Scientific Text Mining and Knowledge Graphs

Meng Jiang and **Jingbo Shang**

University of Notre Dame — mjiang2@nd.edu

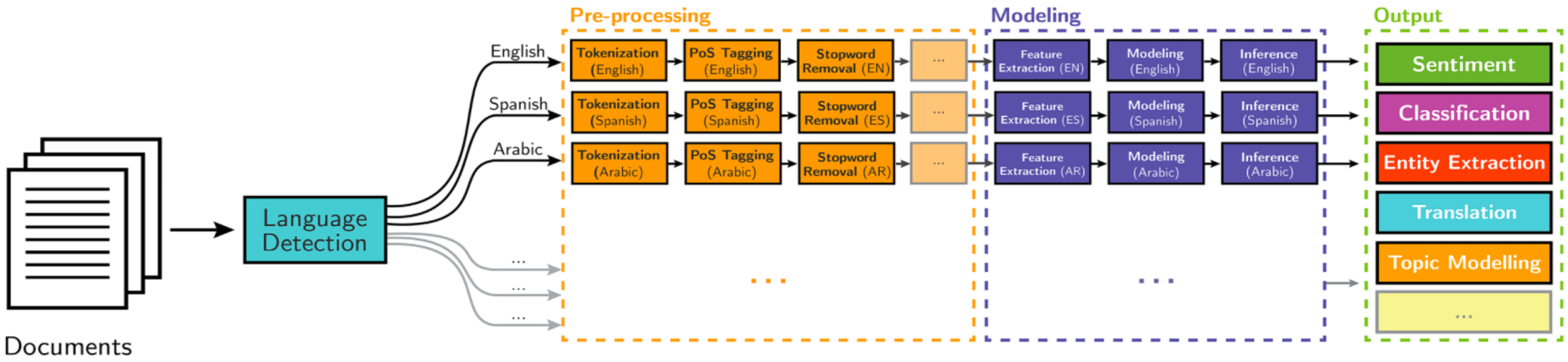
University of California, San Diego — jshang@ucsd.edu

Conclusions: from Unstructured Text to Knowledge

- Mining Structures from Massive Unstructured Text
  - Automated Phrase Mining (AutoPhrase)
  - Automated Entity Typing (AutoNER)
  - Automated Taxonomy Construction (NetTaxo)
- Unique Challenges and Tasks in Sciences
  - Conditions in scientific statements
  - Experimental evidence in tabular data
- The Path: Unstructured Texts → Structures → Knowledge

Future Work: Phrase Mining

- For popular languages with sufficient NLP tools
  - Incorporate more NLP features and structures
  - Incorporate contextualized representations to improve the accuracy
- For low-/zero- resource languages
  - Better unsupervised method



Future Work: Named Entity Recognition

- Improve the distant supervision
  - Can we do better than string match?
  - Can we integrate the phrase mining with NER and let them mutually enhance each other?
- Involve human experts in the loop
  - Given a fixed amount of expert hours, how to build the most reliable NER system?

Future Work: Knowledge Graph Learning

- Taxonomy improvement with dynamic user feedback
- Quality improvement
  - Denoising and cleaning
  - Completion
    - Typing and link prediction: Graph neural networks
    - Inference and reasoning: Reinforcement learning
- Application: Natural language generation
  - Conversational bots/dialogue systems
  - Question answering
  - Summarization

Topics

- Methods for extracting **entities** (methods, research topics, technologies, tasks, materials, metrics, research contributions) and **relationships** from research publications
- Methods for extracting **metadata** about authors, documents, datasets, grants, affiliations and others.
- Methods for quality assessment of **scientific knowledge graphs**
- Methods for the exploration, retrieval and visualization of **scientific knowledge graphs**
- **Scientific claims** identification from textual contents
- **Data models** (e.g., ontologies, vocabularies, schemas) for the description of scholarly data and the linking between scholarly data/software and academic papers that report or cite them
- Automatic or semi-automatic approaches to making sense of **research dynamics**
- Applications for the (semi-)automatic **annotation** of scholarly papers
- Description and use of **provenance** information of scholarly data
- **Theoretical models** describing the rhetorical and argumentative structure of scholarly papers and their application in practice
- **Novel user interfaces** for interaction with paper, metadata, content, software and data
- **Visualization** of related papers or data according to multiple dimensions (semantic similarity of abstracts, keywords, etc.)