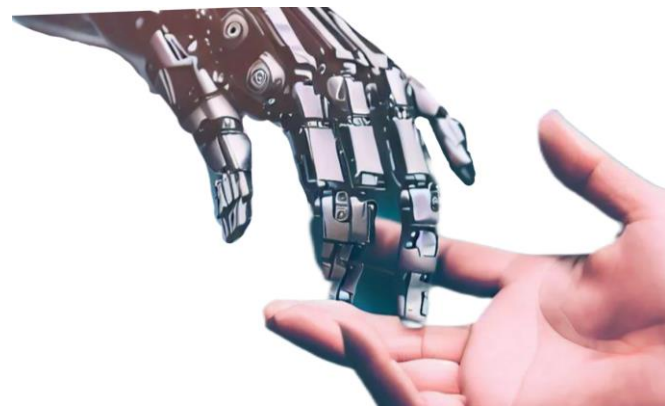




Introduction



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)

The screenshot shows a web browser displaying a document with Named Entity Recognition (NER) results. The text is annotated with colored boxes and labels for various entities. The entities are categorized into PERSON, LOCATION, ORGANIZATION, and DATE. The text is as follows:

Donald John Trump (born June 14, 1946) is the 45th and current president of the United States. Before entering politics, he was a businessman and television personality. He was born and raised in the New York City borough of Queens, and received a B.S. degree in economics from the Wharton School at the University of Pennsylvania. He took charge of his family's real-estate business in 1971, renamed it The Trump Organization, and expanded its operations from Queens and Brooklyn into Manhattan. The company built or renovated skyscrapers, hotels, casinos, and golf courses. He owned the Miss Universe and Miss USA beauty pageants from 1996 to 2015, and produced and hosted The Apprentice, a reality television show, from 2003 to 2015. Forbes estimates his net worth to be \$3.1 billion.

The right side of the screenshot shows a sidebar with a table of entities:

Key	Value
WikiPageID	4548272
Born	1946
Political party	Republican
Spouse	Melania Trump
Parents	Fred Trump, Mary Anne MacLeod
Residence	White House

Below the screenshot, the text "Named Entity Recognition" is displayed in a large, bold, yellow font.



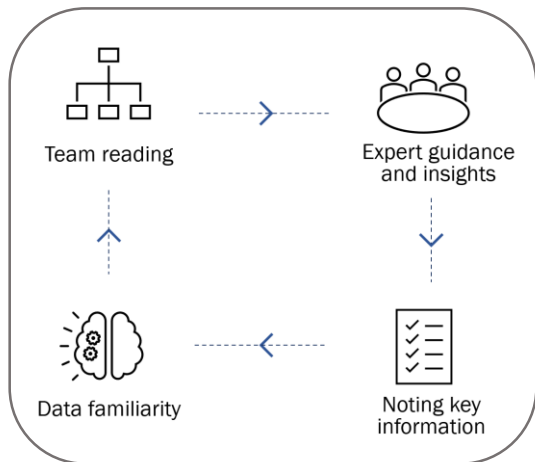
Significance

- 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.

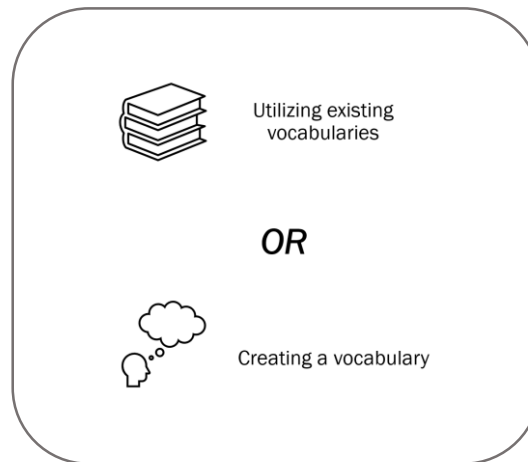


The Five-Step Approach to Manual Annotation

Step 1: Annotator Training



Step 2: Vocabulary Identification

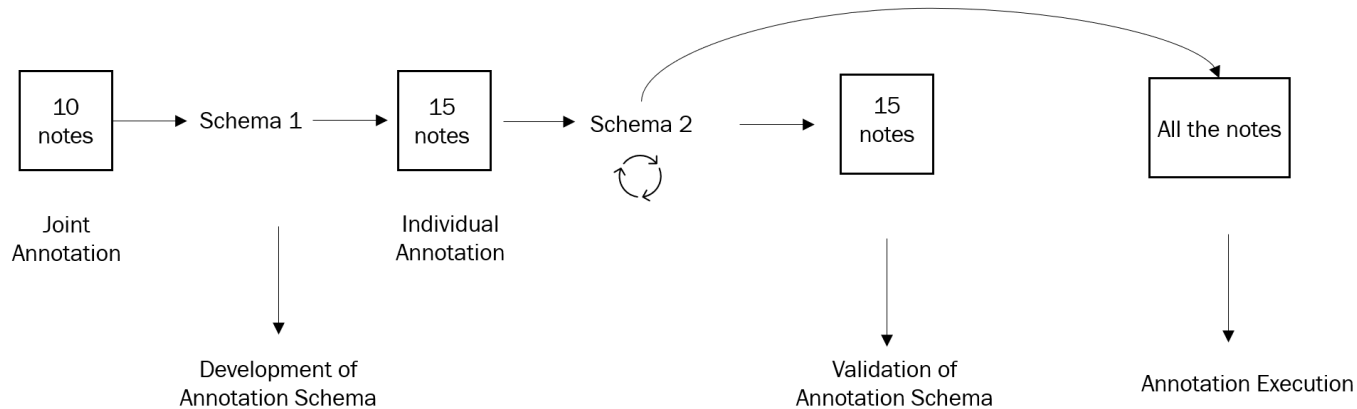




The Five-Step Approach to Manual Annotation

Step 3: Annotation Schema Development

Step 4: Annotation Execution





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.



THANK YOU