得ていることがわかる．中心化を行った場合（*cos+centering*）と中心化を行わなかった場合（*cos*）を比べると，中心化を行った場合の精度が向上していることが確認できる．また，N_{10} 分布の歪度を比べると中心化を行った場合より低い数値を得ていることが確認できる．これらの結果から，中心化がハブの発生を抑制し，対訳抽出の精度を改善していることが予想される．

5.2 対訳抽出の精度とシード対訳対の数の影響

次に，シード対訳対の数と精度の影響に注目する．表 1 より，シード対訳対の数を減らしても，中心化を用いたコサイン類似度（*cos+centering*）は他の手法に比べて良い精度を得ている事がわかる．また，N_{10} 分布の歪度はシード対訳対の数を減らしても，低い数値を保っている事が確認できた．

6 まとめ

本論文では対訳抽出におけるハブの発生とその影響について調査を行った．中心化はハブの発生を抑制する効果があり，対訳抽出の精度を向上させることに繋がった．本論文は対訳抽出の精度がハブに影響を受けることを初めて示した．

今後はトピックモデルや他の類似度尺度を用いた場合の対訳抽出におけるハブの影響を調べたい．

謝辞

なお，本研究の一部は（独）情報通信研究機構の委託研究「知識・言語グリッドに基づくアジア医療交流支援システムの研究開発」の一環として実施した．また，ライフサイエンス辞書の利用を許諾していただいた京都大学の金子周司教授と Wikipedia コンパラブルコーパスを提供して頂いた NAIST の劉暁東氏に深く感謝致します．

参考文献

[1] M. Diab and S. Finch. A statistical word-level translation model for comparable corpora. In *Proceedings of the Conference on Content-Based Multimedia Information Access*, 2000.

[2] L. Eriksson, E. Johansson, N. Kettaneh-Wold, J. Trygg, C. Wikström, and S. Wold. *Multi- and Megavariate Data Analysis*, Vol. Part 1, Basic Principles and Applications. Umetrics, Inc., 2006.

[3] D. H. Fisher and H.-J. Lenz eds. *Learning from Data: Artificial Intelligence and Statistics V: Workshop on Artificial Intelligence and Statistics*. Springer, 1996.

[4] P. Fung and L. Y. Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. ACL '98*, pp. 414–420, 1998.

[5] A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein. Learning bilingual lexicons from monolingual corpora. In *Proc. ACL '08*, pp. 771–779, 2008.

[6] Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

[7] P. Koehn and K. Knight. Learning a translation lexicon from monolingual corpora. In *Proceedings of the ACL-02 workshop on Unsupervised lexical acquisition*, pp. 9–16, 2002.

[8] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. Academic Press, 1979.

[9] D. Mimno and H. Wallach. Polylingual topic models. In *Proc. EMNLP '09*, pp. 880–889, 2009.

[10] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

[11] R. Rapp. Automatic identification of word translations from unrelated English and German corpora. In *Proc. ACL '99*, pp. 519–526, 1999.

[12] I. Suzuki, K. Hara, M. Shimbo, M. Saerens, and K. Fukumizu. Centering similarity measures to reduce hubs. In *Proc. EMNLP '13*, pp. 613–623, 2013.

[13] A. Tamura, T. Watanabe, and E. Sumita. Bilingual lexicon extraction from comparable corpora using label propagation. In *Proc. EMNLP '12*, pp. 24–36, 2012.

[14] I. Vulic and M. Moens. A study on bootstrapping bilingual vector spaces from non-parallel data (and nothing else). In *Proc. EMNLP '13*, pp. 1613–1624, 2013.