Diversity for NLP: how to measure it and how it may help

Estève Louis

2024-06-19

Supervisors: Agata Savary & Thomas Lavergne
Department: STL
Team: SEME

Systems and F1

Low F1	High F1
System appears bad	System appears good

Systems and F1: diversity of gold

	Low F1	High F1
Low diversity **(gold)**	System appears bad, *but is it?*	System appears good, *but is it?*
High diversity **(gold)**	System is *in fact* bad	System is *in fact* good

Systems and F1: diversity of system predictions

	Low F1	High F1
Low diversity **(system predictions)**	System is *in fact* bad	System appears good, *but is it*?
High diversity **(system predictions)**	System appears bad, *but is it*?	System is *in fact* good

Example 1

"I just got of [1] the phone with Hai and he told me how to make [2a] an adjustement [2a] on a day to day basis in regards to incorrect liquidations but he also explained this is just to make the daily P&L #'s right, if nothing were done the month and P&L would still somehow work out [3] because adjustments [2b] would be made [2b].", email-enronsent44_01-0025 (typos from original text)

Example 1

"I just got of [1] the phone with Hai and he told me how to make [2a] an adjustement [2a] on a day to day basis in regards to incorrect liquidations but he also explained this is just to make the daily P&L #'s right, if nothing were done the month and P&L would still somehow work out [3] because adjustments [2b] would be made [2b].", email-enronsent44_01-0025 (typos from original text)

Example 2

"Does this mean that for June for a certain portion of July we should not do anything and just make adjustments [1] on a going forward [2] basis (and assume everything will work out [3] at month end)?", email-enronsent44_01-0026

Flavours of diversity

- Variety: how many types there are
- Balance: how even their distribution is
- Disparity: how fundamentally different they are

Lion-Bouton et al. [2022]

Flavours of diversity

- Variety: how many types there are
- Balance: how even their distribution is
- Disparity: how fundamentally different they are

Lion-Bouton et al. [2022]
$\rightarrow$ information theory to unify variety, balance, and even disparity

[Chao et al., 2014]

$$H^{\text{func}}_{\alpha\neq 1} = \left(\sum_{i,j=1}^{n} d_{ij} \times \left(\frac{p_i p_j}{Q}\right)^{\alpha}\right)^{\frac{1}{1-\alpha}}$$

$$\lim_{\alpha\to 1} H^{\text{func}}_{\alpha} = \left(\sum_{i,j=1}^{n} d_{ij} \times \left(\frac{p_i p_j}{Q}\right) \log_b \left(\frac{p_i p_j}{Q}\right)\right)$$

$$N^{\text{func}}_{\alpha\neq 1} = \left(\frac{H_\alpha}{Q}\right)^{\frac{1}{2}}$$

$$\lim_{\alpha\to 1} N^{\text{func}}_{\alpha} = b^{H^{\text{func}}_{\alpha}}$$

$$Q = \sum_{i,j=1}^{n} d_{ij} p_i p_j$$

Challenges

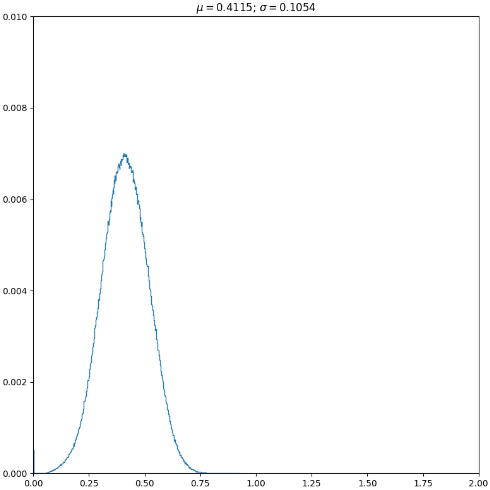
- Contradictory properties / functions

Challenges

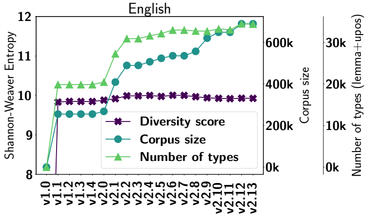
- Contradictory properties / functions
- Coming from ecology $\rightarrow$ balance properties

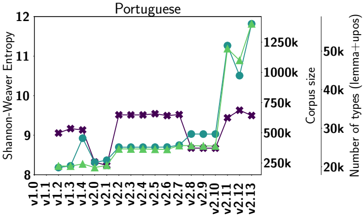
Challenges

- Contradictory properties / functions
- Coming from ecology → balance properties
- Coming from ecology → dimensionality
- Curse of dimensionality


μ = 0.4115; σ = 0.1054
0.010
0.008
0.006
0.004
0.002
0.000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

Use case: Corpus evolution





$$H = -\sum_{i=1}^{n} p_i \log_b (p_i)$$

Use case: Summarisation (with Eve)

- Transformers for summarization

Use case: Summarisation (with Eve)

- Transformers for summarization
- Limited window (1024)

Use case: Summarisation (with Eve)

- Transformers for summarization
- Limited window (1024)
- Maximising diversity in that window

Use case: Summarisation (with Eve)

- Transformers for summarization
- Limited window (1024)
- Maximising diversity in that window
- rouge_1: ≈ 0.16 (LEAD) $\rightarrow 0.34$

Use case: Summarisation (with Eve)

- Transformers for summarization
- Limited window (1024)
- Maximising diversity in that window
- rouge_1: ≈ 0.16 (LEAD) $\rightarrow 0.34$

Marginal increase in computation, and no need for fine-tuning.

[Guo et al., 2023]

- LLM's output is (often) less diverse than its training set

Use case: LLMs

[Guo et al., 2023]

- LLM's output is (often) less diverse than its training set
- These less diverse data may end up used to train other LLMs

Use case: LLMs

[Guo et al., 2023]

- LLM's output is (often) less diverse than its training set
- These less diverse data may end up used to train other LLMs
- Over multiple iterations, decrease in diversity

Use case: LLMs

[Guo et al., 2023]

- LLM's output is (often) less diverse than its training set
- These less diverse data may end up used to train other LLMs
- Over multiple iterations, decrease in diversity

(although the choice of their diversity functions may not be state-of-the-art)

Diversutils, Codabench, and more

- Measuring diversity in practice: diversutils (written in C, but Python interface)

- Measuring diversity in practice: diversutils (written in C, but Python interface)
- Measuring diversity on graphs: diversgraph (written in C) $\rightarrow$ also provides indexing of large amounts of linguistic data

Diversutils, Codabench, and more

- Measuring diversity in practice: diversutils (written in C, but Python interface)
- Measuring diversity on graphs: diversgraph (written in C) $\rightarrow$ also provides indexing of large amounts of linguistic data
- Diversity plugin for Codabench (not yet)

Diversutils, Codabench, and more

- Measuring diversity in practice: diversutils (written in C, but Python interface)
- Measuring diversity on graphs: diversgraph (written in C) $\rightarrow$ also provides indexing of large amounts of linguistic data
- Diversity plugin for Codabench (not yet)
- Morphosyntactic diversity, complexity/information-aware data reduction for tractability (in the works)

References

Anne Chao, Chun-Huo Chiu, and Lou Jost. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45:297–324, 2014. ISSN 1543-592X. URL https://www.jstor.org/stable/24810182. Publisher: Annual Reviews.

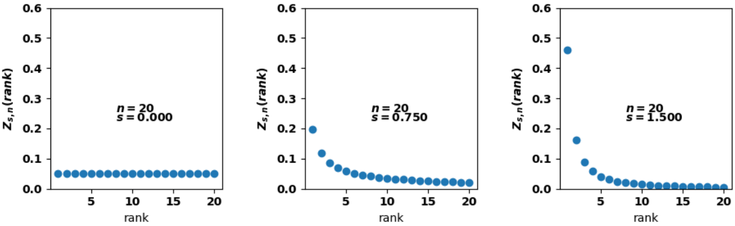
Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. The Curious Decline of Linguistic Diversity: Training Language Models on Synthetic Text, November 2023. URL http://arxiv.org/abs/2311.09807. arXiv:2311.09807 [cs].

Adam Lion-Bouton, Yagmur Ozturk, Agata Savary, and Jean-Yves Antoine. Evaluating Diversity of Multiword Expressions in Annotated Text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3285–3295, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.290.

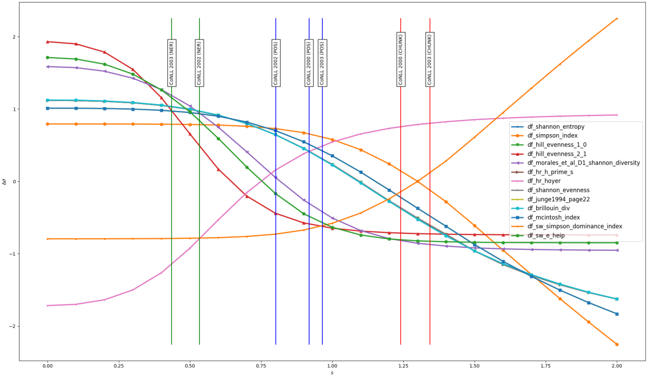
Appendix: Types / items

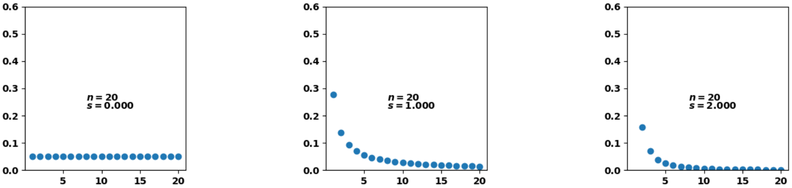
Concept		Possible types	Possible items
Segment annotation	Multi-Word Expressions	MWE classes (VID, LVC, VPC, ...)	Canonical forms
		Canonical forms	All observed forms
		All observed forms	Instances
	Named Entitites	Named entity classes (PER, ORG, LOC, ...)	Standardised forms
		Standardised entities	All observed forms
		All observed forms	Instances
Syntactic dependencies		Dependency type	Instances
		Dependency type + parent element	
		Dependency type + child element	

Appendix: Behaviour depending on Zipfian parameters (1)



Appendix: Behaviour depending on Zipfian parameters (2)





Appendix: Behaviour depending on Zipfian parameters (3)