単語分散表現におけるパラメーター変化の影響：word2vecを用いた事例研究

内田, 諭
九州大学

出版情報：The Institute of Statistical Mathematics cooperative research report. 413, pp.31-42, 2019-03. The Institute of Statistical Mathematics
バージョン：
権利関係：

単語分散表現におけるパラメーター変化の影響：word2vec を用いた事例研究

Effects of Parameter Changes on Word Embeddings: A Case Study Using word2vec

内田　諭
九州大学

Satoru UCHIDA
Kyushu University

1. はじめに

自然言語処理において、単語分散表現（word embedding）は非常に重要な基礎となりつつある。機械翻訳（Cho et al., 2014）、感情分析（Tang et al., 2014）、語彙の平易化（高田他, 2017）、など様々なタスクにおいて精度が向上することが報告されており、今後一層広く用いられるようになると考えられる。

単語分散表現は「意味が類似した単語は類似の文脈に出現する」という分布仮説（Harris, 1954）を理論的土台として開発されたもので、単語の前後の一定スパンに出現する共起語行列を低次元の変数に変換して数値化する手法である。文脈をベクトル化し単語の意味を数値で表現することで、コサイン類似度等を用いてベクトル空間における単語間の近似度等を測定することが可能となる。これにより分布仮説に基づいた基準によって「類義語」（コサイン類似度が高い単語）を抽出することができ、自然言語処理のタスクのみならず、辞書編纂や教育など応用言語学的な利用も可能である。

ベクトル化は単語の表層形を基にして行われるため、多義性が表現できないという問題がしばしば指摘される。これには文章のトピックを付与して分散表現を作成する方法（Fadaee et al., 2017）や依存関係を基に分散表現を作る方法（芦原他, 2018）などが提案されており、多義性を考慮したベクトル化の開発も進められている。

単語の意味を分散表現として表す場合、変数を何次元に圧縮するか、どの程度の幅の単語を文脈とみなすか、学習を何度繰り返すかなど様々なパラメーターを決める必要が

ある。しかしながら、これらの変数はアプリケーションのデフォルト値が用いられるなど、実験において深く考慮されることが少ない。そこで本稿では、広く用いられる単語分散表現のアプリケーションである word2vec (cf. Mikolov et al, 2013)を用いて、「変数の次元」（size パラメーター）と「文脈の幅」（window パラメーター）の設定を段階的に変更し、これらのパラメーターの変化が類義語の抽出結果にどのような影響を与えるかを考察する。

2. 実験条件

2.1 コーパス

内田(2018)は言語研究に word2vec を使う場合の最適なコーパスサイズを FrameNet[1]をベンチマークとして検証し、約 8000 万語〜1 億語で結果が安定することを報告している。本稿ではその基準に従い、Corpus of Contemporary American English[2](COCA)の 2012-2015 年の追加データを用いる。COCA は academic, fiction, magazine, news, spoken の 5 つのジャンルからランダムサンプリングされた均衡コーパスで、各年約 2000 万語を含む。本稿で使用する 2012-2015 年の追加版フルテキストデータは 4 年分あり、合計約 8000 万語からなる COCA のサブコーパスである。全文データは、生テキスト、WLP ファイル（Word, Lemma, POS タグ）、DB ファイルの 3 つの形態で提供されている。以下の実験では、WLP ファイルから Lemma と POS タグの 1 文字目を連結してテキストデータ化し、例えば bought the apples という表層形は buy_v the_a apple_n のように変換したものを入力コーパスとして利用した。word2vec による言語モデルの作成には python のライブラリである gensim を利用して行った。

2.2 パラメータの設定

本稿では size パラメーターについては 100, 200, 300, 400, 500 の 5 パターンを、window パラメーターについては 2, 4, 6, 8, 10 の 5 パターンを設定し、全部で 25 個のモデルを作成した。以下、それぞれのモデル名は「COCA_size の値_window の値.model」で表す。例えば、100 次元のモデルは COCA_100_2.model, COCA_100_4.model, COCA_100_6.model, COCA_100_8.model, COCA_100_10.model の 5 つあり、2 番目の数字が window パラメーターの値を示す。なお、その他のパラメーター(alpha, min_alpha, iter など)についてはデフォルト値を利用した。

---

[1] https://framenet.icsi.berkeley.edu/fndrupal/
[2] https://corpus.byu.edu/coca/

2.3 検証方法

本稿では高頻度語を対象としてそれらとコサイン距離が近い上位 20 語を抽出し、モデル間の一致度を計測した。例えば、big_j に対して、100 次元のモデルで window 幅が 2 と 4 の上位 5 語は表 1 の通りであるが、上位 5 位までの一致数は 4 である。

COCA_100_2.model			COCA_100_4.model		
1	huge_j	0.781	1	huge_j	0.763
2	great_j	0.656	2	great_j	0.668
3	small_j	0.656	3	small_j	0.585
4	*major_j*	0.650	4	large_j	0.575
5	large_j	0.644	5	*tough_j*	0.571

表 1 コサイン類似度が高い単語の例

本稿では英語教育で広く使われている CEFR-J Wordlist[3]（A1, A2, B1, B2 の 4 つにレベル分けされている）を参考に、頻度が高く重要語であると考えられる A1 および A2 レベルからランダムに名詞 10、形容詞 10、動詞 10 の合計 30 単語を抽出した。ただし、スペルが複数あるもの(color/colour など)、非英文字を含むもの（e-mail など)、フレーズであるもの(credit card など)は除外した。検証対象は以下の通りである。丸カッコ内は単語の CEFR レベルを、角カッコ内は本研究で用いたコーパスでの頻度を表す。

agent_n(A2) [5411], anxious_j(A2) [997], big_j(A1) [41017], build_v(A1) [17147], buy_v(A1) [14202], challenge_n(A2) [9790], dirty_j(A1) [1924], embarrassing_j(A2) [653], farm_n(A1) [5180], fast_j(A1) [2063], fun_j(A1) [3731], get_v(A1) [166188], healthy_j(A1) [5816], heart_n(A1) [13292], impossible_j(A2) [3371], increase_v(A2) [10719], leaf_n(A1) [4082], leave_v(A1) [39100], line_n(A1) [21844], manage_v(A2) [6952], mean_v(A1) [44796], mind_v(A2) [2778], note_n(A1) [7161], remain_v(A2) [14521], scientist_n(A1) [7337], ticket_n(A1) [3702], traditional_j(A2) [5983], welcome_v(A1) [4388], wet_j(A2) [2245], yellow_n(A1) [226]

それぞれの単語についてコサイン類似度上位 20 単語を取り出し、行にモデル名、列に各出現単語を変数としてクロス集計し、R (ver. 3.5.1)の dist 関数を用いてモデル間のユ

[3] 『CEFR-J Wordlist Version 1.3』 東京外国語大学投野由紀夫研究室.（http://www.cefr-j.org/download.html）

ークリッド距離を計算した。その距離行列に対して、hclust 関数（method=ward.D2）を使って階層的クラスター分析を全単語、名詞のみ、形容詞のみ、動詞のみの 4 パターンで実施し、plot 関数によってデンドログラムを描画した。

3. 結果

3. 1 全単語を用いた分析

図 1 はすべての単語を対象としたクラスター分析の結果である。この図から 100 次元のモデルを除いては多くの場合で window の幅によってグループが形成されていることがわかる。window 値が異なるモデルがクラスター内に混ざる場合もあるが、COCA_200_6.model が window 値 4 のモデル群に含まれるなど、隣接する window の値を持つグループと距離が近い。また、window の値が小さいモデルが図の左側に、大きいモデルが右側にそれぞれ集まっていることから、window 値の近さはそのままモデルの近さとなって表れていると考えられる。

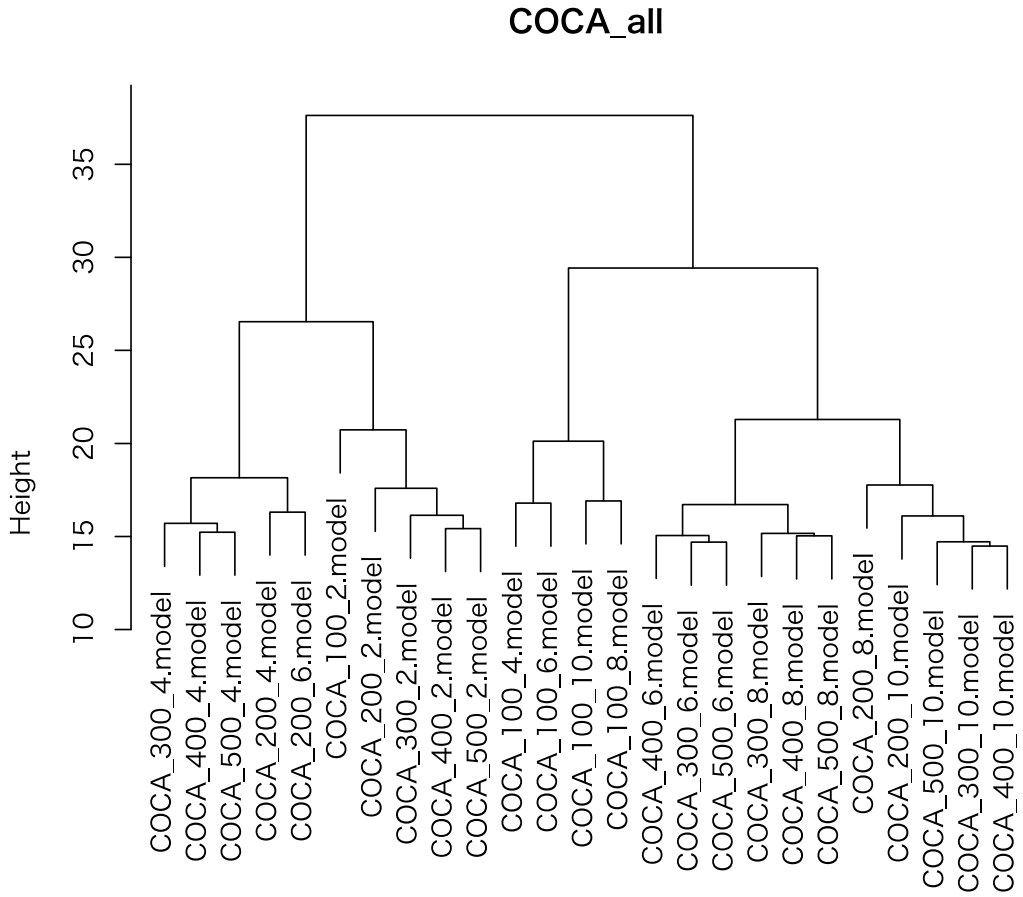

図 1 全単語によるクラスター分析

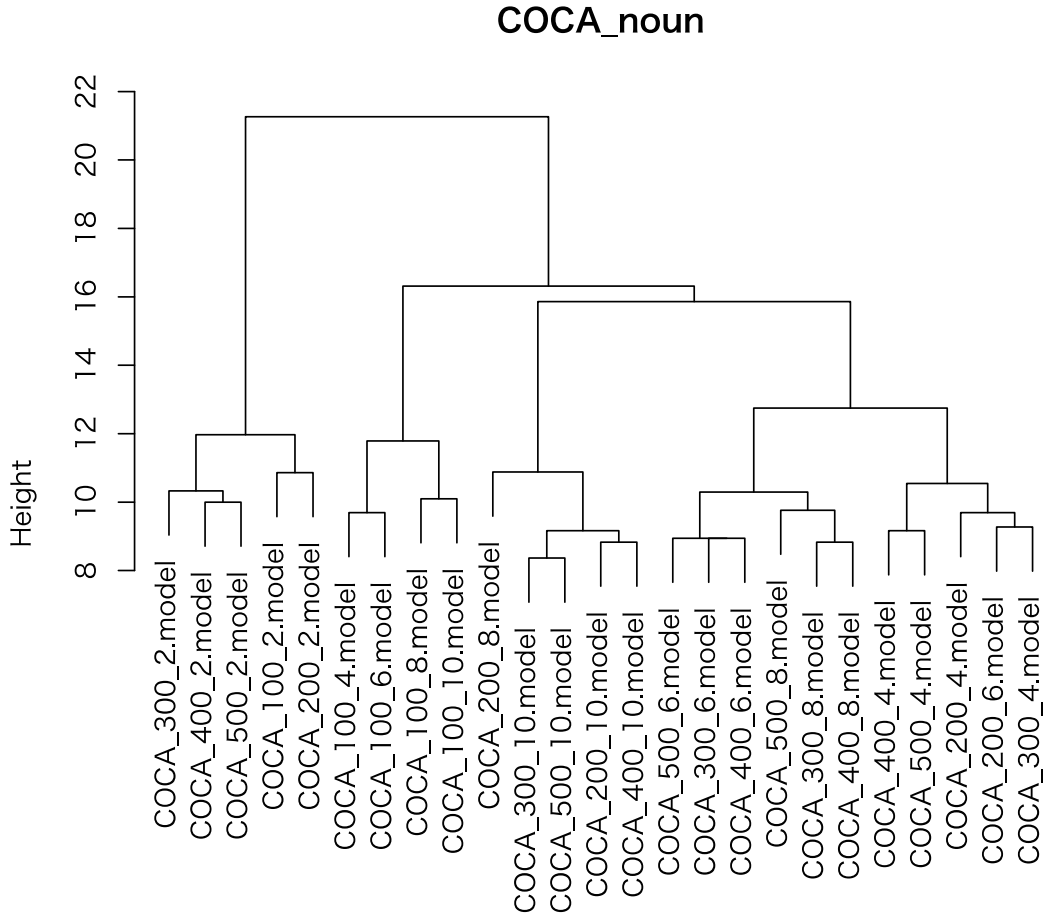

図 2 名詞によるクラスター分析

3.2 名詞を用いた分析

図 2 は名詞のみを抽出してデンドログラムを描画したものである。図 1 と比較して、100 次元のモデルが特殊な振る舞いをしていることは共通しているが、window 値が 2 のグループが独立したグループを形成していることがわかる。これはおそらく名詞の直前に出現する冠詞の影響と考えられる。

以下にすべてのモデルの上位 20 語に共通して出現する単語をリストする。これらの単語は次元・window の幅によらずコサイン類似度が高いことから、ターゲットの単語と非常に近い性質を持つ語群であると考えることができる。例えば note_n に対しては letter_n, notebook_n などが列挙されているが、直感的にもこれらは単語の意味が近いことがわかる。

[全てのモデルに出現する名詞]

【agent_n】 agency_n, informant_n
【challenge_n】 dilemma_n, hurdle_n, issue_n, obstacle_n, problem_n, threat_n
【 farm_n 】 cattle_n, co-op_n, dairy_n, farmer_n, farming_n, farmland_n, feedlot_n, homestead_n, livestock_n, orchard_n, pasture_n, plantation_n, ranch_n
【heart_n】 gut_n, lung_n, soul_n, stomach_n, throat_n
【leaf_n】 bark_n, bough_n, foliage_n, grass_n, petal_n, seedling_n, shrub_n, stalk_n, vine_n
【note_n】 letter_n, notebook_n, query_n, tweet_n
【 scientist_n 】 anthropologist_n, astronomer_n, biologist_n, chemist_n, cosmologist_n, geologist_n, physicist_n, researcher_n
【ticket_n】 coupon_n, discount_n, merchandise_n, vip_n
【yellow_n】 blue-green_j, pink_n, purple_n

3.3 形容詞を用いた分析

図 3 は形容詞のみを対象としたデンドログラムである。この図では特殊な振る舞いをしていると考えられる 100 次元のモデルを除いて考えると、window 値が 2、4 のモデルとその他の値でグループが分かれている。従って形容詞の場合、window 値は 4 以下と 6 以上で違いが出てくると考えられる。

次のリストはすべてのモデルに安定的に出現する形容詞である。例えば、big_j に対して enormous_j や great_j など意味的に近い語や bad_j など意味的には反対だが類似した文脈に出現すると考えられる語が含まれている。

[全てのモデルに出現する形容詞]

【anxious_j】 confused_j, depressed_j, frustrated_j, hesitant_j, impatient_j, nervous_j, worried_j
【big_j】 bad_j, enormous_j, great_j, huge_j, large_j, major_j, real_j, small_j, tough_j
【dirty_j】 filthy_j, smelly_j, ugly_j
【 embarrassing_j 】 bizarre_j, frightening_j, frustrating_j, humiliating_j, outrageous_j, shocking_j
【fast_j】 fast_r, nimble_j, quick_j, slow_j
【fun_j】 amazing_j, awesome_j, enjoyable_j, entertaining_j, exciting_j, fantastic_j, fun_n, funny_j, nice_j, wonderful_j
【healthy_j】 healthful_j, lean_j, nutritious_j, unhealthy_j
【 impossible_j 】 difficult_j, easy_j, hard_j, necessary_j, possible_j, tempting_j, unable_j, unlikely_j
【traditional_j】 conventional_j, established_j, modern_j, secular_j
【wet_j】 damp_j, dry_j, moist_j, soaked_j, soggy_j, warm_j

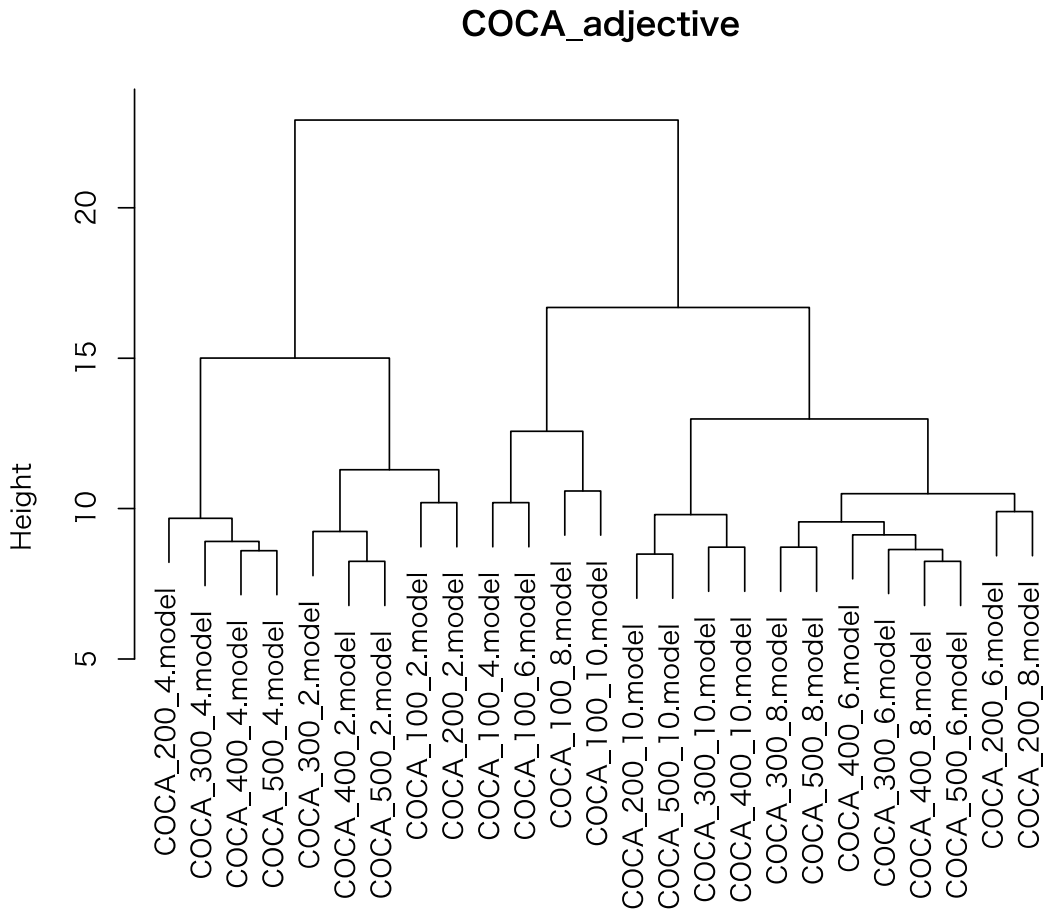

図 3 形容詞によるクラスター分析

3.4 動詞を用いた分析

動詞のみを用いて作成したデンドログラムを図 4 に示す。他の品詞同様、100 次元のモデルが一つのグループを形成するという傾向が観察される。また、window の値に着目すると、値が 2 のモデルが独立したグループになっていることがわかる。これはおそらく名詞の場合と同様に主語や目的語に共起する冠詞の影響だと考えられる。

動詞の全てのモデルに共通して出現する単語を以下に挙げる。これらの単語は特に統語的な振る舞いが似ているものだと考えることができる。例えば、mean_v は assume_v, believe_v, guess_v など that 節をとる単語が類義語として列挙されている。また、increase_v の場合、decrease_v, reduce_v などの反意語が目立つが、これらは increase [decrease, reduce] A from B to C のように同一の構文をとる動詞である（cf.内田 2018）。

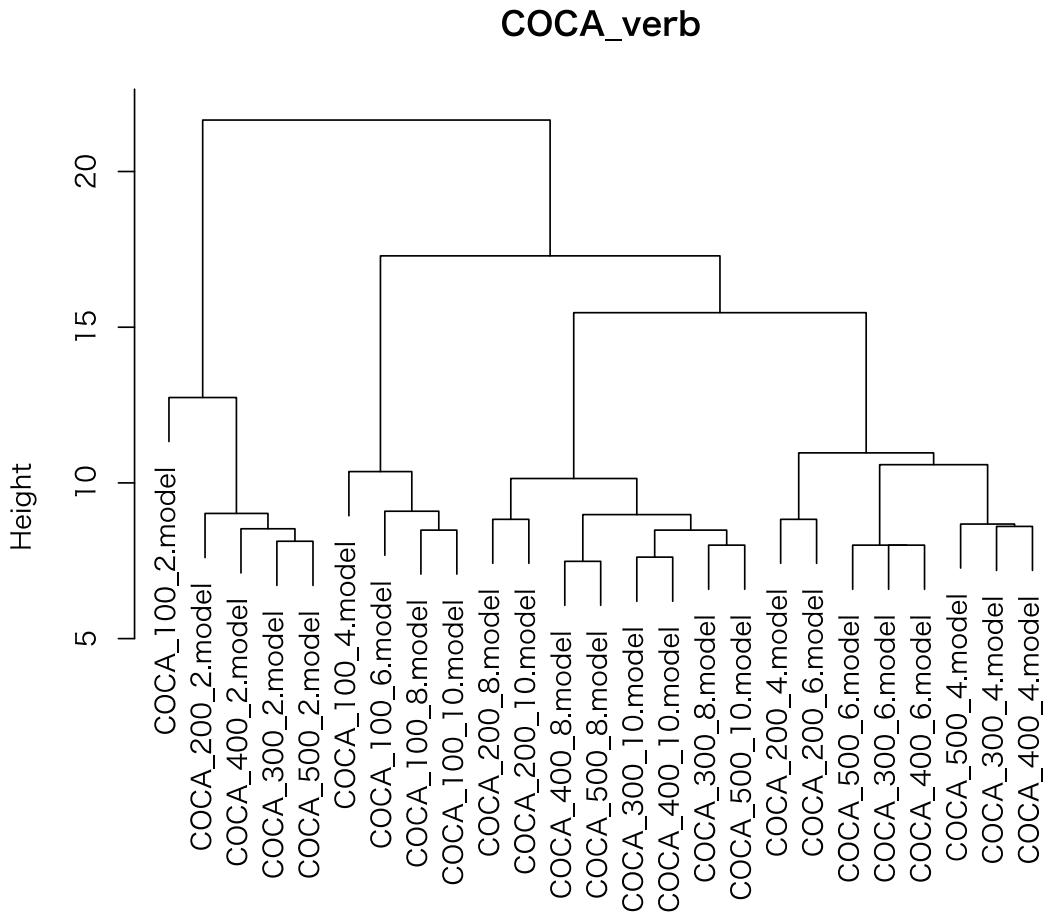

図 4 動詞によるクラスター分析

[全てのモデルに出現する動詞]

【build_v】 construct_v, create_v, develop_v, erect_v, establish_v, rebuild_v

【buy_v】 borrow_v, fetch_v, purchase_v, rent_v, resell_v, sell_v, steal_v, trade_v

【get_v】 come_v, go_v, jump_v, pull_v, put_v, slip_v, sneak_v

【increase_v】 boost_v, decrease_v, decreasing_j, diminish_v, improve_v, lessen_v, maximize_v, reduce_v

【leave_v】 abandon_v, flee_v, return_v, stay_v

【manage_v】 attempt_v, struggle_v, try_v

【mean_v】 assume_v, believe_v, guess_v, hate_v, know_v, like_v, love_v, suppose_v, think_v, understand_v

【mind_v】 bother_v, dare_v, forget_v, know_v, like_v, want_v

【remain_v】 appear_v, become_v, endure_v, linger_v, maintain_v, persist_v, prevail_v, render_v

【welcome_v】 bring_v, invite_v, thank_v, thanks_i, thanks_n, welcome_j

4. 考察

前節での分析の結果、100 次元のモデルは 1 つの独自のグループを形成する傾向があることが明らかになった。このことから、size パラメーターを決定する際は 100 以下で結果が大きく変わる可能性があることを意識する必要があるといえる。size が 200 以上の場合、結果は window パラメーターの影響をより大きく受ける。また、window の値が小さい場合と長い場合で結果が分かれることが示された。特に名詞と動詞では、window 値が 2 の場合に特徴的な性質を示すことが明らかになった。このとき 100 次元のモデルの場合も window 値が 2 のグループに入っており、次元数よりも window 値のほうが分類に効いているということが読み取れる。

window パラメーターの大小がどのように結果に影響するかを確認するため、window の値が 2 の場合と 10 の場合の特徴語を抽出して比較する。ただし、100 次元のモデルは考察対象から除外する。

次のリストは実験対象の語のコサイン類似度が高い上位 20 語について、COCA_200_2.model, COCA_300_2.model, COCA_400_2.model, COCA_500_2.model の全てに出現し、COCA_200_10.model, COCA_300_10.model, COCA_400_10.model, COCA_500_10.model の全てに出現しないものである。これらの語はスパンが短い場合の特徴語とみなすことができる。

［名詞］
【agent_n】 prosecutor_n, regulator_n, supervisor_n
【challenge_n】 drawback_n, pitfall_n, setback_n
【farm_n】 poultry_n, vineyard_n
【heart_n】 midsection_n
【note_n】 notice_n, remark_n, tweets_n
【scientist_n】 archaeologist_n, economist_n, engineer_n, historian_n, observer_n
【ticket_n】 e-book_n, flyer_n, handout_n
【yellow_n】 magenta_n, purplish_j
［形容詞］
【anxious_j】 excited_j
【big_j】 substantial_j
【dirty_j】 creepy_j, slick_j
【embarrassing_j】 confusing_j, distressing_j, insulting_j
【fast_j】 inexpensive_j, smart_j
【fun_j】 fascinating_j, frustrating_j, scary_j
【healthy_j】 decent_j, mature_j, normal_j, sane_j, self-sufficient_j, sober_j, stable_j
【impossible_j】 advisable_j, feasible_j, unwise_j
【traditional_j】 standard_j

【wet_j】 parched_j, ripped_j, sweaty_j
［動詞］
【buy_v】 dump_v, snag_v
【get_v】 drive_v, give_v, haul_v, move_v
【leave_v】 send_v, spare_v
【manage_v】 compensate_v, decide_v, endeavor_v, plan_v, strive_v, unable_j
【mean_v】 hope_v
【mind_v】 tempt_v, understand_v
【remain_v】 be_v
【welcome_v】 beckon_v, flock_v, hurry_v, revert_v, salute_v, send_v

このリストを観察すると、70/72（約 97%）で見出し語と品詞が同じであることがわかる。このことから、window の値が低い場合は文法的に近い振る舞いをする語がリストされる傾向にあるといえるだろう。

一方、COCA_200_10.model, COCA_300_10.model, COCA_400_10.model, COCA_500_10.model の全てに出現し、COCA_200_2.model, COCA_300_2.model, COCA_400_2.model, COCA_500_2.model の全てに出現しないもので、window 幅が広い場合の特徴語と考えられるものは次の通りである。

［名詞］
【 agent_n 】 counterintelligence_n, detective_n, fbi_n, kgb_n, spy_n, undercover_j, unrestricted_j
【challenge_n】 complexity_n, constraint_n, need_n, undertaking_n
【farm_n】 agricultural_j
【heart_n】 breathing_n, chest_n, panic_n, vein_n
【line_n】 slash_n
【note_n】 envelope_n, folder_n, notepad_n, worksheet_n
【scientist_n】 scientific_j
【ticket_n】 $10_m, $20_m, $40_m, $50_m, $75_m, auction_n
【yellow_n】 purple_j, red_n, speckled_j, yellow_j
［形容詞］
【anxious_j】 insecure_j, upset_j
【dirty_j】 bleach_n, discarded_j, disgusting_j, garbage_n, grease_n, scrub_v, stink_v
【fast_j】 furious_j, speed_n, tempo_n
【fun_j】 delightful_j, terrific_j
【healthy_j】 diet_n, diet_v, dietary_j, eating_n, lifestyle_n, nutrition_n, obese_j
【impossible_j】 nonexistent_j, unimaginable_j
【traditional_j】 rigid_j
【wet_j】 frigid_j, spongy_j, sweat_n
［動詞］
【build_v】 builder_n, design_v

【buy_v】 loan_v, retail_v
【get_v】 kick_v, pee_v, shove_v, yank_v
【increase_v】 offset_v, skyrocket_n
【leave_v】 trail_v
【manage_v】 control_v, navigate_v, stabilize_v, sustain_v
【mean_v】 say_v
【mind_v】 ai_f, anymore_r, mention_v
【remain_v】 continue_v
【welcome_v】 congratulations_n, hi_u

window 値が低い場合と比較すると、品詞の一致率が大きく下がり 50/81（約 62%）となった。その一方で、対象語の文脈に出現する単語がリストされる傾向にあることが読み取れる。例えば、heathy_j-diet_n, fast_j-speed_n, farm_n-agricultural_j など単語が持つ性質を表す語がピックアップされていることが観察できるが、これらの単語は healthy diet, fast speed, agricultural farm のように修飾関係にあるものだと考えられる。また、場面的な隣接関係にある単語（welcome_v-hi_u、ticket_n-auction_n, note_n-envelope_n）、位置的な隣接関係にある単語（heart_n-chest_n）、所属関係（agent_n-fbi_n, agent_n-kgb_n）などを表すと考えられる語がリストに含まれている。このように window のパラメーターはリストされる単語に大きな影響を与える可能性があるため、研究の目的に応じて慎重に設定する必要がある。

5. 結語

本論文では word2vec の size および window パラメーターを変化させ、作成した言語モデルの類似度をクラスター分析によって明らかにした。その結果、100 次元のモデルは独自の振る舞いを示し、window の値が低い場合と高い場合でリストされる単語の傾向に違いがあることが明らかになった。すなわち window 値が低い場合は文法的ステータスが近い単語が抽出される傾向を示し、window 値が高い場合は対象語の文脈に出現する単語がリストされることがわかった。以上の結果から、最も標準的なモデルは size=300〜400, window=5〜6 によって生成できると考えられる。

本研究では size と window のパラメーターの変動を実験したため、その他のパラメーターの影響については考慮しておらず、今後の課題の一つである。また、対象とした単語や品詞も限定して行ったため、本稿の結論が強固なものであるかを検証するためにはより大規模な実験を行う必要があると考えている。

参考文献

芦原和樹・梶原智之・荒瀬由紀・内田諭 (2018).「依存構造に基づく単語から語義の分散表現への細分化」『研究報告自然言語処理 (NL)』2018-NL-237(3). pp.1-7.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Learning topic-Sensitive word representations. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. pp.444-447.

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

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.

内田諭(2018).「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」『言語統計を用いた認知言語学研究へのアプローチ』統計数理研究所 pp.41-51.

高田祥平・荒瀬由紀・内田諭(2017).「英語教育支援のための Lexical Simplification: コロケーションスコアを用いたアプローチ」『言語処理学会 第 23 回年次大会発表論文集』pp.939-942.

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