word2vecによる類義語抽出とFrameNetの比較：言語研究のための質的検証

内田, 諭
九州大学

出版情報：言語統計を用いた認知言語学研究へのアプローチ, pp.41-51, 2018-03. 統計数理研究所
バージョン：
権利関係：

word2vec による類義語抽出と FrameNet の比較

―言語研究のための質的検証―

A Comparison of word2vec and FrameNet: A Qualitative Analysis for Linguistic Studies

内田　諭
九州大学

Satoru UCHIDA
Kyushu University

1. はじめに

コンピューターおよびインターネットの進化は、言語研究に対して大きなインパクトを与えている。高速な処理が可能になったことで多変量解析やニューラルネットワークなどの計算量が多いものも手軽に扱うことができるようになった。また、オンラインのリソースに手軽にアクセスできるため、大規模なデータ収集やそれを使った研究も個人ベースでも行うことが可能となった。このような技術的な進歩を背景に、言語処理の分野では単語の意味をベクトルとして表現する word embedding の手法が開発され、word2vec（Mikolov et al., 2013）や fastText（Joulin et al., 2016）などのアプリケーションが利用可能となっている。word embedding の理論的背景には分布仮説（Harris, 1954）があり、意味的に類似する語は周辺の単語の分布も似ていると仮定される。この仮説に基づくと、特定の単語の周辺分布の類似度から、類義語を抽出することが可能となる。この技術は、文書分類、情報検索、質疑応答システムなどで利用され、その性能が向上することが報告されている（Tang et al., 2014 など）。word embedding によって計算された類似度がどの程度有効かという評価の問題に関して議論をする研究（cf. Schnabel et al., 2015）もあるが、言語研究における有用性を検証したものは少ない。

本研究の目的は、word embedding を利用した類義語の抽出結果と言語学的な研究成果の一致性を検証し、言語研究における有用性を検証することである。具体的にはフレー

ム意味論に基づいた辞書である FrameNet の情報と、大規模コーパスに基づいて作成された word2vec のモデルから抽出した類義語のリストを比較し、その一致度および傾向について議論する。また、その結果から言語研究に word embedding を用いる場合に最適なコーパスサイズ等を見積もるための実験を行う。

2. データセット

2.1 コーパスデータ

word embedding を利用して単語のベクトル空間を作成するには一定規模のコーパスが必要となる。本研究では言語研究に広く用いられている Corpus of Contemporary American English[1]（以降 COCA）の全文データを用いることとした。総語数は約 4 億 4000 万語[2]で、academic, fiction, magazine, news, spoken の 5 つのジャンルから構成されている。また、言語研究における最適なコーパスサイズを探るため、表 1 のように COCA のデータを分割してサブコーパスを作成した。

表 1：サブコーパスの概要

コーパス名	語数	概要
COCA_sample	1.7 million	全文データの無料サンプル。ランダム抽出
COCA_2million	2 million	1990 年の academic の前半
COCA_4million	4 million	1990 年の academic
COCA_8million	8 million	1990 年の academic, fiction
COCA_12million	12 million	1990 年の academic, fiction, magazine
COCA_16million	16 million	1990 年の academic, fiction, magazine, news
COCA_20million	20 million	1990 年のテキストデータすべて
COCA_40million	40 million	1990 年＋1991 年
COCA_60million	60 million	1990 年＋1991 年＋1992 年
COCA_80million	80 million	1990 年＋1991 年＋1992 年＋1993 年
COCA_100million	100 million	1990 年＋1991 年＋1992 年＋1993 年＋1994 年
COCA_all	440 million	COCA 全データ

[1] https://corpus.byu.edu/coca/

[2] 2017 年 12 月に COCA のアップデートが行われ、2016 年および 2017 年のデータが追加されたが、本研究のデータには含めない。

2.2 FrameNet

FrameNet はフレーム意味論（Fillmore, 1982, 1985 など）に基づいて作成されたフレームの辞書である（cf. Ruppenhofer et al., 2016）。フレームには認知フレーム（cognitive frame）と言語フレーム（linguistic frame）があり、前者は「理由付け」、「予期」、「想起」などに使用される百科事典的な知識を指すのに対し、後者はフレームを喚起する語が意味的・構文的に共起する要素を指す（cf. Fillmore, 2008, 内田, 2015）。FrameNet は言語フレームを記録したもので、BNC（British National Corpus）などの実例を基に、単語が喚起するフレームと、フレームに伴う意味要素（フレーム要素）およびフレーム要素の実現形などの情報が含まれている。例えば、increase (v)は、Cause_change_of_position_on_a_scale, Change_position_on_a_scale, Biological_mechanisms の 3 つのフレームを喚起すると記述されている。Cause_change_of_position_on_a_scale フレームの記述をみると、このフレームが必須的に伴う要素[3]として、増加や減少を行う主体である Agent や増減する数的要素である Attribute、増減の原因となる Cause などがあることがわかる。また、(1)などの例から Agent は文の主語、Attribute は目的語として出現することがわかるが、その他にどのような実現形のパターンがあるかについても FrameNet では詳細な報告がされている。

(1) [<Agent>We] have <u>increased</u> [<Attribute>the number of businesses in Wales], VAT registrations are up by 25 percent, production industries are up by 60 percent. (FN[4])

さらに、当該のフレームを喚起する語のリストが示されており、その情報を利用することでシソーラス的な使い方が可能となる。Cause_change_of_position_on_a_scale では growth (n), increase (v), lift (v), lower (v), reduce (v), reduction (n)などが記載されている。つまり、これらの単語は言語フレームを介してリンクしており、意味的・構文的に非常に近いものであるいえる。

FrameNet における単語の関連付けの特徴として、反意語を含むことと、異なる品詞を同一フレーム内に含むことが挙げられる。反意語は、意味的には対極の内容を表すが、構造的には共通する部分が多い。例えば、(2)は increase の反意語の reduce を含んでいるが、フレーム要素および構文は(1)と類似していることがわかる。

---

[3] コアフレーム要素と呼ばれ、フレームを弁別する重要な基準の一つである。
[4] FrameNet からの引用であることを表す。

(2) [<Agent>MMT Computing Plc] reduced [<Attribute>its stake in Total Systems Plc] [<Value_2>to under 3%]. (FN)

また、異なる品詞を含むことについて、例えば(2)は[<Agent>MMT computing Plc's] reduction on [<Attribute>…]などのように、同様の内容を名詞構文で表すことができることなどから明らかなように、フレームが意味的に伴う要素の観点からは動詞の場合と共通であることがわかる[5]。

これらの FrameNet の特徴は、word embedding のベンチマークとして都合がよい。前述の通り word embedding の根底には分布仮説があり、意味的に類似する要素が文法的にも類似した形式で出現することはベクトル空間における単語間の距離を縮めることになる。word embedding を利用した類義語の抽出では、しばしば反意語が含まれることが問題視されるが、フレーム意味論の立場からするとこの現象は正当化できるものであり、単語の出現環境を正しく反映した結果であるといえる（ただし、word embedding の実用面での問題は別に議論する必要がある）。また、コーパスの生データでモデルを生成する場合は品詞の区別がないため、文脈が十分に類似していれば、異なる品詞の単語が類義語として抽出される可能性がある。しかしながら、これも同様にフレーム意味論の立場からは意味的に類似する要素が対象語に付随することは合理的であると解釈できるため、妥当性が高いといえる[6]。Chiu et al. (2016)は既存のデータセットが word embedding の結果の評価に適当ではないものが多いことを指摘しており、その原因の一つとして computer, keyboard などの状況的な関連性による単語関係を含むことを指摘しているが、FrameNet ではそのようないわば認知フレーム的なつながりの語は含んでおらず言語フレームに特化しているため、純度の高いデータセットになっているといえる。

3. 評価手法

本研究では word embedding の代表的なアプリケーションである word2vec を使用して各サブコーパスのベクトルモデルから類義語を抽出する。Python の gensim ライブラリを利用し、学習のパラメーターは window=5, size=200, iter=5, min-count=5 とし、学習モデ

[5] 構文的には支援動詞を伴うなど、名詞にユニークな特徴もある。詳細は Ruppenhofer et al. (2016)、内田 (2015)などを参照。
[6] 品詞を考慮して単語の埋め込みを行うことも可能だが、POS タガーの性能に結果が左右される可能性があるため、本研究では行わないこととした。

ルにはCBOWを用いた。それぞれのコーパスについて作成したモデルから、most_similarメソッドを使ってコサイン類似度順に類義語を抽出した。例えば、climb に対してはascend (0.741), jump (0.718), climbing (0.704), clamber (0.703)などがリストされる（括弧内はコサイン係数を表し、値が高いほど類似していることを示す）。

これらのリストに対して、FrameNetの見出し語と一致させるためにelemma[7]を用いて抽出した類義語のレマ化を行った。その際、climbing, climbed などの屈折形についてはclimbにレマ化され、ターゲット語と同一になるためEQUALとコード化し、一致としてカウントした。表2はその例を示したものである。

表2：word2vecとFrameNetの対応付けの例

corpus	target	rank	w2v	lemma	cosine	frame
COCA_all	climb	1	ascend	ascend	0.741	Intentional_traversing
COCA_all	climb	2	jump	jump	0.718	Change_position_on_a_scale
COCA_all	climb	3	climbing	climb	0.704	EQUAL
COCA_all	climb	4	clamber	clamber	0.703	Self_motion
COCA_all	climb	5	crawl	crawl	0.690	Self_motion
COCA_all	climb	6	descend	descend	0.687	#N/A

climbはChange_position_on_a_scale, Intentional_traversing, Self_motionの3つのフレームを喚起する。したがって、これらのいずれかのフレームを喚起する語がリストされていれば、一致とカウントする。例えば、Intentional_traversingフレームにはclimb (v), ascend (v), cut (v), ford (v)が含まれるが、表2の類似度ランク1位のascendはこれと一致するため（つまりclimbもascendもIntentional_traversingフレームを喚起するため）、word2vecの結果とFrameNetが一致していると考えることができる。

本研究では実験の対象として、COCAの頻度ランク3000位以内、FrameNetで関連語が20以上存在することを条件として10単語をランダムに抽出した。その結果、consider, carry, climb, discovery, increase, poor, response, reveal, serious, wonderfulが対象となり、類似度ランク上位20位までを集計の対象としてFrameNetとの一致度を5位以内、10位内、15位以内、20位内の4段階で評価した。

[7] http://www.laurenceanthony.net/software/antconc/よりダウンロード（リストの作成は染谷泰正氏による）。

4. 結果と考察

4.1 実験結果

図 1 および表 3 に実験の結果を示す。図 1 からコーパスサイズが大きくなるにつれて一致度が高くなることが見て取れる。小規模でサイズが類似している COCA_sample（170 万語）と COCA_2milion（200 万語）を比較すると、ランダムサンプリングを行っている COCA_sample のほうがわずかではあるが結果がよいことがわかる。また、COCA_20million（2000 万語）で一度グラフは下降しているが、COCA_40million（4000 万語）では値が大きく向上している。このことから、コーパスサイズが 4000 万語を超えると、一定の結果が得られることが示唆される。さらに COCA_80million と COCA_100million には大きな差はなく、サイズが 4 倍以上である COCA_all でも平均値は大きな上昇を示さない。この結果はコーパスサイズが 8000 万語～1 億語で十分な性能が得られることが示唆される。@5, @10, @15, @20 では@5 が最も結果がよい。このことから word2vec の上位 5 語は特に信頼性が高いということを示している。

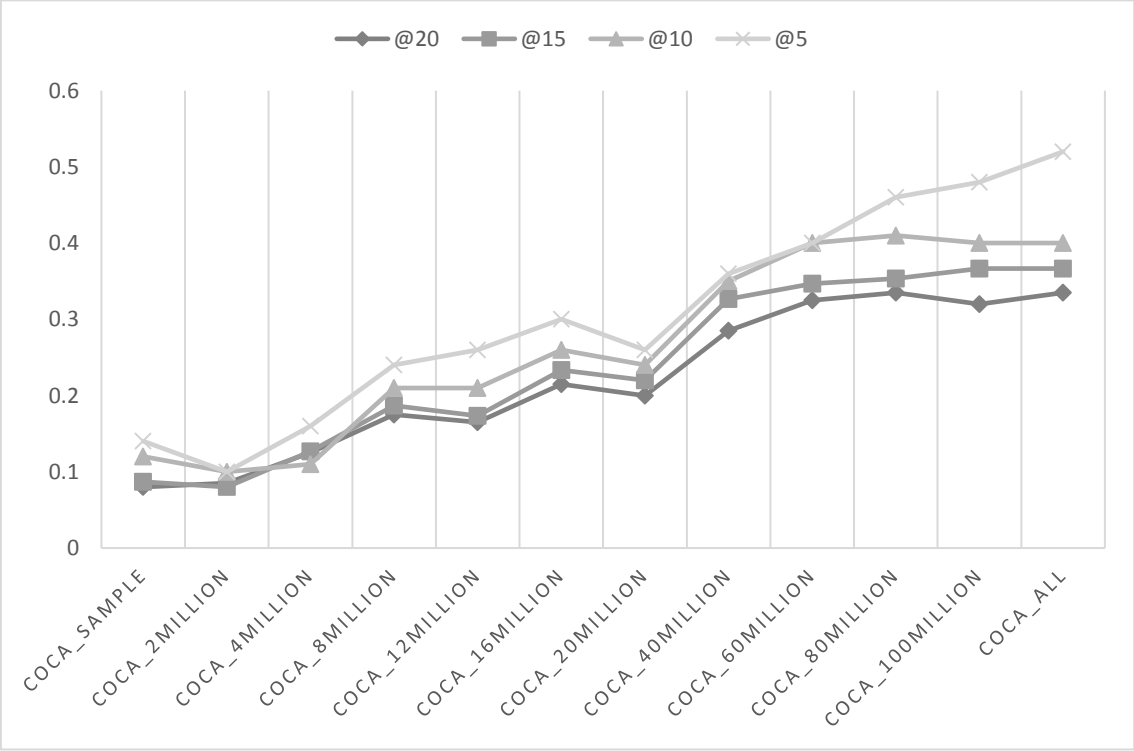

図 1 ：各コーパスの結果と FrameNet の一致度

表 3：各コーパスの結果と FrameNet の一致度の詳細および平均

	@20	@15	@10	@5	average
COCA_sample	0.080	0.087	0.120	0.140	0.107
COCA_2million	0.085	0.080	0.100	0.100	0.091
COCA_4million	0.125	0.127	0.110	0.160	0.130
COCA_8million	0.175	0.187	0.210	0.240	0.203
COCA_12million	0.165	0.173	0.210	0.260	0.202
COCA_16million	0.215	0.233	0.260	0.300	0.252
COCA_20million	0.200	0.220	0.240	0.260	0.230
COCA_40million	0.285	0.327	0.350	0.360	0.330
COCA_60million	0.325	0.347	0.400	0.400	0.368
COCA_80million	0.335	0.353	**0.410**	0.460	0.390
COCA_100million	0.320	**0.367**	0.400	0.480	0.392
COCA_all	**0.335**	**0.367**	0.400	**0.520**	**0.405**
average	0.220	0.239	0.268	**0.307**	0.258

表 4：単語別の一致度

target	@20	@15	@10	@5	average
carry	0.180	0.187	**0.210**	0.160	0.184
climb	0.555	0.567	0.590	**0.660**	0.593
consider	**0.130**	0.107	0.080	0.020	0.084
discovery	0.080	0.093	0.140	**0.260**	0.143
increase	0.650	0.720	0.790	**0.880**	0.760
poor	0.275	0.313	0.320	**0.380**	0.322
response	0.195	0.260	0.380	**0.620**	0.364
reveal	0.240	0.280	**0.340**	**0.340**	0.300
serious	0.100	0.113	0.130	**0.140**	0.121
wonderful	**0.240**	0.227	0.230	0.220	0.229

4.2 考察

前節の結果から 1 億語前後のモデルを利用すると上位 5 語に関しては 5 割近くの単語が FrameNet と一致することが明らかになった。ここでは単語によって傾向に違いがあるかを確かめるため、対象語それぞれについての一致度を検証する。表 4 は各単語の上位 5 位以内、10 位以内、15 位以内、20 位以内における一致度を示したものである。この結果から、単語によって結果に大きな差があることが明らかとなった。特に increase と climb の一致度が高い。これはこれらの単語が喚起するフレームの構文の定形性と特殊性の高さが関与していると考えられる。increase, climb はどちらも数量の変化を表すフレーム（Change_of_position_on_a_scale）を喚起するが、このフレームは(3)の例に見られるように from や to などを定形的に伴うことが多い。

(3) [<Attribute>The rate] increased [<Final_value>to 7.4%], [<Initial_value>from 7.0% in February]. (FN)

このような用法はこのフレームに特徴的なものであるといえる。一方でこのような構文的な特異性が少ないと考えられる serious や wonderful などの形容詞は一致度が低くなっている。

しかしながら、consider は「consider A as B」などのような構文的な特徴があると考えられるが、一致度が非常に低くなっている。COCA_all でも上位 20 位までで一致したのは 3 件（contemplate （Cogitation フレーム）、perceive（Categorization フレーム）、define（Categorization フレーム））のみである。類似度の上位 5 件は recommend, propose, recognize, anticipate, suggest で、これらとは that 節をとるという点で共通している。また、思考や提案、予期などの心理的な内容を表すものである。FrameNet の構築は進行中のため、これらの動詞が将来的に consider と関連付けられる可能性もあるが、このように汎用的な構文をとる動詞については一致度が低くなる傾向にあることが示唆される。

これらの結果を利用して word2vec を用いた FrameNet の拡張支援の可能性も考えられる。FrameNet のエントリーの作成はフレームを出発点とし、ネイティブスピーカーの直感とコーパスデータに拠って進められているため、word2vec の結果が理論的な出発点にはならないが、特定のフレームを喚起する語を探索する補助として機能する可能性がある。特に一致度が高い単語については不一致と判定されたものであっても慎重に検討する価値がある。表 5 は本研究の実験で最も一致度の高かった increase の類義語上位 10 語を示したものである。

表 5：increase の類似度上位 10 語

rank	token	lemma	cosine	frame
1	decrease	decrease	0.888	Cause_change_of_position_on_a_scale
2	increases	increase	0.757	EQUAL
3	reduce	reduce	0.737	Cause_change_of_position_on_a_scale
4	reduction	reduction	0.720	Cause_change_of_position_on_a_scale
5	increased	increase	0.706	EQUAL
6	boost	boost	0.706	#N/A
7	decreases	decrease	0.676	Cause_change_of_position_on_a_scale
8	decline	decline	0.673	Change_position_on_a_scale
9	increasing	increase	0.659	EQUAL
10	uptick	uptick	0.636	#N/A

この結果で不一致（FrameNet にエントリーなし）として判定されたのは boost と uptick であるが、次の例が示すように、Change_position_on_a_scale フレームの用例と非常に似ている用法がある。

(4) That total represented a 9% uptick from the season 5 premiere (9.8 million). (COCA)
(5) For instance all the Big 10 trusts could (if they so chose) fund a 5% dividend increase from reserves for three years … (FN)
(6) A confidential French Senate report recommends that the special forces' strength be boosted by 25 percent, from 3,000 to 4,000. (COCA)
(7) Employment and training programmes will increase by 500,000 places next year, a rise of 50 percent. (FN)

(4)と(5)はともに数量の増減を表す内容で from の前置詞句要素は Change_position_on_a_scale フレームの Initial_value 要素であると見なすことができる。つまり、uptick と increase (n)は同一のフレーム（Change_position_on_a_scale）を喚起していると考えることができるということである。また、(6)と(7)の by の前置詞句は同フレームの Difference 要素であると認定することができる。なお、いずれの場合もフレームの骨格を構成するコアフレーム要素であり、この事実は同一フレームを喚起することの証拠となる。

5. まとめ

本論文では COCA の全文データおよびそれを分割して作成したサブコーパスを用いて word2vec のモデルを生成し、コサイン類似度に基づいて抽出した類義語のリストを、言語理論に基づいて作成された辞書である FrameNet を使用して検証した。その結果、コーパスサイズとしては 8000 万語～1 億語で一定の性能が確保されることが明らかとなり、コサイン類似度の上位 5 語の信頼性が高いことが示された。また、予備的な考察ではあるが、単語別の検証では構文的定形性・特殊性が高いと考えられる語の一致度が高く、汎用的な構文で使用される単語の一致度が低いことが示唆された。さらに word2vec の結果が FrameNet の拡張支援に役立つ可能性が示された。

今後の課題として、フレーム間関係を考慮することが挙げられる。FrameNet のフレームは相互に関連しており、Inheritance, Uses などの様々なフレーム間関係が用意されているが(cf. Ruppenhofer et al., 2016)、フレーム間関係によって関連付けられた単語と word2vec の結果の一致度の検証を進める必要がある。例えば、Change_of_temperature は Change_position_on_a_scale フレームと Inheritance の関係にあるが、このフレームに含まれる heat (v), cool (v)などの語がどのように関係するかは考察の価値がある。また、word2vec のモデルの作成方法として、windows サイズや圧縮次元数の数などの変化が結果にどの程度影響を与えるかということや、品詞を考慮したモデルでの検証も必要だと考えられる。

謝辞

本研究の成果の一部は JSPS 科研費 （15K16798）および KDDI 財団の助成を受けたものである。

参考文献

Chiu, B., Korhonen, A., & Pyysalo, S. (2016). Intrinsic evaluation of word vectors fails to predict extrinsic performance. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP* (pp. 1-6).

Fillmore, C. J. (1982). Frame semantics. In Yang, I. (ed.), *Linguistics in the morning calm: Selected papers from SICOL-1981* (pp.111-137). Seoul: Hanshin.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica* 6 (2) (pp.222-254).

Fillmore, C. J. (2008). The merging of "frames." In R. R. Favretti (Ed.), *Frames, corpora, and knowledge representation* (pp. 1-12). Bologna: Bononia University Press.

Harris, Z. S. (1954). Distributional structure. *Word* 10 (pp. 146–162).

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck, C. R. Johnson, C. Baker and J. Scheffczyk (2016). FrameNet II: Extended Theory and Practice. <https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf> [Accessed: Feb. 20, 2018].

Schnabel, T., Labutov, I., Mimno, D., & Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 298-307).

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* 1 (pp. 1555-1565).

内田諭 (2015). 『フレーム意味論に基づいた対照の接続語の意味記述』花書院.