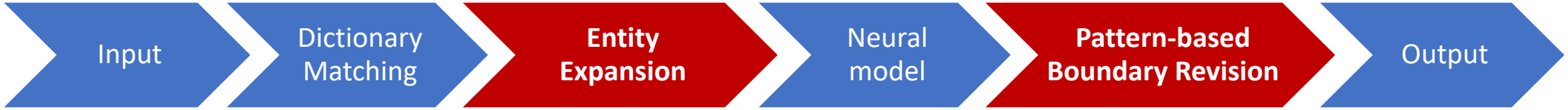
2018 Fall Summary

Xuan Wang

Publications

- **Xuan Wang**, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han, "Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning", Bioinformatics (accepted), Oct. 2018 (to appear 2019)
- **Xuan Wang***, Yu Zhang*, Qi Li, Cathy Wu, and Jiawei Han, "PENNER: Pattern-enhanced Nested Named Entity Recognition in Biomedical Literature", Proc. 2018 Int. Conf. on Bioinformatics and Biomedicine (BIBM'18), Madrid, Spain, Dec. 2018
- **Xuan Wang**, Yu Zhang, Qi Li, Yinyin Chen and Jiawei Han, "Open Information Extraction with Meta-pattern Discovery in Biomedical Literature", in Proc. of 2018 ACM Conf. on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB'18), Washington, DC, August 2018

Automatic biomedical named entity recognition



- Raw corpus
- Entity type dictionaries

- Phrase mining
- Dictionary matching

- Seed clustering
- Seed ensemble
- Feature ensemble

- Tie-or-Break

- Meta-pattern Discovery
- Pattern Expansion

- Entity-tagged corpus

Method	Human Effort	BC5CDR			NCBI-Disease		
		Prec	Rec	F1	Prec	Rec	F1
Supervised Benchmark [16, 22]	Gold Annotations	88.84	85.16	**86.96**	86.11	85.49	**85.80**
SwellShark [9]	Regex Design + Special Case Tuning	86.11	82.39	84.21	81.6	80.1	**80.8**
	Regex Design	84.98	83.49	**84.23**	64.7	69.7	67.1
Dictionary-Match	None	93.93	58.35	71.98	90.59	56.15	69.32
Dictionary-Expansion	None	92.09	63.85	75.42	90.17	56.88	69.76
Fuzzy-LSTM-CRF [36]	None	88.27	76.75	82.11	79.85	67.71	73.28
AutoNER [36]	None	88.96	81.00	84.79	79.42	71.98	75.52
AutoBioNER	None	87.34	84.53	**85.91**	77.98	75.31	**77.58**

Submitted to WWW

Pattern-enhanced named entity recognition

- **Input**: … prethrombin 1 induce heart failure …
- **Initial boundary and typing**: … CHEMICAL_prethrombin | 1 | induce | DISEASE_heart_failure | …
- **Initial score**: score(prothrombin:CHEMICAL) + score(heart failure: DISEASE) + score(CHEMICAL 1 induce DISEASE: PATTERN)
- **Pattern enhanced re-segmentation and retyping**: … CHEMICAL_ prothrombin_1 | induce | DISEASE_heart_failure | …
- **New score**: score(prothrombin 1: CHEMICAL) + score(heart failure: DISEASE) + score(CHEMICAL induce DISEASE: PATTERN)

Others

- Courses:
  - CS447: Natural Language Processing
  - CS598: Machine Learning Theory
- TA for CS412
- Organize the biotext meeting on Fridays