CLEFeHealth 2014 Normalization of Information Extraction Challenge using Multi-model method

Yu-Cheng Liu[1], Lun-Wei Ku[2]

Institue of Information Science, Academia Sinica, Taipei, Taiwan, ROC

brad.ycliu@gmail.com[1],lwku@iis.sinica.edu.tw[2]

Abstract

This work focuses on making clinical documents easier to understand for patients and clinical workers. Normalization values of ten attributes have been predicted by the multi-model method which alternatively uses rule based methods and machine learning methods to solve different attribute problems. Information of text structure, lexical, and grammatical features are used to achieve overall average accuracy 0.787 and 0.849 on training data with run 1 and run 2, respectively. The UMLS CUI tool MetaMap is used to search for CUI category and CRFsuite package is adopted for machine learning method. In this paper, Run 1 is the official method and run 2 is considered as the supplement. Our system achieves overall average accuracy 0.793 on testing data with run 1 methods.

Keywords: multi-model method, MetaMap, CRFsuite

1 Introduction

ShARe/CLEF eHealth 2014 Task 2 extends from task 1 of ShARe/CLEF eHealth 2013 and focuses on Disease/Disorder template filling. It continues the direction of making clinical documents easier to understand for patients and clinical workers [1]. Ten attributes have been proposed by the convention of ShARe/CLEF eHealth 2014. Each of 10 attributes has two types of slot values. One is normalization slot value and the other is lexical cue value. This year our team, ASNLP, joins task 2a, i.e. prediction of normalization slot value.

Many previous works had successful NLP inventions on normalization of medical concepts [2-5]. Hybrid NLP methods, i.e. combining rule based methods and machine learning methods, are widely applied to help solve those problems including clinical entity recognition, and normalization problems when processing medical texts. Text features which include text structure, lexical and grammatical features are revealed helpful for entity processing of clinical documents. The system design of our approach has multi-model conformation which uses alternatively rule based methods and machine learning methods for solving different attribute problems. Some existing NLP packages are also incorporated into the system.

2 Material and methods

2.1 Data

The training corpus, provided by the convention of ShARe/CLEF eHealth 2014 [6] , contains total 298 clinical reports. The corpus consists of four types of clinical reports: discharge summary, ECG report, echo report and radiology report. Each type of the clinical report has 136, 54, 54, and 54 reports, respectively. The testing dataset with 133 reports belongs to one type of clinical report which is discharge summary.

2.2 System design

ShARe/CLEF eHealth 2014 Task 2 proposes two types of slot values: normalization and cue. The system design of our work is a multi-model approach. Ten different models solve ten predictions of different attributes. We have achieved two runs. Run 1 is the official method and run 2 is considered as the supplement. Consequently, Fig. 1 shows the system architecture of run 1. The differences between run 1 and run 2 are on prediction methods of attributes document time and temporal expression. Instead of machine learning methods in run 1, rule based methods are adopted in run 2.

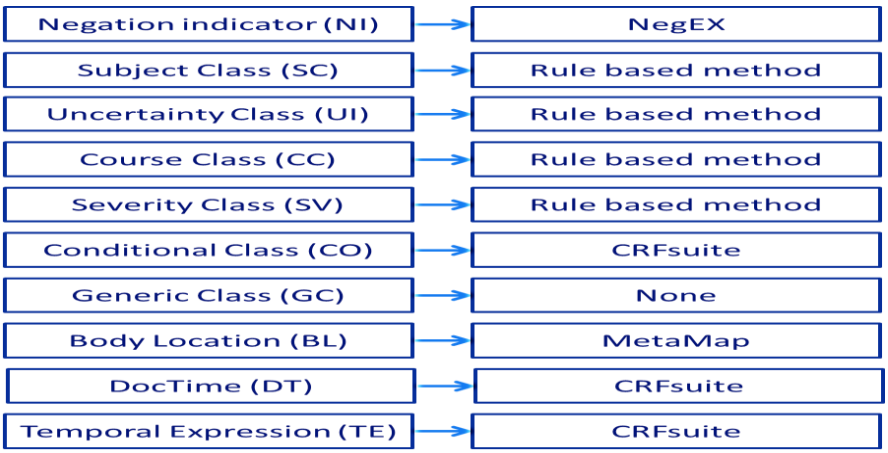

Fig. 1. The system architecture

Some existing methods and systems are incorporated into the system to solve corresponding problems, e.g. for the attribute of negation indicator, NegEX[7] package is used to determine if the specified disease/disorder(DD) entity has negative expression in the sentence. No new keywords/cues and rules are added to NegEX. MetaMap[8] system is applied to find classes of Unified Medical Language System concept unique identifiers (UMLS CUI) of DDs. CRFsuite package is taken as a method of machine learning.

In training data, we have observed that subject class and document time class have highly correlation with section information. As a result, the subject class is judged by

relative position of DDs considering location of family history section, physical exam section or social history. The rules of subject class determination are shown as Fig. 2. For predictions of document time class, instead of machine learning methods in run1, we use rule based method to identify whether DDs locate in history section.

Rule 1:	If DDs in FamilyHistorySect→ SubjectClass is "family_member"
Rule 2:	(If PhysicalExamSect after FamilyHistorySect) AND (DDs in PhysicalExamSect) → SubjectClass is "family_member "
Rule 3:	(If PhysicalExamSect before FamilyHistorySect) AND (DDs in PhysicalExamSect) → SubjectClass is "patient"
Rule 4:	If DDs in SocialHistorySect→ SubjectClass is "other"
Rule 5:	DDs Not satisfy rules 1, 2, 3, 4 → SubjectClass is "patient"

Fig. 2. Rules for subject class determination.

Rule based method is applied to solve uncertainty class, course class, severity class and temporal expression problems. Corresponding class keywords and phrases are collected from the training data and simple string matching method is applied. For attributes DocTime and temporal expression in run 1, we use CRFsuite package[9] as the predictor and 19 features are generated by the package. Although the features were used to solve chunking problems, they included word position and syntactic information. We consider they may then solve the normalization problems. Syntactic information is generated by Stanford parser[10].

Rules of DDs conditional class can be rarely concluded due to complex expressions. However, machine learning methods are suitable for dealing with this kind of problems. CRFsuite package, thus, is adopted as the predictor method. Part-of-speech, lexicon features and word position features are incorporated into twenty three features in total. In addition to 19 features which are generated by CRFsuite package, 4 key words "while", "when", "at" and "on" are applied. In these 23 features, Five-fold cross validation and three-fold cross validation are used to tune the parameters of the predictor for different types of clinical reports. Due to the different sample sizes, five-fold cross-validation adapts to reports of discharge summary type while three-fold cross-validation adapts to report types which are ECHO, ECG and radiology.

2.3 Data analysis

We propose a ratio, called occurrence contrast, to show if a word can be distinguishable for a class in an attribute. The formula is shown in the eq. (1). It suggests larger occurrence contrast and more distinguishable of a word for a class. It helps us to find key words of a class in an attribute.

$$occurence\ contrast = \frac{\text{occurence of a word in a class}}{occurence\ of\ a\ word\ in\ the\ alternative\ class} \quad (1)$$

3 Results and Discussion

We have accomplished two runs for ShARe/CLEF eHealth 2014 Task 2a. D Table 1 shows the evaluation of run 1 on accuracy for each attribute. We treat multiple class prediction as binary class prediction, i.e. all predictions only contain class "known" and "unknown". Our system achieves accuracy 0.793 on testing data with run 1 (official results). By contrast with evaluation results of program eval_t2a.pl, shown on column Run 1 of Table 2, two results are similar except attributes conditional class and time expression. At run 1, attributes Conditional, Document Time and Temporal Expression are dealt with machine learning methods. Other attributes are dealt with rule-based methods. At run 2, we replace machine learning methods with rule based methods on attributes Document Time and Temporal Expression. As the results show, rule-based methods perform better than machine learning methods on most of attributes. However, there are higher precision on prediction of conditional class attribute with machine learning method. Therefore, we apply rule-based method to solve Document Time and Temporal expression class problems and got improvement on accuracy with training data at run 2. The overall accuracy rates are 0.787 and 0.849 resulted by run 1 and run 2, respectively.

By statistics of occurrence contrast, mentioned in method section, figure 3 displays the distribution of occurrence contrast of each word in collected lexicon for attribute uncertainty indicator. It suggests key words "might", "suggests", "perhaps" are the first three distinguishable words for uncertain indicator. With the same method for attribute severity class, we can find that words "acute", "flash" and "severe" are the first three distinguishable words on the server class. On the slight class, "minimally", "minimal" and "slightly" are the most distinguishable words. "Mildly", "moderately" and "mild" are the most distinguishable words on the moderate class. Obviously, the completeness of collected lexicon would affect prediction results. Thus the lack of completeness of collected lexicon often leads to prediction errors. On the other hand, in applying our method we have problems performing string matching. We match string beyond the DDs terms. As a result, words, contained in DDs terms, do not be matched and lead to prediction errors.

From Table 3, it is shown that data distribution of each attribute is skew. It implies that we can set default value as the majority during prediction processes. From column Run 1 in Table 2, it shows high prediction accuracy on attributes conditional class and temporal expression. Those are the results of setting default values with majority, observed from training data. Therefore, F-measure would be suggested more discriminative on system performance than accuracy. However, prediction accuracy is still important for evaluation of a system. It can reveal the balance of evaluation for prediction accuracy of positive and negative samples. F-measure can reveal prediction accuracy of positive samples. Our system appears low F-measure on average at most of attributes of predictions.

Table 1. Prediction accuracy of all attributes for training data (evaluated with our method)

Document Type	DISCHARGE SUMMARY	ECG REPORT	ECHO REPORT	RADIOLOGY REPORT
Negation Indicator	0.920	0.970	0.981	0.897
Subject Class	0.904	1.000	1.000	1.000
Uncertainty Indicator	0.898	0.864	0.884	0.787
Course Class	0.935	0.859	0.956	0.897
Severity Class	0.897	0.935	0.752	0.905
Conditional Class	0.472	0.651	0.585	0.346
Generic Class	X	X	X	X
Body Location	0.550	0.337	0.208	0.471
DocTime Class	0.152	0.534	0.405	0.017
Temporal Expression	0.079	0.555	0.404	0.028
Average	0.645	0.745	0.686	0.594

Table 2. Prediction accuracy of all attributes for training data (evaluated with program eval_t2a.pl)

Document Type	DISCHARGE SUMMARY	
Run	Run 1	Run 2
Negation Indicator	0.920	0.924
Subject Class	0.904	0.913
Uncertainty Indicator	0.898	0.895
Course Class	0.935	0.937
Severity Class	0.900	0.900
Conditional Class	0.937	0.937
Generic Class	1.000	1.000
Body Location	0.522	0.522
DocTime Class	0.005	0.580
Temporal Expression	0.839	0.878
Overall Accuracy	0.787	0.849

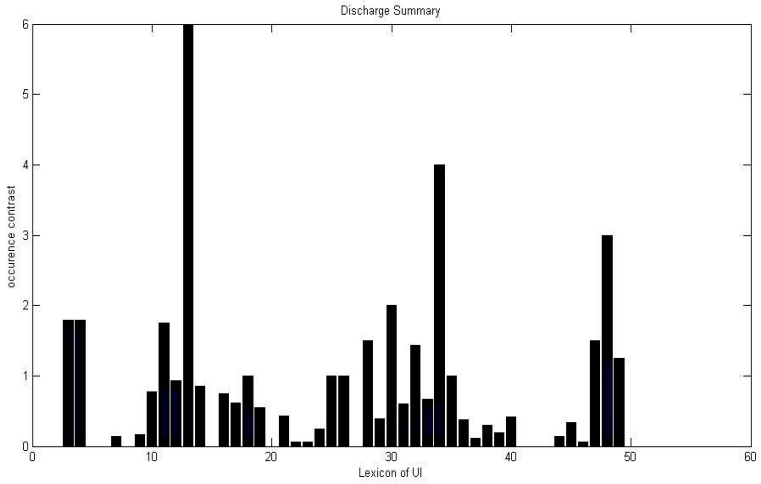

Fig. 3. The distribution of occurrence contrast of each word in collected lexicon for attribute uncertainty indicator.

(…. Ellipsis)
Family History:
Father died of MI at age 69.

Physical Exam:
PE: T 97.6 BP 142/70 R arm, 150/70 P 42-64 R 16 Sat 92% RA G: Elderly female, **NAD**
HEENT: MMM, anicteric
Neck: **JVD** diff to assess
Lungs: +end exp **rhonchi** bilaterally upper lung zones
CV: RRR, S1S2, distant heart sounds, +2/6 **systolic murmur at apex**
Abd: Soft, **NT**, **ND**, BS+
Ext: trace bilateral lower ext edema; R groin small hematoma, no bruits
Nails: No bed abnormalities, lunulas present, no splinters, pulses
Rectal: guiac neg

Pertinent Results:
(…. Ellipsis)

Fig. 4. A paragraph in the file "00211-027889-DISCHARGE_SUMMARY.txt" (Bold words are DDs terms).

Structural information of documents is useful for attributes subject class and document time predictions. Lexical information has advantage over other features for predictions of attributes uncertainty indicator, severity class, course class and time expression. Subject classes have correlation with locations of DDs on the medical text, i.e. locations of DDs have correlations with those sections of family history, physical examination and social history. In Table 1 and Table 2 subject class, uncertainty indicator, course class and severity class can be predicted reasonably with lexicon and simple rules from training data. Lexical features can help increase prediction

accuracy of conditional class. Total collected words and phrases in lexicons of uncertainty indicator, course class and severity class are 50, 12 and 22. It suggests that words and phrases used in clinical text for these three attributes are limited. Many of date time expression can be captured by regular expression, "\[**[\d\w]*-[\d\w]*-[\d\w]***\]". However, determining time and duration classes is relatively difficult by using combination of preposition and temporal terms. Grammatical features, i.e. syntactic and part-of-speech features, are less helpful with the predictions of most of attributes. Fig. 4 shows a segment of a discharged summary. DDs terms often contain in short descriptions and lack information about subjects, time, conditions and so on. However, grammatical features have positive impact on predictions of negation indicator, subject class and time expression. The predictor for attribute body location needs to be further developed or using more analytic tools to investigate the class of UMLS CUI of DDs.

4 Conclusion

We have introduced a system for disease/disorder template filling on normalization. Rule based methods outperform machine learning methods in terms of prediction accuracy. However, machine learning NLP methods have higher precision than rule based NLP methods. Most of attribute values can be captured by using simple rules from four types of medical text. Discharge summary has more complex clinical descriptions about disease/disorder than the other three types of medical text.

Most of distributions of attributes have data skew phenomena. As a result, it achieves high accuracy on predictions. Rule based NLP methods have high prediction recall and machine learning NLP methods have high prediction precision. Hence, combining rule based NLP methods and machine learning NLP methods should have reasonable effects on normalization problems. This had been similarly reported in other study[2]. We will continue looking for useful methods and features, e.g. words, symbols, position, and context features, to improve our system.

Acknowledgement

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC 102-2221-E-001-026.

Table 3. Data distribution of each attribute in discharged summary.

DISCHARGE_SUMMARY			
Negation Indicator			
yes	no		
1860	8012		
Subject Class			
donor_other	family_member	other	patient
1	72	14	9785
Uncertainty Indicator			
no	yes		
9265	607		
Course Class			
changed	decreased	improved	unmarked
7	148	74	9300
no	null	resolved	
1	2	56	
worsened	increased		
58	226		
Severity Class			
moderate	severe	slight	unmarked
239	325	93	9215
Conditional Class			
TRUE	FALSE		
603	9269		
Generic Class			
FALSE			
9872			
Body Location			
CUI-less	Cui-less	Cui-less	CUI
3271	2	1	6598
DocTime Class			
AFTER	BEFORE	BEFORE_OVERLAPS	OVERLAP
484	1378	2697	5263
UNKNOWN	unknown		
29	21		
Temporal Expression			
DATE	DURATION	TIME	none
1123	137	60	8552

5 Reference

1. L Kelly, L Goeuriot, G Leroy, H Suominen, T Schreck, DL Mowery, S Velupillai, WW Chapamn, G Zuccon, J Palotti.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. Springer-Verlag. (2014)
2. YC Chang, HJ Dai, JC Wu, Chen JM, Tsai RT, Hsu WL: TEMPTING system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. J Biomed Inform 46, S54-S62 (2013)
3. Yonghui Wu, Buzhou Tang, Min Jiang, Sungrim Moon, Joshua C. Denny and Hua Xu: Clinical Acronym/Abbreviation Normalization using a Hybrid Approach. In: CLEF. (2013)
4. James Gung.: Using Relations for Identi cation and Normalization of Disorders: Team CLEAR in the ShARe/CLEF 2013 eHealth Evaluation Lab. In: CLEF. (2013)
5. Jon D. Patrick, Leila Safari, Ying Ou.: ShARe/CLEF eHealth 2013 Normalization of Acronyms/Abbreviations Challenge. In: CLEF. (2013)
6. Elhadad N, Chapman WW, O'Gorman T, Palmer M, Savova G.: The ShARe Schema for the Syntactic and Semantic Annotation of Clinical Texts., Under Review.
7. Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper and Bruce G. Buchanan: A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 34, p.301-p.310 (2001)
8. Alan A. Aronson: Effective mapping of biomedical text to the UMLS Metathesaurus:the Metamap program. In: AMIA, pp. p.17-21. (2001)
9. CRFsuite package: http://www.chokkan.org/software/crfsuite/
10. The Stanford Parser: http://nlp.stanford.edu/software/lex-parser.shtml