形式的類推関係を用いた単語表現再学習

松岡　仁　　ルパージュ　イヴ
早稲田大学大学院　情報生産システム研究科
jinmatsuoka@akane.waseda.jp　yves.lepage@waseda.jp

1　はじめに

深層学習 (deep learning) により意味的かつ統語的な単語表現 (word representation) を自動で学習可能となり自然言語処理 (Natural Language Processing, NLP) の様々なタスクで成功を収めている。特に機械翻訳 (Machine Translation, MT) ではモデル毎に様々な単語表現学習が提案されている。例えば、言語モデルでは NNLM [1] や RNNLM [5] や単言語や双言語を用いた単語表現などがある。特に単語表現を学習するツールとして word2vec [6] が有名であり非常に扱い易く高速な処理を行う。

単語表現を学習するにはその単語の近傍にある単語の統計情報を用いる。基盤となる考えは分布仮説 (distributional hypothesis, [2])（似ている文章で用いられる単語は似ている）に依存する。例えば、「今日は雨です。」と「明日は晴れです。」では「今日」と「明日」、「雨」と「晴れ」の単語同士似ている事を意味する。しかし、「雨」と「晴れ」は反意関係 (antonymy) であり似ているとは考えにくい。（ここでの「似ている」というのは類義語関係 (synonymy) と解釈する）。単語表現では 2 次元上に圧縮した空間でこれらの差異を見ることができる。さらに、学習された単語表現から句や文の意味表現を獲得することも可能である（構成性の原理）。

本研究では学習された単語表現を形式的類推関係を用いて再学習手法を提案する。特徴として単語より広い表現（句や文など）を用いて単語表現の再学習を行わず、単語同士の関係に着目して単語表現の再学習を行う。学習された単語表現をタスクに合わせて再度学習を行うことで、より意味的かつ統語的な単語表現の獲得を目指す。

本論文の構成は 2章で関連研究を紹介し、3章で形式的と意味的類推関係を定義し、再学習のための定式化を行う。4章で実験に用いる学習データと再学習に必要となるデータの構築方法を説明し、結果と考察を示す。

2　関連研究

単語表現は一種のベクトル空間モデル (Vector Space Model, VSM) である。単語の共起情報を用いて次元削減を用いて表現される。次元削減には特異値分解 (Singular Value Decomposition, SVD) や非負値行列因数分解 (Non-negative Matrix Factorization, NMF) などの行列演算手法や確率モデル（トピックモデル）の確率的潜在意味索引付け (Probabilistic Latent Semantic Indexing, PLSI) や潜在ディリクレ配分 (Latent Dirichlet Allocation, LDA) などの手法が用いられた。近年では深層学習でのニューラルネットワークを用いた手法が台頭している。

ニューラルネットワークを用いた単語表現では基本的に入力層、出力層、隠れ層の 3 層ニュラルネットワークから構成される。隠れ層で入力層の次元を削減し、隠れ層の次元から出力層を復元する。このとき、入力層と出力層の差が最小となるように逆伝播法を用いて学習される。word2vec では skip-gram モデル（入力層では対象とする単語となり、出力層はその単語の近傍にある単語群）と処理速度を向上のため negative サンプリングが用いられる。

単語表現を再学習した研究では、椿らは文の意味表現を単語から構成し、その逆の演算（分解）からそれらの単語表現の再学習を行った [8]。彼らは文の意味的類似度を測るタスクで利用したが、本論文では単語間の関係を測るためのタスク（類推推論）で利用し、意味的かつ統語的な要素の強度を再評価する。

3　類推関係

類推関係 (proportional analogy) は 4 項で表現され、表記としては $A : B :: C : D$ となり、意味は「A と B の関係は C と D の関係と同じ」である。類推関係は一種の知識処理であり、大学入学試験（Scholastic Aptitude Test, SAT）などで利用されている。

類推関係には文字列と記号による類推関係（形式的類推関係）と単語間の関係による類推関係（意味的類推関係）が存在する。そのため、我々は3.1章で形式的類推関係、3.2章で意味的類推関係を紹介する。形式的類推関係では意味的類推関係を一部表現することができ、逆も同様である。それらの事実を踏まえて3.3章で単語表現再学習手法を提案する。

3.1 形式的類推関係

形式的類推関係は文字列とその文字列で用いられる記号に着目した関係である [3]。以下に形式的類推関係の例を示す。

動物 : 動物園 :: 保育 : 保育園

上記の例では「動物」と「動物園」、「保育」と「保育園」は両方共「園」が接尾となり形式的類推関係である。以下に形式的類推関係 $A : B :: C : D$ となる 3 つの制約を示す。

$$\text{edit\_distance}(A, B) = \text{edit\_distance}(C, D) \quad (1)$$

$$\text{edit\_distance}(A, C) = \text{edit\_distance}(B, D) \quad (2)$$

$$|A|_x - |B|_x = |C|_x - |D|_x, \forall x \quad (3)$$

ここでの edit_distance は挿入と削除のみを考慮し、それらのコストが 1 での編集距離である。最後の制約は $|A|$ は A の文字列において用いられる記号 x の濃度である。これらの 3 つの制約を満たすとき、formal_score($A : B :: C : D$) を 1 とし、それ以外を 0 で定義する。

3.2 意味的類推関係

意味的類推関係は単語間の意味に着目した関係である [7]。以下に形式的類推関係の例を示す。

フランス : パリ :: 日本 : 東京

上記の例では「フランス」の首都は「パリ」であり、「日本」の首都は「東京」という同一関係にある。形式的類推関係では文字列とその記号を用いるが、意味的類推関係では学習されたベクトルによる単語表現を用いる。意味的類推関係 $A : B :: C : D$ は D を求めたいとするなら、

$$\mathbf{w}_D \approx \mathbf{w}_B - \mathbf{w}_A + \mathbf{w}_C \quad (4)$$

で定式化できる。$\mathbf{w}$ はベクトル化された単語表現である。つまり、A と B と C の簡易な演算のみで D を表現することができる。WordNet などの知識資源から単語間の意味関係（上位/下位、全体-部分、反意関係など）や単語間の統語語彙パターンから意味的類推関係を構築する事ができる。

3.3 単語表現再学習

形式的類推関係と意味的類推関係を用いて学習された単語表現を再学習手法を提案する。そのため、式 4の意味的類推関係を以下のように再定義する。

$$\text{semantic\_score}(A : B :: C : D) = \exp(-\frac{(\mathbf{W}^T \phi(A : B :: C : D))^2}{2s^2}) \quad (5)$$

ここでの $\mathbf{w}$ は $V \times 1$ 次元の学習された単語表現であり、$\mathbf{W}$ は $V \times 1$ 次元の特徴ベクトルである。本式は平均 0 で分散 s^2 のガウス基底関数を用いた。つまり、指数部分が 0 になるほど値としては 1 に近くなる。ϕ は意味的類推関係で用いた考えを定式化した素性関数であり、以下に示す。

$$\phi(A : B :: C : D) = \mathbf{w}_B - \mathbf{w}_A - \mathbf{w}_D + \mathbf{w}_C \quad (6)$$

この素性関数は前章での式 4から導出できる。意味的類推関係での素性関数と特徴ベクトルの内積が 0 に近づく程スコアは 1 に近づく。逆に内積が 0 から遠くなる程スコアは 0 に近づく。つまり、0 から 1 の値域で semantic_score を定義できる。

formal_score（3.1参照) と式 5を用いて L2 正則化を用いたコスト関数 $J(S)$ を以下のように定義する。

$$\frac{1}{2}||\text{formal\_score} - \text{semantic\_score}||^2 + \frac{1}{2}\mathbf{S}^T\mathbf{S} \quad (7)$$

このコスト関数を最小化するために確率的勾配降下法 (Stocastic Gradient Decent, SGD) を用いて、それぞれのパラメータを以下のように学習する。

$$S_{n+1} = S_n - \eta \times \frac{dJ(S)}{dS} \quad (8)$$

ここでのパラメータは $S = \{\mathbf{w}_A, \mathbf{w}_B, \mathbf{w}_C, \mathbf{w}_D, \mathbf{W}\}$ 単語表現は高次元かつ密なベクトルであることを前提とするので L2 正則化を用いた。本実験では、学習率 η を 0.1、分散を 0.01、イテレーション数を 150 回に設定した。学習率はイテレーション毎に指数的に減衰させる。

4 評価実験

本実験では単語表現を word2vec[1] で 300 次元に設定し事前に学習を行う。また、単語表現を学習するため

[1] https://code.google.com/p/word2vec/

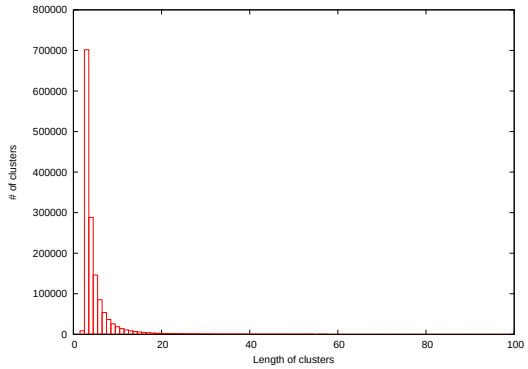

図 1: 得られた形式的類推クラスター数

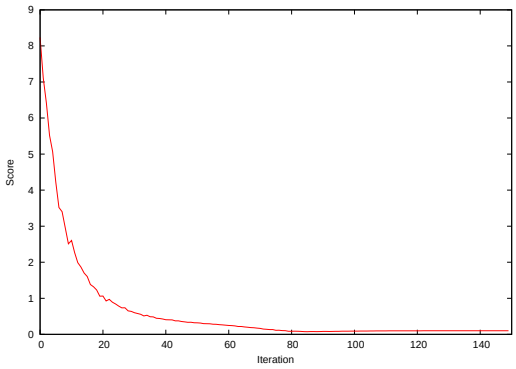

図 2: コスト関数の変化

のコーパスとして text8[2]を用いた。

4.1章で再学習に必要な形式的類推関係をコーパスから構築する方法を述べ、4.2章で類推推論タスクでの実験結果と考察を述べる。

4.1 再学習データ構築

再学習のための形式的類推関係を構築するには次の 4 つのステップから構成される。

- ランダムに単語をサンプリング
- 形式的類推クラスタリング
- 形式的類推クラスターから形式的類推関係を抽出
- 形式的類推関係から正例と負例の作成

形式的類推関係を構築するには $O(n^4)$ の計算量を必要とする。そのため、我々は形式的類推クラスタリングの手法を用いて、計算量を $O(n^2)$ を削減する [4]。得られた形式的類推関係のクラスターを用いて、そこから形式的類推関係を構築する。しかし、単語数 n は膨大であるため単語をランダムにサンプリングを行う。本実験ではサンプリング数を 1 万に設定した。得られた形式的類推クラスターの分布を図 1に示す。ここでの横軸は形式的類推クラスターに含まれる単語ペアのサイズである。図 1から形式的類推クラスターに含まれる単語ペアのサイズが 3 のとき最も多く得られた。得られたサイズ 3 での形式的類推クラスターは以下に示す。

stage : dispute	shop : studio
stages : disputes	shops : studios
staged : disputed	hopes : outside

[2] http://mattmahoney.net/dc/text8.zip

形式的類推クラスタリングでは計算量を削減のために近似処理を行うため形式類推クラスターに含まれる単語ペアは形式的類推関係であるとは言い切れない。そのため得られた形式的類推クラスターから3.1章での 3 つの制約を満たす形式的類推関係を抽出する。得られた形式的類推クラスターから形式的類推関係 3,687 を抽出できた。

提案手法では正例と負例の形式的類推関係を想定している。我々は正例の形式的類推関係を以下のように定義した。

$$A : B :: C : D \quad B : A :: D : C$$

次に負例の形式的類推関係を以下のように定義した。

$$B : A :: C : D \quad A : B :: D : C$$

負例の形式的類推関係はランダムにサンプリングする手法も考えられるが組み合わせ爆発が行うので本実験では用いない。我々は形式的と意味的類推関係は一種の平行四辺形のような関係で表現されると考えている。そのため、それぞれの設定する単語の位置もまた重要である事を考慮して正例と負例の形式的類推関係を定義した。正例と負例を含めた形式的類推関係の数は 14,748 である。

4.2 結果

第一に提案手法の単語再学習のコスト関数が減少しているかを調べる。本実験ではコスト関数のイテレーション数を 150 回に設定した。図 1に各イテレーション毎のコスト関数の変化を示す。

評価データとして word2vec で用意されている類推推論のベンチマークを用いて、それに含まれる 2 万の類推関係を評価する。評価するためのスクリプト

は wod2vec にある compute-accuracy のプログラムを用いるが類推関係の候補数の設定は全ての単語とする。word2vec で学習した単語表現のみ（baseline）と形式的類推関係を用いた単語再学習手法（ours）を各カテゴリーによる正解率で比較する。表 1に実験結果を示す。平均した解の正解率は微小ながら提案手法が上回った。

表 1: ベースラインと提案手法とのベンチマークによる比較

Category	baseline	ours
capital-common-countries	74.1	**74.3**
capital-world	44.0	44.0
currency	20.5	20.5
city-in-state	42.5	42.5
family	**58.6**	58.3
adjective-to-adverb	11.0	**11.1**
opposite	16.0	16.0
comparative	64.3	**64.6**
superlative	**22.4**	22.2
present-participle	34.5	**34.8**
nationality-adjective	77.5	**77.6**
past-tense	37.6	37.6
plural	53.3	**53.6**
plural-verbs	31.7	**32.2**
Avg	42.0	**42.1**

baseline より上回ったカテゴリーの comparative や plural などでは形式的類推関係の要素が強いので他のカテゴリーよりも上昇する割合が大きくなった。例えば、comparative でのカテゴリーでは bright : brighter :: fast : faster のような形式的類推関係が多く存在する。また、plural のカテゴリーでは banana : bananas :: bird : birds のような形式的類推関係が多く存在する。しかし、superlative のカテゴリーでは cold : coldest :: cool : coolest のように形式的類推関係が存在するにも関わらず精度は減少した。再学習データでは superlative での形式的類推関係があまり存在しないために減少したと考えられる。

これらの結果から、事前に構築した形式的類推関係に似ている関係が含まれていたので再学習された単語表現に反映された考えられる。また、提案手法でのコスト関数で特徴ベクトルを考慮することで形式的類推関係に影響を及ぼす次元の要素を知る事ができた。

5 おわりに

本論文では形式的類推関係を用いた単語表現の再学習手法を提案した。再学習のために必要となる形式的類推関係を構築するために、形式的類推クラスタリングを用いて計算量の削減を行い、得られた形式的類推クラスターから正例と負例の形式的類推関係を抽出した。本実験では形式的類推関係に必要な次元の要素と類推推論のベンチマークでの精度は微小ながら向上を確認できた。

将来課題として形式的類推関係を構築する際にランダムなサンプリングを行ったが全データから意味的類推関係を構築する。また、学習された単語表現が類推推論のタスク以外での有用性を示す。

参考文献

[1] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. A neural probabilistic language model. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

[2] Zellig S. Harris. Distributional structure. *Word*, 10:146–162, 1954.

[3] Yves Lepage. Languages of analogical strings. In *Proceedings of 18th COLING*, volume 1, pages 488–494, Stroudsburg, PA, USA, 2000.

[4] Yves Lepage. Analogies between binary images: Application to Chinese characters. In *Computational Approaches to Analogical Reasoning: Current Trends*, pages 25–57. Springer, 2014.

[5] Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and J Cernocky. RNNLM - recurrent neural network language modeling toolkit. In *Proceedings of Automatic Speech Recognition and Understanding Workshop*, pages 196–201, 2011.

[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[7] Peter D. Turney. Measuring semantic similarity by latent relational analysis. In *Proceedings of IJCAI*, pages 1136–1141, 2005.

[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, pages 1015–1018, 3 月 2014.