ORIGINAL RESEARCH PAPER
Volume - 13 | Issue - 08 | August - 2024 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

A NOVEL ML BASED APPROACH FOR CLINICAL INFORMATION EXTRACTION BASED ON LEXICAL FEATURES AND CANDIDATE PAIRS

Computer Science

Vani K B Research Scholar, Dept of CSE, VIT., Puttur, India

Lokesh M. R Professor, Dept of CSE, VIT., Puttur, India

ABSTRACT

EMRs, or electronic medical records, are the main resources that make it possible to reuse clinical data. Since EMR data is free text, information must be extracted using natural language processing techniques. The main element that makes a difference in utilizing the valuable data found in the Information Extraction Task is its semantically focused functionality. Systems that use natural language processing (NLP) have the ability to carry out IE tasks. Such an NLP system would, however, produce inconsistent outcomes and always leave room for development. There may be room for improvement in the Clinical Information Extraction Process because the semantics of information extraction in EMRs will be dependent on temporal expression. Therefore, in this work, we offer an improved descriptive methodology that makes use of lexical-based clinical text extraction.

KEYWORDS

Candidate pairs, Temporal , HeidelTime, Tags, EMRs, Temporal

INTRODUCTION

With the rapid adoption of Electronic Medical Records(EMRs), it is being possible to support automated clinical decision making systems at the point of care by harvesting the information and knowledge from EMRs and thus thereby enable the secondary use of EMRs for Clinical and Translational research. However, much of the EMR data is in free-text form. Free text is a more natural and expressive method to document the clinical events and facilitate the communication among the care team in the health care environment. The prime critical factor for leveraging such a usefulness of EMRs in the Health Care Environment is Information Extraction (IE) tasks. IE is commonly referred as a special area in empirical natural language processing and it refers to automatic extraction of concepts, entities, events and their relation and associated attributes.

Most IE system are Expert Based System that consists of patterns defining lexical, synctatatic and semantic constraints. Most IE applications involve the subtasks such as concept or named entity recognition that identifies concept mentions or entity names for the text, co-reference resolution that associates mention or names referencing to the same entity and relation extractions that identifies the inherent and implicit relations between concepts, entities and attributes. NLP focuses on developing the computational model for understanding the language. An NLP system can include syntactic processing modules such as tokenizer, sentence detection, POS tagging etc., and /or semantic processing modules. Thus an IE application is an NLP system with semantic processing modules for extracting predefined types of information from text.

In the clinical domain NLP system have to identifies and extract temporal information from electronically implemented clinical narratives and reports. Temporal information means time based expression, events and associated relations between them in the clinical text. Unlike Newswire text, clinical text records are complicated to annotate and extract information from them. Over the years many methodologies were developed to extract temporal information the clinical narratives. However, all those methods were attempted using annotated corpora.

Although few researchers showed ways to extract temporal information out of unannotated corpora, there is wide scope for enforcing the novel ideas and further redefine the process using machine learning algorithms to get better and improved results in temporal information extraction from the clinical text. In this research work to be carried out, it is expected to adopt enhanced lexical feature set such as medical dictionary data and semantic ontology's to extract the clinical event with improved results and then make use of enhanced hypothesis to generate candidate pairs of events and expression to enhance the temporal relation classification and thus to obtain improved classical performance.

Background Of The Work

Clinical decision making systems or automated systems at point of care in the state of the art of health care facilities, makes use of knowledge that can be derived from clinical text [1]. With the introduction of TIMEX extraction in MUC-6, TIMEML with annotation guidelines and corpora was developed for the processing of natural language text. The development of Corpora facilitates the researchers to identify and extract events and TIMEX operations from in a text documents.

The existence of clinical text in the recent electronics medical records(EMRs) has received a great level of recognition due to the need for analysis of progressions of disease and effect of treatment procedure. However, EMRs clinical text poses technical challenges due to nature of the text. Clinical Temporal information representation emerged as the most important representation form since all clinical events and medical information are noted with time stamps and indirectly follow a temporal order in discharge summaries, doctors notes and nurse narratives. Thus loss of chronological order of clinical events could have negative impact on disease diagnosis.

Research works has been reported on various topics of temporal reasoning and representation in the clinical data [19, 46, 48, 50]. To exploit the extracted temporal information in medical research and care, few studies have been initiated in natural language processing (NLP) on clinical text (applications by utilizing temporal information) [5, 20, 42]. Clinical decision support systems [5] and detection of adverse drug events [20] have become possible due to the availability of temporal information.

To accelerate the extraction of temporal information, the I2B2 challenge5 provides an annotated corpus on shared temporal relation tasks [32, 33, 35]. Followed by I2B2, SemEval challenge 2015 [6], and SemEval challenge 2016 [7] made efforts to generate an annotated corpus to expedite the researches of temporal information extraction. Bethard et al. [6] created this annotation corpus on clinical text based on domain-specific THYME-TimeML guidelines.6 This THYME-TimeML guidelines are adopted from ISO-TimeML guidelines [29]. Semantic information is added to the initial document, which is called as annotation [14, 30]. This is useful for various kinds of information extraction. To extract temporal expressions, the well-established HeidelTime has proven very powerful [35]. HeidelTime is a Java-based multilingual, cross-domain temporal expressions tagging system for newswire text data. Tang et al. developed a hybrid method by combining the conditional random fields (CRFs) and support vector machine (SVM) methods [35], and reported the extraction of the temporal events and event features by using this method. This hybrid method is an extension of [25] for the named entity recognition (NER) tagger system and Tang et al. used this method to extract the temporal events and the partial event features [35]. Various combinations of features (Lexical and UMLS (unified medical language system)) were utilized for event extraction in the NER tagger system.

A pipeline approach was developed by Jindal et al. to extract events and expressions from clinical narratives [16]. They extracted the time spans in the first stage, and afterwards, the types of events were parsed. An integer quadratic program (IQP) was implemented for the temporal event extraction and HeidelTime was adopted to extract the temporal expressions [16]. However, the pipeline approach has some limitations. Firstly, performance decreases in event extraction was caused by drop in accuracy of finding the attributes. Secondly, in comparison to previous systems, the performance of the feature extraction also degraded [33, 35]. Lastly, developing hypothesis and rules are complex and takes more time.

In case of temporal relation classification, various candidate generation and relation classification approaches have been attempted in existing studies. Chambers et al. [9] used Naive Bayes classifier to classify the temporal relation between the events from the Time Bank corpus. Chang et al. developed a hybrid system, which consists of rule-based and maximum entropy (ME)-based approaches to generate the candidate pairs [47]. Finally, they have proposed an algorithm that integrates the candidate pairs from two approaches. Even though they have developed the generation process of the candidate pairs by using the ME method, the processing steps are not very transparent and developing rules for each category by analyzing the clinical text is ambiguous and time-taking. In their most recent work, pathology notes of patients with colon cancer and brain cancer from the SemEval 2017 challenge are studied [21].

In contrast to the previous work, Tang et al. addressed the candidate generation problem with a formulated hypothesis based on the dependency parsing approach [35]. In that study, they have used different strategies for each category of temporal relation (relation between events and section times, intra-sentence, and inter-sentence). The hypothesis and approach used to generate the candidate pairs for inter-sentence and intra-sentence temporal relation classifier, respectively, fail to capture some potential candidate pairs. At last Gandhimathi and Tu-Bao-ho proposed semi supervised framework to extract temporal expression from unannotated clinical data set and a novel hypothesis to candidate pair generation for effective temporal relation classification [50]. However still the same work if had done with enhanced lexical feature set and an expanded candidate pair generation heuristic, a better and improved performance can be expected while also out performing all the previous work.

Proposed Methodology

The novel methodology to clinical temporal information extraction consists of the following subtasks with the newly adopted ideas and computations embedded in them .

a) Temporal Expressions Extraction
b) Temporal Event Extraction.
c) Temporal Relation Extraction

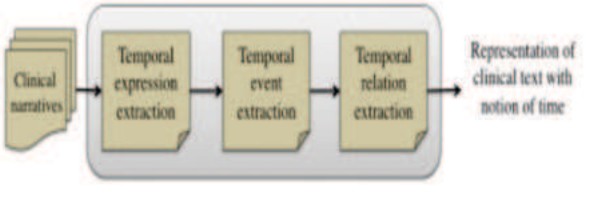

Fig1. Prposed Methodology

Temporal Expression Extraction

Following steps will be realized to constitute extraction of temporal expressions

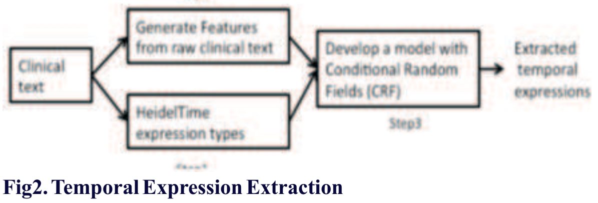

Fig2. Temporal Expression Extraction

Step 1: Clinical TempEval 2016 raw text of I2B2 data set will be used generate Lexical features set Using NLP package Genia Tagger System.
Step 2: Clinical TempEval 2016 raw text of I2B2 data set will be used generate HiedelTime tags.
Step 3: Model is to be developed with Conditional Random Fields (CRF) which extracts clinical temporal expression on any new raw clinical text.

Temporal Events Extraction.

Following sub task will have to be realized to achieve event extraction.

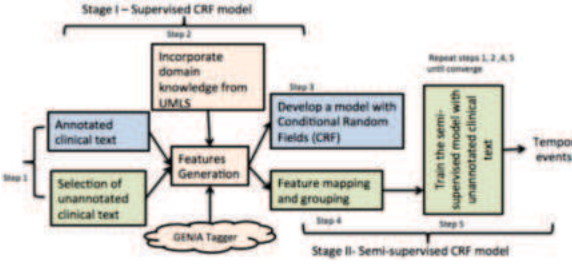

Fig2. Temporal Events extraction

Step1: Both annotated and Un-annotated clinical text data set is used to generate feature set using UMLS, Medical Dictionary System and Semantic Ontology like CNTRO 2 for temporal events and relations.
Step2: All Generated feature are used build a CRF model and clustering of the similar features.
Step3: The CRF Model is also trained on Un-annotated clinical text
Step4: All the above steps will be executed until the convergence of the model to extract temporal Events is met.

Temporal Relation Extraction

Following subtasks will have to be realized to achieve improved relation classification out of generated candidate pairs.

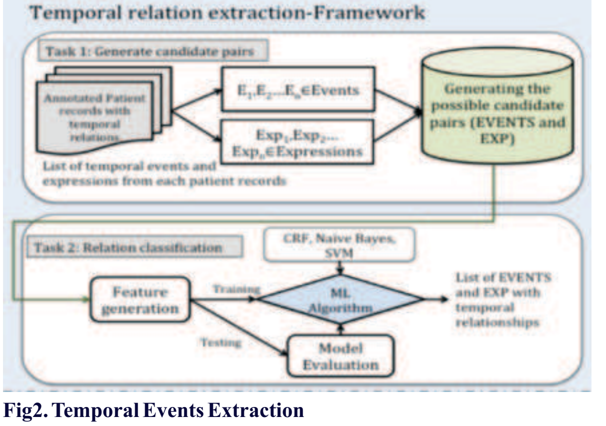

Fig2. Temporal Events Extraction

Step1: Candidate pairs from the list of extracted events and expression is to be generated with the Dependency Parsing Approach.
Step2: Existing Candidate generation heuristic techniques are enhanced by incorporating a new assumption that temporal relationships exist between events or expression as well as events and expression which will have events types such as (PROBLEM, TEST, TREATMENT, EVIDETIAL, DATE, FOLLOW UP, DOSAGE) in both intra sentence and inter-sentence sentence as within a window size of 3 sentences. Event Type identification and classification will have to be done based on domain knowledge system and HeidelTime. The hypothesis which will be formulated is based expansion of the previous work in the field by incorporating the new event types and considering combination of new event type with any other existing in both intra-sentence and inter-sentence with window size equal to 3.
Step 3: Naïve Byes Classifier will be used to train for the classification of temporal classifier

Evaluation Of Proposed Methodology

The proposed methodology will be evaluated against all the existing work on the same using F-Measure as a Metric and the analysis of the proposed methodology will be shown via the plot of output of each sub task against data set chosen and last work in the field.

EXPERIMENTAL PLAN

For all the tasks mentioned the proposed methodology Data sets such as Clinical TemEval2016 or I2B2 will be used. NLP Packages with Python tool kit will be used for Dependency Parsing. HeidelTime system will be Tagging and then for training of classifier R tool or Python can be used.

CONCLUSION

In this work we proposed an enhanced methodology to extract clinical text form EMR. Detailed analysis of semantics and measures to improvise information extraction through incorporation lexical feature set and heuristic candidate pairs. A significant Improvement in terms F-measure of the proposed methodology over any the existing towards the clinical temporal information extraction is expected and analysis of the proposed methodology with various parameter can be made further.

REFERENCES

[1] Aggarwal CC, Zhai C (2012) A survey of text clustering algorithms. In: Mining text data, pp 77–128. Springer

[2] Agrawal A (2009) Medication errors: prevention using information technology systems. Br J Clin Pharmacol 67(6):681–686

[3] Allen JF (1983) Maintaining knowledge about temporal intervals. Commun ACM 26(11):832–843

[4] Ambit H, Gonzalo C (2016) Clinical narrative analytics challenges. In: Proceedings Rough Sets: International Joint Conference, IJCRS 2016, Santiago de Chile, Chile, October 7–11, 2016, vol 9920. Springer, p 23

[5] Augusto JC (2005) Temporal reasoning for decision support in medicine. Artif Intell 33(1):1–24

[6] Bethard S, Derczynski L, Savova G, Savova G, Pustejovsky J, Verhagen M (2015) Semeval-2015 task 6: clinical tempeval. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp 806–814

[7] Bethard S, Savova G, Chen WT, Derczynski L, Pustejovsky J, Verhagen M (2016) Semeval-2016 task 12: clinical tempeval. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics, pp 962–972

[8] Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(Jan):993–1022

[9] Chambers N, Wang S, Jurafsky D (2007) Classifying temporal relations between events. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, ACL '07, pp 173-176. Association for Computational Linguistics, Stroudsburg, PA, USA

[10] Dubois S, Kale DC, Shah N, Jung K (2017) Learning effective representations from clinical notes. arXiv: 1705.07025

[11] Feldman K, Hazekamp N, Chawla NV (2016) Mining the clinical narrative: all text are not equal. In: 2016 IEEE international conference on healthcare informatics (ICHI), pp 271–280. IEEE

[12] Galescu L, Nate B (2012) A corpus of clinical narratives annotated with temporal information. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pp 715–720

[13] Grishman R, Sundheim B (1996) Message understanding conference-6: a brief history. In: COLING 1996 Volume 1: The 16th international conference on computational linguistics, vol 1

[14] Styler WF IV, Bethard S, Finan S, Palmer M, Pradhan S, de Groen PC, Erickson B, Miller T, Lin C, Savova G, Pustejovsky J (2014) Temporal annotation in the clinical domain. Trans Assoc Comput Linguist 2:143–154

[15] Jiao F, Wang S, Lee CH, Greiner R, Schuurmans D (2006) Semi-supervised conditional random fields for improved sequence segmentation and labeling. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics, pp 209–216. Association for computational linguistics

[16] Jindal P, Roth D (2013) Extraction of events and temporal expressions from clinical narratives. J Biomed Inform 46, Supplement(0):S13 – S19. https://doi.org/10.1016/j.jbi.2013.08.010. 2012 i2b2 {NLP} challenge on temporal relations in clinical data

[17] Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY (2002) An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24(7):881– 892

[18] Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th international conference on machine learning, ICML, vol 1, pp 282–289

[19] Zhou L, Friedman C, Parsons S, Hripcsak G (2005) System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. Am Med Inform Assoc 2005:869 20. Liu Y, LePendu P, Iyer S, Shah NH (2012) Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Summits Transl Sci Proc 2012:47–56

[20] Long Y, Li Z, Wang X, Li C (2017) XJNLP at SemEval-2017 Task 12: clinical temporal information extraction with a hybrid model. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp 1014–1018

[21] Mani I (2004) Recent developments in temporal information extraction. In: Proceedings of the international conference on recent advances in natural language processing (RANLP'03), pp 45–60

[22] Martinho R (2015) Text mining applied to electronic medical records. Int J E-Health Med Commun 6(3):1–18 Journal of Healthcare Informatics Research

[23] Miller TA, Bethard S, Dligach D, Lin C, Savova GK (2015) Extracting time expressions from clinical text, pp 81–91

[24] Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC, Xu H (2011) A study of machine learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc 18(5):601–606

[25] Moharasan G, Ho TB (2016) A semi-supervised approach for temporal information extraction from clinical text. In: 2016 IEEE RIVF international conference on computing & communication technologies, research, innovation, and vision for the future (RIVF), pp 7–12. IEEE

[26] Moharasan G, Ho TB (2017) Extraction of temporal events from clinical text using semi-supervised conditional random fields. In: International conference on data mining and big data, pp 409–421. Springer

[27] Pustejovsky J, Hanks P, Sauri R, See A, Gaizauskas R, Setzer A, Radev D, Sundheim B, Day D, Ferro L et al (2003) The timebank corpus. In: Corpus linguistics, vol 2003, p 40

[28] Pustejovsky J, Lee K, Bunt H, Romary L (2010) Iso-timeml: an international standard for semantic annotation. LREC 2010

[29] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Setzer A, Roberts I (2008) Semantic annotation of clinical text: the CLEF corpus. In: Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining, pp 19–26

[30] Sohn S, Wagholikar K, Li D, Jonnalagaddaa S, Tao C, Elayavilli RK, Liu H (2013) Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. JAMIA 20(5):836–842

[31] Sun W, Rumshisky A, Uzuner O (2013) Annotating temporal information in clinical narratives. J Biomed Inform 46:s5–s12

[32] Sun W, Rumshisky A, Uzuner O (2013) Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J Am Med Inform Assoc 20(5):806–813

[33] Sun W, Rumshisky A, Uzuner O (2013) Temporal reasoning over clinical text: the state of the art. J Am Med Inform Assoc 20(5):814–819

[34] Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H (2013) A hybrid system for temporal information extraction from clinical text. J Am Med Inform Assoc 20(5):828–835

[35] Tao C, Filannino M, Uzuner O (2017) Prescription extraction using CRFs and word embeddings. ¨ Journal of biomedical informatics 72:60–66

[36] Trivedi G, Pham P, Chapman W, Hwa R, Wiebe J, Hochheiser H (2017) An interactive tool for natural language processing on clinical text. arXiv: 1707.01890

[37] UzZaman N, Llorens H, Allen J, Derczynski L, Verhagen M, Pustejovsky J (2012) Tempeval-3: Evaluating events, time expressions, and temporal relations. arXiv: 1206.5333

[38] Verhagen M, Gaizauskas R, Schilder F, Hepple M, Moszkowicz J, Pustejovsky J (2009) The tempeval challenge: identifying temporal relations in text. Lang Resour Eval 43(2):161–179

[39] Verhagen M, Sauri R, Caselli T, Pustejovsky J (2010) SemEval-2010 task 13: TempEval-2. In: Proceedings of the 5th international workshop on semantic evaluation, pp 57–62. Association for Computational Linguistics

[40] Vilain MB, Kautz HA (1986) Constraint propagation algorithms for temporal reasoning. In: Aaai, vol 86, pp 377–382

[41] Wang CC, Chien MN, Huang CH, Liu L (2007) A rule-based disease diagnostic system using a temporal relationship model. In: 4th international conference on fuzzy systems and knowledge discovery, 2007. FSKD 2007. vol 4, pp 109–115. IEEE

[42] Wang Y, Rastegar-Mojarad M, Elayavilli RK, Liu S, Liu H (2016) An ensemble model of clinical information extraction and information retrieval for clinical decision support. In: TREC

[43] Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S et al (2017) Clinical information extraction applications: a literature review. Journal of biomedical informatics 45. Wong KF, Xia Y, Li W, Yuan C (2005) An overview of temporal information extraction. Int J Comput Process Orient Lang 18(02):137–152

[44] Lin YK, Chen H, Brown RA (2013) Medtime: a temporal information extraction system for clinical narratives. J Biomed Inform 46: s20–s28

[45] Chang YC, Dai HJ, Wu JC, Chen JM, Tsai RT, Hsu WL (2013) Tempting system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. J Biomed Inform 46: s54–s62 Journal of Healthcare Informatics Research

[46] Zhou L, Hripcsak G (2007) Temporal reasoning with medical data-a review with emphasis on medical natural language processing. J Biomed Inform 40(2):183–202

[47] Zhu X (2005) Semi-supervised learning literature survey. World 10:10

[48] Gandhimathi and Tu-Bao-ho (2019) journal of health care informatics Springer Nature Switzerland.