接尾辞を持つ単語の語義定義文とその分散表現の分析

須山晃平　佐々木稔
茨城大学 工学部 情報工学科
{18t4050x, minoru.sasaki.01}@vc.ibaraki.ac.jp

・概要

単語の語義定義文は、使用頻度が高い単語については精度が高いが、使用頻度が低いものについては精度が低いという課題がある。この論文では、自然言語処理モデルの一つである Sentence-BERT[1]モデルを用いて、接尾辞を持つ単語 godhood と意味的類似性の高い単語を導出した。また、膨大な語彙を持つ英語版 Wiktionary[2]の語義定義文を用いることで、意味的類似性の検索の精度の向上を図った。

・1 はじめに

本研究では、英語版ウィクショナリーの記事の内容を機械可読にした辞書である ENGLAWI を用いて、接尾辞を持つ英単語の語義定義文において、分散表現を取ることで定義文の精度について分析した。

・1.1 研究の背景と目的

Sajous ら[3]は、ENGLAWI の定義文で学習した FastText モデル[4]と LSTM アーキテクチャ[5]の 2 つのモデルを用いて、英単語 godhood と意味的類似性の高い上位 10 個の単語をそれぞれ求めた。図 1 は、Sajous らが行った実験によって求められた、2 つのモデルにおける godhood の上位 10 個の近傍語である。図 1 に示したように、この実験により得られた単語の中には、接尾辞"-hood"を含む単語が FastText モデルには 5 個、LSTM アーキテクチャでは 3 個存在しているが、この中には fatherhood や motherhood など、godhood とは意味的類似性のない単語が含まれていた。"-hood"のような接尾辞は、語基に別の意味を添加するものであるため、単語の意味においては補助的な役割を果たす。ENGLAWI の語義定義文では、接尾辞を持つ単語の語義定義文に特徴的な記述の形式が存在することがある。表 1 は、godhood を含む 5 つの接尾辞"-hood"を含む単語とそれらの単語の ENGLAWI における語義定義文を示したものである。これらの単語は、接尾辞"-hood"に対応する記述(The state of being)による影響を強く受けたために、同じ接尾辞を持つ単語の類似度が高くなるという問題があった。

FastText ENGLAWI	LSTM ENGLAWI
fatherhood	deityhood
demigod	blessedness
motherhood	divineness
selfhood	paganity
triune	angelhood
childhood	deathlessness
nirvana	fathership
bodhisattva	worshipability
manhood	creatorship
incarnation	buddhahood

図 1：godhood の上位 10 個の近傍語[3]

この問題に対処するために、本論文では、高精度で文章の意味的類似性の検索が可能な Sentence-BERT モデルを用いることを提案する。この Sentence-BERT モデルを用いて、ENGLAWI の語義定

表 1：接尾辞”-hood”を含む語の語義定義文

単語	ENGLAWI の語義定義文
godhood	The state of being a god; divinity
fatherhood	The state of being a father.
motherhood	The state of being a mother.
childhood	The state of being a child.
angelhood	The state of being an angel.

義文から godhood との意味的類似性が高い単語を求めることにより、意味的類似性の検索の精度が向上するかどうかを研究した。

また、接尾辞を含む低頻度の単語について、ENGLAWI の定義文の分散表現を求め、その値を他の文章と比較・分析する。

・2 関連研究および関連手法

・2.1 Sentence-BERT

Sentence-BERT は、自然言語処理モデルである BERT [6]を改良したモデルであり、2019 年に Nils Reimers と Iryna Gurevych により提案された[1]。この Sentence-BERT は、学習済みの BERT モデルに Siamese Network を組み合わせてファインチューニングすることで文章の分散表現を求める手法である。複数の文章を用いた意味的類似性の検索などのタスクにおいて、BERT モデルを使用すると精度が低いという課題が存在する。Sentence-BERT では、この課題に対処するために、Siamese Network と Triplet 損失関数を用いることで、2 つの文章の類似度を効率的に求めることができるようになる。また、Sentence-BERT を用いることによって、BERT を用いた文章間の類似性判定の精度を維持しながら、文章の意味的類似性の検索が可能である[7]。

・2.2 Wiktionary

Wiktionary は、2002 年に開始した、多言語に対応したオンライン上の辞書である。ある 1 つの単語や熟語、フレーズを見出し語とする記事があり、各記事には、見出し語の語義定義文のほか、語源や発音、例文などが収録されている。Wiktionary には、180 以上の言語に対応したヴァージョンがあり、その中で記事数が最多である英語版 Wiktionary には 700 万以上の記事がある。

・2.3 ENGLAWI

ENGLAWI[i]は、2020 年に Franck Sajous らにより提案されたもので、英語版 Wiktionary の記事の内容を XML で暗号化することにより、機械が記事の内容を読み取ることを可能にした辞書である。ENGLAWI における単語の語義定義文を利用して意味的類似性の検索をすることにより、Wiktionary 上の語義定義文における意味的類似性の検索が可能となる。

・3 提案方法

本研究では、Sentence-BERT モデルを用いて、ENGLAWI の語義定義文の分散表現を求めることにより、godhood と意味的類似性の高い上位 10 個の単語を求めた。1 つの見出し語が複数の語義定義文を持つ場合は、各定義文と godhood の定義文とのコサイン類似度を求め、すべての類似度の値を平均したものをその見出し語の類似度とした。

・4 実験

・4.1 実験方法

Omkiran Malepati [8]が提案したプログラムを基に、ENGLAWI に収録されたすべての見出し語におけるすべての語義定義文の分散表現を Sentence-BERT モデルを用いて抽出し、godhood の語義定義文 (The state of being a god; divinity)の分散表現とのコサイン類似度を取ることで、godhood と他の ENGLAWI の見出し語との意味的類似性を求めた。見出し語が複数の定義文を持つ場合は、その見出し語のすべての定義文との類似度をそれぞれ取ったものを平均した。

・4.2 実行結果

表 2 は、Sentence-BERT モデルを用いて求められた、godhood と意味的類似性の高い上位 10 個の見出

し語と godhood とのコサイン類似度を表したものである。表 3 は、表 2 で示された 10 個の単語の ENGLAWI における語義定義文である。表 3 において、godlike は、”Having characteristics of a god.”と”Characteristics of a god.”の 2 つの語義定義文を持っており、表 1 における godlike のコサイン類似度の値は、2 つの語義定義文それぞれと godhood の語義定義文との類似度の平均を取ったものである。

表 2：godhood の上位 10 個の近傍語

順位	見出し語	類似度
1	deityhood	0.9413002
2	theomonism	0.91305542
3	godship	0.90239114
4	divineness	0.89775801
5	skydaddy	0.89645851
6	divinelike	0.88855153
7	shechinah	0.88607669
8	yazata	0.88598311
9	godlike	0.88481197
10	ens entium	0.8823793

表 3：ENGLAWI の語義定義文（Sentence-BERT）

見出し語	語義定義文
deityhood	The state of being a deity; divinity
theomonism	a monism that recognizes the existence of God
godship	The condition of being a god; divinity; especially as a jocular epithet
divineness	The quality of being divine; divinity
skydaddy	A god (especially, God).
divinelike	Characteristic of divinity.
shechinah	the presence of God
yazata	A divinity.
godlike	Having characteristics of a god. Characteristics of a god.
ens entium	The ‘being of beings'; God.

・5 考察

表 4 は、図 1 で示した、FastText と LSTM を用いて導出された godhood の近傍語のうち表 1 にない 16 個の単語の ENGLAWI における定義文である。

表 4：ENGLAWI の語義定義文（FastText, LSTM）

単語	語義定義文
demigod	A half-god or hero; the offspring of a deity and a mortal.
selfhood	The quality of being self-centered or egocentric; selfishness.
triune	Threefold, having three components that are both separate and united; said especially of the Trinity of Christian doctrine.
nirvana	State of paradise; heightened or great pleasure.
bodhisattva	A person who has taken specific lay or monastic vows and who is on the road to perfect knowledge; specifically, one who foregoes personal nirvana in order to help others achieve enlightenment.
manhood	The state of being man as a human being.
incarnation	The state of being incarnated.
deityhood	The state of being a deity; divinity
blessedness	The state or condition of being blessed, holy.
divineness	The quality of being divine; divinity.
paganity	The state of being a pagan; paganism.
deathless-ness	The state of being deathless; eternity; immortality.
fathership	The state of being a father; fatherhood; paternity.
worship-ability	Capability of being worshiped; worthiness of veneration.
creatorship	State or condition of a creator.
buddhahood	The state of being spiritually enlightened by the Buddhist teachings.

表 2 の語義定義文で示したように、Sentence-BERT モデルを用いて導き出された godhood の上位 10 個の近傍語の語義定義文には、表 4 で示した語義定義文の多くとは異なり、”god”や”divinity”といった単語が含まれている。これらの単語は、godhood の ENGLAWI の語義定義文である”The state of being a god; divinity”において元の単語 (god) の意味を表す部分 (a god; divinity) に含まれている。これは、godhood の語義定義文において、「～の状態」などの意味を持つ接尾辞”-hood”の意味を表す部分 (The state of being) よりも、元の単語の意味を表す部分が重視され、接尾辞による影響を受けなかったことを意味する。したがって、Sajous ら[3]が 2 つのモデルを用いて行われた実験の結果よりも、意味的類似性の精度が向上しているといえる。

表 5 は、表 2 で示した godhood の上位 10 個の近傍語について、Wikipedia の単語の利用頻度のファイル (enwiki-2022-08-29.txt[ii]) を基に、それぞれの見出し語の Wikipedia 上における出現回数を表したものである。ただし、出現回数が「≦2」となっている見出し語は、Wikipedia における出現回数が 2 回以下である、すなわち、単語の利用頻度のファイルに記載がない単語である。また、ens entium は 2 単語で構成される見出し語のため、出現回数がより少ない entium の出現回数を示した。

表 5 に示したように、godhood の上位 10 個の近傍語のうちの 7 個の単語で出現回数が 1 桁であった。このことから、膨大な語彙を持つ英語版の Wiktionary を用いることにより、利用頻度の高い単語だけでなく、低頻度の単語に対しても高い精度で意味的類似性を求めることができる。

表 5：見出し語の出現回数

見出し語	出現回数（回）
deityhood	≦2
thomonism	≦2
godship	8
divineness	6
skydaddy	≦2
divinelike	≦2
shechinah	84
yazata	134
godlike	675
ens entium	3

・6 まとめ

本研究では、自然言語処理モデルである Sentence-BERT を用いて、ENGLAWI の語義定義文の分散表現から定義文どうしのコサイン類似度を求めることにより、godhood と意味的類似性の高い単語を求めた。実験の結果、”god”や”divinity”といった単語が語義定義文に含まれる単語が上位 10 単語を占め、接尾辞”-hood”を含む単語が 1 つのみであったことから、意味的類似性の検索の精度が、低頻度語を含めて向上したことが確認された。

今後の課題として、本研究で用いた ENGLAWI は、2017 年 6 月 1 日現在の英語版 Wiktionary のデータを基に作成されているので、現在の語義定義文が ENGLAWI に記載された内容と異なる見出し語が存在するため、より新しい語義定義文を用いることで意味的類似性の検索の精度が向上するかどうかなどが挙げられる。

---

[i] 現在の ENGLAWI は以下のページからダウンロード可能である。
http://redac.univ-tlse2.fr/lexiques/englawi.html

[ii] 2022 年 8 月 29 日現在の Wikipedia のすべての記事の中で計 3 回以上出現する単語 274 万語余の単語と出現回数をまとめたテキストファイルである。以下のページからダウンロード可能である。
https://github.com/IlyaSemenov/wikipedia-word-frequency/blob/master/results/enwiki-2022-08-29.txt

・参考文献

[1] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 3982-3992, 2019.

[2] Wiktionary, the free dictionary, 2002. https://en.wiktionary.org/wiki/Wiktionary:Main_Page, 2022-06 閲覧

[3] Franck Sajous, Basilio Calderlone, Nabil Hathout. ENGLAWI: From Human- to Machine-Readable Dictionary. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 3016-3026, 2020.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5:135-146, 2017.

[5] Sepp Hocheriter, Jürgen Schmidhuber. Long Short-Term Memory. Neural Computation, 9 (8):1735-1780, 1997.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186, 2019.

[7] enzo. 【入門】Sentence BERT, 2022. https://zenn.dev/en2enzo2/articles/a574b52bb8d116

[8] Omkiran Malepati. Sentence Similarity with BERT, 2021. https://medium.com/@omkiran/sentence-similarity-with-bert-49bcd250c1bc