Received 16 August 2017
Accepted 30 August 2017

Mining medical related temporal information from patients' self-description

Lichao Zhu
School of Management and Economics,
Beijing Institute of Technology, Beijing, China, and
Hangzhou Yang and Zhijun Yan
Beijing Institute of Technology, Beijing, China

Abstract

Purpose – The purpose of this paper is to develop a new method to extract medical temporal information from online health communities.

Design/methodology/approach – The authors trained a conditional random-filed model for the extraction of temporal expressions. The temporal relation identification is considered as a classification task and several support vector machine classifiers are built in the proposed method. For the model training, the authors extracted some high-level semantic features including co-reference relationship of medical concepts and the semantic similarity among words.

Findings – For the extraction of TIMEX, the authors find that well-formatted expressions are easy to recognize, and the main challenge is the relative TIMEX such as "three days after onset". It also shows the same difficulty for normalization of absolute date or well-formatted duration, whereas frequency is easier to be normalized. For the identification of DocTimeRel, the result is fairly well, and the relation is difficult to identify when it involves a relative TIMEX or a hypothetical concept.

Originality/value – The authors proposed a new method to extract temporal information from the online clinical data and evaluated the usefulness of different level of syntactic features in this task.

Keywords Support vector machine, Co-reference, Conditional random field, Temporal information extraction, Word embedding

Paper type Technical paper

1. Introduction

In recent years, with the popularity of online health communities, more and more people are taking advantage of internet to obtain medical information or even professional help for personal medical problems. Online medical data contain great potential value and are growing in an exponential rate which has been attracting scholars to carry out much research work in many aspects, including the automatic medical entity identification and

This paper was funded by National Natural Science Foundation of China (Award: 71572013, 71272057) and Beijing Natural Science Foundation (Award: 9152015).

International Journal of Crowd Science
Vol. 1 No. 2, 2017
pp. 110-120
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-08-2017-0018

entity relationship extraction (Martnez *et al.*, 2016; Nie *et al.*, 2014), the monitoring of adverse drug reactions from social media (Azadeh *et al.*, 2015; Ruwen *et al.*, 2016), the automatic diagnosis of diseases using online medical QA data (Nie *et al.*, 2015). However, existing studies barely considered temporal information when extracting useful medical knowledge from online medical data. Medical time-line is an essential factor in diagnosis and treatment.

The timing information of medical event can help to track the development of patients' medical condition, determine the causes of diseases, detect the roles of drugs and adverse reactions, which can improve the disease cure effect and predict disease progression. Therefore, extracting temporal information from online medical data is a critical step for online health information extraction.

For temporal information extraction in clinical domain, existing studies mainly based on electronic medical records (EMR). The extraction process mainly includes two aspects, the extraction of event and temporal expression and the recognition of temporal relation. The temporal relation can be divided into three types: relation between an event or a temporal expression and the creation time of document, relation between the event and event and relation between event and temporal expression.

However, there are significantly differences between EMR data and online health communities' data. First, online data contain much noisier information which would impact the extraction result seriously. Besides, people usually use diverse expressions rather than professional medical terms in internet for the same concept. Furthermore, existing methods developed for EMR data are quite limited in dealing with relative temporal expression. Aiming for solving these problems, this paper proposes a novel approach that contains a corpus filtering system and a robust co-reference identification system to extract temporal information from online medical data.

The rest of the paper is organized as follows. Section 2 will introduce the related research on temporal information extraction. Then Section 3 will describe the entire framework of our method followed by the presentation of evaluation and results in Section 4. Finally, we discuss major research findings and limitations of this study in Section 5.

2. Related work

2.1 Temporal information extraction in general domain

The temporal information recognition in general domain used news corpus and includes two sub-tasks: event (EVENT) and temporal expression (TIMEX) recognition and temporal relation identification (TLINK). The definition of EVENT, TIMEX and TLINK are mainly based on the definition in TimeML (www.timeml.org/). EVENT mainly refers to a verb word, and the TLINK includes 13 types, such as "before" overlap "after" and so on.

For TIMEX recognition, machine learning method, rule systems and the combination of both have achieved similar results (Filannino *et al.*, 2013; Bethard, 2017; Strtgen *et al.*, 2013). In the aspect of the normalization of time phrase, the method of making rules is more effective than machine learning. As for EVENT recognition, recent studies are mostly based on machine learning methods, including conditional random fields (CRF), support vector machine (SVM), maximum entropy (MaxEnt) and logistic regression. For temporal relations identification, most existing studies are based on machine learning methods or the combination of machine learning methods and rules-based methods. About the feature engineering, Laokulrat *et al.* (2013) introduced semantic information at the sentence level through deep parsing and achieved a top result.

2.2 Temporal information extraction in clinical domain

Temporal information identification in the clinical domain mainly contains two sub-tasks also (Sun *et al.*, 2013b). The first step is the EVENT and TIMEX extraction, including extraction and recognition properties of the phrase. The second step is to identify the temporal relations between those extracted entities.

2.2.1 EVENT and TIMEX extraction. The definition of EVENT in the medical area is quite different with that in the common areas, and contains more medical specific information. The EVENT includes six types of entities which are medical problem, examination, treatment, department, occurrence and evidential. Most of current studies that focus on the extraction of event use CRF as the tagging model and SVM model for the recognition of event attributes. Lee *et al.* (2016) used HMM-SVM model for event recognition and SVM model for attributes recognition of EVENT. The model applied basic features such as lexical features, morphological features and syntactic features. Many researchers have tried to use external information resources to improve the performance. Roberts *et al.* (2013) uses Brown clustering to make use of the external text resource to find the similarity between phrases. Lin *et al.* (2013) uses Wikipedia and MetaMap to extract semantic features of medical terms.

For the extraction and normalization of TIMEX, current approaches usually combine the output of rule-based system and machine learning models such as SVM and MaxEnt (Roberts *et al.*, 2013; Lin *et al.*, 2013; Xu *et al.*, 2013). Tang *et al.* (2013) and Grouin *et al.* (2013) acquired competitive results by applying existing temporal annotation systems such as HeidelTime (Strtgen and Gertz, 2010) and SUTIME (Chang and Manning, 2012).

2.2.2 TLINK classification. Numerous machine-learning methods are adopted for temporal relation classification including MaxEnt, Bayes, SVM and CRF. Some studies also explored the use of new machine learning models, such as neural network learning model (Li and Huang, 2016) and deep learning framework (Chikka, 2016). Fries (2016) used recurrent neural network and DeepDive (Zhang, 2015) framework for the extraction of event and TIMEX and the TLINK although the result in the existing studies is not the best; it is still an acceptable result in the condition of without much feature engineering. Besides, some studies also incorporate heuristic and rule-based components. For example, Cherry *et al.* (2013) divides the task into four sub-tasks: the relationship between EVENT and document creation time, event/TIMEX in a same sentence, event/TIMEX in different sentence and relations referred by the causal relationship. Tang *et al.* (2013) uses a heuristic algorithm to identify candidate phrases for a TLINK. Chang *et al.* (2013) integrates the results produced by MaxEnt with the results generated by the rules. Nikfarjam *et al.* (2013) integrates the results of the SVM and the results of rule system.

Based on our observation, there are still two challenges remain unsolved. The first one is the normalization of relative temporal expression, such as “the week before the operation”. Further, although the relationship between event and document time and the time relationship in the same sentence are easy to identify, it is difficult to identify the temporal relations between EVENT/TIMEX in different sentences.

3. Method

To address these problems of existing approaches, we proposed an integrative method which incorporate high-level semantic features. Through word embedding, we could grasp semantic similarity among words which may not appear in training data. As for the relative temporal expression, we mined the co-reference relation between concepts to infer the absolute time point of the relative TIMEX. The pro- posed method mainly includes the following steps: word segmentation and tagging, corpus filtering, high-level semantic feature construction, TIMEX extraction and TLINK identification. The whole process of our method is showed in Figure 1.

3.1 Word segmentation and part-of-speech tagging

We first removed online information that is less than 100 characters, and the invalid characters such as HTML tags, extra spaces and other special symbols. Then, we performed word segmentation using a package and part of speech tagging using Stanford part-of-speech (POS) tagger. The following example shows a segmented Chinese sentence with POS tags: Segmented sentence: "In August, MRI scan showed that the shape of both legs was basically symmetrical".

Result of POS tagging: "In/IN August/NNP/PU MRI/NNP scan/VB showed/VBD that/IN the/DT shape/NN of/IN both/CC legs/NNS was/VBD basically/RB symmetrical/JJ".

Among these POS tags, "NNP" means proper noun, singular; "NNS" means noun, plural; "VB" means verb, base form; "CC" means coordinating conjunction; "PU" represent a punctuation, "JJ" means adjective and "RB" means adverb.

3.2 Corpus filtering

The corpus we try to utilize is the objective description of a patient's own condition in the online environment. Patients' description in online communities usually contains irrelevant or subjective contents. Those irrelevant contents include the description of patients' daily life or expressions about worrying, which usually do not contain useful medical information. The subjective contents in the online communities refer to a patient's subjective questions, such as if the patient can continue to take chemotherapy under certain condition. These unreliable expressions would increase the difficulty of further identification.

To obtain higher quality corpus, we conduct a further screening to filter the irrelevant or subjective contents from the original corpus. We filtered the corpus in the sentence level and classify useless sentences using the bag of words model and SVM classifier. We used tokens only as basic features and add a keyword list as another feature which consists of words that often appear in useless sentence.

At the same time, we removed the information of which the length is less than 200 characters considering that short information usually contains little medical information and cannot form a medical timeline.

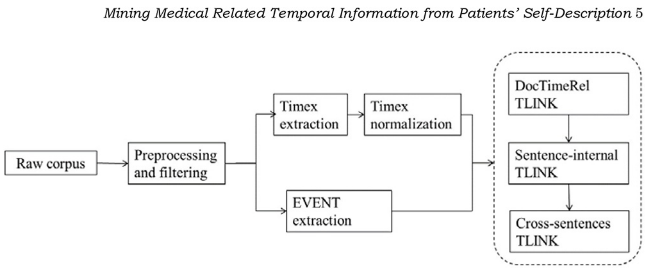

Figure 1.
Overview of our medical temporal information extraction method

3.3 High-level semantic features

We constructed two kinds of high-level semantic features. The first one is semantic similarity among words and the other is co-reference relation among medical concepts.

The supervised tagging model often struggle to get correct results when it comes to the word that does not or seldom appear in the training set. To solve this problem, we introduce the use of word vectors. The word vector is a distributed representation of words and each word is mapped to a lower dimensional vector with real values through a multi-layer neural network. As a result, the distance between vectors represents the semantic similarity between two words. If two word vectors have a short distance, we can infer that these two words often appear in similar contexts.

We use the Google's Open Toolkit word2vec (https://code.google.com/p/word2vec) as the training algorithm. All the unlabeled corpus which has not been filtered out is used as the training corpus of the word vector. The dimension of the result vector is set to 100 and 200. Besides, we apply the word vector in our method through two different ways. The first is to use the clustering result of word vectors as an extra feature for the training of CRF model. The other is to use the embedded word vector to represent each token when classify temporal relations.

Meanwhile, from our observation, the co-reference relation is quite useful when classify the temporal relations. So we defined a simple rule to identify the possible co-reference relation. If two events existing in adjacent sentence share the same type and the latter one is a part of the first one in word composition, then there is a co-reference relation between these two events.

3.4 TIMEX extraction

We use the CRF++ as tag model for TIMEX extraction. We use the context tokens and POS tag as the basic features. Besides, considering the unique characteristics of Chinese temporal expressions, we also include a series of Chinese characters as extra features which show temporal meaning explicitly, including: "year, month, day, week, etc". In addition, digit often appear in temporal expressions and it is easy to get confused with the indicators of medical test, so we limit the digit in the range 1-31 as another kind of features. All the features we used are showed in Table I. For example, the original sentence is "one month after the operation, we took the chemotherapy in May". If the current word is "chemotherapy" then the context words in a window of [−2,2] would be [took, the, in, May]. The value for feature "temporal character" and "temporal digit" would be "no" both.

The temporal expressions may share some similarities in medical area and other domains especially for some common date or duration expressions. So we use other existing corpus as supplement training corpus to improve the accuracy of recognition. This paper makes use of the data provided in seminal 2010 (Verhagen *et al.*, 2010) which is a news corpus collected from People's Daily as an additional training corpus.

Table I. Feature for TIMEX extraction

Feature	Description
Current and context tokens	Current token and the tokens in a window of [−2, 2]
POS tag	The POS tag of current and context tokens
Temporal character	Current token if contains temporal character
Temporal digit	Current token if contains temporal digit

3.5 TIMEX normalization

Current studies about TIMEX normalization mostly use rule-based methods, which normally summarize the regular pattern of temporal phrases through grammar analysis.

First, we make use of the rule system in HeidelTime (Strtgen, 2013) annotation systems and supplement some extra rules. Each rule is composed of two parts: the regex expression for matching and the corresponding normalized value. Here is a concrete example: regex: "October of 2016" is normalized as: "2016-10" Second, we analyze the expressions that cannot be matched by the rules. These irregular expressions can be divided into two classes. The first class is the relative temporal expression which often involves an event entity like "five days prior to the onset".

We need to determine the local time of the "onset" first which could be inferred by the TLINK with other entities, so this kind of expressions would be solved after TLINK identification. The second class is about the inexact temporal expressions, which do not indicate a specific time point, such as "at that time" "before" and "later" For this kind of expressions we use word vector-based clustering to match known value expressions.

3.6 TLINK classification

As there are three kinds of relations to be extracted in total, we use SVM model to construct several classifiers to classify the TLINKS. Depending on the relative positions of entities, TLINK could be divided into three categories: TLINK with document creation time (DocTimeRel), TLINK within same sentence and TLINK in different sentences. In this paper, we introduce the identification of DocTimeRel.

Before the identification of DocTimeRel, we need to identify the creation time of a document itself. In this paper, the creation time of each document is the date that the patient posted it on website. For classification, in addition to the basic features, we also consider the attributes of entities, and we find that the first verb and the last word of a sentence are useful in some situation. The features we use for this task are showed in Table II. For example, the original sentence is "the day before yesterday had a fever, now taking treatment in hospital" Assuming the current event is "fever" the entity words would be [fever] and the event type is "problem" the context tokens in a window of [-2,2] would be [had, a, now]. The TIMEX would be [the day before yesterday] and corresponding value is "2016-12-19" (the DocTime is "2016-12-21").

4. Evaluation

4.1 Data acquisition

We crawled online patient consultation posts of two departments: pediatrics and oncology from "www.haodf.com" generated between September and December in 2016. The origin

Feature	Description
Event tokens	Each token contained in the event
Event type	Event type
Context tokens	Tokens in a window of [−2, 2]
POS tag	The POS tag of context tokens
TIMEX	The type and value of the first Date or TimeTimex in current sentence
Context of TIMEX	In a window of [−1, 1]

Table II.
Feature for DocTimeRel

data contains more than 30,000 records, and we got a corpus of 8,600 records after preprocessing and filtering.

We follow the definition provided by Sun *et al.* (2013a), and three graduate students independently annotated the EVENT/TIMEX and TLINK in those records. Then, they discussed disagreed results, and only the annotations that all three reviewers had agreed upon at the end were included in the correct annotation sets. To increase the tagging efficiency, we separated the whole work into several sub-tasks: tagging of entity phrase, tagging of TIMEX normalization value and tagging of TLINK. Besides, we built a java application for the first task and applied the multi-purpose annotation environment toolkit for the second and third tasks. Both of these software can display each document entirety and record the tagging information automatically which help to reduce the tagging time.

Before the submission of this paper, we annotated 3612 TIMEXs, 1131 EVENTs and 1131 corresponding DocTimeRel TLINKs. A completed annotated sentence is showed in Figure 2.

4.2 TIMEX extraction and normalization

4.2.1 TIMEX extraction. The annotated corpus was divided into training data and test data in a proportion of 2:1. For the extraction of temporal expressions, we first train the CRF model based on the combination of our own corpus and the news corpus. Then we use the trained CRF model to tag the test data automatically and then we use the normalization rules to modify the result. The final result is showed in Table III. The last column shows the number of each type of TIMEX, and the last row shows the weighted average of each

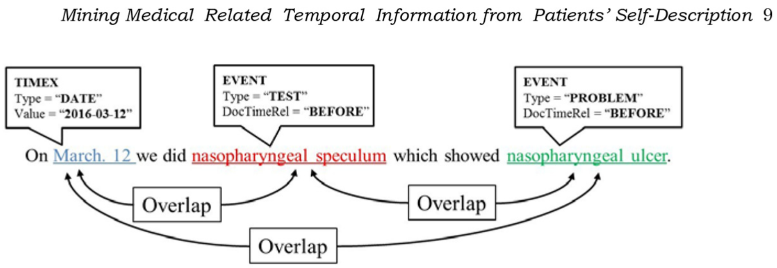
Mining Medical Related Temporal Information from Patients' Self-Description 9
TIMEX
Type = "DATE"
Value = "2016-03-12"
EVENT
Type = "TEST"
DocTimeRel = "BEFORE"
EVENT
Type = "PROBLEM"
DocTimeRel = "BEFORE"
On March. 12 we did nasopharyngeal speculum which showed nasopharyngeal ulcer.
Overlap
Overlap
Overlap

Figure 2.
Example of a sentence with annotated event, temporal expression and temporal relation

Note: Mining medical-related temporal information from Patients' Self-Description 9

Table III.
Result of TIMEX extraction

TIMEX type	Precision	Recall	F1	No. of instances
Date	0.7941	0.5684	0.6626	2660
Duration	0.8892	0.7083	0.8293	672
Frequency	0.8566	0.2623	0.4016	168
Time	0.9321	0.7501	0.8571	112
Weighted Avg.	0.8483	0.5814	0.6899	3612

measure. From the result, we can find that for four types of TIMEX, the precision measures are all higher than the recall measure, but the recall of type: "frequency" is much lower than the precision which may be resulted from the unbalanced distribution of frequency TIMEX in training and testing data.

4.2.2 TIMEX normalization. After observing several hundreds of TIMEX instances, we constructed a rule system with 106 rules in total, more detail of the rules distribution is showed in Table IV. The last column showed the percentage of TIMEX which could be matched using the normalization rules. The matching results show that the proportion of the four types of TIMEX can be normalized by the rules are about 70 per cent, which indicates that a considerable part of the TIMEXs are expressed in infrequent patterns.

4.3 DocTimeRel identification. Because for each event there must be a temporal relation with the DocTime, we trained three classifiers for each type of TLINK. The TLINK of "before" which occupies a large part of all TLINK shows the best F1 score. The result is showed in Table V. As the amount of annotated corpus of TLINK is quite limited for now, the identification result is far from ideal output. Besides, the TLINKS of overlap and after only occupy a small amount of total; the result would be improved much more with sufficient training data.

5. Conclusion

For the extraction of TIMEX, we find that well-formatted expressions are easy to recognize, and the main challenge is the relative TIMEX such as "three days after onset". It also shows the same difficulty for normalization of absolute date or well formatted duration, whereas frequency is easier to be normalized. For the identification of DocTimeRel, the result is fairly well, and the relation is difficult to identify when it involves a relative TIMEX or a hypothetic concept.

Rule Type	No. of rules	Correctly matched (%)
Date	68	58.9
Duration	25	65.3
Frequency	8	54.1
Time	5	68.5
Overall	106	61.8

Table IV. Result of TIMEX normalization

Relation type	Precision	Recall	F1	No. of instances
Before	0.780	0.565	0.655	1,002
Overlap	0.273	0.360	0.310	75
After	0.250	0.462	0.324	54
Weighted average	0.597	0.505	0.535	1,131

Table V. Result of DocTimeRel identification

There are several limitations for this study. First of all, we are still working about the extraction of event and the TLINK within a sentence or cross sentences which are important to acquire a complete clinical timeline. Second, the effect of the high-level features should be tested in the construction of machine learning models. Finally, building an inference system would be a potential solution for a further improvement of our method as we used some rule based system in our current research.

References

Azadeh, N., Abeed, S., Karen, O., Rachel, G. and Graciela, G. (2015), "Pharma- covigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features", *Journal of the American Medical Informatics Association Jamia*, Vol. 22 No. 3, pp. 671-681.

Bethard, S. (2017), Cleartk-timeml: A minimalist approach to tempeval 2013. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia*, pp. 10-14.

Chang, A.X. and Manning, C.D. (2012), "Sutime: a library for recognizing and normalizing time expressions", *Lrec*, Vol. 9 No. 1, pp. 3735-3740.

Chang, Y.C., Dai, H.J., Wu, J.C., Chen, J.M., Tsai, R.T. and Hsu, W.L. (2013), "Tempting system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries", *Journal of Biomedical Informatics*, Vol. 46, pp. S54-S62.

Cherry, C., Zhu, X., Martin, J. and Bruijn, B.D. (2013), "La recherche du temps perdu: extracting temporal relations from medical text in the 2012 i2b2 nlp challenge", *Journal of the American Medical Informatics Association*, Vol. 20 No. 5, pp. 843-848.

Chikka, V.R. (2016), "Cde-iiith at semeval-2016 task 12: Extraction of temporal information from clinical documents using machine learning techniques", *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), Association for Computational Linguistics, San Diego, CA*, pp. 1237-1240.

Filannino, M., Brown, G. and Nenadic, G. (2013), "Mantime: Temporal expression identification and normalization in the tempeval-3 challenge", *Second Joint Conference on Lexical and Computational Semantics*, Computer Science.

Fries, J.A. (2016), "Brundlefly at semeval-2016 task 12: Recurrent neural networks vs. joint inference for clinical temporal information extraction", *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), Association for Computational Linguistics, Computer Science*, pp. 1274-1279.

Grouin, C., Grabar, N., Hamon, T., Rosset, S., Tannier, X. and Zweigenbaum, P. (2013), "Eventual situations for timeline extraction from clinical reports", *Journal of the American Medical Informatics Association Jamia*, Vol. 20 No. 5, pp. 820-827.

Laokulrat, N., Miwa, M., Tsuruoka, Y. and Chikayama, T. (2013), Uttime: Temporal relation classification using deep syntactic features. *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Association for Computational Linguistics, Atlanta, Georgia*, pp. 88-92.

Lee, H.J., Xu, H., Wang, J., Zhang, Y., Moon, S., Xu, J. and Wu, Y. (2016), "Uthealth at semeval-2016 task 12: an end-to-end system for temporal information extraction from clinical notes"*Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, Association for Computational Linguistics, San Diego, CA, pp. 1292-1297.

Li, P. and Huang, H. (2016), "Uta dlnlp at semeval-2016 task 12: Deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports", *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), Association for Computational Linguistics, San Diego, CA*, pp. 1268-1273.

Lin, Y.K., Chen, H. and Brown, R.A. (2013), "Medtime: a temporal information extraction system for clinical narratives", *Journal of Biomedical Informatics, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), Association for Computational Linguistics, San Diego, CA*, Vol. 46, pp. S20-S28.

Martnez, P., Martnez, J.L., Segura-Bedmar, I., Moreno-Schneider, J., Luna, A. and Revert, R. (2016), "Turning user generated health-related content into actionable knowledge through text analytics services", *Computers in Industry*, Vol. 78, pp. 43-56.

Nie, L., Wang, M., Zhang, L., Yan, S., Zhang, B. and Chua, T.S. (2015), "Disease inference from health-related questions via sparse deep learning", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27 No. 8, pp. 2107-2119.

Nie, L., Zhao, Y.L., Akbari, M., Shen, J. and Chua, T.S. (2014), "Bridging the vocabulary gap between health seekers and healthcare knowledge", *IEEE TransActions on Knowledge & Data Engineering*, Vol. 27 No. 2, pp. 396-409.

Nikfarjam, A., Emadzadeh, E. and Gonzalez, G. (2013), "Towards generating a patient's timeline: Extracting temporal relationships from clinical notes", *Journal of Biomedical Informatics*, Vol. 46, pp. S40-S47.

Roberts, K., Rink, B. and Harabagiu, S.M. (2013), "A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text", *Journal of American Medical Informatics Association Jamia*, Vol. 20 No. 5, pp. 867-875.

Ruwen, B., Leocadie, V.H., Thomas, H., Hans-Joachim, K., Oliver, B., Holger, P. and Jan, H. (2016), "Openvigil FDA inspection of U.S. American adverse drug events pharmacovigilance data and novel clinical applications", *Plos One*, Vol. 11 No. 6, p. e0157753.

Strtgen, J. (2013), "Multilingual and cross-domain temporal tagging", *Language Resources & Evaluation*, Vol. 47 No. 2, pp. 269-298.

Strtgen, J. and Gertz, M. (2010), "Heideltime: High quality rule-based extraction and normalization of temporal expressions", *Proceedings of the 5th International Workshop on Semantic Evaluation*, pp. 321-324.

Strtgen, J., Zell, J. and Gertz, M. (2013), Heideltime: Tuning English and developing Spanish resources for tempeval-3, *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval'13, Atlanta, Georgia*, pp. 15–19.

Sun, W., Rumshisky, A. and Uzuner, O. (2013a), "Evaluating temporal relations in clinical text: 2012 i2b2 challenge", *Journal of the American Medical Informatics Association Jamia*, Vol. 20 No. 5, pp. 806-813.

Sun, W., Rumshisky, A. and Uzuner, O. (2013b), "Temporal reasoning over clinical text: the state of the art", *Journal of the American Medical Informatics Association Jamia*, Vol. 20 No. 5, pp. 814-819.

Tang, B., Wu, Y., Jiang, M., Chen, Y., Denny, J.C. and Xu, H. (2013), "A hybrid system for temporal information extraction from clinical text", *Journal of the American Medical Informatics Association Jamia*, Vol. 20 No. 5, pp. 828-835.

Verhagen, M., Roser, Caselli, T. and Pustejovsky, J. (2010), "Semeval-2010 task 13: tempeval-2", *Proceedings of the 5th international workshop on semantic evaluation*, pp. 57-62.

Xu, Y., Wang, Y., Liu, T., Tsujii, J. and Chang, E.I. (2013), "An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge", *Journal of the American Medical Informatics Association Jamia*, Vol. 20 No. 5, pp. 849-858.

Zhang, C. (2015), "Deepdive: A data management system for automatic knowledge base construction", *Dissertations & ThesesGradworks*.

About the authors

Lichao Zhu is a master’s student in Beijing Institute of Technology, School of Management and Economics. His research interests include e-commerce and medical big data analysis.

Hangzhou Yang is a PhD student in Beijing Institute of Technology, School of Management and Economics. His research interests include e-commerce, medical big data analysis and health-care management.

Professor Zhijun Yan is a Professor of Beijing Institute of Technology, School of Management and Economics. His research interests include e-commerce, health-care management, medical big data analysis, social network analysis. Zhijun Yan is the corresponding author and can be contacted at: yanzhijun@bit.edu.cn