Rule-based Approach for Information Extraction from Chinese Free Text

Xiao Yu *(xiyu7940@uni.sydney.edu.au)* *Supervisor: Josiah Poon*
School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

1. Introduction & Motivation

Information Extraction (IE) is an essential task in Text Mining since it can transform unstructured or semi-structured information into structured data, which are easier for researchers to analyse. IE from Chinese will be a hot topic, since China has been developing extremely fast these years, and vast quantities of information needs processing efficiently.

2. Research Questions

- Can generalised extraction patterns be used in Chinese IE?
- Is this approach able to deal with texts from multiple domains?

3. Methodologies

Generalised Extraction Patterns

We have developed 17 Generalised Extraction Patterns and successfully applied them in the Chinese IE tasks in two different domains (TCM and Personnel Transfer). The patterns include various combinations of information to be extracted, tagging words, connector, etc. In addition, we have compiled patterns of *Composition*, *Keyword* as well as *Parallel.* These patterns are very effective, since they are generalised and can be applied in various domains in IE, especially in dealing with IE from Chinese free text.

For example, we can apply the same pattern for extracting the target information from three different domains as followings.

Sample sentence	ITBE 1	ITBE 2	Domain
从北京到上海	北京	上海	Path
包含工资和奖金	工资	奖金	Salary
选举张三担任销售总监	张三	销售总监	Appointment

(ITBE: Information to be extracted)

Integrating the patterns into rules with GATE

For applying our patterns in IE, we need to integrate the patterns into generalised rules. After the research, we chose GATE to be our testing platform since it supports IE from Chinese, and some previous Chinese IE tasks have been done by applying it. We extended the *ANNIE Gazetteer* and built our rules in *ANNIE NE Transducer* using *JAPE*. For validating the extracting performance, we applied the *Corpus Quality Assurance*, and calculated the Precision, Recall and F1.0 score by comparing the extracted results with our Gold Standard.

4. Application in the Personnel Transfer

Normally there are two main domains in a Personnel Transfer, which are personnel appointment and dismissal. For the event of "appointment", we want to extract the new person names along with their new organizations and positions. For the event of "dismissal", we want to get the person names, their past organisations and positions. We confirmed the tagging words for our patterns after analysing 109 announcement of Personnel Transfer, and applied them in our generalised rules.

Performance Evaluation

- Evaluation data

36 announcement of personnel transfer downloaded from the website of "CPC news".

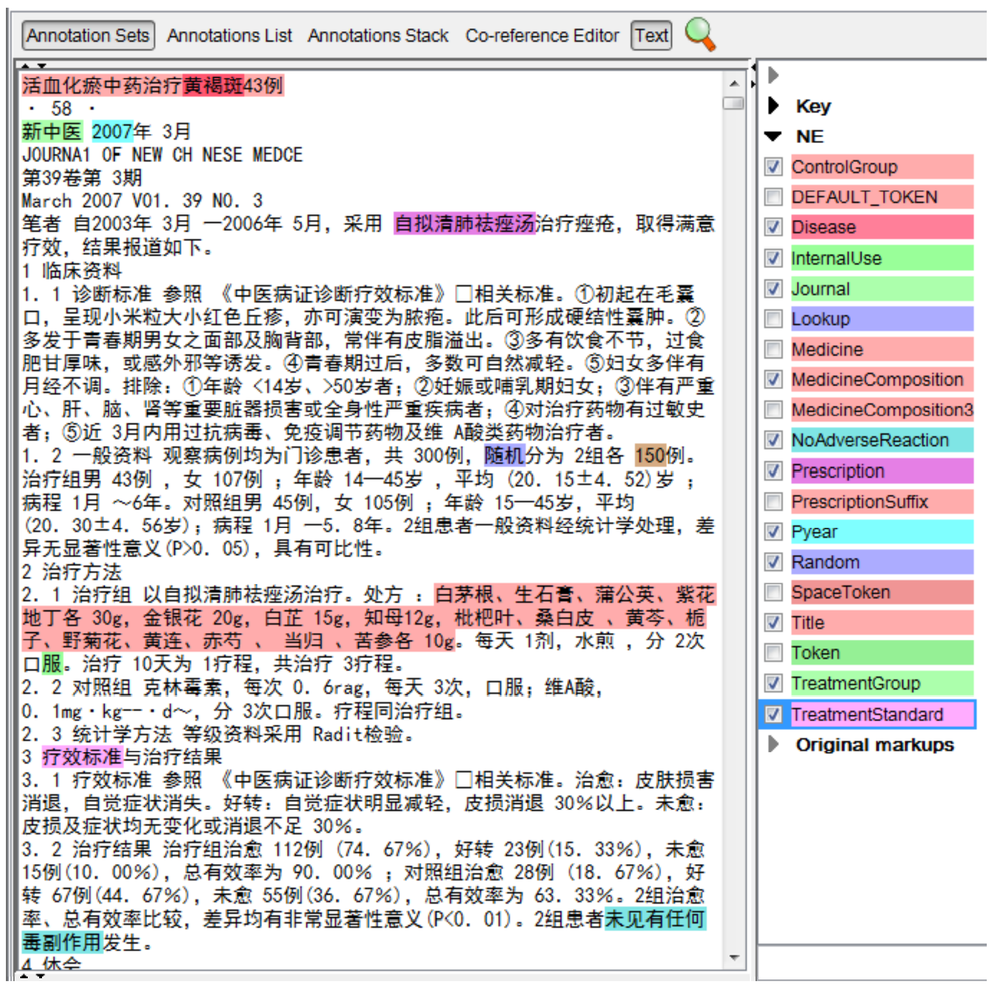
- Evaluation methods

Precision, Recall and F 1.0-score strict

Annotation	Prec. B/A	Rec. B/A	F1.0-s.
Date	1.00	1.00	1.00
NewOrganization	0.91	0.86	0.88
NewPerson	0.96	0.72	0.82
NewPosition	0.95	0.78	0.86
OldOrganization	1.00	0.91	0.95
OldPerson	1.00	0.69	0.82
OldPosition	1.00	0.88	0.94
Source	0.97	0.97	0.97
Macro summary	0.97	0.85	0.90
Micro summary	0.96	0.82	0.88

***Table 1**: Results of applying patterns in Personnel Event*

5. Application in the TCM



After discussion with the experts from SIRC-TCM, we confirmed to extract 14 items from each clinical literature. The information we want to extract is title, journal name, publish year, disease, prescription, taking method, medicine composition, sizes of treatment group and control group, whether use random allocation for allocating groups, whether have treatment standard, and whether have adverse reaction. We confirmed the tagging words for our patterns after analysing 177 clinical literatures from various journals, and applied them in our generalised rules.

Performance Evaluation

•Evaluation data

51 Chinese clinical literatures downloaded from the website of CQVIP as suggested from the experts of SIRC-TCM.

•Evaluation methods

Use Precision, Recall and F 1.0-score strict for evaluation. In addition, the experts from SIRC-TCM could help us to evaluate the results.

Annotation	Prec. B/A	Rec. B/A	F1.0-s.	Annotation	Prec. B/A	Rec. B/A	F1.0-s.
Adverse Reaction	0.75	1.00	0.86	NoAdverse Reaction	0.69	0.90	0.78
Control Group	0.97	0.97	0.97	Prescription	0.91	0.88	0.89
Disease	0.86	0.85	0.85	Pyear	0.86	0.96	0.91
External Use	0.80	0.67	0.73	Random	0.94	1.00	0.97
Internal Use	0.97	0.67	0.79	Title	0.80	0.80	0.80
Journal	1.00	0.83	0.91	Treatment Group	0.94	0.97	0.96
MeCompo sition	0.80	0.78	0.79	Treatment Standard	1.00	1.00	1.00

Annotation	Prec. B/A	Rec. B/A	F1.0-s.
Macro summary	0.88	0.88	0.87
Micro summary	0.89	0.87	0.88

***Table 2&3**: Results of applying patterns in TCM*

6. Conclusion

We have tackled the problem of IE from free Chinese text using a generalised extraction pattern based approach. The experiments achieved very promising results, which can prove that our IE systems are very effective and can be successfully adapted into two different domains.

Acknowledgements

THIS RESEARCH IS SPONSORED BY

This work was supported by the Shanghai Innovation Research Center of Traditional Chinese Medicine.

SIRC/TCM