P6-26

言語は等しく複雑か？: 多義語埋め込表現による形式–意味対応の複雑性

中山拓人（慶應義塾大学）email: tnakayama.a5ling@gmail.com

Keio University 1858 CALAMVS GLADIO FORTIOR

1. 目的

- 「言語の等複雑性」に関して，形式-意味の対応関係に注目した計測法を用いて，多言語間の複雑性比較を行う.

2. 先行研究

- 言語の等複雑性については，未だ統一的見解が無い
  - 言語の複雑性をベクトルとして比較した結果，有意差が見られなかった[1]→**言語は等複雑である**.
  - 言語間における形式系列のエントロピー差が，異なるコーパスでも保存される[2]→**言語は等複雑でない**.
- これまでの研究では，形式面に注目するものがほとんどであった.

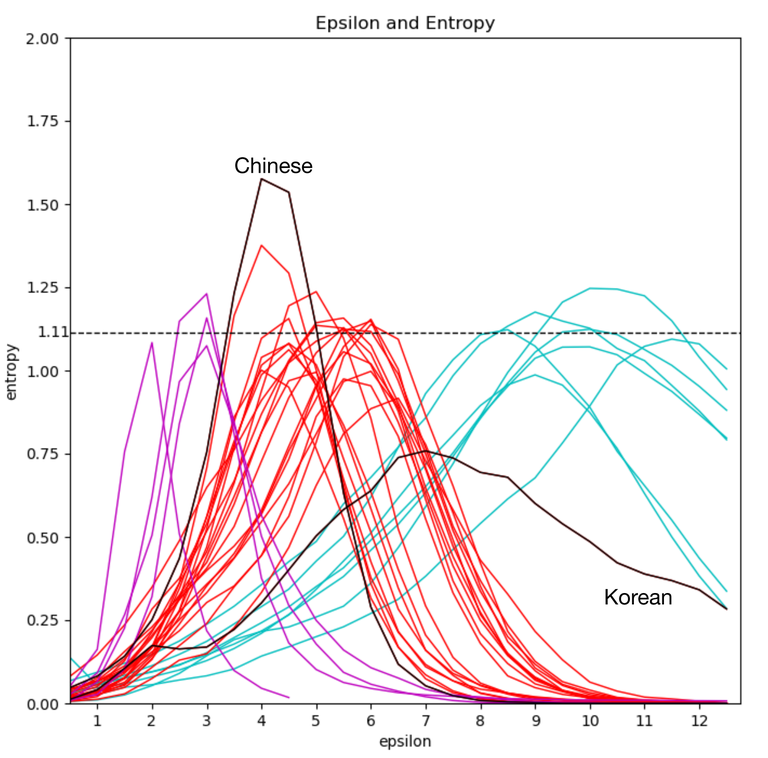
3.方法論

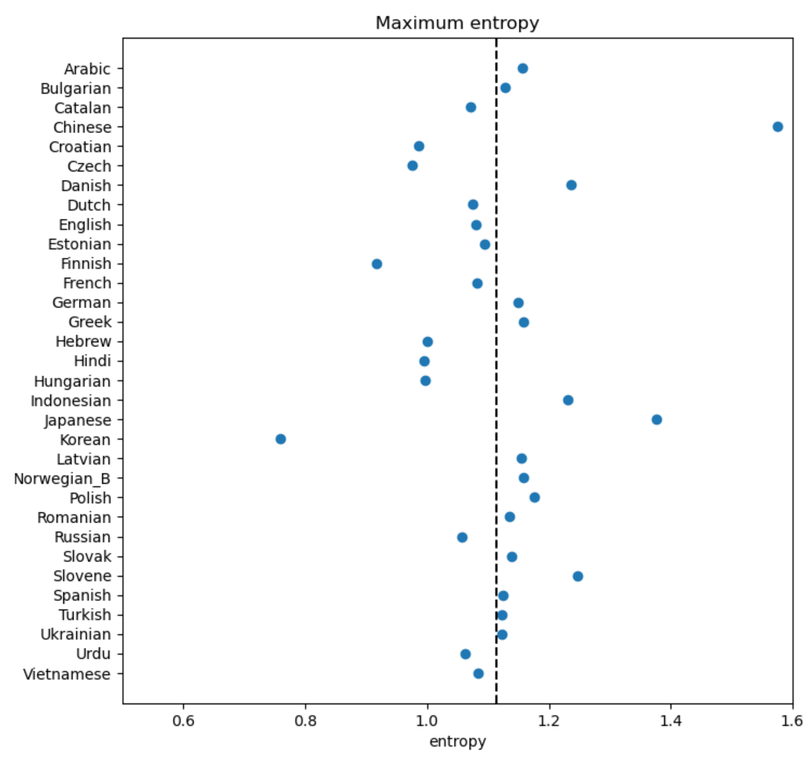
1. ある言語について，コーパスからランダムサンプリングした文から，特定のトークンを含む文を抽出し，そのトークンに対して文脈を考慮した埋め込み表現を得る.
2. 得られた埋め込み表現から，クラスタリングにより，そのトークンの語義数 (n) の推定を行う.
3. 各語義で使用される確率 ($p(m_i)$) から，そのトークンについてのシャノンエントロピー (H_t) [3]を得る.

$$H_t = -\sum_{i=1}^{n} p(m_i)\ \log_2\ p(m_i)$$

4. 1~3を，その言語において対象となるトークン全てについて行い，その平均を得る.
5. これを対象となる言語全てで行い，その平均エントロピーを比較する.

4.結果





- 対象としたトークンの単位は，単語とした．単語分割には，sentence-transformerの多言語学習済みのモデル [4]を用いた.
- 埋め込み表現については，多言語について学習済みのモデル [5][6]を用いた.
- クラスタリングはDBSCANを用いた．最低サンプル数は2，半径の値を0.5から13まで0.5間隔で計算した.
- 最大エントロピーとなる半径には，各言語で差が見られる.
- 各言語の最大エントロピーは，1.1付近に集中しているが，中国語と韓国語は大きく離れた値を示す.

- Dutch
- Indonesian
- Norwegian_B
- Vietnamese

- Arabic
- Bulgarian
- Chinese
- Czech
- Danish
- English
- Finnish
- French
- German
- Greek
- Hebrew
- Hindi
- Hungarian
- Japanese
- Korean
- Latvian
- Romanian
- Russian
- Slovak
- Ukrainian
- Urdu

- Catalan
- Croatian
- Estonian
- Polish
- Slovene
- Spanish
- Turkish

5. 考察

- 最低サンプル数の設定から，最大エントロピーを取ることは，クラスター数とその成員数が最大となることを意味し，その値が概ね一致している言語が多いと言える
  →**言語は概ね等複雑である**
- 外れ値となる言語 (中国語，韓国語) がある
  →**言語は完璧な等複雑ではない**

6. 結語

- 多くの言語の最大エントロピーが概ね同様の値を示していることから，「言語の等複雑性」はある程度正しいが，一方で外れ値を示す言語もあるため，強い性質とは言えないことも示唆される.
- 最大エントロピーを取る半径が言語によって異なる原因については，追調査が必要である.

謝辞: 本研究は，JST 次世代研究者挑戦的研究プログラ ム JPMJSP2123 の支援を受けたものです.

参考文献:

[1] Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardz˘ic´. Complexity trade-offs and equi-complexity in natural languages: a meta-analysis. **Linguistics Vanguard**, Vol. 9, pp. 9–25, 2022.
[2] Alexander Koplenig, Sascha Wolfer, and Peter Meyer. A large quantitative analysis of written language challenges the idea that all languages are equally complex. **Scientific Reports**, Vol. 13, , 2023.
[3] Claude E. Shannon. A mathematical theory of communication. **The Bell System Technical Journal**, Vol. 27, pp. 623–656, 1948.
[4] Nils Reimers, and Iryna Gurevich. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. https://aclanthology.org/D19-1410
[5] Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. Towards better UD parsing: Deep contextualized word embed- dings, ensemble, and treebank concatenation. In **Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**, pp. 55–64, Brussels, Bel- gium, 2018. Association for Computational Linguistics.
[6]Murhaf Fare, Andrey Kutuzov, Stephan Oepen, and Erik Vell- dal. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In **Proceedings of the 21st Nordic Conference on Computational Linguistics**, pp. 271– 276, Gothenburg, Sweden, 2017. Association for Computational Linguistics.

GitHub: https://github.com/takuto-nakayama/form_meaning_pairing/tree/main