Analogy comprehension between psychological experiments and word embedding models

浅川 伸一 *1　岡 龍之介 *2　楠見 孝 *3
Shin Asakawa　Ryunosuke Oka　Takashi Kusumi

*1東京女子大学
Tokyo womens' Christian university

*2京都大学教育学研究科
Kyoto University, School of Education

*3京都大学教育学部
Kyoto University

Word embedding models such as word2vec (Mikolov,2013) and GloVe (Pennington et al.2014) became to be widely recognized, while similarity judgment of analogy and comprehension could be described on a Euclidian space. We tried to compare between word embedding models and human performances. Despite that several models were proposed to deal with adjust each domain in order to improve their performance, it has still remained unsolved the relation between human judgment and understanding of analogy and their performance of word embedding models. We investivated the relationship between them. Our results suggest that human understanding of analogy might be understood in terms of word embedding spaces. We disscussed the further possibilities to understannd the semantic space in human in these models.

1. はじめに

語義の分散仮説は言語の基本課題の一つである。たとえば文献によれば "words that occur in the same contexts tend to have similar meanings" [Harris 54] "a word is characterized by the company it keeps" [Firth 57] などの記述が見られる。単語埋め込みモデルは自動翻訳，推薦，チャットボット，極性分析，など広範な言語課題で用いられている。Mikolov ら [Mikolov 13] は $\mathbb{R}^d$ に単語を埋め込むことで，テキストコーパス内の類似の文脈内で単語が共起するときにその内積が最大になるように学習を行った。分布仮説 [Harris 54, Firth 57] と呼ばれる古典的な概念を実装したとみなすことが可能である。

典型的には，埋め込み方法の目的は，埋め込み空間におけるそれらの類似性がその意味または機能を反映するような方法で記号オブジェクト（単語，物体，対象，概念）を編成することである。この目的のため，対象物の類似性は，通常それらの距離または埋め込み空間内の内積 (cosine 類似度) によって測定される。

Mikolov らが単語埋め込みモデルの性能評価に用いた意味類推課題の例を挙げれば，たとえば $\mathrm{vec}(\text{ロンドン}, \text{イギリス}) \simeq \mathrm{vec}(\text{バグダッド}, \text{イラク})$ のごとくである。a:ロンドン, a^*:イギリス, b:バグダッド, b^*:イラク (この場合の正解語) とすれば，以下のように定式化できる。Mikolov らは cosine 類似度 ($\mathrm{Sim}_{\cos}(\mu,\nu) = \cos(\mu,\nu) / (|\mu||\nu|)$) を用いた。これに対し Levy & Goldberg[Levy 14] は以下のような派生手法を提案した。

PAIRDIRECTION:

$$\arg\max_{b^* \in V} (\cos(b^*, b - a + a^*)) \tag{1}$$

3COSADD:

$$\arg\max_{b^* \in V} (\cos(b^* - b, a^* - a)) \tag{2}$$

3COSMUL:

$$\arg\max_{b^* \in V} \left(\frac{\cos(b^*, b)\cos(b^*, a^*)}{\cos(b^*, a) + \epsilon} \right) \tag{3}$$

Contact: Shin Asakawa, Tokyo women's Christian university, Zempukuji 2-chome, 6, 1, Suginami, 1678585, Tokyo, asakawa@ieee.org

一方，Che ら [Che 17] は Euclid 距離を用いることを提案した。

EUC:

$$\arg\max_{b^* \in V} \left(1 - \frac{|(a^* - a) - (b^* - b)|}{|a^* - a| + |b^* - b|} \right) \tag{4}$$

Che らは更に a, a^*, b, b^* の推論には 3 つの別問題を考えることが可能であり，cosine 類似に基づく推論が成り立たない場合があることを示した。*knowing:knew::selling:sold* の場合を表 1 に示した。

表 1: knowing:knew::selling:sold の推論課題結果 [Che 17]Tab.1 より

ターゲット語	knowing		sold	
予測語	thought	0.573	**sold**	**0.568**
	know	0.504	sell	0.535
	wanted	0.494	bought	0.528
	konwing	**0.489**	buying	0.486

表 1 中**太字**は正解語を示している。このように同一単語で推論方向を変化させると正解に至らない場合が存在する。

認知心理学の文献では，心理実験結果に基づく単語の意味空間は Euclid 的であることが示唆されている (図 1)。

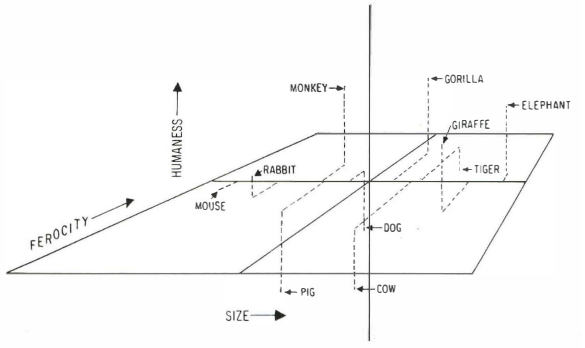

図 1: 心理実験による動物概念の布置 [Rumelhart 73] より

Rumelhart & Abrahamson は意味空間内の単語意味空間 (図 1) で人間の意味判断が Luce の選択公理 [Luce 59] に従うと仮定すれば心理実験結果を説明可能であることを示した。類推課題において単語 b^* が選択される確率は以下の式 5 に従う:

$$Pr\left(b^* \mid a, a^*, b\right) = \frac{v\left(d_i\right)}{\sum_j v\left(d_j\right)}, \tag{5}$$

ここで d_i は各単語間の距離である。意味判断過程に距離の応じた関数 $d \sim \exp\left(-\alpha \mathrm{Sim}\left(a, b\right)\right)$ を考えれば，式 5 はソフトマックス関数に等しくなる。Rumelhart & Abrahamson は式 5 が心理実験結果を説明可能であるとした。加えて Sternberg & Gardner[Sternberg 83] は，異なる単語類推課題である，系列補間課題，分類課題を提案した (図 2(b) および (c))。Sternberg & Gardner 課題を考慮に入れるとすれば，単語埋め込みモデルの評価に際して用いられるテスト課題に基いて性能を評価することの意味は再考する必要があるとも解釈できよう。

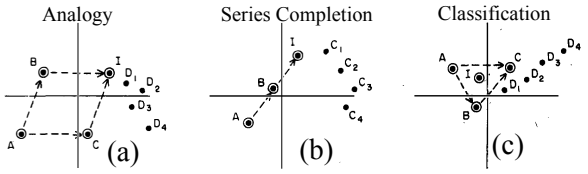

図 2: Stenberg [Sternberg 83]Fig.1 より

このような事実から，単語埋め込みモデルにおける推論について検討する必要があると考える。

2. 実験

2.1 中本・楠見の日本語の比喩理解

中本・楠見 [中本 04] の心理実験による日本語の比喩理解結果 (以下 NK04 と略記する) データと単語埋め込みモデルとを比較した。NK04 は 120 語対からなる日本語の比喩表現の心理評価データである。例えば「笑顔は花のようだ」のような比喩表現が用いられた。評定は，理解可能性，構成語類似性，独創性，面白さの 4 項目であり，9 段階評定である。単語埋め込みモデルとしては，2017 年 7 月時点の日本語ウィキペディア [*1] を mecab + NEologd[*2] によって分かち書きし，word2vec[*3] により単語埋め込みベクトル化した [*4]。Skip-gram を使用し，ベクトル化した際のパラメータは，埋め込みベクトル次元:200，ウィンドウ幅:20，負例サンプリング:20 とした。出現頻度 5 回以上の単語のみを考慮することとし，総語彙数 180,543 語を得た。NK04 に用いられる単語で「水晶細工」の 1 単語だけ単語埋め込みモデルの訓練に用いた日本語ウィキペディアの語群に該当単語が存在しなかった。そのため以下では「水晶細工」を含む比喩文データを削除した全 119 単語対について結果を示した。

NK04 による単語対の関係を図 3, 4 に示した。図中のシンボルは原論文でのクラスタを示している。図 3 は横軸が cosine 距離であり，縦軸が Euclid 距離である。図中には双方への回帰直線も示した。

*1 https://dumps.wikimedia.org/jawiki/latest/
*2 https://github.com/neologd/mecab-ipadic-neologd
*3 https://code.google.com/archive/p/word2vec/
*4 https://github.com/ShinAsakawa/2017jpa

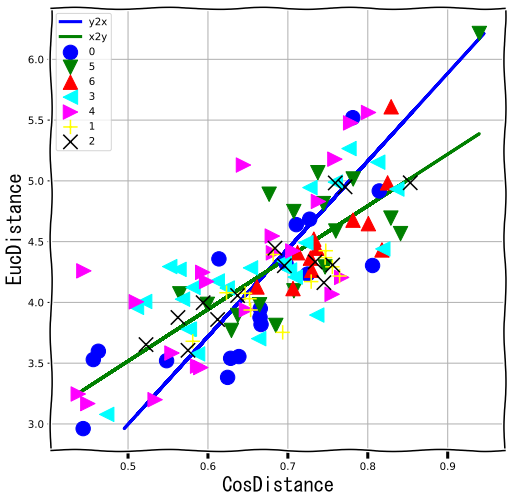

図 3: NK04 と cosine(x) and Euclid(y) distance

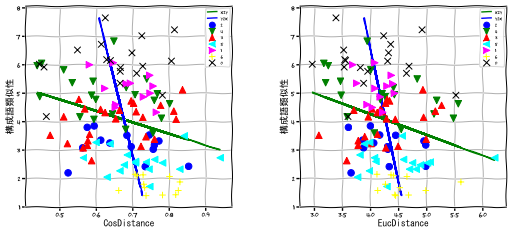

図 4: NK04:comparison human comprehension

図 4 は心理実験データが類似度評定であり単語埋め込みモデルでは距離を算出したため回帰直線も右下がりの傾向を示している。全体的な傾向は人間の判断と一致しているものの，単語埋め込みモデルと人間の類似性判断には尚乖離があることが推察される。

NK04 による心理実験結果と単語埋め込みモデルによる cosine 距離，及び Euclid 距離の相関係数を表 2 に示した。先の図 3 と図 4 とを考慮すれば，相関係数間の差異のみに基いて優劣を判断するのは難しいと考えられる。

表 2: NK04 の類似性評価と word2vec(wikipedia.ja) 間の相関係数

	NK04	cosine
cosine	-0.276	
Euclid	-0.279	0.768

日本語ウィキペディアは，英語版ウィキペディアと比較すると総量で 10GB 程度の差があり既存の単語埋め込みモデルが十分な性能を示すことができないという考え方もありうる。そこで次節以降では，英語版ウィキペディアを用いて人間の類似性判断との相違を検討した。

2.2 SimLex999

SimLex999[Hill 14][*5] は 999 単語対からなるデータセットである。NK04 と同様の手続きにより SimLex の 999 単語対を人間が類似性判断したデータセットを単語埋め込みモデルを用いて予測することを試みた。単語埋め込みベクトルには Google の公開している訓練済の単語埋め込みモデル GoogleNews-vectors-negative300.bin.gz[*6] を用いた。Google 版の単語埋め込みモデルは語彙数 3,000,000 であった。

NK04 の図 3 に対応する図が 5 であり，同じく図 4 のそれへの対応が図 5 である。NK04 の相関係数表 2 に対応する表が表 3 である。

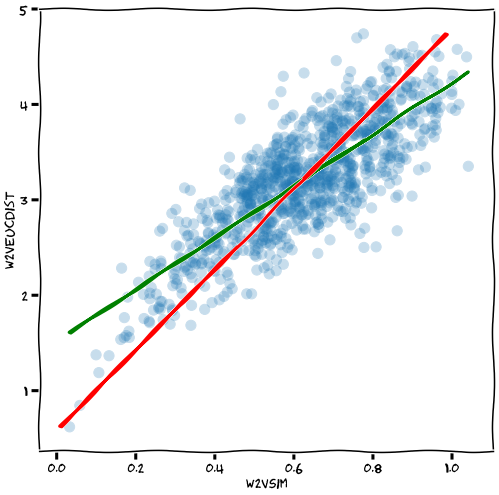

図 5: SimLex:cosine vs Euclid

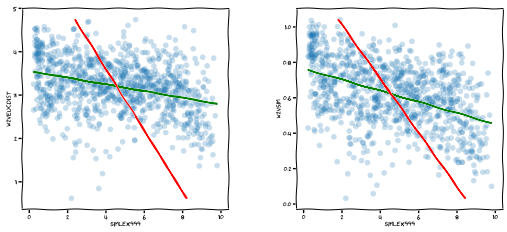

図 6: SimLex:comparison human comprehension

表 3: SimLex999 correlations

	SimLex	cosine
cosine	-0.453	
Euclid	-0.330	0.802

NK04 の結果と比較すれば，cosine 距離，Euclid 距離両者とも人間の類似性判断との相関係数が大きく，かつ，cosine, Euclid 距離測度間の相関係数も絶対値が大きくなった。データセットの語彙対が 999/119 と 8 倍以上，ウィキペディアの語彙数では 3,000,000/1,805,463 と 1.6 倍であった。評価データ数と訓練データ数と差異が精度向上に関与する可能性は考えられるものの，図 5 からは cosine 距離と Euclid 距離との乖離は保たれていた。このため人間が行う類似性判断との乖離は埋められずに残されていると見做すことが可能であろう。

2.3 MEN

MEN[*7] は人間による類似性判断データセットである [Bruni 12]。Amazon Mechanical Turk (MTurk)[*8] を通じて収集された。データ数は 3000 語対である。用いた単語埋め込みモデルは SimLex999 と同じく Google 版の訓練済データを用いることとした。先と同様の手法により図 7, 8, 表 4 を作成した。

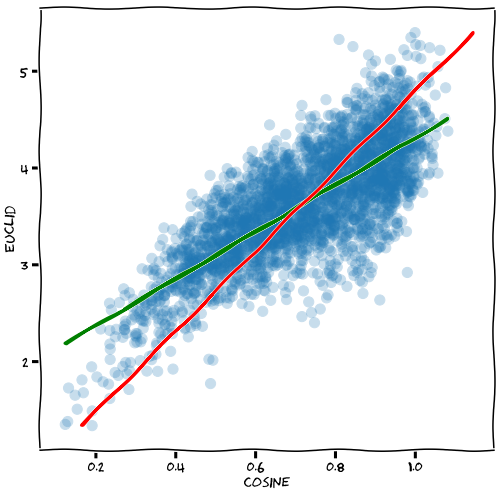

図 7: MEN:cosine vs Euclid

cosine, Euclid 距離間の相関係数は SimLex ほど高くないが，両距離とも人間の類似度評定との関連がか高くなった，単語埋め込みモデルは SimLex999 と MEN とで共通であるから，相違が生じたとすれば，データセットの持つ性質である可能性が考えられる。しかし，本データだけでは得られた結果の相違を十分に説明する証拠とは断言できないように思われる。

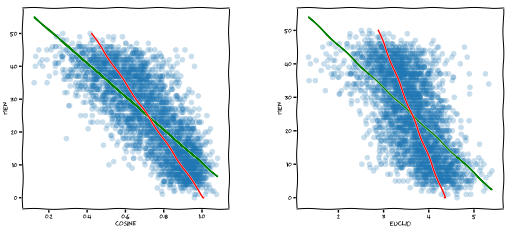

図 8: MEN:comparison human comprehension

*5 https://www.cl.cam.ac.uk/~fh295/simlex.html
*6 https://github.com/mmihaltz/word2vec-GoogleNews-vectors
*7 http://clic.cimec.unitn.it/~elia.bruni/MEN
*8 https://www.mturk.com/

表 4: MEN correlations

	MEN	cosine
cosine	-0.770	
Euclid	-0.612	0.766

3. 考察

人間による単語の類似性判断を説明するため，単語埋め込みモデルが用いることが可能か否かを検討するために 3 つのデータセットを用いて評価した。得られたデータの差異は，言語間の差異に帰すものであるか，訓練データ量であるのか，データセットの持つ性質によるのかを本研究だけから判断することはできない。しかし，単語埋め込みモデルが自動翻訳，文章理解，生成など自然言語処理技術に活用される基礎技術となった今日，単語埋め込みモデルを活用しない方が不自然であるとさえ感じられる。

本研究の目的の一つは，日本語比喩理解のための心理学的モデルとしての単語埋め込みモデルの可能性を検討することであった。比喩理解のためには Kintsch[Kintsch 01] の古典的モデルでも単語表象に対して，意味空間内での隣接語の相互作用が仮定されている。このことは可塑的フィッテイングと見做すことも可能であるが，一方で入力層の one-hot vector を単語埋め込みモデルを用いて次元圧縮し，上位層であるリカレントニューラルネットワーク層への入力とする意味では，言語モデルであれ翻訳モデルであれ，本研究で用いた比喩理解と同様に，入力表現を変形させることが言語理解過程であると見做すことも可能であろう。

4. まとめと展望

人間の意味理解，語彙の類似性判断と単語埋め込みモデルの関連を検討した。示した結果は初歩的な結果の域を出ていないが，人間が行っている文章理解の解明とその実装へ向けて意味モデルの果たす役割は大きいと考える。人間の示す意味理解を深く理解しなければ，実用的な知的システムを構築することはできない。例えば，mixup[Zhang 17] のようなデータ拡張を自然言語処理でも行うための基礎技術としても単語埋め込みモデルを精緻化していく必要があると考えられる。

文献

[Bruni 12] Bruni, E., Boleda, G., Baroni, M., and Tran, N.-K.: Distributional Semantics in Technicolor, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pp. 136–145, Jeju, Republic of Korea (2012)

[Che 17] Che, X., Ring, N., Raschkowski, W., Yang, H., and Meinel, C.: Traversal-Fre eWord Vector Evaluation in Analogy Space, in Hwa, R. and Riedel, S. eds., *Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark (2017)

[Firth 57] Firth, J. R.: *A synopsis of linguistic theory 1930-55.*, Vol. 1952-59, The Philological Society, Oxford (1957)

[Harris 54] Harris, Z. S.: Distributional Structure, *Word*, Vol. 10, No. 2-3, pp. 146–162 (1954)

[Hill 14] Hill, F., Reichart, R., and Korhonen, A.: SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation, *arXiv preprint* (2014)

[中本 04] 中本 敬子, 楠見 孝比喩材料文の心理的特性と分類—基準表作成の試み—, 読書科学, Vol. 43, No. 1, pp. 1–10 (2004)

[Kintsch 01] Kintsch, W.: Predication, *Cognitive Science*, Vol. 25, pp. 173–202 (2001)

[Levy 14] Levy, O. and Goldberg, Y.: Linguistic Regularities in Sparse and Explicit Word Representations, in *Proceedings of the Eighteeth Conference on Computational Language Learning*, pp. 171–180, Baltimore, Maryland, USA (2014)

[Luce 59] Luce, R. D.: *Individual Choice Behavior*, Wiley, New York, USA (1959)

[Mikolov 13] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, in Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. eds., *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, Curran Associates, Inc. (2013)

[Rumelhart 73] Rumelhart, D. E. and Abrahamson, A. A.: A Model for Analogical Reasoning, *Cognitive Psychology*, Vol. 5, pp. 1–28 (1973)

[Sternberg 83] Sternberg, R. J. and Gardner, M. K.: Unities in Inductive Reasoning, *Journal of Experimental Psychology: General*, Vol. 112, No. 1, pp. 80–116 (1983)

[Zhang 17] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D.: *mixup*: Beyond Empirical Risk Minimization, *arXiv preprint* (2017)