分散表現を用いた動詞・フレーズの含意関係認識

高津 弘明　　小林 哲則
早稲田大学 理工学術院
takatsu@pcl.cs.waseda.ac.jp　　koba@waseda.jp

1 はじめに

動詞間およびフレーズ間の含意関係を判定する含意認識器の素性に分散表現を導入する。

含意認識 (テキスト含意認識, 含意関係認識,Recognizing Textual Entailment;RTE) は、質問応答や文書要約などの深い自然言語理解を必要とするタスクにおいて重要な技術である。例えば、質問応答において「C 国の石油消費量は A 国についで世界第二位である」という知識から「世界で一番石油の消費量が多い国は A 国だ」という含意関係を推論できれば、「世界で一番石油の消費量が多い国はどこですか？」という質問に答えることができる。含意認識に関する研究の目標は、このような文間での含意関係を認識することである。しかしながら、現状このような高度な含意認識を十分な精度で行えるまでに至っていない。

そこで、問題をより単純化して動詞やフレーズ単位での含意関係を扱う場合が多い。動詞間での含意関係を扱った研究として橋本ら (2009,2011) の研究 [1] [9] がある。彼らは、条件付き確率に基づく方向付き類似度尺度を提案し、従来手法よりも高精度に動詞含意知識を獲得できることを示した。また、フレーズ間での含意関係を扱った研究として泉ら (2014) の研究 [2] がある。彼女らは、「同義」「含意」「反義 (さらに、「属性反義」「経時反義」「視点反義」に分類される)」「無関係」という 4 つの意味関係を付与した述部意味関係コーパス [1] を作成した。その他のフレーズ単位での研究として Kloetzer ら (2013) の研究 [3] がある。彼らは、クラス制約付き二項テンプレート間での含意関係を扱っている。

本研究では、動詞間およびフレーズ間の含意認識を行うための識別関数の素性に分散表現を導入する。分散表現には、似た意味の単語は似たベクトル構造を持つという性質がある。また、分散表現が表す「意味」には、同義性や含意性も含まれ、意味的な包含関係を判定する含意認識において有力な素性になることが期待できる。

本論文の構成は次の通りである。まず、第 2 章で最近の含意関係の研究とそこでの含意の定義について説明し、第 3 章で分散表現について簡単に説明する。そして、第 4 章で分散表現を用いた動詞間での含意認識結果を示し、そこで学習した含意認識器を動詞含意知識の獲得に応用した結果を第 5 章で述べる。また、第 6 章でフレーズ間での含意認識結果を示し、第 7 章で抽象化したフレーズ間での含意認識結果を示し、第 8 章で抽象化したフレーズ間での含意矛盾認識結果を示す。最後に、第 9 章でまとめと今後の展望について述べる。

2 含意関係の定義

動詞間の含意関係を扱った橋本ら (2009,2011) の研究 [1] [9] において、動詞 1 が動詞 2 を含意する (以降、動詞 1 → 動詞 2) とは「動詞 1 の表す事態が成立するなら、同時かそれ以前に、動詞 2 の表す事態も成立している」ことを意味する。フレーズ間の含意関係を扱った泉ら (2014) の研究 [2] において、含意の定義は「どちらか一方の述部がもう一方の述部の意味を包含していること」となっている。ここで注意すべき点は、「借りる → 貸す」[2] など、橋本らの定義で含意とされるものの中には泉らの定義において視点反義 [3](e.g. 「お金を貸す」と「お金を借りる」) とされるものも含まれていることである。

テキスト間での含意関係を扱った研究として横手ら (2013) の研究 [10] がある。彼らはテキスト含意関係認識のための意味類似度変換とその獲得法について提案した。ここでの含意の定義はテキスト T と仮説 H が与えられたとき、T から H が推論可能であるならば、T は H を含意しているというものである。テキストレベルでの含意関係を扱う場合、ほとんどがこの定義に従う。

本研究では、主に動詞の含意関係を対象としているため、橋本らの定義に従う。

3 分散表現

分散表現とは、単語やフレーズを固定長のベクトルで表現したもので、そのベクトルは単語やフレーズが持つ言語的な性質を含み、類似する単語やフレーズは類似したベクトルを持つことが知られている。

従来の方法 (1-of-K 符号化) で単語をベクトル化しようとした場合、語彙数が数万あったとすると一つの単語は数万次元のベクトルで表現されることになる。そのため、1-of-K ベクトルの足しあわせである Bag-of-Words を素性として機械学習を行おうとすると、膨大なメモリと計算時間が必要となる。従来の含意獲得に関する研究の多くが教師無し学習で行われているのもこのことが原因の一つとして考えられる。

Mikolov らは単語の分散表現を学習させるために Continuous Bag-of-Words(CBOW) と Skip-gram というモデルを提案した [4]。CBOW では、注目している単語 w_t の前後 k 単語を文脈とし、文脈の Bag-of-Words 表現を入力として注目している単語 w_t を出力するようなニューラルネットワークを学習する。一方、Skip-gram では、注目している単語 w_t を入力として、文脈中の一単語 w_{t+k} を推定する問題となる。これらのモデルの学習方法として階層的ソフトマックスとネガティブサンプリングがある。階層的ソフトマックスでは、各単語にハフマン符号を振り、階層的なグループを作る。そして、ロジスティック回帰を階層的なグループに対して適用することでソフトマックスを近似的に解く。一方、ネガティブサンプリングでは、出力層で正解ニューロン以外のニューロンを

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?PredicateEvalSet

[2] 『動詞含意関係データベース』上では「作用反作用関係」に含まれる

[3] 定義：「格構造が全く同じだと真逆の意味を表すが、格を交代することで同義になる」

更新しない代わりに、ランダムに 5 個程度の偽入力を選び、その偽入力で正解出力の出る確率が下がるように学習する。ネガティブサンプリングを使う利点として、階層的ソフトマックスよりも計算速度が速いことの他に、出現頻度の高い語に強い、隠れ層の次元が低い場合に強いという特徴がある。また、高頻度語のサブサンプリングを行うことによって学習の高速化が期待できる。なお、Mikolov らは高頻度語のサブサンプリングを行うことによってベクトルの足し引きを行い単語を推論する実験において成績が向上したと報告している [5]。

学習の結果得られたベクトルは、似た意味の単語同士が似たベクトル構造を持つ。さらに、異なる言語でも同じ意味の単語同士は似たベクトル構造になることが確認されており、機械翻訳の精度向上に使えるのではないかという指摘もされている [6]。また、西尾はこのベクトルが表す「意味」について考察する実験を行っており、ベクトルの距離が互いに最も近い単語のペアを出力したところ、高頻度語において 30%が類義語で、22%が同一単語の表記ゆれ、13%が関連語、9%が数値ペア、8%が対義語であったと報告している [11]。このことから、含意関係のある単語が近くに分布していると同時に、矛盾関係にある単語も近くに分布していることが分かる。そのため、分散表現は含意関係があるかどうかの識別には有効であるが、含意関係と矛盾関係の識別に適用するのは難しいと考えられる。実際、8 節で行った含意矛盾認識において、矛盾関係と含意関係間の誤識別は比較的多かった。

4 動詞間の含意関係認識

4.1 素性

含意認識の素性として、分散表現以外に次の 2 つの素性を提案する。次のスコアを計算するために名詞のクラスタリングを行う。名詞のクラスタリングには、鳥澤の方法 [7] [8] を利用する。この方法では、次のような名詞と動詞の係り受け関係に関する隠れクラスを仮定する。

$$P(n,c,v) = \sum_{k=1}^{K} P(\langle c,v\rangle|z_k)P(n|z_k)P(z_k) \quad (1)$$

ここで、v は動詞を表し、n は v と助詞 c を介して係り受け関係にある名詞で、K クラスにクラスタリングすることを考える (以降、$t = \langle c,v\rangle$ をテンプレートと呼ぶ)。各パラメータ $P(t|z_k), P(n|z_k), P(z_k)$ は EM アルゴリズムにより計算する。今回の実験では $K = 500$ とした。学習データは $\{(t_i, n_i, f_i)\}_{i=1}^{L}$ からなる。ただし、f_i は係り受け関係 (t_i, n_i) がコーパス上で現れた回数を表す。コーパスとして Wikipedia と GSK の『京都大学格フレーム (Ver 1.0)』[4] を使用した。Wikipedia の方では、KNP[5] を使用して係り受け関係を求めた。格フレームコーパスの方では、格フレーム情報から係り受け関係を抽出した。ただし、同じ述語 (態タイプは区別する) の格フレームは統合した。

以降の表記において、含意する側の動詞を v_l、含意される側の動詞を v_r とし、助詞 c_l, c_r を伴った動詞 (テンプレート) をそれぞれ l, r とする。

4.1.1 分布スコア

含意獲得の従来研究では、分布類似度に基づいた方法がよく用いられる。そこで用いられる含意関係らしさを表す尺度の多くが似た意味の単語は似た文脈に現れるという分布仮説を仮定している。本研究でも似た意味の単語は似た格構造を有していると仮定して「分布スコア」を導入する。

各動詞の { ガ格, ヲ格, 二格 }(= $\mathcal{C}$) をそれぞれ 500 個の名詞クラスの分布で表し、各動詞のそれぞれの格間で分布の積をとり、その結果を二つの動詞の間に成り立つ特徴とし、このことをスコア関数 $Score_{dist}$ で次のように表現する。したがって、含意スコアに基づく素性の次元数は、「ガ格」「ヲ格」「二格」の各名詞クラスについて $Score_{dist}$ を計算した 1500 次元となる。ただし、$\mathcal{N}(v,c)$ は動詞 v が格 c において項に取る名詞の集合を表し、$f(n|v,c)$ は動詞 v の格 c における名詞 n の出現頻度を表すものとする。

$$Score_{dist}(z_k, c|v_l, v_r) = Score_{case}(z_k, c|v_l) \times Score_{case}(z_k, c|v_r)/Z_d \quad (2)$$

$$Z_d = \sum_{k=1}^{K}\sum_{c\in\mathcal{C}} Score_{dist}(z_k, c|v_l, v_r) \quad (3)$$

$$Score_{case}(z_k, c|v) = \sum_{n\in\mathcal{N}(v,c)} f(n|v,c)\cdot P(z_k|n) \quad (4)$$

$$= \sum_{n\in\mathcal{N}(v,c)} f(n|v,c)\cdot \frac{P(n|z_k)P(z_k)}{P(n)} \quad (5)$$

4.1.2 含意スコア

橋本ら (2009,2011) は、動詞の含意知識獲得に有効な尺度として条件付き確率に基づく方向付き類似度を提案した [1] [9]。そこで定められているスコア関数 $Score$ の根幹に当たる $Score_{base}$ は次のように定義されている。ただし、N_l, N_r はそれぞれのテンプレートが項に取る名詞の集合とする。

$$Score_{base}(l,r) = \sum_{n\in N_l\cap N_r} P(r|n)P(n|l) \quad (6)$$

このような含意の方向性を考慮した条件付き確率に基づくスコア関数を利用することで、低頻度語を過大評価しがちな相互情報量の使用を避けつつ高い精度での動詞含意知識の獲得を可能にした。

この知見に習い、次のような「含意スコア」を導入する。含意スコアに基づく素性の次元数は、「ガ格」「ヲ格」「二格」の各名詞クラスについて $Score_{ent}$ を計算した 1500 次元となる。

$$Score_{ent}(z_k, c|v_l, v_r) = P(\langle c, v_r\rangle|z_k)P(z_k|\langle c, v_l\rangle)/Z_e \quad (7)$$

$$= P(r|z_k)\cdot \frac{P(l|z_k)P(z_k)}{P(l)}\Big/ Z_e \quad (8)$$

$$Z_e = \sum_{k=1}^{K}\sum_{c\in\mathcal{C}} Score_{ent}(z_k, c|v_l, v_r) \quad (9)$$

4.2 実験設定

動詞間の含意関係コーパスとして『動詞含意関係データベース (Version 1.3.1)』[6] を使用した。このデータベースは橋本ら (2009,2011) の手法 [1] [9] をもとに獲得した結果を人手で整備したものである。

[4] http://www.gsk.or.jp/catalog/gsk2008-b/
[5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[6] https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#A-2

このデータベース上のデータの内「含意が成り立つ類義/上位下位関係」「文字列上包含関係にあり、含意が成り立つ類義/上位下位関係」「前提関係」「作用反作用関係」を正例として使用し、「含意、反義、予測関係ではない関連語ペア」「文字列上包含関係にあるが、含意、反義、予測関係ではない関連語ペア」「反義関係」「予測関係」を負例として使用した。

動詞の分散表現を学習するために Wikipedia をコーパスとして使用した。ただし、接頭辞や接尾辞を伴う名詞や連続する名詞は結合して一つの単語として扱う。また、数量を表す名詞は意味属性 ⟨ 数量 ⟩ として汎化し、含意関係コーパス中で “N” と記されている数量表現は ⟨ 数量 ⟩ に置き換えることで表記の統一を行った。なお、ここでは動詞の基本形のみを対象とする。

分散表現の学習においてベクトルサイズは 200 次元に定め、モデルには CBOW を、学習アルゴリズムにはネガティブサンプリングを使用した。また、文脈のウィンドウサイズは 7 に設定しサブサンプリングなしで学習を行った。形態素解析には JUMAN[7] を使用した。

識別関数として RandomForest と SVM を使用し、10 分割交差検定で評価を行った。ただし、RandomForest の決定木の数は 200 個とし、乱雑さの評価基準としてジニ不純度を用いた。また、SVM のコストパラメータ C は 0.1 に設定した。さらに、不均衡データに対して比率に基づいた補正を行っている。

4.3 実験結果

素性として分散表現の結合 (動詞 1 の分散表現に動詞 2 の分散表現を結合したベクトル,400 次元) のみを使用した場合の結果を表 1 に示す (正例 21674 個, 負例 32056 個)。分散表現の結合に加えて分布スコアと含意スコアを加えたベクトル (3400 次元) を素性とし、スコア計算に Wikipedia の係り受け関係を使用した場合の結果を表 2 に示す (正例 10682 個, 負例 12308 個)。スコア計算に格フレームコーパスを用いた場合の結果を表 3 に示す (正例 7371 個, 負例 20809 個)。

表 1: 分散表現の結合

	含意あり ⇒ 含意あり	含意なし ⇒ 含意なし	正解率
RandomForest	64.60	86.76	75.68
SVM (Linear)	81.64	53.37	67.50
SVM (Poly:3)	89.72	42.31	66.01
SVM (RBF)	80.28	55.17	67.72

表 2: 分散表現の結合+分布スコア+含意スコア (Wiki)

	含意あり ⇒ 含意あり	含意なし ⇒ 含意なし	正解率
RandomForest	80.85	79.07	**79.96**
SVM (Linear)	79.96	66.87	73.14
SVM (Poly:3)	97.52	21.58	59.55
SVM (RBF)	87.93	50.21	69.07

表 3: 分散表現の結合+分布スコア+含意スコア (CF)

	含意あり ⇒ 含意あり	含意なし ⇒ 含意なし	正解率
RandomForest	46.43	95.85	71.14
SVM (Linear)	75.84	73.68	74.76
SVM (Poly:3)	95.91	21.67	58.79
SVM (RBF)	83.66	48.82	66.24

[7] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

5 動詞含意知識の獲得

動詞やフレーズの含意獲得に関する研究の多くが、教師なし学習によるものである。含意獲得では、スコア関数を定義して動詞やフレーズのペアに対して適用し、順位付けを行い、ランキング上位に含意関係を有するペアがたくさん集まることを期待している。しかし、その中には含意関係のないペアもいくつか含まれているため、そのようなペアを取り除くことができれば、より純度の高い含意知識を獲得できると考えられる。そこで、ここでは従来の含意関係獲得手法でランキングされた結果に対して含意認識を行うことによって高精度に含意知識を獲得することを試みた。

Wikipedia 上の動詞 (異なり数 140,430 語) のペアに対して、橋本らの *Score* 関数を用いてランキングを行った。その結果に対して含意認識器を適用してフィルタリングを行った。今回は分散表現の結合のみを素性として学習した RandomForest を含意認識器として使用した。そして、両データの上位 1,000 件からランダムに 200 個取り出し、筆者らを除く 3 人の評価者に含意関係の判定を依頼した。評価者に提示した含意の定義は「動詞 1 の表す事態が成立するなら、同時かそれ以前に、動詞 2 の表す事態も成立している」というもので『動詞含意関係データベース』の定義に従った。作業条件は橋本らと同様、知らない単語は辞書やインターネットを使用してその意味を確認してもよいものとし、それでもなお意味が分からない場合は不正解とした。複数の語義を持つ動詞に関しては、そのいずれかの語義について含意関係が認められれば正解とした。なお、評価者間の Fleiss' Kappa は 0.572 で適度に一致していると言える。

Acc1 は 1 人以上が正解と判定した場合の正解率を表し、Acc2 は 2 人以上が正解と判定した場合の正解率を、Acc3 は 3 人とも正解と判定した場合の正解率を表す。この結果からフィルタリング後の方が上位 1,000 件に含まれる含意関係ペアの割合が高いことが分かる。なお、フィルタリング後の 1,000 位の動詞ペアの元の順位は 1479 位である。

表 4: 動詞含意知識の獲得精度

	Acc1	Acc2	Acc3
フィルタリング前	85.2	74.3	67.5
フィルタリング後	89.5	80.0	72.0

6 フレーズ間の含意関係認識

Wikipedia 上で (名詞, 助詞, 述語) の順に出現しているペアを結合し、一つのフレーズとして分散表現を学習した。ただし、述語のフォーマットは『京都大学格フレーム』に合わせて、述語タイプ (e.g. 動詞, 形容詞)、態タイプ (e.g. 使役, 受身) を付与し、さらに、肯定か否定かも区別して学習を行った。学習データには『述部意味関係コーパス』を使用した。ここで、「含意」「同義」「同義 (左右反転)」を正例とし、「無関係」「含意 (左右反転)」「反義」「反義 (左右反転)」を負例とした。ただし、「視点反義」は除いた。正例の数は 7651 個で、負例の数は 4179 個である。

フレーズの分散表現の結合 (400 次元) を素性とした場合の結果を表 5 に示す。

表 5: 分散表現の結合

	含意あり ⇒ 含意あり	含意なし ⇒ 含意なし	正解率
RandomForest	95.76	9.261	52.51
SVM (Linear)	48.62	62.34	**55.48**
SVM (Poly:3)	90.37	10.67	50.52
SVM (RBF)	60.08	40.15	50.12

7 抽象化フレーズ間の含意関係認識

名詞部分を抽象化したフレーズ (e.g. X を殴る → X を攻撃する) 間での含意関係を扱う。

7.1 素性

フレーズを単位とする場合、格が定まっているので、分布スコアと含意スコアの計算にはその格を使用する。したがって、次のように各スコア関数を再定義する。素性の次元数はその格における各名詞クラスについてのスコアとなるため、それぞれ 500 次元である。

$$Score_{dist}(z_k|\langle c_l, v_l\rangle, \langle c_r, v_r\rangle) = Score_{case}(z_k, c_l|v_l) \times Score_{case}(z_k, c_r|v_r)/Z_d \quad (10)$$

$$Z_d = \sum_{k=1}^{K} Score_{dist}(z_k|\langle c_l, v_l\rangle, \langle c_r, v_r\rangle) \quad (11)$$

$$Score_{ent}(z_k|\langle c_l, v_l\rangle, \langle c_r, v_r\rangle) = P(\langle c_r, v_r\rangle|z_k)P(z_k|\langle c_l, v_l\rangle)/Z_e \quad (12)$$

$$Z_e = \sum_{k=1}^{K} Score_{ent}(z_k|\langle c_l, v_l\rangle, \langle c_r, v_r\rangle) \quad (13)$$

7.2 実験

分散表現は抽象化したフレーズに関して学習した。また、学習データ上のフレーズの名詞部分も抽象化している。ただし、分布スコアと含意スコアの計算には格フレームコーパスを用いた。正例の数は 10040 個で、負例の数は 5870 個である。

分散表現の結合 (400 次元) を素性とした場合の結果を表 6 に示す。さらに、分布スコアと含意スコアを加えたベクトル (1400 次元) を素性とした場合の結果を表 7 に示す。この結果から抽象化したフレーズ間の方が認識精度が良いことが分かる。これは、抽象化した方がコーパス上での出現頻度が高くなり、より正確な意味を反映した分散表現を学習できたためだと考えられる。

表 6: 分散表現の結合

	含意あり ⇒ 含意あり	含意なし ⇒ 含意なし	正解率
RandomForest	86.13	45.06	65.59
SVM (Linear)	60.48	69.34	64.91
SVM (Poly:3)	43.65	84.82	64.23
SVM (RBF)	63.38	65.91	64.64

表 7: 分散表現の結合+分布スコア+含意スコア

	含意あり ⇒ 含意あり	含意なし ⇒ 含意なし	正解率
RandomForest	86.18	52.91	69.55
SVM (Linear)	72.63	71.53	**72.08**
SVM (Poly:3)	25.55	93.31	59.43
SVM (RBF)	58.62	68.78	63.70

8 抽象化フレーズ間の含意矛盾認識

学習データには『述部意味コーパス』を使用した。ここで、「含意」「同義」「同義 (左右反転)」を含意関係とし、「反義」「反義 (左右反転)」を反義関係、「無関係」「含意 (左右反転)」を無関係とした。含意関係の数は 9921 個で、反義関係の数は 1417 個、無関係の数は 4402 個である。識別関数の設定条件は 4.2 節と同じである。評価は 10 分割交差検定で行った。

分散表現の結合 (400 次元) を素性とした場合の結果を表 8 に示す。さらに、分布スコアと含意スコアを加えたベクトル (1400 次元) を素性とした場合の結果を表 9 に示す。

表 8: 分散表現の結合

	含意 ⇒ 含意	反義 ⇒ 反義	無関係 ⇒ 無関係	正解率
RandomForest	93.99	34.44	37.78	55.40
SVM (Linear)	74.84	46.22	45.07	55.38
SVM (Poly:3)	36.18	83.84	44.46	54.82
SVM (RBF)	56.93	70.36	51.64	56.31

表 9: 分散表現の結合+分布スコア+含意スコア

	含意 ⇒ 含意	反義 ⇒ 反義	無関係 ⇒ 無関係	正解率
RandomForest	94.59	35.93	31.22	53.91
SVM (Linear)	77.78	49.85	57.21	**61.61**
SVM (Poly:3)	16.23	39.55	82.87	46.22
SVM (RBF)	39.46	73.53	46.41	53.13

9 おわりに

分散表現を用いた動詞およびフレーズの含意関係認識を行った。実験では、動詞間の含意認識において、分散表現のみを使用した場合約 76%の精度で含意関係を判定することができ、さらに提案した素性を加えることで約 80%の精度で判定することができた。また、抽象化フレーズ間の含意認識では、分散表現のみを使用した場合約 66%の精度で含意関係を判定することができ、さらに提案した素性を加えることで約 72%の精度で判定することができた。動詞含意知識の獲得実験では、従来手法でランキングした結果よりもその結果に含意認識を施して得られた結果の方がランキング上位に占める含意関係ペアの割合が高いことを確認した。

今後はより含意認識に有効な分散表現の使い方について検討するとともに、格フレームごとに述語を区別するなどして多義性の問題にも挑戦したいと考えている。

参考文献

[1] Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Masaki Murata, and Jun'ichi Kazama, "Large-Scale Verb Entailment Acquisition from the Web", EMNLP 2009: Conference on Empirical Methods in Natural Language Processing, Poster, pp.1172–1181. 2009. 8.

[2] Tomoko Izumi, Tomohide Shibata, Hisako Asano, Yoshihiro Matsuo and Sadao Kurohashi, "Constructing a Corpus of Japanese Predicate Phrases for Synonym/Antonym Relations", Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014), Reykjavik, Iceland, (2014.5).

[3] J. Kloetzer, S. De Saeger, K. Torisawa, M. Sano,C. Hashimoto, and J. Gotoh, "Large-scale acquisition of entailment pattern pairs", In Information Processing Society of Japan (IPSJ) Kansai-Branch Convention, 2013.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space", In Proceedings of Workshop at ICLR, 2013.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality", In Proceedings of NIPS, 2013.

[6] Tomas Mikolov, Quoc V Le, Ilya Sutskever, "Exploiting similarities among languages for machine translation", arXiv preprint arXiv:1309.4168, 2013.

[7] Kentaro Torisawa,"An Unsuperveised Method for Canonicalization of Japanese Postpositions", in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001), pp. 211-218, Tokyo, Japan, December, 2001.

[8] Jun'ichi Kazama, Kentaro Torisawa, "Inducing Gazetteers for Named Entity Recognition by Large-scale Clustering of Dependency Relations", In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT), pp. 407-415, Columbus, Ohio, USA, June, 2008.

[9] 橋本力, 鳥澤健太郎, 黒田航, デサーガ・ステイン, 村田真樹, 風間淳一, "WWW からの大規模動詞含意知識の獲得", 情報処理学会論文誌, Volume 52, Number 1, pp.293–307, 2011.

[10] 横手健一, ボレガラ・ダヌシカ, 石塚 満, "テキスト含意認識に有効な意味類似度変換及びその獲得法", 人工知能学会誌, Vol.28, No.2, pp.220-229 (2013.2).

[11] 西尾泰和, "word2vec による自然言語処理", 株式会社オライリー・ジャパン, 初版第 1 刷, 2014.