人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第10回)
SIG-AM-10-04

同義語判定問題を用いた語義ベクトルの評価の検討
ーSkip-gram モデルで獲得した語義ベクトルを例としてー

Evaluation of Word Vectors by Synonym Identification
- Skip-gram Word Vectors as an Example -

城光 英彰[1]　松田 源立[1]　山口 和紀[1]

Hideaki Joko[1], Yoshitatsu Matsuda[1], Kazunori Yamaguchi[2]

[1]東京大学総合文化研究科

[1]Graduate School of Arts and Sciences, the University of Tokyo

Abstract: Automatic synonym acquisition is an important problem in the field of document retrieval and data mining using natural language data. In this paper, we conducted two experiments to identify the properties of word vectors acquired by the Skip-gram model, related to the synonym identification. In the first experiment, we confirmed that the cosine similarity of a synonym pair is significantly higher than that of a non-synonym pair. In the second experiment, we show that only a limited number of components of word vectors are needed for discriminating synonym pairs from non-synonym pairs.

1 はじめに

自然言語処理において人間のような意味処理を実現する上で，言い換え表現の獲得は中心的な課題とされている[1]．そのような言い換え表現獲得を含む汎用な意味処理を実現する一つのアプローチとして，意味の基本単位である「単語の意味」について着目することは有用であると考えられる[2]．例えば，文書検索において「東京大学」を検索する際に，「東京大学」だけではなく「東大」や「UT」などを含めた文書も検索対象としたいと考えている場合，「東京大学」の同義語として「東大」や「UT」を獲得しておく必要がある．また，Web 上の文章を用いたデータマイニングなどにおいても，同じ意味を表す単語の違いにより生じるデータスパースネスを解消する上で，同義語の獲得は重要な課題となる．人手により同義語辞書を作成するアプローチも考えられるが，次々と生まれる新語に対応することが困難であることなど問題点が多く，人手による網羅的な同義語辞書の作成は現実的ではない．このような理由から，同義語の自動推定は重要な課題であると考えられる．

同義語の自動推定には様々な手法が存在する．笠原ら[2]は，国語辞典を用いて，

1. 見出し語に対して語義文より特徴行列を作成する．
2. 特徴行列から大規模なシソーラスを用いて属性行列を作成する．

という処理で属性行列を生成しておき，個々の刺激語が与えられたら，

1. 属性行列の語彙に対して，刺激語と単語親密度が高い語のみを検索対象として絞り込む．
2. 1.で検索対象となった語と属性行列を用いて求めた類似度の高い語を結果とする．

という手法により，刺激語の同義語を推定した．吉田ら[3]は同義語の抽出手法として，

1. コーパスにおいて，検索文字列に隣接する文字列を検索する．
2. 得られた文字列から適切な文脈(文字列)を選択する．
3. 文脈に隣接する文字列を検索する．

という処理により，検索時に実用に耐えうる速度で実行できる同義語の抽出手法を提案している．

これらの手法では，同義語の推定に，何らかの単語の素性を使用している．例えば，[2]では，VSM(ベクトル空間モデル)を使用しており，[3]では隣接文字列を使用している．

「同じ文脈に現れる単語は類似した意味を持つ」という分布仮説(distributional hypothesis)[4]や，実際に文脈情報が同義語判定に有用であるとの報告 [5]から，同義語判定においては文脈情報を活用するこ

とは重要であると考えられる．一方で，吉田ら[3]のように単語の出現頻度のみを用いた VSM は，語順を無視し(必然的に，周辺文脈を無視し)ているが，特異値分解などの手法が適用可能であり，スパースネスなどの問題を緩和できるという利点を持つ．

近年，分布仮説に基づきニューラルネットワーク的な手法を用いて単語の”意味”を表すベクトル(語義ベクトル)を求める Skip-gram モデルが提案された[6]．Skip-gram モデルで得られた語義ベクトルは，加法構成性(後述)を持つことやコサイン類似度により単語の意味の類似度が計算できることが報告されている．Skip-gram モデルで求めた語義ベクトルは，VSM と異なり周辺文脈を考慮に入れており，その語義ベクトルを用いれば，同義語を，従来手法より高精度に判定できる可能性がある．しかし，Skip-gram モデルで求めた語義ベクトルの性質については定量的な分析も少なく，どのように利用すれば同義語の判定に効果的に利用できるかは明らかになっていない．そこで，我々は Skip-gram モデルで求めた語義ベクトルの性質を明らかにするためにいくつか実験を行った．本論文では，その実験の結果を報告する．

本論文は以下のような構成となっている．2 節では Skip-gram モデルについて説明する． 3 節では今回行った実験の内容とその結果を述べる．最後に 4 節でまとめと今後の課題を示す．

2 Skip-gram モデル

ここでは Skip-gram モデル[7]について概説する．まず基本的なモデルについて述べ，次にその近似である階層的 softmax モデルについて述べる．最後に，現在までに報告されている特徴を述べる．

Skip-gram モデルは，ニューラルネットワーク的な手法を用いて，コーパスの文脈情報から，各単語の語義ベクトルを学習する手法の一種である．Skip-gram モデルでは，ある単語w_tが文章内の位置 t に存在した場合，その周囲の単語 w_{t+j} $(j \neq 0)$の発生確率$\mathrm{p}(w_{t+j}|w_t)$を以下の式で与える．

$$\mathrm{p}(w_{t+j}|w_t) \propto \exp\left({v'_{w_{t+j}}}^{T} v_{w_t}\right)$$

ここで，ニューラルネットワークモデル的に言えば，v_wはある入力単語wに依存した入力用ベクトル，v'_wはある周辺単語wの出力確率を計算するための出力用ベクトルである．出力確率は，入力用ベクトルと出力用ベクトルの内積に依存し，内積が大きい程確率は高くなる．本論文では，わかりやすさのため，v_wを単語の語義ベクトル，v'_wを文脈ベクトルと呼ぶことにする．なお，確率分布は 1 に正規化されるので，語彙に含まれるすべての単語wでの正規化により，$\mathrm{p}(w_{t+j}|w_t)$は以下で与えられる．

$$\mathrm{p}(w_{t+j}|w_t) = \frac{\exp\left({v'_{w_{t+j}}}^{T} v_{w_t}\right)}{\sum_w \exp\left({v'_w}^{T} v_{w_t}\right)}$$

さらに$\mathrm{p}(w_{t+j}|w_t)$から，あるコーパスが与えられたときの尤度関数ℓを以下のように定義する．

$$\ell = \sum_{t=1}^{T} \sum_{-c \le j \le c,\ j \neq 0} \log \mathrm{p}(w_{t+j}|w_t)$$

ここで T はコーパスのサイズ，c は事前に与えられる文脈窓のサイズである．実際のコーパスを利用して，ℓを最大化する語義ベクトルv_wおよび文脈ベクトルv'_wを求めることが，Skip-gram モデルにおける学習である．

本来のモデルは上記の通りであるが，尤度関数ℓをこのままの形で最大化することは，計算量等の問題で困難であるため，実際にはいくつかの近似が用いられる．ここでは[7]で採用されている近似である階層的 softmax モデルについて述べる．階層的 softmax モデルでは，文脈の計算において，まず単語群を事前に二分木構造に整理しておく．二分木構造としては様々な候補がありうるが，実験的には，頻度に基づく手法である Huffman 木が有効であることが知られており，[7]でも Huffman 木が用いられている．二分木完成後，文脈ベクトルを，各分岐ノードのみに割り当てる．木構造の葉に相当する実際の周辺単語には文脈ベクトルは割り当てられず，これにより推定すべきパラメータ数は大幅に減少する．あるノード k とある単語 w が与えられた場合，そのノードで分岐の右(right)と左(left)を辿る確率を以下で定義する．

$$\mathrm{p}(\mathrm{right}|\mathrm{k},\mathrm{w}) = \frac{1}{1+\exp(-{v'_k}^{T} v_w)}$$

$$\mathrm{p}(\mathrm{left}|\mathrm{k},\mathrm{w}) = \frac{1}{1+\exp({v'_k}^{T} v_w)}$$

ここで，ある周辺単語が与えられたとすると，二分木内にはその単語に辿りつく唯一のパスが存在する．そして，そのパスは，根ノードから，葉に辿りつくまでに，順番に左右どちらを選ぶかで表現される．従って，そのパスを辿る確率は，$\mathrm{p}(\mathrm{right}|\mathrm{k},\mathrm{w})$もしくは$\mathrm{p}(\mathrm{right}|\mathrm{k},\mathrm{w})$をパスに沿って積算することによって与えられる．以上が階層的 softmax モデルによる近似の原理である．実際の学習においては，確率的勾配法に基づくバックプロパゲーションアルゴリズムが利用される．最大化を効率的に行うため，それ以外にもサンプリング等でいくつかの技法が用いられている．学習アルゴリズムの詳細は[7]を参考にさ

れたい．

最後に，Skip-gram モデルの特徴について述べる．Skip-gram モデルにより学習される単語の語義ベクトルは，単語間の何らかの関係性を学習したものであることが報告されている[7]．具体的には，単語”Berlin”の語義ベクトルから，”Germany”の語義ベクトルを減じ，”France”の語義ベクトルを加算すると，”Paris”の語義ベクトルと近いベクトルとなることが知られている．これは，「国家」の「首都」という単語間の関係が，語義ベクトル空間の中で，加法的な関係として抽出されていることを示しており，加法構成性と呼ばれている．このような関係性が語義ベクトルに埋め込まれていることから，単語の関係の主要なものの一つである「同義関係」も語義ベクトルに埋め込まれていることが期待される．しかし，Skip-gram モデルの実験的/理論的な性質には未解明の部分が多いため，本論文では「同義関係」に絞って挙動の詳細な分析を行う．

3 実験

3.1 実験目的

本実験の目的は同義語推定に関係する，Skip-gram モデルの性質を明らかにすることである．そのために，ここでは 2 種類の実験を行った．

実験 1

Skip-gram モデルに関する研究では，意味の類似度を語義ベクトルのコサイン類似度で測るのが一般的である([6], [8]). 同義語は意味の”距離”が近い単語であることから，同義対のコサイン類似度は非同義語対と比較して高い値をとるはずである．そこで，実験 1 では，同義対のコサイン類似度が非同義語対のものよりも高いことを確認する．同義対のコサイン類似度の値には幅があることが予想されるため，特定の閾値を設けて同義語か非同義語かを判定してその精度を見るのではなく，類似度の分布を用いて比較を行った．

実験 2

上で述べたように，語義ベクトルのコサイン類似度により意味の類似度を測る事が一般的に行われているが，語義ベクトルの全ての成分が意味の距離に等しく影響を与えているとは考えにくい．そこで実験 2 では，同義語推定においてベクトルの各成分が与える影響の違いを調査した．なお，実験には線形 SVM を使用したが，これは各成分が与える影響の解釈しやすさを考慮してのことである．

3.2 データ

語義ベクトル作成において用いたコーパスとしては，日本語 Wikipedia データ[1](2Gbytes)を MeCab[2]を用いて基本形出力でわかち書き行った後に，出現回数が 30 万回以上の高頻度語と 100 回未満の低頻度語を除いた 78274 語を使用した．Skip-gram[3]モデルでは，語義ベクトル(および文脈ベクトル)の次元は 200，文脈の広さ c は 5 として，階層的 softmax モデルを用いて学習を行った．求められた語義ベクトルには L2 ノルムが 1 となるよう正規化処理を施してから実験 1 と実験 2 で使用した．

実験 1 及び実験 2 において，同義語の教師例としては Wordnet 同義対データベース[4]を使用した．同義語データベースの中で語義ベクトルが獲得されている 8373 対を正例とし，語義ベクトルが獲得されている 78274 単語の中からランダムに選んだ 8373 対を負例とした．実験 2 においては，これらの単語対に対応する 200 次元の語義ベクトル 2 つを，論文[9]を参考にして，図 1 のように結合し 400 次元のベクトルとして線形 SVM を適用した．論文[9]では，ベクトルの成分毎の和や差を用いた分析も行っているが，結合したベクトルは，和や差のベクトルよりも情報が多く，元のベクトルとの要素の対応も単純で解釈しやすいことから，本実験では結合したベクトルを用いた実験のみを行った．

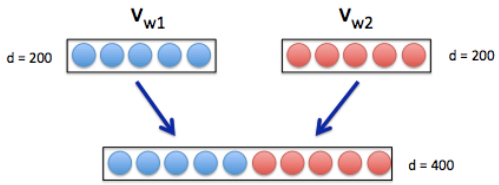

図 1: 実験 2 で用いたベクトルの構成

[1] http://dumps.wikimedia.org/jawiki/ (accessed 2015-5-12).
[2] http://taku910.github.io/mecab/ (accessed 2015-5-29).
[3] https://code.google.com/p/word2vec/ にて Google が公開しているものを使用した．(accessed 2015-2-15).
[4] http://nlpwww.nict.go.jp/wn-ja/jpn/downloads.html にて NICT が提供する，Wordnet[10]を元に作成された同義対データベースである．(accessed 2015-6-6).

3.3 実験 1:コサイン類似度の分布

実験 1 では正例と負例のコサイン類似度の分布の比較を行った．結果を図 2 に示す．平均値については，正例では 0.258，負例では 0.075 となった．正例と負例の分布について，等分散の仮定の下で，右片側 2 標本 t 検定を適用したところ，t 統計量は 71.7 となり，p 値はほぼ 0 であった．これにより，正例と負例でコサイン類似度が有意に異なることが確認できた．

表 1 にコサイン類似度が上位 10 位の同義対, 及び下位 10 位の同義対を示す. コサイン類似度が上位である同義対と比較すると，下位である同義対には一方の単語が複数語義を持つもの(例えば，サークル，ポイント)が多く見受けられる．また，下位 10 例の同義対においては, すべて, 一方が外来語(カタカナ)であり，もう一方が和語や漢語(平仮名や漢字)である．逆に，類似度の上位 10 例は，外来語同士か和語/漢語同士が対になっていることから，外来語と和語/漢語は同義であっても周囲に異なった単語分布を生む可能性が高いと考えられる．

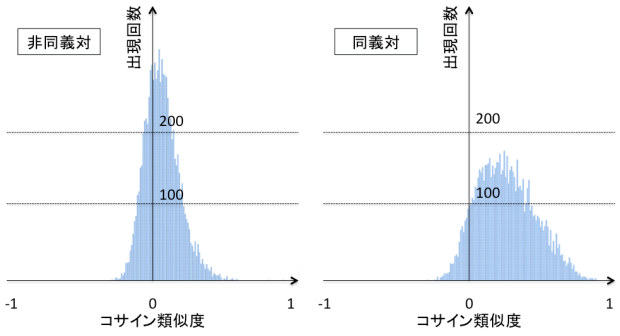

図 2: 同義対(右)と非同義語対(左)の
コサイン類似度の分布

3.4 実験 2:SVM による分類

実験 2 として線形 SVM を用いて同義語の判定を行った．ここでは，分類の性能を上げることが目的ではなく，語義ベクトルの各成分の影響を明確にすることが目的であるため，線形カーネルを用いた．10 分割交差検定により求めた分類の結果は，正答率が 92.59%，精度と再現率は 0.913 と 0.942，F 値は 0.927 となった．また，Confusion matrix を表 2 に示す．これらの結果より，一定程度の分類性能を持っていることが分かることから，意味のある超平面が構成されていると考えられる．

表 1: 同義対とそのコサイン類似度
(上位 10 例と下位 10 例)

同義対	コサイン類似度
ウェブブラウザ, ブラウザ	*0.88946885*
相打ち, 相討ち	*0.87708294*
サイト, ホームページ	*0.8724321*
ウェブサイト, サイト	*0.87111324*
反乱, 叛乱	*0.86343205*
敵意, 敵愾心	*0.85473984*
ウェブサイト, ホームページ	*0.8524012*
吃水, 喫水	*0.8515739*
憤慨, 激怒	*0.8491848*
考え, 考え方	*0.8447531*
・	・
・	・
・	・
キャリヤ, 経歴	*-0.19684665*
サイン, 兆	*-0.1994007*
キー, 緒	*-0.21134022*
サム, 和	*-0.21272291*
サークル, 丸	*-0.21344633*
ルール, 定則	*-0.21581507*
ノース, 子	*-0.21692836*
ポイント, 地	*-0.22948295*
ハイム, 家作	*-0.2397215*
ラック, 幸	*-0.2841633*

次に線形 SVM により算出された 400 次元の各成分の重みの度数分布を図 3 に示す．図 3 から，単語ベクトルの全ての成分が同義語決定に等しく影響を与えているのではなく，少数の成分が大きな影響を与えていることが分かる．

表 2: Confusion matrix

		予測されたクラス	
		正例	負例
実際のクラス	正例	7888	485
	負例	756	7617

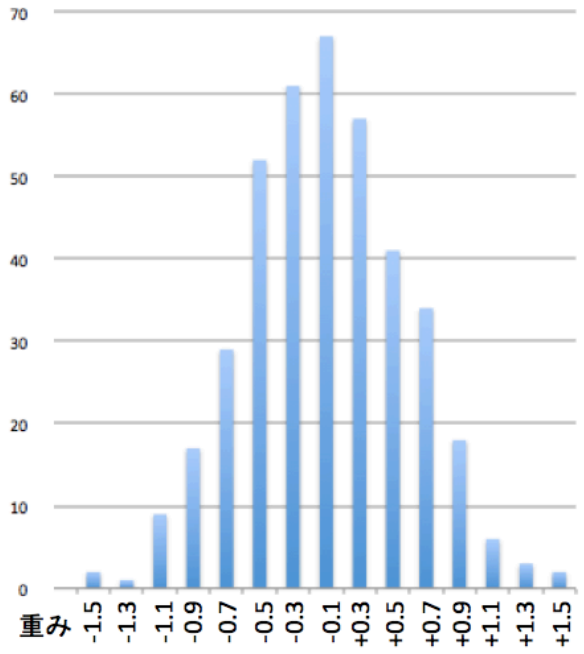

図 3: 線形 SVM の重みの度数分布

4 まとめと今後の課題

本論文では，同義語推定への応用を視野に入れ，Skip-gram モデルを用いて得られた語義ベクトルの性質を明らかにするための 2 つの実験を行った．実験 1 では，同義語のコサイン類似度が非同義語のコサイン類似度より有意に高いことが確認された．同義語であるにもかかわらずコサイン類似度が低いものには，どちらか一方の単語が複数語義を持つものや，一方が外来語で他方が和語/漢語であるものが見られた．単語が複数語義を持つ問題は，単語のもつ語義ごとに語義ベクトルを生成する Skip-gram モデルの拡張(例えば，[11])などを用いることで解決できる可能性がある．また，外来語と和語/漢語の問題は，外来語自体はどの言語にもあるが，今回用いた Wordnet の同義語対データベースが英語の同義語対の翻訳であることに起因する可能性がある．この点については英語の Wordnet で実験することで日本語 Wordnet や翻訳による影響がどの程度あるかを検討することができるであろう．実験 2 の結果からは，語義ベクトルには同義語判定において重要な役割を果たす成分と，それほど影響を与えない成分が存在することが明らかとなった．今後は，同義語の語義ベクトルと線形 SVM の重みの対応を検討することで，各成分の担う意味を明らかにし，これらを進めることで，同義語の推定手法を構築していきたい．

参考文献

[1] 乾健太郎: 自然言語処理と言い換え，日本語学，Vol. 26, No. 11, pp. 50–19, (2007).

[2] 笠原要，稲子希望，加藤恒昭: テキストデータを用いた類義語の自動作成，人工知能学会論文誌，Vol. 18, No. 4, pp. 221-232, (2003).

[3] 吉田稔，中川裕志，寺田昭: コーパス検索支援のための動的同義語候補抽出，人工知能学会論文誌(Web), Vol. 25, No. 1, pp. 122-132, (2010).

[4] Harris Zellig: Distributional structure, Word, Vol. 10, No. 23, pp. 146-162, (1954).

[5] Hagiwara, Masato, Yasuhiro Ogawa, Katsuhiko Toyama: Selection of Effective Contextual Information for Automatic Synonym Acquisition, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pp. 353–360, (2006).

[6] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean: Efficient Estimation of Word Representations in Vector Space, Workshop at International Conference on Learning Representations (ICLR), (2013)

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean: Distributed Representations of Words and Phrases and their Compositionality, Advances in Neural Information Processing Systems (NIPS), (2013).

[8] 椿真史, Kevin Duh, 新保仁，松本裕治: 文の意味構成に伴う高次元空間の最適化と単語表現学習，言語処理学会第 20 回年次大会発表論文集，pp. 1015-1018, (2014)

[9] Julie Weeds, Daoud Clarke, Jeremy Reffin, David Weir, Bill Keller: Learning to Distinguish Hypernyms and Co-Hyponyms, Proceedings of the 25th International Conference on Computational Linguistics, (2014).

[10] Francis Bond, Timothy Baldwin, Richard Fothergill, Kiyotaka Uchimoto Japanese SemCor: A Sense-tagged Corpus of Japanese, The 6th International Conference of the Global WordNet Association (GWC-2012), Matsue, (2012).

[11] Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, Andrew McCallum: Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space, EMNLP, (2014).