

Research Progress

Collaborators: Liangming Pan & Min-Ren Kan

anyahui.120@gmail.com

2018/5/25

TPDL 2018^[1] - Task

- Mention Extraction
 - SCI (Single, Concrete, and within-course entities)

YES: $\langle m_1 \rangle$ Is it just me or were some questions on Quiz 2 a surprise? There were a few questions that were not discussed in the lesson plan.

YES: $\langle m_2 \rangle$ Hello, I just would like to note that on 12:30 in the answer to question 3 in the lecture 2.4 it says that the network is deadlock-free, whereas ...

NO: $\langle m_3 \rangle$ The last item, that is “Probability Models for Customer-Base Analysis.pdf”, in the Resources > Additional Readings by Week section for Week 3 is not accessible.

NO: $\langle m_4 \rangle$ I’m working on the programming assignment for ML, week 2. I successfully submitted answers to the obligatory questions.

NO: $\langle m_5 \rangle$ At around 5:00 in the lecture, we see that the regularization term in the cost function is summed from 1 to L-1. Shouldn’t this be 2 to L?

NO: $\langle m_6 \rangle$ Hello. I wanted to use “e” as a number for ex.2/week3. It didn’t work, and I didn’t find useful help with “help exponent”.

Fig. 4. Actual resource mentions in our 1,087 sample sized dataset, illustrating the variety of expressions. Our Wikification currently handles the first two mentions.

TPDL 2018^[1] - Task

- Mention Extraction
 - Coverage Low

Table 2. Mention extraction coverage.

Annotator ID	# of Posts	# of posts identified as having mentions	# Extracted by our Wikifier	# Correct	Coverage
Annotator 1	1,087	156	5	5	14.4%
Annotator 2	1,087	175	5	5	16.1%
Overall	1,087	196 (Union)	5	5	18.0%

[1] The MUIR Framework: Cross-Linking MOOC Resources to Enhance Discussion Forums. (Submitted to TPDL 2018.)

Mention Extraction for MOOC Forums

- How about extracting mentions automatically?
- How about covering all kinds of actual mentions?

Mention Extraction - Challenge

- MOOC Uniques
 - Temporary.
 - Time intervals between posts are frequent. And the mentions distribution will bias with different intervals.
 - Consistency
 - The posts and comments in the same thread is usually discussing one problem or a subset of problems.
 - Roles
 - Staff, instructors, student, university_admin, data_coordinator and not_enrolled. The distribution of mentions in posts from different roles is different.
 - Upvotes
 - Upvotes for posts can help to examine the quality of post to introduce or solve a target problem.

Data Annotation

- 11679 posts
- ~9000 annotated

Model Construction

- LSTM
 - Precision: 50%~55%
 - Recall: 35%~40%
- CRF
 - Precision: 65%~70%
 - Recall: 55%~60%
- LSTM-CRF
 - Precision: 65%~70%
 - Recall: 65%~70%

To do list

- Data Annotation
- Error Analysis
- Model construction