Towards Tracing Knowledge Flows in Martial Arts

BIOGRAPHICAL DATA
AND INTERPERSONAL CONTACTS

Yumeng Hou
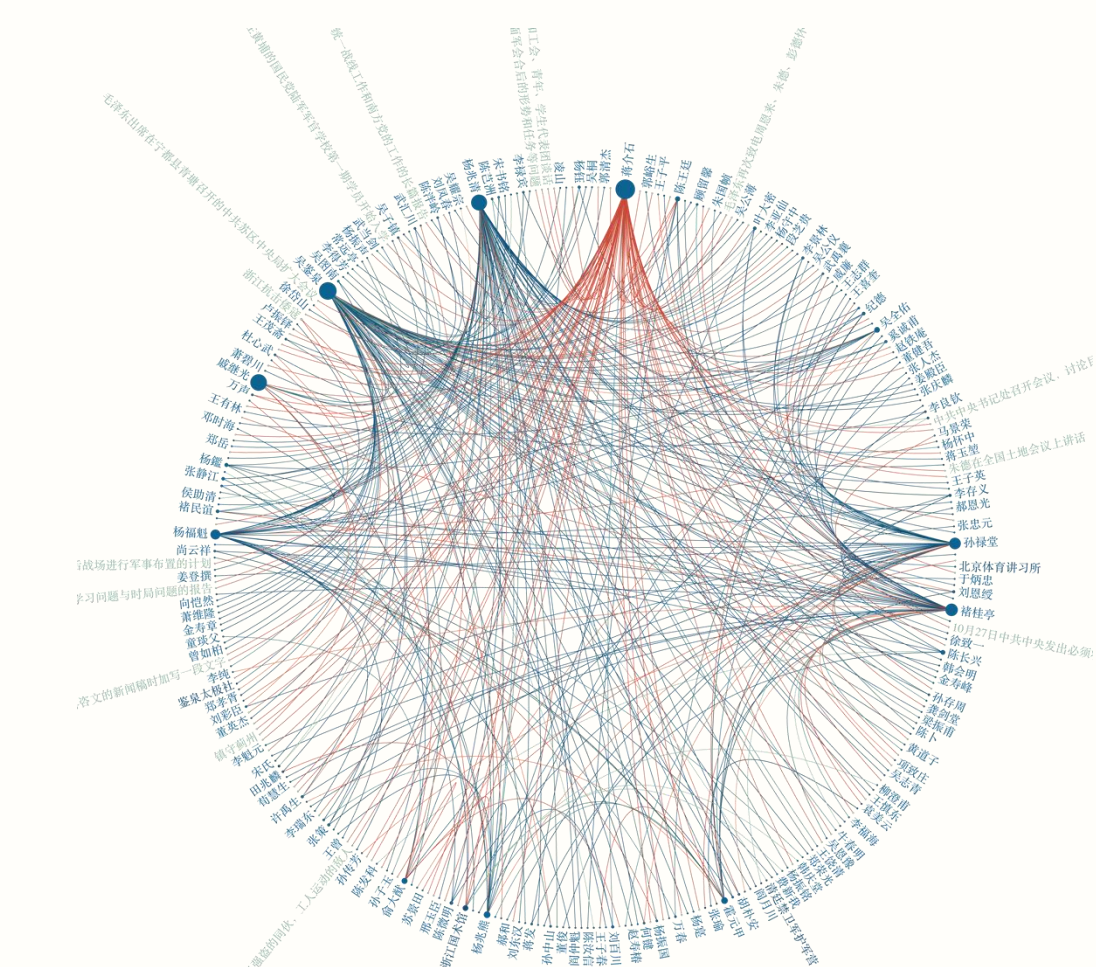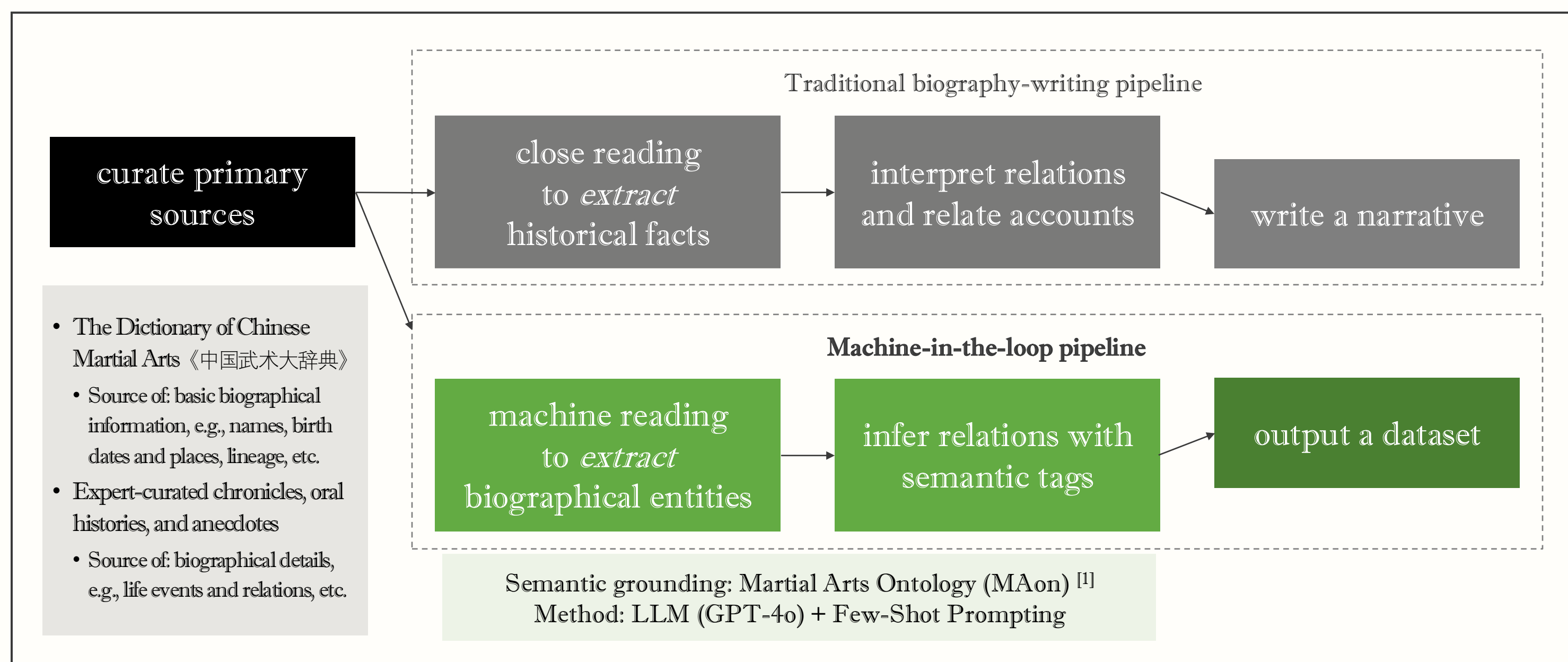| yumeng.hou@nus.edu.sg
| yumenghou.com

Context

- Martial arts are forms of embodied knowledge passed on through interpersonal interaction. The relationships among practitioners play a crucial role in knowledge transmission.

- In Chinese history, martial artists were rarely scholars. Historical materials about them are sparse and often contain ambiguous or even contradictory accounts.

- Scholars have attempted to write biographies based on archival research, yet the process remains selectively condensed, interpretive, and inevitably biased.

PROPOSITION

Construct evidence-based biographical and relational datasets through entity extraction across multiple sources.
Provide a graph-data foundation for detecting contact patterns and contradictory relational properties.

Traditional biography-writing pipeline

curate primary sources

close reading to *extract* historical facts → interpret relations and relate accounts → write a narrative

- The Dictionary of Chinese Martial Arts《中国武术大辞典》
  - Source of: basic biographical information, e.g., names, birth dates and places, lineage, etc.
- Expert-curated chronicles, oral histories, and anecdotes
  - Source of: biographical details, e.g., life events and relations, etc.

Machine-in-the-loop pipeline

machine reading to *extract* biographical entities → infer relations with semantic tags → output a dataset

Semantic grounding: Martial Arts Ontology (MAon) [1]
Method: LLM (GPT-4o) + Few-Shot Prompting

Sample chord diagram visualising categorized relations among martial artists associated with Taiji (Tai Chi) transmission.

[1] Martial Arts Ontology (MAon) is available at https://purl.org/maont/techCorpus .

Lessons Learned

- **Normalisation** is essential when processing heterogeneous sources. Use **reference tables** to standardise appellations, time expressions, regional names, etc.

- With tuned few-shot instructions, LLMs perform more effectively in extracting relations from nuanced Chinese texts. However, they still tend to **confuse relational direction** when faced with less-common or inherently ambiguous expressions.
  - **Potential solutions**:
    - Fast but naïve: identify frequently confusing terms and regulate them using indicator-based or rule-based extraction.
    - Scalable but costly: train supervised models by feeding corrected directional relations along with their source excerpts.
    - Balanced approach: use multiple models to cross-check results and flag contradictory inferences for expert review.
- **A cautionary note:**
  - Domain experts play a pivotal role in curating core sources, but **they are often trained to rely on materials that reflect mainstream narratives**. This (un)intentionally produces patterns that appear more consistent and less contradictory or critical.
- **Good practice to emphasise:**
  - **Preserve source excerpts alongside each extraction**. This enhances explainable machine inference, human verification and overall transparency, and can also serve as secondary training data should further training or fine-tuning become necessary.