サプライザルを利用した日本語の流暢性フィルタリングの試み

田村 鴻希 土井 惟成 西田 直人 Junjie Chen 谷中 瞳東京大学

tamura-k@tkl.iis.u-tokyo.ac.jp doi@cl.rcast.u-tokyo.ac.jp
nawta@g.ecc.u-tokyo.ac.jp {christopher, hyanaka}@is.s.u-tokyo.ac.jp

概要

機械翻訳の分野では、Web サイトから大量の対訳を自動収集し、対訳コーパスを作成する技術について研究が進められている。しかし、自動収集した対訳コーパスには不適切な対訳が多く含まれることがあり、不適切な対訳は対訳コーパスを用いて学習した機械翻訳モデルの品質に影響を与えうる。このような問題を解決するため、近年では、対訳コーパスから不適切な対訳を取り除く手法に関する研究が追められている。ここで、不適切な対訳の一つとして流暢性の低い対訳が考えられる。本研究では、単語や文のサプライザルが読みやすさを決める主要因であるとするサプライザル理論に基づき、サプライザルが流暢性と相関するという仮説を立て、その検証を行う。実験では、サプライザルが流暢性と相関するという有意な証拠は得られなかった。

1 はじめに

現在の機械翻訳では、対訳コーパスを利用して教師ありの手法で学習する、ニューラル機械翻訳(以下、NMT)[1]が主流である。NMTモデルの学習には大規模な対訳コーパスが必要であり、これを作成するためにWebサイトから大量の対訳を自動収集して対訳コーパスを作成する研究が行われている[2,3]。しかし、自動収集した対訳コーパスには不適切な対訳が多く含まれてしまうという問題がある。NMTモデルは、学習時の対訳コーパスの質と量に大きな影響を受ける傾向があり、不適切な対訳を含む文を学習に用いると翻訳精度が低くなることが知られている[4]。このような問題を解決するため、近年では、対訳コーパスから不適切な対訳を取り除く手法に関する研究が進められている[5]。

不適切な対訳の例として、流暢性の低い文を含む 対訳が挙げられる。一般社団法人日本翻訳連盟の JTF 翻訳品質評価ガイドライン[6]では、翻訳に含まれるエラーの度合いで評価指標の基準を定めた JTF 翻訳品質評価モデルを提供している。「流暢さ」¹⁾は JTF 翻訳品質評価モデルのエラー・カテゴリーの一つであり、「あるテキストが形式として整っているかの程度」と定義されている。流暢性が低い対訳を NMT モデルの学習に用いると翻訳の品質や精度に影響を与える可能性があるため、対訳コーパスから流暢性の低い文を取り除く手法が求められる。

本研究では、流暢性の低い和文を取り除く手法として、文の読み時間との相関が指摘されている、サプライザル理論に基づくサプライザル [7,8,9] の利用を検討する。サプライザルは、対象の単語に対して、その文脈における負の対数確率により、文の処理に関する負荷をモデル化した指標である。このサプライザルが流暢性と相関するという仮説を立て、流暢性が多岐にわたる和文を収集し、サプライザルの高低から流暢性の傾向を分析した。

実験は、機械的評価による実験と、人手評価による実験から構成される。前者では、英文と和文のサプライザルの分布を分析する。また、これらのサプライザルが他の言語的なパラメータと相関関係を持つか分析する。後者では、英文と和文のサプライザルの差に基づいて作成したデータセットの文の流暢性を複数の作業者が人手で評価し、サプライザルと流暢性が相関関係を持つか分析する。

2 関連研究

サプライザルは、心理言語学での活用だけでなく、対話の容認性評価 [10]、小論文の自動採点 [11]、英作文のレベル判定 [12]、Code-Switching の判定 [13] など自然言語処理の応用タスクの評価指標としても用いられている。既存研究を踏まえて、本研究ではサプライザルを対訳コーパスの流暢性フィルタリン

¹⁾ 本稿における「流暢性」と「流暢さ」は、同義と見なす。

グの指標として適用できるか分析を行う。

流暢性を自動評価する試みは今までにも行われている [14,15] が、サプライザルとの相関を分析する既存研究は存在しない。本研究と近い研究としては、サプライザルとの相関が報告されている読み時間を自然言語生成の文評価に用いるもの [16] が存在するが、人手での流暢性評価との相関を示すものではない。

3 実験用データセット

3.1 日英対訳コーパス

本 研 究 で 利 用 す る 対 訳 コ ー パ ス と し て、JParaCrawl[3] の English-Japanese training set (v3.0) [17] (以下, JParaCrawl v3.0) を採用する。この理由として、JParaCrawl v3.0 は、和文と英文の掲載元 Web サイトのドメインが併記されていることや、和文と英文のアライメントの精度が十分高いと考えられることが挙げられる。

3.2 ドメインによる整理

本研究では、英語圏のドメインの Web サイトに記載されている英文は、ネイティブによって書かれており、流暢性は十分に高い、また、その英文に紐付いている和文は、必ずしも流暢性が高いとは言えない、という仮説のもと、英文が英語圏のドメインのWeb サイトに掲載されている対訳を収集した。具体的には、英文が掲載されていた Web サイトのトップドメインが、主な英語圏の国別コードトップレベルドメイン等である対訳を抽出した。対象とした国別コードトップレベルドメインと、その内訳は付録の表7に示す。なお、英文と和文が重複している対訳は削除し、1つの対訳とした。この結果、128,602 対訳のデータセット(以下、本データセット)を作成した。

3.3 サプライザルの計算

サプライザルとは、単語や文のサプライザル(– *log*(単語や文 | 先行文脈))が読みやすさを決める主要な要因であるとするサプライザル理論において用いられている指標である [7,8].

本データセットの和文と英文を対象に、各文のサプライザルを算出した。サプライザルは、 Kuribayashi et al.[18] のコード²⁾により、トークンご

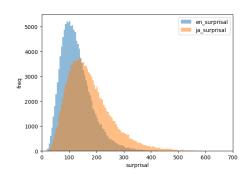


図1 和文と英文のサプライザルの分布

とのサプライザルを算出し、その総和を指標として 用いた。なお、和文のサプライザルの計算モデルに は Japanese LSTM LM、英文のサプライザルの計算 モデルには English Transformer-sm LM を用いた³⁾。

4 実験 1:機械的評価による実験

実験1ではサプライザルが言語的なパラメータと どのような相関を持つか機械的に分析する。

4.1 サプライザルの分布

和文と英文のサプライザルの分布を図1に示す。 また、和文と英文のサプライザルに対してt検定を 行ったところ、和文サプライザルの平均値の方が大 きい(p値は0.001未満)ことがわかった。このこと から、和文と英文では、サプライザルの分布が異な るといえる。

4.2 サプライザルと言語的パラメータの 相関

本研究では、言語的パラメータとして、和文と英文の一致度を対象に分析するため、日英機械翻訳モデルによる和文の英訳文と参考訳(英文の原文)のBLEU スコア [19] と、日英の分散表現のコサイン類似度(以下、USE スコア)を用いた。日英機械翻訳モデルは、opus-mt-ja-en ⁴⁾を使用した。また、テキストの分散表現の算出にはTensorFlowのEmbeddingメソッド⁵⁾を使用し、分散表現のモデルにはTensorFlowのuniversal-sentence-encoder-multilingual⁶⁾を使用した。あわせて、サプライザルの算出時の日英のトークン数を比較の対象と

²⁾ https://github.com/kuribayashi4/surprisal_reading_

time_en_ja

³⁾ https://github.com/kuribayashi4/surprisal_reading_ time_en_ja/releases/tag/v0.2

⁴⁾ https://huggingface.co/Helsinki-NLP/opus-mt-ja-en

⁵⁾ https://www.tensorflow.org/api_docs/python/tf/keras/ layers/Embedding

⁶⁾ https://tfhub.dev/google

表1 サプライザル、トークン、言語的パラメータ間の相関係数(p値はいずれも 0.001 未満)

	Ja Surprisal	Ja Token	En Surprisal	En Token	BLEU Score
Ja Token	0.957				
En Surprisal	0.846	0.843			
En Token	0.875	0.883	0.934		
BLEU Score	-0.120	-0.101	-0.184	-0.132	
USE Score	0.017	0.028	-0.012	-0.018	0.290

した。

言語的パラメータとの相関分析の結果を、表1に示す。なお、p値はいずれも0.001未満であった。まず、和文と英文のサプライザルは、トークン数に強く相関することがわかる。一方で、和文と英文の一致度に関連すると考えられる、BLEU スコアとUSE スコアはいずれも、和文と英文のサプライザルとの相関は認められなかった。ただし、BLEU スコアとUSE スコアは互いに一定程度の相関が認められることから、文の一致度の尺度としては十分に機能していると考えられる。このことから、機械的評価からは、和文と英文のサプライザルは文の一致度とは相関しないことが考えられる。

5 実験 2: 人手評価による実験

実験2では、本データセットを対象に、サプライザルの傾向が異なる3種類のデータセットを作成し、それぞれに含まれる和文の流暢性を人手で評価し、その傾向の差異について分析する。

5.1 流暢性評価データセットの作成

まず、和文と英文のサプライザルの傾向の差異を踏まえ、下記のとおりサプライザルの順位の差に基づく3種類のデータセットをそれぞれ100文対ずつ作成した。なお、順位はサプライザルが高いほど小さい値をとる。

- (英文サプライザル順位–和文サプライザル順位) の値が低い 100 文対 (en_high)
- (和文サプライザル順位-英文サプライザル順位) の値が低い 100 文対 (ja_high)
- 両サプライザルの順位差が小さい順に並べた 100 文対 (sim)

ただし、en_high と ja_high については、翻訳のエラーによる外れ値の影響を軽減するため、サプライザル順位差の上位 5%を除外した残りのデータからそれぞれ上位 100 文対を選んだ。これらの各 100 文対を対象に、(1) 対訳エラーの判定と (2) 流暢性評価を行った。具体的には、それぞれの文対に対して

表2 流暢性の評価基準

⇒π: / π *	++ v/H:
= 11/. / (1111)	_B_(TE.

- 5 日本語として完璧である
- 4 日本語母語話者が言っていても違和感がない程度
- 3 非日本語母語話者が言っていても違和感がない程度
- 2 日本語の構文として不適切である部分が多い
- 1 理解不能である

3人の作業者を割り当て、(1)-(2)を文対ごとに手作業で実施した.作業者は、いずれも日本語を母語とする情報系の大学院生であり、作業時間は合計 10.8時間(一人当たり平均 3.6時間)だった。以下では、それぞれの作業手順の詳細を述べる。

(1) 対訳エラーの判定 各作業者は、各データセットの文対に対して、実際に対訳として成立しているか否かを判定する。流暢性によらず翻訳エラーを含む文対は、対訳データセットの流暢性を評価する際にノイズとなるため、それぞれのエラータイプを分類したうえで流暢性評価の対象から外す。

以下に、本研究で用いたエラータイプの分類と、 各エラーの内容を記述する。

- アライメントエラー:全く別の文が割り当てられている場合
- テキスト抽出エラー:途中で文が切れていたり、複数の文が混ざっている場合
- 翻訳エラー: 誤訳の程度が非常に大きい場合
- 英文過多:英文の各要素に対応する和文の要素 が著しく欠けている場合
- 和文過多:和文の各要素に対応する英文の要素 が著しく欠けている場合

なお、翻訳エラーの基準には、特許庁の機械翻訳に関する調査報告書「特許文献機械翻訳の品質評価手法に関する調査」(平成25年度)⁷⁾の「内容の伝達レベルの評価」を使用し、5段階評価のうち2以下(2はいくつかの重要情報が正確に伝達されている、1は文意がわからない、もしくは正確に伝達されている重要情報はほとんどない)であると作業者が判断したものを翻訳エラーとしている。

https://www.jpo.go.jp/system/laws/sesaku/ kikaihonyaku/document/kikai_honyaku/h25_01.pdf

表3 流暢性評価データセットの各作業者がエラー判定した文対の数の平均

データセット	アライメントエラー	テキスト抽出エラー	翻訳エラー	和文過多	英文過多
en_high	4.33	0	4.33	1.00	9.67
ja_high	4.33	0	3.67	4.67	0
sim	2.00	1.00	1.67	1.33	0.33

表 4 作業者間の流暢性評価の一致度(Cohen's Kappa)

作業者ペア	一致度
1, 2	0.438
1, 3	0.695
2, 3	0.572

表 5 作業者間の流暢性評価に対する Kruskal-Wallis 検定

作業者	p 1但
1	1.60×10^{-5}
3	0.0671

(2) 流暢性の判定 (1) の手順で除外しなかった文 対を対象に、それぞれの和文の流暢性を判定する。 流暢性評価の基準は、表 2 に示すとおり、長我部ら による「流暢性の評価基準」[20] を利用した。なお、 表 2 の各評価に対応する、全作業者の評価が一致し た例文を、付録の表 8 に記載する。

5.2 流暢性人手評価の統計情報

流暢性評価データセットで各作業者がエラー判定した文対の数の平均と、その内訳を表3に示す。表3から、英文のサプライザルが高い文対では英文からの翻訳抜けが多く、また和文のサプライザルが高い文対では和訳に余分な情報が入りやすい傾向がみられる。これは、表1から読み取れるサプライザルとトークン数との相関を裏付ける結果である。

流暢性評価データセットでの、作業者間の流暢性評価の Cohen's Kappa[21] による一致度を表 4 に示す。なお、エラー判定の場合も含めて一致度を計算するため、対訳エラーと判定された文対は評価ラベルを 0 として一致度を計算した。また、流暢性評価の差の大きさを考慮するため、各文対の評価の差をそれぞれ二乗して重みづけした [22]。

表4に示した一致度をみると、作業者2の一致度 が低く、作業結果が一貫していないと考えられる。 したがって、本研究での流暢性評価では作業者1お よび作業者3の結果のみを用いる。

5.3 サプライザルと流暢性人手評価の相関

作業者1と作業者3の、各データセットでそれぞれエラーと判定しなかった文対への流暢性評価に対するKruskal-Wallis 検定[23]の結果、および文対への平均評価値をそれぞれ表5と表6に示す。表5か

表6 各作業者がそれぞれエラー判定しなかった文対での 平均評価値

作業者	en_high	ja_high	sim
1	4.54	3.92	4.38
3	4.4	4.16	4.45

らは、有意水準 5%の下で、作業者 1 の結果では有意に各データセットの評価値の分布に差があるといえる。しかし、作業者 3 の結果では帰無仮説を棄却できず、作業者によってサプライザルと流暢性の主観評価の関係にはばらつきがあることが考えられる。ただし、現在の p 値がデータ数の少なさに依存する可能性を注記する。また、平均評価値からは、評価値の分布に差があるとすればどちらの作業者の場合も和文のサプライザルが高いデータセットでの評価値の低さに原因を求められる。

本実験の結果からはサプライザルと流暢性の相関関係は定かにはならなかった。サプライザルと流暢性評価の相関関係の個人間のばらつきは、流暢性の判断基準として(サプライザルと相関する)読解時間に重点を置いているかにばらつきがあることを示しており、流暢性評価の基準は個人差が大きいことを示唆している。この結果、および一致度の低い作業者の存在は、より個人差の少ない流暢性評価の基準、もしくは評価手法の必要性を示唆している。また、現在の手法において精度の高い結果を得るためには多人数、およびより多い文対での実験が必要といえる。

6 おわりに

本研究では、サプライザルが流暢性と相関するという仮説を立て、機械的評価と人手評価の2種類の実験で検証を行った。実験の結果、サプライザルが流暢性の低い和文を取り除くためのパラメータとして有意であることは確認できなかった。今後の課題としては、人手評価の作業者を増やし、評価する文対の数を増やすこと、また、流暢性やエラー判定の作業者間での判定基準をより詳細に定めることによって作業者間での評価基準のぶれやその影響を減らし、より信頼性の高い統計とすることが挙げられる。

謝辞

本研究は JSPS 科研費 JP20K19868 の助成を受けた ものです.

参考文献

- [1] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In **Proceedings** of the 27th International Conference on Neural Information Processing Systems Volume 2, NIPS'14, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.
- [2] Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks, pp. 118–119, Dublin, Ireland, August 2019. European Association for Machine Translation.
- [3] Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Proceedings of The 12th Language Resources and Evaluation Conference, pp. 3603– 3609, Marseille, France, May 2020. European Language Resources Association.
- [4] Huda Khayrallah and Philipp Koehn. On the impact of various types of noise on neural machine translation. In Proceedings of the 2nd Workshop on Neural Machine Translation and Generation, pp. 74–83, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [5] Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In Proceedings of the Fifth Conference on Machine Translation, November 2020.
- [6] 一般社団法人日本翻訳連盟. JTF 翻訳品質評価ガイドライン. 2018.
- [7] John Hale. A probabilistic Earley parser as a psycholinguistic model. In Second Meeting of the North American Chapter of the Association for Computational Linguistics, 2001.
- [8] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, p. 1126–1177, 2008.
- [9] Irene Fernandez Monsalve, Stefan L. Frank, and Gabriella Vigliocco. Lexical surprisal as a general predictor of reading time. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 398–408, Avignon, France, April 2012. Association for Computational Linguistics.
- [10] Investigating perception of spoken dialogue acceptability through surprisal. In Hanseok Ko and John H. L. Hansen, editors, **Proceedings of Interspeech 2022**, pp. 4506– 4510. ISCA, September 2022.
- [11] Gaurav Kharkwal and Smaranda Muresan. Surprisal as a predictor of essay quality. In Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 54–60, Baltimore, Maryland, June 2014. Association for Computational Linguis-

- tics.
- [12] 正頼林, 遼平笹野, 大也高村, 学奥村. 誤り傾向と文の容認性に着目した英作文のレベル判定. 第 226 回情報処理学会自然言語処理研究会, No. 7, pp. 1–7, 2016.
- [13] Jesús Calvillo, Le Fang, Jeremy Cole, and David Reitter. Surprisal Predicts Code-Switching in Chinese-English Bilingual Text. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4029–4039, Online, January 2020. Association for Computational Linguistics.
- [14] Takehiko Yoshimi, Katsunori Kotani, Takeshi Kutsumi, Ichiko Sata, and Hitoshi Isahara. Automatic evaluation of the fluency of english-to-japanese machine translation using word alignment. 2010.
- [15] Katharina Kann, Sascha Rothe, and Katja Filippova. Sentence-Level Fluency Evaluation: References Help, But Can Be Spared! In Proceedings of the 22nd Conference on Computational Natural Language Learning, pp. 313–323, Brussels, Belgium, October 2018. Association for Computational Linguistics.
- [16] Sina Zarrieß, Sebastian Loth, and David Schlangen. Reading Times Predict the Quality of Generated Text Above and Beyond Human Ratings. In Proceedings of the 15th European Workshop on Natural Language Generation (ENLG), pp. 38–47, Brighton, UK, September 2015. Association for Computational Linguistics.
- [17] Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6704–6710, Marseille, France, June 2022. European Language Resources Association.
- [18] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5203–5217, Online, August 2021. Association for Computational Linguistics.
- [19] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. pp. 311–318, 2002.
- [20] 長我部恭行, 甲斐優人, 石井奏人, 荻野天翔, 黒澤道希, 小町守. 機械翻訳に対する文間文脈を考慮した評価と分析. pp. 1703-1706, 3 2019.
- [21] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement, Vol. 20, No. 1, pp. 37–46, April 1960. Publisher: SAGE Publications Inc.
- [22] Jacob Cohen. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, Vol. 70, No. 4, pp. 213–220, October 1968.
- [23] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, Vol. 47, No. 260, pp. 583–621, 1952

A JParaCrawl の前処理

JParaCrawl v3.0 は非常に大規模なコーパスであるため、前処理として、下記の条件を元に対訳のフィルタリングを行った。フィルタリングの結果、約 2574 万文対から、約 985 万文対(約 38.3%)が得られた。

- 英文がピリオドで終わっている.
- 和文が句点で終わっている.
- 和文に全角又は半角スペースを含まない.
- 対訳のスコアが 0.7 以上.

B ドメインの内訳

イタリア、フランスは英語を公用語としていないが、話者数の多さを考慮して含めている。

	表 7 ドメインの内訳	
ドメイン	国名	対訳数
.edu	アメリカ(教育機関)	26,688
.uk	イギリス	24,041
.gov	アメリカ(政府)	15,674
.it	イタリア	12,588
.fr	フランス	12,019
.eu	欧州連合(EU)	11,971
.au	オーストラリア	8,896
.us	アメリカ	5,345
.ca	カナダ	3,519
.sg	シンガポール	3,456
.in	インド	2,471
.nz	ニュージーランド	1,466
.ph	フィリピン	375
.pk	パキスタン	93
合計		128,602

C 流暢性の評価点数ごとの例文

表8 流暢性の評価点数ごとの例文 (JParaCrawl v3.0 より引用)

評価 例文

- 5 京丹後市は条例で国際交流の推進をうたっており、国際的な視野を持つ人材の育成を目指している。
- 4 調査では、大手プラットフォームひとつで 157 万 6,600 人の労働者が供給過剰であることがわかっている。
- 3 プレートは、全体の保護面全体に影響の分布を保証するドロップワイヤーシステムに参加しています。 プロフェッショナルラーニングコミュニティ内では、各学年レベルとコンテンツ領域が共同学習チーム (CLT) として協力し、研究、スキル、専門的な開発、ツール、リソース、およびさまざまなサポートスタッフ
- 2 (強力な教育コーチチームやさまざまな特別教育の専門家とサービスプロバイダーの-私たちの目標を果たすために: すべての子供が意味のある、本物の方法ですべての必要なスキルを習得することを保証すること。
- 1 米国の研究では、女性の最大50%が生涯に尿路感染症を起万人になると示されています。