SWP

SS 2023

Frederick Riemenschneider

Heidelberg University

18.04.2023

1 Authorship Attribution

2 EvaCun 2023

3 References

Authorship Attribution

Authorship Attribution

- identifying the author of a given document
- old but unsolved task
- highly relevant for ancient languages

Challenges

- inconsistent and insufficient evaluation
- dataset diversity
- no agreed "best method"

Idea

1. (re)implement (an) Authorship Attribution approach(es)
2. evaluate on established dataset
3. evaluate on a novel dataset containing ancient texts
4. explainability (?)

Goals

- address an unsolved research question
- develop and refine a methodology
- esoteric methodology, rigorous evaluation

Pointers

- overview: Tyo et al. 2022
- dataset, challenges: Kestemont et al. 2020
- common approaches, baselines, ideas: Altakrori et al. 2021
- how not to: Yamshchikov et al. 2022
- ideas: Uchendu et al. 2020

EvaCun 2023

EvaCun 2023

- Sumerian and Akkadian: cuneiform writing system
- Sumerian: earliest recorded written language
- shared task with three subtasks
    - Akkadian (cuneiform) to English
    - Akkadian (transcriped) to English
    - Sumerian (transcripted) to English

Idea

- take care of transcription
  - or let ByT5 (Xue et al. 2022) take care of it
- implement classical MT approach
- ensembling

Goals

- win shared task

Goals

- implement and refine established approach
- solve low-resource problems
- "classical NLP project"

Pointers

- shared task:
  https://digitalpasts.github.io/EvaCUN/Home/
- MT overview: Stahlberg 2020
- low-resource MT overview: Haddow et al. 2022
- promising model: Xue et al. 2022

References

References

📄 Altakrori, Malik, Jackie Chi Kit Cheung, and Benjamin C. M. Fung (Nov. 2021). "The Topic Confusion Task: A Novel Evaluation Scenario for Authorship Attribution". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4242–4256. DOI: 10.18653/v1/2021.findings-emnlp.359. URL: https://aclanthology.org/2021.findings-emnlp.359.

📄 Haddow, Barry et al. (Sept. 2022). "Survey of Low-Resource Machine Translation". In: *Computational Linguistics* 48.3, pp. 673–732. DOI: 10.1162/coli_a_00446. URL: https://aclanthology.org/2022.cl-3.6.

📄 Kestemont, Mike et al. (2020). "Overview of the Cross-Domain Authorship Verification Task at PAN 2020.". In: *CLEF (Working Notes)*.

📄 Stahlberg, Felix (2020). "Neural machine translation: A review". In: *Journal of Artificial Intelligence Research* 69, pp. 343–418.

📄 Tyo, Jacob, Bhuwan Dhingra, and Zachary C Lipton (2022). "On the state of the art in authorship attribution and authorship verification". In: *arXiv preprint arXiv:2209.06869.*

📄 Uchendu, Adaku et al. (Nov. 2020). "Authorship Attribution for Neural Text Generation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8384–8395. DOI: 10.18653/v1/2020.emnlp-main.673. URL: https://aclanthology.org/2020.emnlp-main.673.

📄 Xue, Linting et al. (2022). "ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models". In: *Transactions of the Association for Computational Linguistics* 10, pp. 291–306. DOI: 10.1162/tacl_a_00461. URL: https://aclanthology.org/2022.tacl-1.17.

📄 Yamshchikov, Ivan et al. (Dec. 2022). "BERT in Plutarch's Shadows". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6071–6080. URL: https://aclanthology.org/2022.emnlp-main.407.