サプライザルを用いた 名言の改変しやすさの指標についての基礎的検討

伊藤 薫 [1]
[1] 九州大学
ito@flc.kyushu-u.ac.jp

概要

本研究は，名言の安定性や改変のしやすさを検討するため，言語モデル BERT を用いてサプライザルを測定し，名言と対照文を比較した．本研究で用いるサプライザルの計算には，研究対象の性質上前方文脈のみでなく後方文脈も考慮している．測定の際は文の 1 語のみをマスクし，BERT に正解語の確率を出力させた．名言は全体的に単語や文の予測が容易であり，特に内容語において対照文との差が大きいことが明らかになった．これは名言がまとまりのある安定した表現であり，1 語のみ改変する場合は内容語の置換が容易であることを示唆している．

1 はじめに

名言は基本的に固定された表現だが，名言を引用する際の誤りや意図的な改変によりバリエーションが生まれる．このうち，意図的な改変には暗示引用やパロディ，パスティーシュ [1] などレトリックと関わるものが含まれており，これらは何らかの形で元表現との関連を維持しつつ別の表現を生み出す．言語学では構文論との関連でこのような改変・拡張について研究されており [2, 3]，コーパスを用いた定型表現からの拡張の研究 [3] では内容語の拡張（学問に王道なし→ダイエットに王道なし），句の省略（陰で糸を引く→糸を引く），態の変化（真綿で首を絞める→真綿で首を絞められる），助詞の省略・拡張など，様々なパターンが観察されている．

本研究では言語学的関心に基づき，名言の安定性（意図しない変容の少なさ）や意図的な改変のしやすさについて示唆を得ることを目的とするが，それを直接観察することは容易ではない．そこで，言語モデルを用いてサプライザル [4] を測定し，単語や文の予測しやすさを定量化する．サプライザルと改変のしやすさの関係は 2.2 で詳述する．そして，名言でない対照文と予測しやすさを比較することで，名言の性質についての知見を得ることを目的とする．結果として，名言はまとまりのある安定した表現であり，1 語のみ改変する場合は内容語の置換が容易であるという示唆が得られた．

2 方法

2.1 データ

名言のデータは Huggingface で公開されている English Historical Quotes[1)] を使用した．english_historical_quotes にはインターネット上のオープンアクセスアーカイブから収集され，データ作成者によってクリーニングされた 24,022 件の名言が収録されている．これらの名言との比較対象として，Gutenberg Books[2)] を使用した．Gutenberg Books は Gutenberg Project[3)] 上で公開されている 74,329 冊の英語書籍から抽出された 97,646,390 パラグラフからなるデータセットである．

次に，名言との比較対象を得るため，Gutenberg Books データセットを用いて English Historical Quotes の名言集 $Q = \{q_1, q_2, ..., q_N\}$ に含まれる q_i との比較対象となる文 g_i を選び，対照文集 $G = \{g_1, g_2, ..., g_N\}$ を構築した．g_i は次のような手順で抽出した．まず，Gutenberg Books から q_i と語数の等しい文を抽出し，その文に含まれる語に対し品詞を付与した．この文の品詞ベクトルと q_i の品詞ベクトルのコサイン類似度を算出し，コサイン類似度が最大のものを g_i として選んだ．品詞付与には spaCy 3.8.11, モデルは en_core_web_sm を用いた．Gutenberg Books には聖書が含まれるが，聖書には名言として引用される文も多いため，書籍タイトルによりフィルタリン

1) https://huggingface.co/datasets/m-ric/english_historical_quotes
2) https://huggingface.co/datasets/Navanjana/Gutenberg_books
3) https://www.gutenberg.org/

グして抽出対象からは除外した．また，同じ書籍から多数の用例が抽出されることを避けるため，1 冊あたりの収録上限を 25 件に設定した．著者の計算資源の制限により，1 つの名言に対する文の候補は 60,000 文としており，この中に語数が一致し，かつ，すでに G に収録されていない文がなければ，その名言は Q から除外した．結果，18,940 件の q_i と g_i のペアが得られた．

得られたペアを確認したところ，Gutenberg Books は英語の本から構築されているにも関わらず英語以外の文が G に含まれていたため，fasttext 0.9.2 の言語識別モデルを用いて英語以外の文を除外した．この操作により，最終的に 17,431 ペア (もとの名言 24,022 件のうち約 72.6%) が得られた．

2.2 指標

本研究では，名言の改変しやすさの 1 側面を捉える指標のベースとしてサプライザル (surprisal) [4] を考える．一般的にサプライザルは次のように定義される．

$$\text{Surprisal} = -\log P(x|\text{CONTEXT}) \tag{1}$$

よく使われる式として，ある時点までの単語が与えられたときに，それらを文脈として次の単語を予測するときの難易度を測るものが挙げられる．

$$\text{Surprisal} = -\log P(w_i|w_1, w_2, ..., w_{i-1}) \tag{2}$$

本研究の関心は名言の改変のしやすさであるが，どの単語を置き換えやすいかを直接測るのは困難である．そこで，代理の指標として文 $s = (w_1, w_2, ..., w_n)$ に含まれる単語 w_i をマスクし `[MASK]` に置き換えて `[MASK]` の単語を予想する際のサプライザルを考える．これは，ある名言の他の単語が与えられているとき，マスク部分に入る単語に様々な候補が上がる場合は，その部分を入れ替えやすいという想定に基づく．今回は解釈の明快さを重視し，1 度にマスクする単語の数は 1 つとした．これを数式で表すと以下のようになる．

$$\text{Surprisal} = -\log P(w_i|w_1, w_2, ..., w_{i-1}, w_{i+1}, ..., w_n) \tag{3}$$

式 (2) との違いは，(2) では前方文脈のみが与えられているのに対し，(3) では前方・後方の文脈が与えられている点である．以降，本研究のサプライザルは (3) を用いて計算する．なお，プログラミングの簡便さのため，底は自然対数を使用した．実際の計算には BERT[5][4] を用いた．なお，BERT はサブワード分割を行うため，w_i が分割される場合は単語の予想しやすさはサブワードに対して得られるサプライザルの総和とした．文全体のサプライザルには文に含まれる単語全てについてのサプライザルの平均または中央値を用いる．

3 結果

3.1 品詞別のサプライザル

まず，各コーパス内での品詞種別による予測しやすさの傾向を見るため，品詞ごとに集計したサプライザルの分布を図 1 に箱ひげ図で示す．図中の濃い青色は English Historical Quotes, 水色は Gutenberg Books を表す．また，本研究の関心は名言中の単語の予測しやすさであるため，English Historical Quotes においてサプライザルの中央値が低い順に左から箱ひげ図を描画した．

図からは，内容語に比べて機能語のサプライザルが低いことが読み取れる．具体的には PART(不変化詞詞)，ADP（前置詞），DET（決定詞），PRON（代名詞），AUX（助動詞），SCONJ（従属接続詞），CCONJ（等位接続詞）のサプライザルが低く，ADV（副詞），VERB（動詞），PROPN（固有名詞），ADJ（形容詞），NOUN（名詞）のサプライザルが高かった．この傾向は English Historical Quotes, Gutenberg Books ともに共通しているが，Gutenberg Books では機能語グループ内で従属接続詞のサプライザルが若干高く，内容語グループでは固有名詞のサプライザルが大幅に高かった．

3.2 品詞についての名言・対照文間比較

各コーパスにおける品詞の予測しやすさの差を図 2, 3 に示す．図 2 は各品詞について，English Historical Quotes の平均サプライザルから Gutenberg Books の平均サプライザルを引いた値， 図 3 は中央値について同様にの操作を行い得られた値を図示している．図 2 で平均の差を見ると，助動詞は名言の方が対照文よりもサプライザルが高い結果となった．他の機能語である等位接続詞，不変化詞，決定詞ではほぼ両者の差がなく，代名詞，前置詞，従属接続詞においても対照文におけるサプライザルの方が高かったが，比較的差が抑えられている．一方，内容語である副詞，動詞，形容詞，名詞，固有名詞

4) Huggingface の bert-base-uncased を使用した．

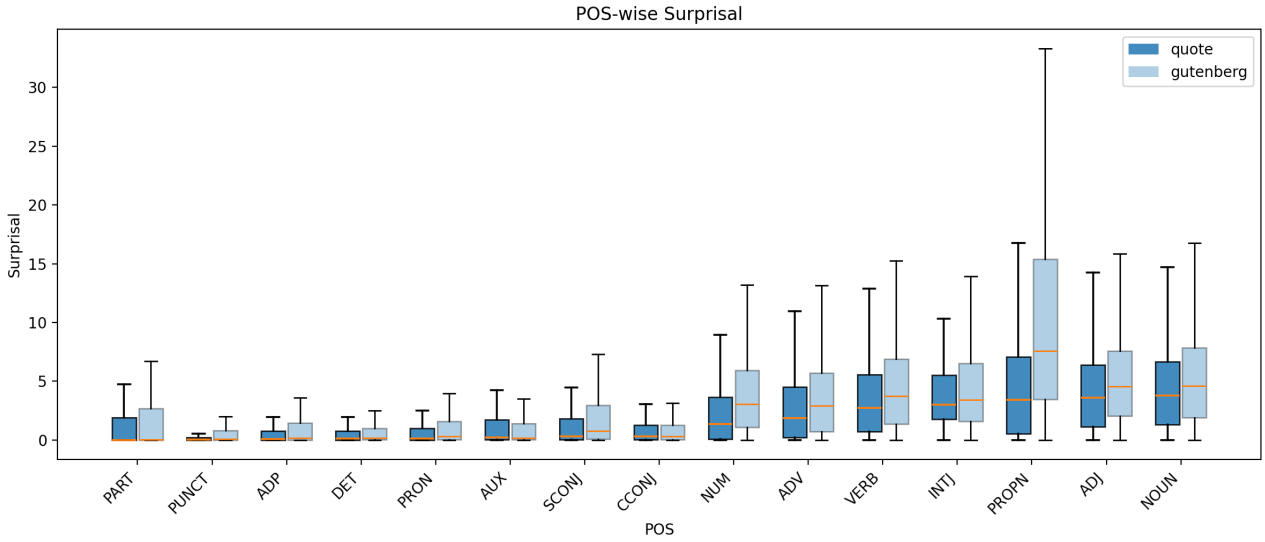

図 1　品詞別サプライザルの分布

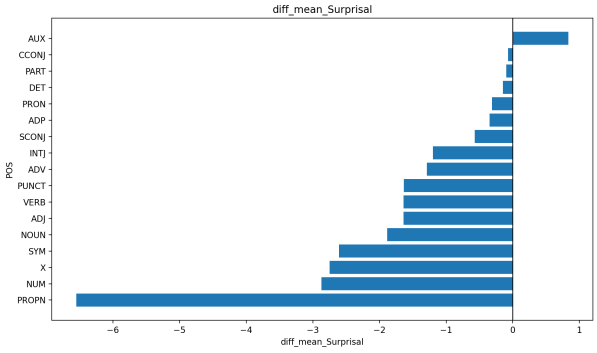

図 2　品詞別サプライザル平均の名言集・対照文集間差

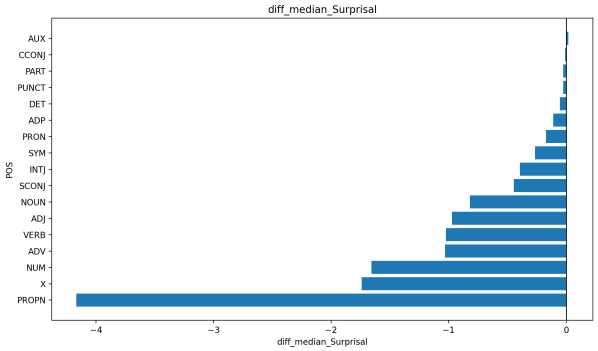

図 3　品詞別サプライザル中央値の名言集・対照文集間差

は機能語と比較して大幅に対照文におけるサプライザルが高い結果となった．この傾向は中央値を示した図 3 においてもほぼ同様であるが，平均の場合に比べて助動詞のサプライザルの差はほとんど見られない．

3.3　文についての名言・対照文間比較

最後に，名言・対照文における文単位の予測しやすさを示す．図 4 は文をデータポイントとし，その文に含まれる各単語のサプライザル平均の分布を示したもの，図 5 は中央値に関して同様に分布をしめしたものである．両図ともに青色は English Historical Quotes, オレンジ色は Gutenberg Books を表す．また，縦軸は頻度そのものではなく密度を示していることに注意されたい．

図 4 では分布のピークが名言集の方が左に寄っており，対照文の方がロングテールの分布となっている．図 5 では分布のピークこそ名言集・対照文集間でほぼ差がないものの，対照文集の方がロングテールの分布となっており，全体的に名言の単語の方が対照文よりも予測しやすいことを示している．

4　考察

4.1　品詞別のサプライザル

機能語は内容語に比べてサプライザルが低く予測しやすい傾向にあるが，これには様々な理由が考えられる．まず，機能語は closed class なので語彙数が限られており，品詞さえ特定できれば正解となる語が上位に来やすいということが考えられる．また，今回は 1 語のみを空所としたため文全体の構造がほぼ決定されており，文法的要素を担う機能語はあまり改変の余地がなかった可能性が考えられる．

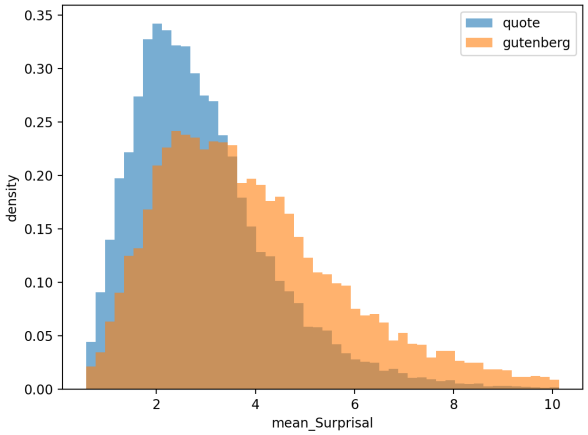

図 4　名言集・対照文集間の文サプライザル平均の分布

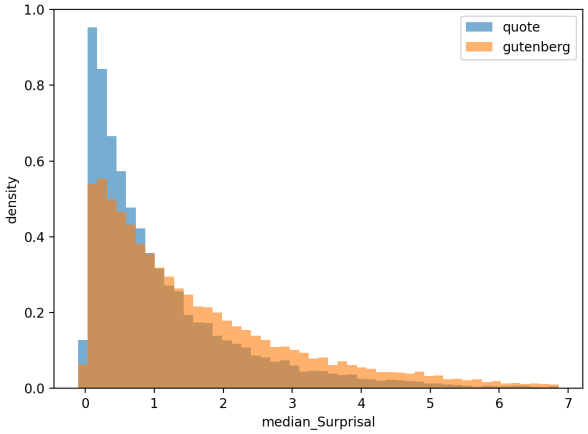

図 5　名言集・対照文集間の文サプライザル中央値の分布

4.2　品詞についての名言・対照文間比較

前節で見た通り，同一コーパス内で品詞別に見た場合には機能語の方が予測しやすく内容語の方が予測しづらいという結果が得られたが，コーパス間の差では全体的にこの傾向が強まっていると思われる．つまり，全体的に名言よりも対照文においてマスクされた単語の予測は難しいが，予測が比較的容易な機能語はコーパス間の差があまり見られないのに対し，予測が難しい内容語は対照文において予測がより難しくなっている．

個別の傾向としては，助動詞と固有名詞が特異な振る舞いをしている．助動詞については，コーパスの平均の差と中央値の差の乖離が大きいことから，名言に含まれる一部の助動詞が予測しづらい可能性がある．固有名詞に関しては一般的に予測が難しいと考えられるが，名言においては少なくとも対照文よりも予測が容易になっている．これは名言が 1 つのまとまりとして機能し，安定していることを示唆している．

4.3　文についての名言・対照文間比較

図 4, 図 5 から，文レベルでも全体的に名言の方が予測しやすいことが示唆される．この傾向は平均で見た場合も中央値で見た場合も同じであるため，文に含まれる少数の予測が難しい単語によって対照文の文レベルの予測難易度が上がっているわけではなく，名言では文内の単語が全体的に予測しやすくなっていると思われる．このことも名言のまとまりや安定性を示唆している．

4.4　本研究の限界

付録の表 1 に示したように，対照文の文平均サプライザルが高いものには URL や Gutenberg Project の書式で強調を示すためのアンダーバー，目次と思われる要素が含まれている．これらがサプライザル（特に文平均サプライザル）を上げている可能性があり，対照文の選定方法をより工夫する必要がある．

5　おわりに

本研究から，名言はある単語を周りの単語から予測しやすく，まとまったユニットであると考えられることがサプライザルの観点からも示唆された．また，内容語のサプライザルが高いことは，先行研究で指摘されている拡張パターンである内容語の拡張（置換）が行われることとも一致する．一方で，4.1 でも示した通り，機能語は語彙数が限られているため，BERT がうまく品詞を絞り込むことで元の語を予測する確率が高くなっている可能性もあるため，今後の精査が求められる．

今後の課題としては，本研究ではサプライザルを改変しやすさの指標として用いたが，サプライザルはあくまでも予測の難しさを表す指標であって改変しやすさの指標ではないため，今後はコーパスを用いた調査結果との比較などを通してこれらの関係を詳しく検討する必要がある．また，今回は 1 語のみをマスクしたが，複数の単語をマスクする場合や一部が脱落する場合など，より実際の改変と近い設定で実験することが望ましい．将来的には，どのように名言との関係を保ったまま表現を改変するかについて，定量的な研究につながれば幸いである．

謝辞

本研究は JSPS 科研費 23K12164 の助成を受けたものです.

参考文献

[1] 佐藤信夫, 松尾大, 佐々木健一. レトリック事典. 大修館書店, 東京, 2006.

[2] 山梨正明. 認知構文論 : 文法のゲシュタルト性. 大修館書店, 2009.

[3] 土屋智之. 言語と慣習性：ことわざ・慣用表現とその拡張用法の実態. ひつじ書房, 2020.

[4] John Hale. A probabilistic earley parser as a psycholinguistic model. In **Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies**, NAACL '01, p. 1–8, USA, 2001. Association for Computational Linguistics.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

A 付録

名言・対照文の実例を平均サプライザルが高い群，中程度の群，低い群に分けて 10 件ずつ下表に示す．

表 1 名言・対照文の実例

コーパス	群	サプライザル	テキスト
quote	low	0.20	If you think you can do it, you can.
quote	low	0.23	Just do what you do best.
quote	low	0.24	Next time I see you, remind me not to talk to you.
quote	low	0.26	Food for the body is not enough. There must be food for the soul.
quote	low	0.27	You can fool all the people some of the time, and some of the people all the time, but you cannot fool all the people all the time.
quote	low	0.27	To be successful, a woman has to be much better at her job than a man.
quote	low	0.28	When they are alone they want to be with others, and when they are with others they want to be alone. After all, human beings are like that.
quote	low	0.28	Most of the time I spend when I get up in the morning is trying to figure out what is going to happen.
quote	low	0.30	Things do not happen. Things are made to happen.
quote	low	0.30	To know that we know what we know, and to know that we do not know what we do not know, that is true knowledge.
quote	mid	2.68	Method is more important than strength, when you wish to control your enemies. By dropping golden beads near a snake, a crow once managed To have a passer-by kill the snake for the beads.
quote	mid	2.68	Trust and belief are two prime considerations. You must not allow yourself to be opinionated.
quote	mid	2.68	Lovers may be - and indeed generally are - enemies, but they never can be friends, because there must always be a spice of jealousy and a something of Self in all their speculations.
quote	mid	2.68	Thank God men cannot fly, and lay waste the sky as well as the earth.
quote	mid	2.68	Words derive their power from the original word.
quote	mid	2.68	In every man's heart there is a secret nerve that answers to the vibrations of beauty.
quote	mid	2.68	Education is the best provision for old age.
quote	mid	2.68	I think that a man should not live beyond the age when he begins to deteriorate, when the flame that lighted the brightest moment of his life has weakened.
quote	mid	2.68	The men who have guided the destiny of the United States have found the strength for their tasks by going to their knees. This private unity of public men and their God is an enduring source of reassurance for the people of America.
quote	mid	2.68	Insofar as international law is observed, it provides us with stability and order and with a means of predicting the behavior of those with whom we have reciprocal legal obligations.
quote	high	13.53	By indignities men come to dignities.
quote	high	13.74	Inactivity is death.
quote	high	13.82	Audacity augments courage hesitation, fear.
quote	high	13.87	Wisdom comes by disillusionment.
quote	high	13.98	Marches alone won't bring integration when human respect is disintegratin'
quote	high	14.09	Exuberance is beauty.
quote	high	14.29	Poetry is life distilled.
quote	high	16.47	Nature hates calculators.
quote	high	17.80	Mechanization best serves mediocrity.
quote	high	19.30	Fear clogs faith liberates.
gutenberg	low	0.08	Had he told them the truth they would have laughed at him.
gutenberg	low	0.08	according to the Apostles' Creed.
gutenberg	low	0.11	And that is all you have to say?
gutenberg	low	0.15	She felt as if she had lived a long, long time.
gutenberg	low	0.16	What has he been doing?
gutenberg	low	0.18	What _is_ the matter with you, and what _are_ you doing?
gutenberg	low	0.18	What was it that you wanted me to do?
gutenberg	low	0.19	What could have gone wrong, she wondered.
gutenberg	low	0.19	She knew what she had to do, and she did it.
gutenberg	low	0.20	We had known each other years and years, and in spite of our differences we had a good deal in common.
gutenberg	mid	3.58	They would be all right after a night's real rest.
gutenberg	mid	3.58	One summer he went roving about the British Isles and there he fell in with a man named Asmund Ashenside, who also was a great champion and had worsted many vikings and men of war.
gutenberg	mid	3.58	And indeed there was a strange mustiness in everything.
gutenberg	mid	3.58	I have used these sparingly, and all extracts from them have been subjected to her censorship.
gutenberg	mid	3.58	Some at least of the others I possessed; and finding much entertainment in our commerce, I did not suffer my advantages to rust.
gutenberg	mid	3.58	When I was in Spaceland I heard that your sailors have very similar experiences while they traverse your seas and discern some distant island or coast lying on the horizon.
gutenberg	mid	3.58	He had a considerable independence besides two good livings—and he was not in the least addicted to locking up his daughters.
gutenberg	mid	3.58	What if they're soldiers?
gutenberg	mid	3.58	This, naturally, was declared by several voices to give the thing the utmost price, and our friend, with quiet art, prepared his triumph by turning his eyes over the rest of us and going on: "It' s beyond everything.
gutenberg	mid	3.58	Retaining enough for his own use (he uses a good deal, because every day he does the work of five or six men), he distributes the inexhaustible remainder among those who most need it.
gutenberg	high	16.01	Slow Navigation.—Borrowing Things.—Boarding the Wreck.—The Plotters.—Hunting for the Boat.
gutenberg	high	16.02	Pudd'nhead Wins His Name.
gutenberg	high	16.16	"Phantastes from 'their fount' all shapes deriving, In new habiliments can quickly dight."
gutenberg	high	16.75	Ahab and Pip CHAPTER CXXXI.— The Hat CHAPTER CXXXII.— The Pequod meets the Delight CHAPTER CXXXIII.—The Symphony CHAPTER CXXXIV.—The Chase.
gutenberg	high	17.25	Else, to install TeX Live manually, go to http://mirror.ctan.org/systems/texlive/tlnet and download the file install-tl-unx.tar.gz.
gutenberg	high	18.40	From http://gnuwin32.sourceforge.net/packages/groff.htm download _Complete package, except sources - Setup_ (groff-1.20.1-setup.exe) and run it.
gutenberg	high	18.94	Schwartau explains complex technology facilely and without condescension.
gutenberg	high	20.72	Go to http://mirror.ctan.org/systems/texlive/tlnetanddownload the file install-tl.zip.
gutenberg	high	24.56	From http://pypi.python.org/pypi/setuptools download setuptools-0.6c11.win32-py2.7.exe (or any newer version) and run it.
gutenberg	high	41.85	Long installation instructions can be found at http://www.tug.org/texlive/doc/texlive-en/texlive-en.html#x1-150003