Detecting Herb Activity in Chinese Literature Using Ontology and Rule-Based Approach

Heng Zhong
Supervisor's Name: Josiah Poon
School of Information Technologies

MOTIVATION AND CONTRIBUTION

Motivation

- Increasing demand of digitalising printed documents and storing them in a database.
- Current performance of Chinese relation extraction is less not as satisfiying comparing to other languages

Contribution

- Activity detection application using ontology and rule-based approach
- Ontology for TCM domain
- Set of rules for activity detection in TCM literatures
- Set of pre-processed TCM documents with more than 800 activities annotated

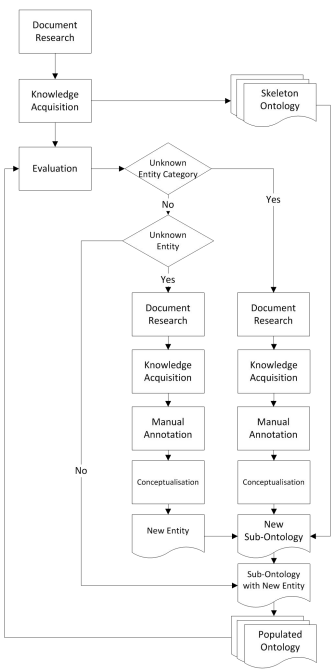
PROJECT GOAL

The goal of the project is to discover segments of sentences that has a herb activity, and annotating the pair of entities that participate in the activity.

- In the current scope of the project, 10 types of herb activties are the focused for detection.

ONTOLOGY

Building The Ontology



The Ontology

21 Bottom level object classes, e.g. TCM theoretical knowledge, human components, to TCM pathology knowledge, etc.

19 Object properties describing TCM relations and activities, e.g NourishOrgan, CleanseHeat, etc.

Currently the ontology contains more than 110 herb entities, 34 symptoms, 10 organs, 8 syndromes and etc,.

RULES FOR ACTIVITY DETECTION

Rules

A sentence segment that has a herb activity should:

1. Contains both entities (herb and participant)
2. Considers a "Comma + Herb Value" pattern as a new sentence segment.
3. Not contain more than n_{token} tokens between the entity pair
4. Not contain more than n_{verb} verbs

(n_{token} and n_{verb} are parametres for the rules and can be changed)

Example

- 党参大补元气、益气生津，麦冬润肺滋水、清心泻热，五味子润肺生津，固摄肾关，三药配伍，肺肾心脾四脏兼治，肺气布津。
- According to the second rule, sentence is splitted up into three segments:

  1.党参大补元气、益气生津

  2.麦冬润肺滋水、清心泻热

  3.五味子润肺生津，固摄肾关，三药配伍，肺肾心脾四脏兼治，肺气布津

Herb Entity	Participant Entity (Entity Class)
党参	元气 (Qi)
党参	气 (Qi)
党参	津 (Body Fluid)
麦冬	肺 (Organ)
麦冬	水 (Five Element)
麦冬	心 (Organ)
麦冬	热 (Syndrome)
五味子	肺 (Organ)
五味子	津 (Body Fluid)
五味子	肾 (Organ)

(n_{token} = 10 and n_{verb} = 4)

- The herb "五味子" is not annotated to have activities with organs "肺" (second occurrence), "肾" (second occurrence), "心", "脾", "肺" (third occurrence), Qi flow "气" and body fluid "津" (second occurence), because of excessive number of tokens or/and verbs between these entities and the herb.

LIMITATION AND FUTURE RESEARCH DIRECTIONS

Limitation

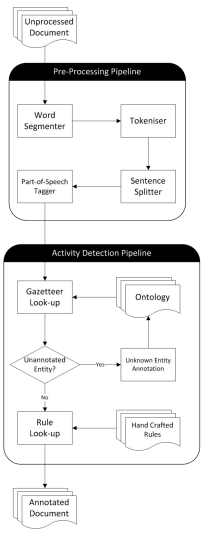
- Activity detection remains in sentence scope
- Detecting herb activities only
- No further processing after detection
- Large amount of manual work (annotation, ontology population, etc.)

Future Research

- Activity detection/extraction in paragraph/document scope
- Saving detected activities into database
- Detecting/extracting more complex activities/relations
- Automated ontology population

ACTIVITY DETECTION SYSTEM

System Overview



The whole system consists of two pipelines

- **Pre-processing pipeline** takes raw text documents as input, split the input into multiple sentences that consists of multiple tokens, each token is assigned a part-of-speech tag.
- **Activity detection pipeline** takes annotated tokens as its input. Search through the ontology gazetteer to match entities, and uses the crafted rules to identify sentence segments that descibe an activity.

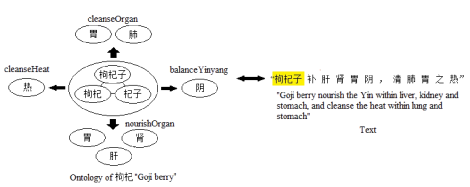
RESULTS AND EVALUATION

Baseline

For the baseline, only a conventional gazetteer of entity names is used, and only the first rule is used for activity detection. The system reached an micro F-Score of 71.1%, with a high recall value of 98%.

Performance with Ontology

- Instead of a plain gazetteer list, an ontology based gazetter is used. Though having the same performance as a conventional gazetteer, using ontology can provide better semantic definition, and create more relation between entities than using a conventional gazetteer.



Performance with Ontology and Rule

- By applying the designed rules to the dataset, the micro F-score reached 87.5%, the recall dropped to 87.1%, however the precision rose to 87.9%.

(n_{token} = 21 and n_{verb} = 7)