International Journal of Biomedical Data Mining

Li et al., Int J Biomed Data Min 2018, 7:2
DOI: 10.4172/2090-4924.1000136

Research Article Open Access

CIES: Clinic Temporal Information Extraction System

Zhijing Li[1*], Yu Long[1], Xuan Wang[1], Qinghua Zheng[2] and Chen Li[1*]

[1]Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Tech R and D, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China
[2]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China

Abstract

Temporal information extraction is important to understand the text in clinical documents. Extracting temporal courses of clinical events from a patient's electronic health records may present structured information of the patient's symptoms, diagnosis, and treatments etc. in the actually occurring order. We have developed the clinic temporal information extractor, an easy-to-use standalone application, which can automatically extract such information and mitigate the difficulty of analyzing tremendous clinical data. This extractor is able to extract temporal expressions and events with their attributes and relations. The system has been tested on the colon cancer data and brain cancer data of Semeval 2017 Clinical TempEval data and achieves the top performance comparable to the state of the art. It achieves 0.62 F-measure for the temporal expression extraction, 0.73 for event extraction and 0.59 for document time relation extraction.

Keywords: Clinical temporal information extraction; Data mining; Machine learning

Abbreviations: CIES: Clinic Temporal Information Extraction System; UIMA: Unstructured Information Management Architecture; POS: Part-of-speech; CRF: Conditional Random Field; SVM: Support Vector Machine; DR: Document-time Relations; RNN: Recurrent Neural Networks; XML: eXtensible Markup Language

Background

Extraction and interpretation of temporal information from the clinical text are essential for clinical practitioners and researchers. Extracting temporal information from unstructured clinical narratives is an important step towards the accurate construction of a patient timeline over the course of clinical care. We here present CIES, which extracts time, event expressions, their attributes, and relations. CIES employs a hybrid architecture utilizing machine learning and syntactic rules. The system has been evaluated on medical narratives from Clinical TempEval's THYME corpus and achieves state-of-the-art performance for temporal and relational information extraction. Its event extraction has a comparable performance in comparison with the current best result.

Many academic researchers, who want to use clinical information, utilize the SemEval Task 12 (Clinical TempEval) database. Researchers have explored ways to extract temporal information from clinical text [1], developed an approach based on linear and structural (HMM) support vector machines using lexical, morphological, syntactic, discourse, and word representation features. Velupillai [2] developed a pipeline based on ClearTK and SVM with lexical features to extract TIMEX3 and EVENT mentions. Most of the participants of these challenges used CRF and SVM for the event and time expression extraction with features including the information gathered from different resources like UMLS (Unified Medical Language System), the output of TARSQI toolkit, Brown Clustering, Wikipedia and Metamap [3]. Those machine-learning methods may be a kind of complex that cost much time to run. However, they can be not only flexible but also convenient when compared to the handcrafting label. Others also used some rule-based methods, which are fast but not flexible enough. It seems that the combination of those two methods may gain a better result. Since in I2b2 2012 temporal challenge, all top performing teams used a combination of supervised classification and rule-based methods for extracting temporal information and relations [4]. Besides THYME corpus, there have been other efforts in clinical temporal annotation including works by Roberts [5], Savova [6,7] and so on. Recently, interest in temporal processing has moved forward in two directions: cross-document timeline extraction [3] and domain adaptation [4,8] To our knowledge, no ready-to-use system is provided to the users to extract such complete temporal information from clinical data. The

	Attribute	Coloncancer Train	Braincancer Train	Coloncancer Dev
EVENT	Documents	293	30	147
	aspectual	546	51	246
	evidential	2,206	85	1,314
	N/A	36,185	2,421	19,414
	most	96	2	45
	little	143	18	65
	N/A	38,698	2,537	20,864
	positive	34,832	2,386	18,795
	negative	4,105	171	2,179
	actual	35,781	2,172	22,647
	hedged	889	81	443
	hypothetical	1,656	88	829
	generic	611	216	611
TIMEX	Date	2,588	204	1,422
	Duration	434	29	200
	Pre-Post Exp	313	37	172
	Set	218	13	116
	Quantifier	162	9	109
	Time	118	58	59

Table 1: Different time and event attributes in the THYME3 corpus.

***Corresponding authors:** Zhijing Li, Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Tech. R and D, Xi'an Jiaotong University, China, E-mail: tokyojackson@126.com

Chen Li, Shaanxi Province Key Laboratory of Satellite and Terrestrial Network Tech. R and D, Xi'an Jiaotong University, China, E-mail: cli@xjtu.edu.cn

Received Ocober 22, 2018; **Accepted** Ocober 31, 2018; **Published** October 01, 2018

Citation: Li Z, Long Y, Wang X, Zheng Q, Li C (2018) CIES: Clinic Temporal Information Extraction System. Int J Biomed Data Min 7: 136. doi: 10.4172/2090-4924.1000136

extracted temporal information may well present a structured result to the users and could be incorporated into other analysis systems (Table 1).

Implementation

The CIES extractor utilizes a pipeline to extract the time and event expressions along with document time relations from the clinic texts. Figure 1 (A) shows the workflow of the system.

The pipeline adopts the Unstructured Information Management Architecture (UIMA) framework to incorporate the workflow of all processes.

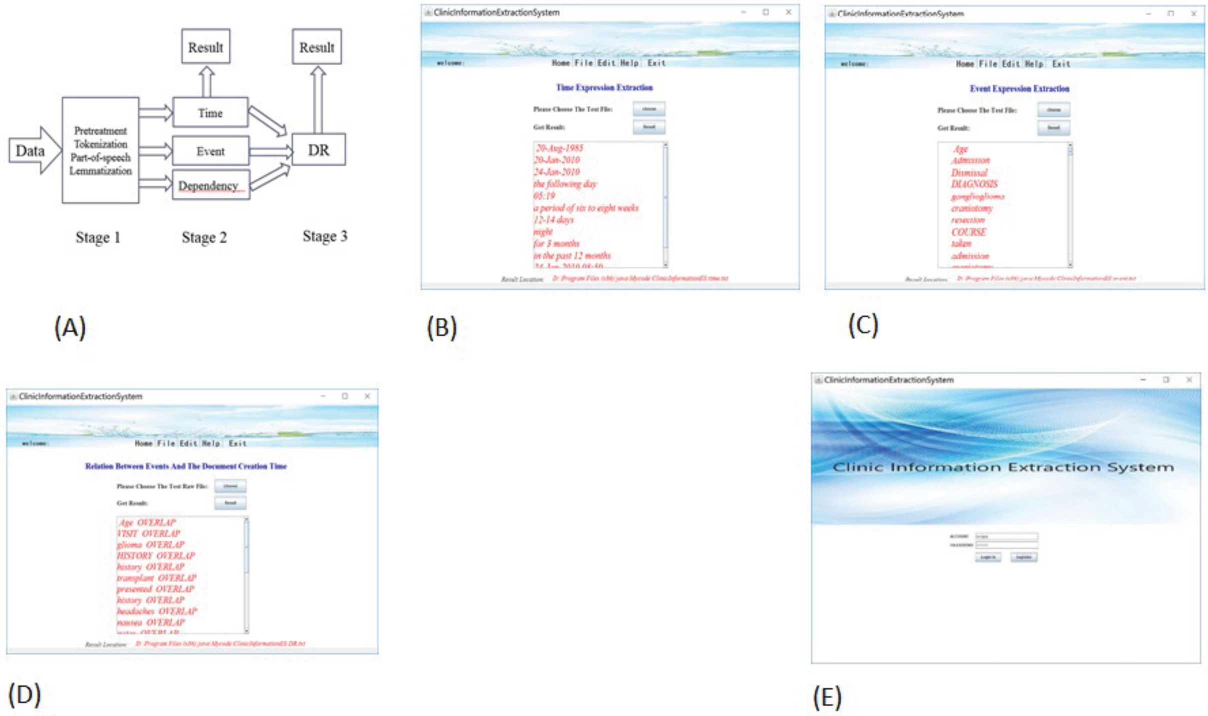
For text preprocessing including tokenization, part of speech tagging and lemmatization, the system uses the Stanford coreNLP toolkit. A hybrid method of machine learning and rules using lexical and Part-Of-Speech (POS) information is designed for extracting time expression. In the machine learning method, the Conditional Random Field (CRF) algorithm was used to extract time expressions. The features we used are some of the manually crafted linguistic feature combinations (tokenization, parts of speech, prototype, N-gram) that are specifically implemented using the CRF++ system. As for the rule-based method, we use the regular expressions to extract the standard format of the time expressions (e.g. 2013-09-09, 23-Jan-1993). Since the manually crafted linguistic features cannot reflect the relevant information about these standard formats, temporal expressions are extracted by using these rules. The results of the time expression extracted by CIES are shown in Figure 1 (B).

We extract medical events from the clinical text, first, we make a colon corpus about colon cancer which comes from training data. The classifier we choose to extract the event expression is CRF++, then we train a CRF classifier to complete prediction as the way we get the time expression. Finally, we can remove the events which exist in the clinical test data from the prediction result. The major features we used for training the SVM (Support Vector Machine) classifier are pos, lemma, and N-gram. We also extract all the features by using the Stanford coreNLP package. The results of the time expression extracted by CLES are shown in Figure 1 (C).

Document-time relations (DR) are specific attributes of EVENTs indicating their temporal relation with the document creation time. There are 3 different types of DRs, namely, BEFORE, AFTER, OVERLAP and BEFORE/AFTER, Some of the words belong to both BEFORE and AFTER. For identifying the DR attribute types, the RNN (Recurrent Neural Networks) classifier was used. The classifier was trained for each DR type using a set of features, the main features of identifying DR are word embeddings and dependency, verb tense and the models in the sentence are also indicative of the sentence tense and can help in identifying the document-time relation. The results of the DR extracted by CIES are shown in Figure 1 (D).

Description

The CLES extractor aims to extract the time and event expressions and the document time relations from the clinic texts. We show the main interface of our system in Figure 1 (E). Users should register accounts and passwords, and they can change it anytime under the Help function. After users enter the system, they can find the introduction of the system so that users can know the system better. To do the extractions, the model was needed to train, the given colon cancer XML- files (extensible Markup Language) were used and all contained elements were put into an internal data structure. For a detailed format of the XML file, users can refer to the available website given in this paper. To get further information and annotation, the CIES is queried via the golden annotations for each element in the document (entity, id, span, type, properties, etc.). In the first step of the extraction,


Result
Result
Data
Pretreatment
Tokenization
Part-of-speech
Lemmatization
Time
Event
Dependency
DR
Stage 1
Stage 2
Stage 3
(A)
ClinicInformationExtractionSystem
welcome:
Home File Edit Help Exit
Time Expression Extraction
Please Choose The Text File:
Get Result:
Result Location:
(B)
ClinicInformationExtractionSystem
welcome:
Home File Edit Help Exit
Event Expression Extraction
Please Choose The Text File:
Get Result:
Result Location:
(C)
ClinicInformationExtractionSystem
welcome:
Home File Edit Help Exit
Relation Between Events And The Document Creation Time
Please Choose The Text Raw File:
Get Result:
Result Location:
(D)
ClinicInformationExtractionSystem
Clinic Information Extraction System
(E)

20-Aug-1985
20-Jan-2010
24-Jan-2010
the following day
05:19
a period of six to eight weeks
12-14 days
night
for 3 months
in the past 12 months

Age
Admission
Dismissal
DIAGNOSIS
gangliogliomas
craniotomy
resection
COURSE
taken
admission

Age OVERLAP
VISIT OVERLAP
glioma OVERLAP
HISTORY OVERLAP
history OVERLAP
transplant OVERLAP
presented OVERLAP
history OVERLAP
headaches OVERLAP
nausea OVERLAP

Figure 1(A): The system workflow, **(B):** The time expression extraction in CIES, **(C):** The time expression extraction in CIES, **(D):** The DR extraction in CIES, and **(E):** Main interface of the system.

Subtask	System	P	R	F
TIMEX3_SPAN	CIES	0.72	0.55	0.62
	LIMSI	0.51	0.67	0.58
EVENTS_SPAN	CIES	0.68	0.80	0.73
	LIMSI	0.69	0.85	0.76
DR	CIES	0.70	0.51	0.59
	LIMSI	0.53	0.66	0.59

Table 2: Results for each subtask.

the users may choose to let CIES extractor to extract information automatically under the FILE button, our system has three functions-TIMEX3, EVENTs, DR. Each function will pop up the corresponding window. As for TIME and EVENTS functions, the extractor has one input pathway, and users could browse their computers and choose the text file, then press the execute button, thus, the extractor will extract the time and event expressions and users can check the results through the window. For identifying the DR attribute types, users should also browse the input file and choose the text file, CIES will extract the event expressions first and then identify their DR attribute types, CIES will display the results in the window according to the event expression extraction result. We also have the additional function, this function can identify the DR attribute types directly without extracting event expressions. Users only need to input the extracted event expressions and can get the DR attribute types.

Results

Datasets

We use THYME corpus for training and evaluating the methods, which consists of clinical and pathology notes of patients with colon cancer and brain cancer from Mayo Clinic. The THYME corpus is split into training, development, and test sets based on the patient number, with 50% in training and 25% each in development and test sets. The training data about colon cancer contains 3,833-time expressions and 38,890 events, the development data contains 2,078-time expressions and 20,974 events. The training data about brain cancer contains 350-time expressions and 2,557 events.

Result and analysis

We compare our results with the best results (LIMSI) posted on the SemEval2017 web site. Table 2 shows that the approach achieves relatively good results comparing with the previous best system, even better than the best result, that means CIES could extract the clinical information relatively effectively. Our system still has a lot of room for improvement. There are some options to improve the system, ranging from fine-tuning the pre-processing phase in order to avoid offset misalignments, to the generation of better features for the event and DR extraction. In future work, we aim to implement all the improvements mentioned above.

Conclusions

CIES extractor is a stand-alone application with a graphical user interface that runs on windows system for which a Java™ virtual machine is available. To our knowledge, there is no other extractor is available to extract clinical formatted files to visible output formats in different domains, and has so many functions, using various features and methods. Furthermore, CIES extractor is simple, easy-to-use and comes with a powerful command-line and graphical user interface. It allows a quick and easy usage of files from the SemEval Task 12 (Clinical TempEval) database in a wide range of other applications. The extracted temporal information may well present a structured result to the users and could be incorporated into other analysis systems. Our system also has a lot of space for improvement, we will also continue to strive to improve results and develop more new functions.

Availability and Requirements

Project name: CIES (Clinic Temporal Information Extraction System)

Project homepage: https://github.com/tokyojackson/CIES

Operating system(s): Platform independent

Programming language: Java

Other requirements: Java 1.3.1 or higher

License: GNU GPL

Any restrictions to use by non-academics: license needed

Declarations

Ethics approval and consent to participate: All the data we used were provided by Semeval 2017 Task 12: Clinical TempEval organizers. The organizers of Semeval 2017 agreed to our use of the data. We signed a confidentiality agreement with MAYO clinic and promise not to divulge data.

Consent for publication: Not applicable

Availability of Data and Materials

The data that support the findings of this study are available from [MAYO clinic] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [MAYO clinic].

Competing interests: The authors declare that they have no competing interests.

Funding

This work has been supported by "The Fundamental Theory and Applications of Big Data with Knowledge Engineering" under the National Key Research and Development Program of China with grant number 2016YFB1000903; Project of China Knowledge Centre for Engineering Science and Technology; Innovation team of Ministry of Education (IRT_17R86); Innovative Research Group of the Nation Natural Science Foundation of China (61721002); Ministry of Education-Research Foundation of China Mobile Communication Corp(MCM20160404); Professor Chen Li's Recruitment Program for Young Professionals of "the Thousand Talents Plan".

Authors' Contributions

ZJL was a major contributor in writing the manuscript and CL made the modifications. All authors read and approved the final manuscript.

Acknowledgments

Thank the MAYO clinic for providing us the clinic data and give us great help in scientific research.

References

1. Christopher D, Mihai S, John B, Jenny F, Steven J, et al. (2016) The stanford core NLP natural language processing toolkit. In proceedings of the 52nd annual meeting of the assoc for computational linguistics: System demonstrations 55-60.
2. Rodriguez P, Wiles J, Elman JL (1999) A RNN that learns to count. Connection Sci 11:5-40.
3. Jain A, Zamir AR, Savarese S, Saxena A (2015) Structural RNN: Deep learning on spatio temporal graphs. Computer Sci.
4. Namikawa J, Tani J (2008) A model for learning to segment temporal sequences, utilizing a mixture of RNN experts together with adaptive variance. Neural Networks 21:1466-1475.
5. Alan RA, Franc ML (2010) An overview of Meta Map: historical perspective and recent advances. J of the American Med Informatics Assoc. 17:229-236.

6. Steven B, Leon D, James P, Marc V(2015) Semeval 2015 task 6: Clinical tempeval. In Proceedings of the 9th International workshop on semantic evaluation. Assoc for Computational Linguistics.
7. Wang WJ, Liao YF, Chen SH (2002) RNN-based prosodic modeling for mandarin speech and its application to speech-to-text conversion. Speech Communication 36:247-265.
8. Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, et al. (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. Computer Sci 3:1724-1734.