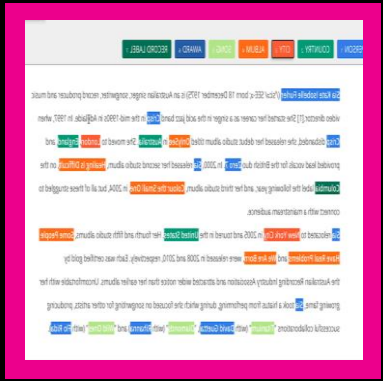
MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA



@twitterhandle

A five-step approach to manually annotating unstructured data into training dataset for natural language processing

Yunshu Zhu*, Ting Song, Zhenyu Zhang, Cedar Yin, Ping Yu

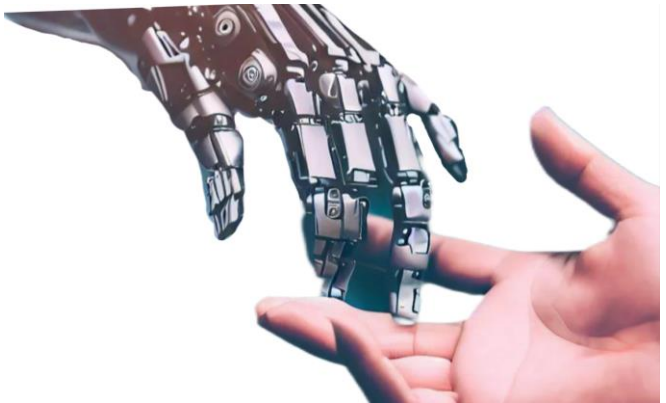
*PhD candidate

University of Wollongong

AIDH IMIA

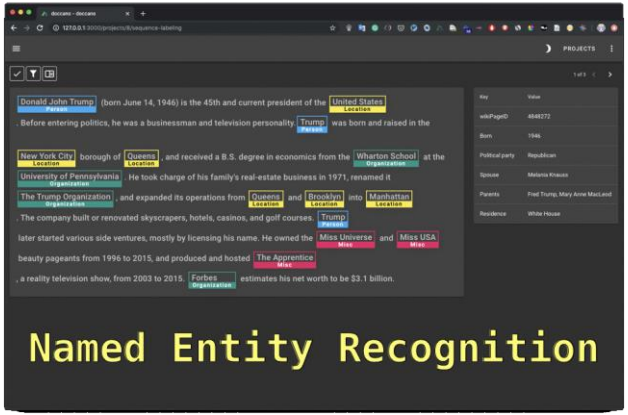
Introduction

Unlocking NLP Potential: The Power of Annotation



Objective

- to present *an approach for manually annotating unstructured data* into a training dataset specifically for natural language processing (NLP)

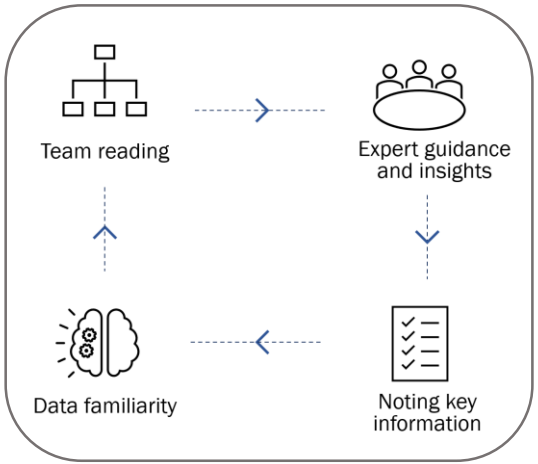


Significance

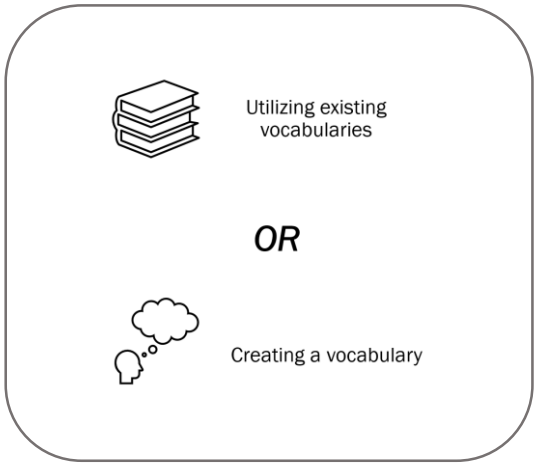
- Overcome **challenges of working with unstructured data** in NLP applications.
- Create **high-quality annotated training datasets** to enhance NLP model performance.
- Equip practitioners and researchers with **practical knowledge and tools for effective manual annotation**.
- Empower researchers to **contribute to NLP advancement** through annotated datasets.

The Five-Step Approach to Manual Annotation

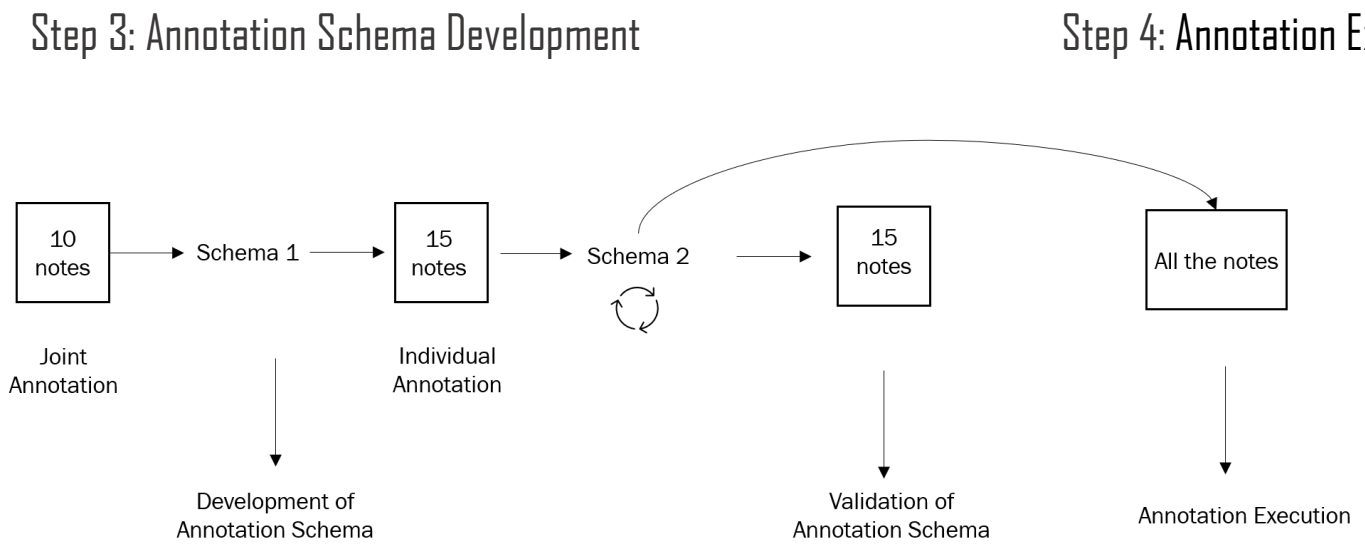
Step 1: Annotator Training



Step 2: Vocabulary Identification



The Five-Step Approach to Manual Annotation



The Five-Step Approach to Manual Annotation

Step 5: Result Validation

e.g., Inter-Annotator Agreement (Kappa statistics)

F-score can be used to calculate the precision and recall value

Case Study

1 Data

4,445 residents' data from 40 residential aged care facilities (RACF).

2 Objective

To extract agitated behaviors from residents.

3 Results

Identified 67 agitated behaviors.
F score: accuracy rate of 96%.

Conclusion

- **Summary:** The five-step approach provides a systematic and effective way to manually annotate unstructured data for NLP.
- **Significance:** improved model accuracy, better generalization, and the ability to address specific NLP tasks with tailored annotations.
- **Limitation:** The annotation task was time-consuming and had the potential to cause annotator fatigue.

References

- J. Park, S.C. You, E. Jeong, C. Weng, D. Park, J. Roh, D.Y. Lee, J.Y. Cheong, J.W. Choi, and M. Kang, A framework (SOCRATex) for hierarchical annotation of unstructured electronic health records and integration into a standardized medical database: development and usability study, JMIR medical informatics 9 (2021), e23983.
- Q. Wei, A. Franklin, T. Cohen, and H. Xu, Clinical text annotation–what factors are associated with the cost of time?, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2018, p. 1552.
- W.W. Chapman and J.N. Dowling, Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports, Journal of biomedical informatics 39 (2006), 196-208.
- F. Altalhi, A. Altalhi, Z. Magliah, Z. Abushal, A. Althaqafi, A. Falemban, E. Cheema, I. Dehele, and M. Ali, Development and evaluation of clinical reasoning using 'think aloud'approach in pharmacy undergraduates–A mixed-methods study, Saudi Pharmaceutical Journal 29 (2021), 1250-1257.
- Z. Zhang, P. Yu, H.C. Chang, S.K. Lau, C. Tao, N. Wang, M. Yin, and C. Deng, Developing an ontology for representing the domain knowledge specific to non-pharmacological treatment for agitation in dementia, Alzheimer's & Dementia: Translational Research & Clinical Interventions 6 (2020), e12061.

MEDINFO23

8 – 12 JULY 2023 | SYDNEY, AUSTRALIA

THANK YOU