言語一般の計測を目指して: サブワードと分散意味論に基づく言語の複雑性計測

中山拓人
慶應義塾大学大学院・日本学術振興会特別研究員 DC
tnakayama.a5ling@gmail.com

概要

本研究は，「あらゆる言語は等しく複雑である」という，いわゆる言語の等複雑性について，単位形式系列当たりの語義数とその出現確率の偏りを通して，言語の複雑性を計測を試みる．このために，BERT を用いたサブワードへの分割や分散表現を利用している．12 言語[1)]を対象とし，ランダムサンプリングした Wikipedia の 100 記事をデータとして，各言語につき 4 回の分析を行った．

結果として，全体的には単位形式系列当たりの語義数に大きな差が見られなかった．一方で 4 回の分析に渡って，各言語はそれぞれ非常に近い値を示すことが多かった．このことから言語の複雑性は，非常に狭い領域に収まり理ながら，その領域内で各言語が特有の傾向を持つことが示唆される．

1 はじめに

「あらゆる言語は等しく複雑である」という，いわゆる言語の等複雑性については，20 世紀中に多くの言語学者が言及しており，それ以来，言語学者の間で広く受け入れられてきた．しかし，この主張は確固たる証拠に基づいて提唱されたわけではなく，経験的に信じられてきた一種のドグマであると指摘する声もある [1]．近年では，コンピュータ技術の進歩により，大規模データを用いた計算的手法による言語の複雑性の計測が数多く行われており [2][3]，この分野の議論が特に活発化している．一方で，「あらゆる言語は等しく複雑か？」という問いについては，依然として明確な合意には至っていない．

言語の複雑性研究における大きな課題の 1 つとして，複数の言語間で妥当な比較が可能な計測手法が未確立である点が挙げられる．例えば，単語を基準にした複雑性を考えようとすると，分かち書きをしない言語では単語の境界が明確でない場合がある．また，そのような言語で何らかの方法で “単語” を定義したとしても，それに基づいて得られた複雑性の数値を他言語と比較することが妥当であるかは自明ではない．さらに，言語の複雑性研究では，形式的側面に比べて意味的側面の研究があまり進んでいないことも課題である．これは意味の側面を考慮しようとすると，コンピュータによる自動処理を基盤とした大規模な調査が，より難しくなるためである．この課題の解決を目指し，本研究は形式–意味対応の複雑性を言語の複雑性の一要素と捉え，その言語間比較を試みる．具体的な実装としては，BERT によるサブワードトークンへの分割とその分散表現を利用することで，形式意味対応の複雑性を計測した．

以下では，第 2 節で，言語の等複雑性がこれまでどのように扱われてきたかを概観し，その問題点を指摘する．第 3 節で，本研究が行った分析の手法を説明し，第 4 節で，分析の結果，及び考察を行う．

2 先行研究

言語の等複雑性に関する最初期の言及としてよく挙げられるのが, Edward Sapir[4] による次の言葉である: “When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam.” また，その後 Charles Hockett[5] による次の発言も広く知られている: “Impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other.” この Hockett の発言は，形態論的複雑性と統語論的複雑性の間にトレードオフ関係が存在するという議論の端緒ともなった．しかし，“impressionistically” という表現が示す通り，

1) 英語，フランス語，フィンランド語，ヘブライ語，ヒンディー語，ハンガリー語，カザフ語，ロシア語，スワヒリ語，タガログ語，トルコ語，ウルドゥー語である．後述の通り，単語を分ち書きする表記システムを持つ言語を対象とした．

この時点ではその主張は経験的観察に基づくものであり，明確な証拠に裏付けられていたわけではなかった．

その後，言語の特定の領域間でのトレードオフ関係については，[6] によっても言及されている．彼女らは音素と音節の間のトレードオフ関係に注目し，音節中の音素数が多い言語ほど文中の音節数が少なく，逆に文中の音節数が多い言語ほど音節中の音素数が少なくなることを指摘した．このように，言語の 1 つの領域で複雑性が増すと，それを補うかのように別の領域で複雑性が減少する傾向，または性質を「言語の等複雑性」と総称している．

形態論的複雑さと統語的複雑さのトレードオフ関係について，[7][8] は情報理論の視点からコルモゴロフ複雑性を取り入れた観点を提案している．具体的には，テキストファイルを zip ファイル化した際に，どれだけ圧縮されるのかを見ることで，コルモゴロフ複雑性を近似する手法である．また，一部をランダムに削除したテキストファイルを zip ファイル化した際の圧縮率を見ることで，形態論的，及び統語的複雑性も計測することができるとしている．例えば，文字単位でランダム削除を行ったテキストファイルの圧縮率と，元テキストファイルの圧縮率を比較した時，その差が大きいほど元テキストの文字列が規則的であったことを意味するため，形態論的に単純であることを意味する．また，単語単位でランダム削除を行った場合では，圧縮率の差が大きいほど元テキストファイルの単語系列が規則的出会ったことを意味するため，統語的に単純であることを意味するとしている．[2] はこの尺度を用いて英語の通時的変化および多言語間の共時的比較を行い，形態論的複雑性と統語的複雑性の間にトレードオフ関係が存在することを示した．

また，シャノンの情報エントロピーを用いた研究も広く行われている．[9] は，エントロピーを使って単語選択の予測不能性を計算し，言語間の差が非常に小さいことを明らかにした．一方で，[3] は，形式の出現頻度に基づくエントロピーを多言語間で比較する分析を，複数のコーパスに渡って行った．結果として，あるコーパスで観察されたエントロピーの順位が，別の子＝パスでも同様に観察されることを示した．このことから，言語間の複雑性には有意な違いがある，即ち，言語の等複雑性は誤った言説であるという可能性が示唆された．

このように計算的手法を用いた複雑性の計測は進展しているものの，その多くは言語の形式的な側面に限られており，意味の側面を考慮した複雑性の計測はほとんど行われていない．考えられる原因の 1 つとして，コンピュータ処理の難しさが挙げられる．意味の側面を考慮するためには，文字列のみを処理する場合と比べて，非常に大きな計算リソースを使用する必要があるため，大規模な分析の実施のハードルが高い．

また，意味というものの複雑さに関する概念が，不足していることも原因の 1 つであると考えられる．形式系列の複雑さに関しては，上記の通りさまざまな知見が蓄積されており，別分野からの概念を借用することも，比較的容易に行える．その一方で，形式系列の複雑さに関してでさえ，どのような尺度が最も妥当であるのかに関しては，議論の余地がまだあるため，意味の側面に関する分析は遅れをとっている状況である．加えて，情報理論において，歴史的に意味の側面が捨象されてきたことも影響していると考えられる．Shannon は次のように述べており，意味が情報理論の枠組みで扱われない理由を説明している [10].

> Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

3 方法論

本研究は，ある形式系列が，幾つの語義に対応しており，さらにその中の特定の語義として使用される確率がどれほどかを見ることで，意味の側面を考慮した上での言語の複雑性計測を試みる．これを実装するためには，一定の基準で分割された単位形式系列，そしてその各々に対応する語義数を推定することの，2 点が必要となる．本研究の実装の実装を以下で説明する．Algorithm 1[2] は，擬似コードによる実装の概要である．

2) 実際に分析に使用したコードは，https://github.com/takuto-nakayama/subword-polysemy/tree/tsne にて，公開している．

Algorithm 1

Input: T := n random articles from Wikipedia
Output: H := Shannon entropy

for $i \leftarrow 1$ to n **do**
  $Token \leftarrow$ BertTokenizer(T_i)
  $Type \leftarrow set(Token)$
  for each $type_j \in Type$ **do**
    for each $token_k \in Token$ **do**
      if $type_j = token_k$ **then**
        $emb_k \leftarrow$ BertEmbedder($token_k$)
        dict $\leftarrow \{type_j{:}[emb_k, \dots]\}$
      end if
    end for
    $tsne_j \leftarrow$ tSNE(dict[$type_j$])
    $clusters_j, n\_meanings \leftarrow$ DBSCAN($tsne_j$)
    for $l \leftarrow 1$ to $n\_meanings$ **do**
      $p_l \leftarrow \frac{\mathrm{num}(clusters_{jl})}{\mathrm{num}(tsne_j)}$
    end for
    $H_j \leftarrow \sum p_l \log_2 p_l$
  end for
  $H \leftarrow \sum H_j$
end for
return H

3.1 データ

対象とする言語は，表 1 に示した 12 言語である．後述のサブワード分割に関して，単語を基準に分割するため，分ち書きにより明確に「単語」という単位が定義されている言語と，そうでない言語で差が生まれる可能性がある．そのため本研究では，対象を単語の分ち書きを行う言語に限定した．

使用したデータは，ランダムサンプリングした Wikipedia の記事である．本研究では，各言語ランダムに抽出された 100 記事を対象に，同様の分析を計 4 回ずつ行った．

3.2 サブワード分割・分散表現

多言語に関して事前学習を行った BERT の学習済みモデル [11][3] を用いて，得られたテキストデータのサブワード分割，及び分散表現の計算を行った．BERT モデルを使用する理由は，後述の語義数推定のために，文脈を考慮した分散表現をトークンごとに集め，それをタイプにおける使用頻度の偏りの計算に利用することを念頭に置いているからである．サブワードの単位は WordPiece であり，分散表現は，モデルの隠れ層の最終層を用いた．最終層の分散表現は，768 次元のベクトルとして表されており，この後のクラスタリングには次元が大きすぎるため，tSNE により次元圧縮を行った．

3) https://huggingface.co/google-bert/bert-base-multilingual-cased

3.3 語義数の推定

各トークンに対する分散表現を得た後，それをタイプごとにクラスタリングを行うことで，各サブワードにおける語義数の推定を行う．クラスタリングによって得られたクラスター数が語義数に相当し，各クラスターの成員数を元に，使用頻度の分布を観察できる．ここで得た分散表現は文脈の情報を考慮したものであるため，同音異義語の様な意味差だけでなく，より細かい意味の差も同時に考えることができる．本研究では，事前にクラスター数を決定しない手法が望ましいため，探索的にクラスターを見るつける DBSCAN を用いた．

3.4 シャノンエントロピー

シャノンの情報理論におけるエントロピーは，対象の確率を元に，ある事象の発生がどれだけ発生しづらいかを，定量化した値である．一般的に，i 番目の事象が発生する確率が p_i である時，エントロピー H は，以下の方程式で表される．

$$H = \sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

これを各タイプのクラスタリング結果に対して使用することで，そのタイプがどれだけ単一/複数の語義と結びついているかが得られる．最終的に，全てのタイプの値の平均を，その言語の値として比較に使用する．

4 結果・考察

結果は図 1 に示す通りであった．図 1 は，縦軸に各言語を，横軸にエントロピーを置いており，4 回の分析で得られた値の内，最大のものに関して順に並べている．

まず全体を見ると，最も大きい値が英語の ≈ 0.410 である．エントロピーは底が 2 である対数をとっているものであるから，即ち，サブワードにつき約 1.329 個の意味が対応していることになる．反対に

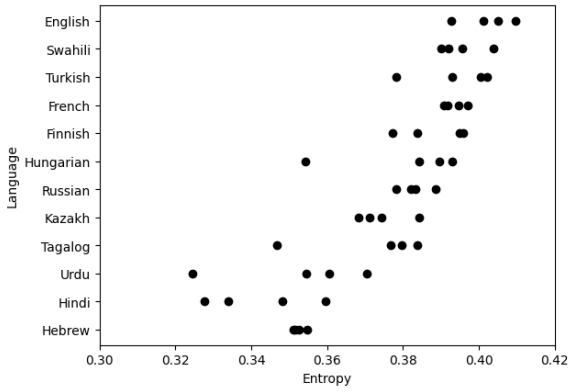

図 1

最も小さかった値が，ウルドゥー語の ≈ 0.324 であり，こちらはサブワード当たり平均で約 1.252 個の意味が対応していることになる．語義数の差を見ると，非常に狭い領域に全ての言語が収まっていると言える．

一方で，各言語の取る値に目を向けると，それぞれが非常に近い値を取るものが多くあることがわかる．このことから，言語の複雑性は，全体的には非常に狭い領域に収まっているが，同時にその狭い領域内では，各言語が特有の傾向を持っている，ということが示唆される．

5 結語

本研究では，意味の側面を考慮した言語の複雑性計測を目指し，BERT モデルを活用することで，単位形式系列当たりの語義数について，12 言語間の比較を行った．全体的には，この対応関係の複雑性に関して，大きな差が見られなかったが，各言語はそれぞれ非常に近い値を示すことが多かった，このことから示唆されたのは，言語の複雑性は非常に狭い領域に収ま理ながら，その領域内で各言語が特有の傾向を持つ可能性であった．

さらなる追研究として，より多くの言語を対象にすることが考えられる．特に本研究は，対象とする言語を単語を分ち書きするものに限定していたため，そうでない表記体系を持つ言語でも同様の傾向が見られるのか，分析する必要がある．

それに対して，今回行った分析の手法が，単語を分ち書きしない言語へもそのまま応用可能であるかは自明でないことが，課題として挙げられる．サブワードへの分割に際して，表記体系の差が異なる言語同士の妥当な比較に，どれだけ影響を与えるのかに関しても，より議論が必要になると考えられる．

謝辞

本研究は JSPS 科研費 JP24KJ193 の助成を受けたものです．

参考文献

[1] Fenk-Oczlon, Gertraud and Fenk, August. "Complexity trade-offs do not prove the equal complexity hypothesis." *Poznan Studies in Contemporary Linguistics*, vol. 50, no. 2, De Gruyter Mouton, 2014, pp. 145–155. URL: https://doi.org/10.1515/psicl-2014-0010.

[2] Ehret, Katharina and Szmrecsanyi, Benedikt. "An information-theoretic approach to assess linguistic complexity." In Baechler, Raffaela and Seiler, Guido (Eds.), *Complexity, isolation, and variation*, vol. 57, De Gruyter, 2016, pp. 71–94. URL: https://doi.org/10.1515/9783110348965-004.

[3] Koplenig, Alexander, Wolfer, Sascha, and Meyer, Peter. "A large quantitative analysis of written language challenges the idea that all languages are equally complex." *Scientific Reports*, vol. 13, no. 1, 2023, p. 15351. URL: https://doi.org/10.1038/s41598-023-42327-3.

[4] Sapir, Edward. *An introduction to the study of speech*. Citeseer, 1921.

[5] Hockett, Charles F. *A course in modern linguistics*. Macmillan, 1958.

[6] Fenk-Oczlon, G and Fenk, A. "The mean length of propositions is 7 plus minus 2 syllables—but the position of languages within this range is not accidental." *Cognition, information processing, and motivation*, 1985, pp. 355–359.

[7] Juola, Patrick. "Measuring linguistic complexity: The morphological tier." *Journal of Quantitative Linguistics*, vol. 5, no. 3, Routledge, 1998, pp. 206–213. URL: https://doi.org/10.1080/09296179808590128.

[8] Juola, Patrick. "Assessing linguistic complexity." In Miestamo, Matti, Sinnemäki, Kaius, and Karlsson, Fred (Eds.), *Language Complexity: Typology, Contact, Change*, 2008, pp. 89–108. URL: https://doi.org/10.1075/slcs.94.07juo.

[9] Bentz, Christian, Alikaniotis, Dimitrios, Cysouw, Michael, and Ferrer-i-Cancho, Ramon. "The Entropy of Words—Learnability and Expressivity across More than 1000 Languages." *Entropy*, vol. 19, no. 6, Multidisciplinary Digital Publishing Institute, 2017, p. 275. URL: https://doi.org/10.3390/e19060275.

[10] Shannon, C E. "A mathematical theory of communication." *The Bell System Technical Journal*, vol. 27, no. 3, Nokia Bell Labs, 1948, pp. 379–423.

[11] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Burstein, Jill, Doran, Christy, and Solorio, Thamar (Eds.), *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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://api.semanticscholar.org/CorpusID:52967399.