P4-23

2025/03/11
NLP2025@出島メッセ長崎

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）email: tnakayama.a5ling@gmail.com

Keio University
1858
CALAMVS GLADIO FORTIOR

1. 目的

- 形式-意味対応の複雑性(=多義性)を言語の複雑性の1つの側面と捉え，その言語間比較を試みる
- 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して，特定の言語に依存しない単位を基準とした計測を試みる

2. 背景

- 言語の等複雑性は20世紀以来，言語学者の間で受容されてきた[1][2]一方で，経験的に信じられてきた一種のドグマであると指摘されている[3]
- 妥当な言語間比較に関して，総意が得られていない
- これまでの計測手法の対象は，言語の形式的な側面に限られており，意味の側面は考慮されてこなかった

3. 方法論

Wikipedia記事をランダム抽出

- 1言語につき，100記事の分析を10回ずつ行った
- 空白で分ち書きを行う
64言語を対象とした

↓

WordPiece[4]にトークナイズ

↓

mBERT[5]による埋め込み表現

BERTによって得られる文脈を考慮した埋め込み表現を，クラスタリングすることで語義数を推定する

↓

DBSCAN[6]によるクラスタリング

↓

Shannon[7]のエントロピー

$$-\frac{1}{N}\sum_{i=1}^{n} p_i \log_2 p_i$$

N: 総サブワード数
n: あるサブワードの語義数
p_i: i番目の語義で使われる確率

4. 結果

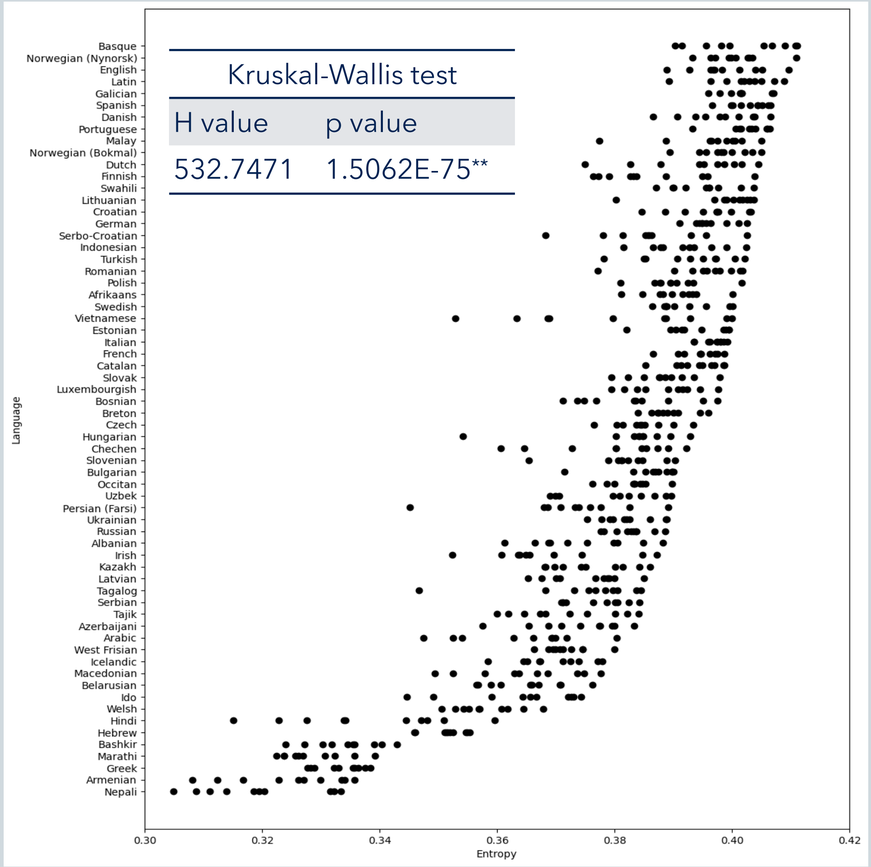

図1. 各言語のエントロピー

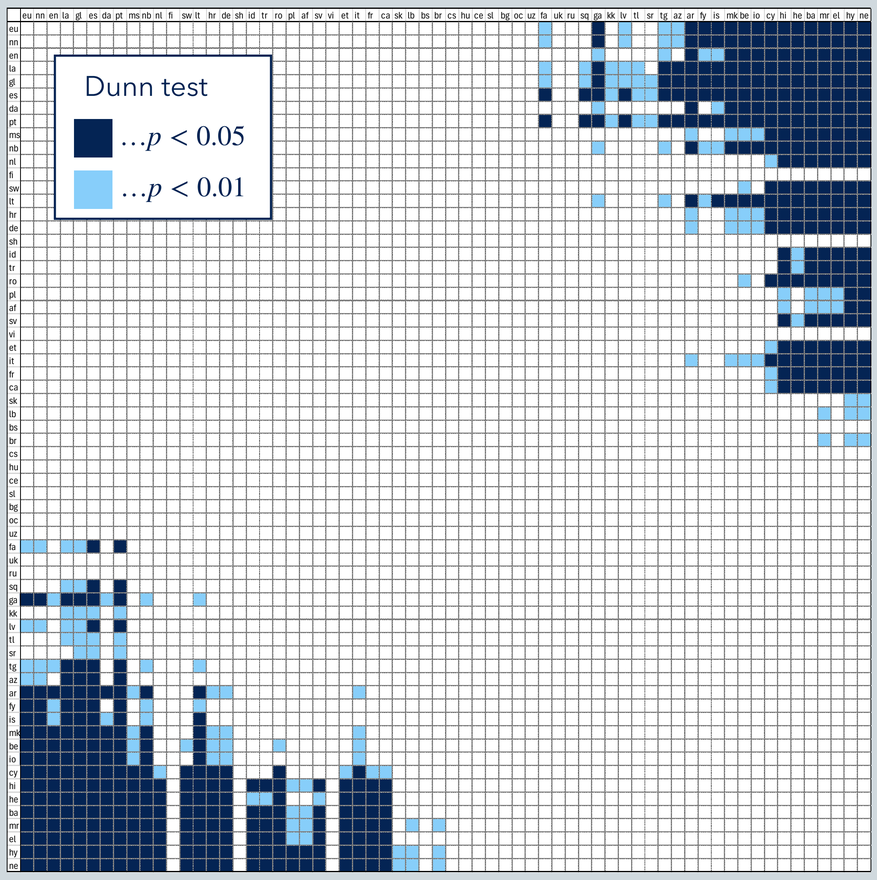

図2. 各言語間の分布の有意差

【図1】

- バスク語が最大
(≈ 0.411, ≈ 1.329/sw)
- ネパール語が最小
(≈ 0.305, ≈ 1.235/sw)
- 狭い範囲に全ての言語が収まっている
- 一方で，Kruskal-Wallis検定では有意差あり

【図2】

- Dunn検定により各言語間の有意差を見ると，上位と下位の言語間に有意差あり

5. 考察

- 「多義性が狭い範囲に収まっている」→等複雑性を支持
- 「その範囲内で有意差が見られる」→等複雑性を不支持

→これまでの研究が，言語の等複雑性に関して統一的な見解が得られない要因の可能性

6. 結論

- 各言語は狭い多義性の範囲に収まっている事が示唆された
- 同時に，その範囲において各言語間には差異があることも示唆された

【今後の課題】

- 分ち書きを行わない言語への適用可能性を探る
- 語義を離散的に扱っていることの妥当性が自明ではない

謝辞: 本研究は JSPS 科研費 JP24KJ1938 の助成を受けたものである.

参考文献:

[1] Edward Sapir. **An introduction to the study of speech**. Citeseer, 1921.
[2] Charles Hockett. **A course in modern linguistics**. Macmillan, 1958.
[3] Gertraud Fenk-Oczlon and August Fenk. Complexity trade-offs do not prove the equal complexity hypothesis. **Poznan Studies in Contemporary Linguistics**, Vol. 50, No. 2, pp. 145–155, 2014.
[4] HuggingFace, **WordPiece tokenization**. [Online]. Available: https://huggingface.co/learn/nlp-course/en/chapter6/6.
[5] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of naacL-HLT**, Vol. 1, p. 2. Minneapolis, Minnesota, 2019.
[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiawei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**, pp. 226–231, 1996.
[7] Claude Shannon. A mathematical theory of communication. **The Bell System Technical Journal**, Vol. 27, No. 3, pp. 379–423, 1948.