Journal of Engineering and Applied Sciences 11 (3): 408-413, 2016
ISSN: 1816-949X
© Medwell Journals, 2016

# Linguistic Rule-Based Methods for the Extraction of Medical Summaries to Benefit Patients Progression Tracking

Nurfadhlina Mohd Sharef and Mahda Noura
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia,
Seri Kembangan, Malaysia

**Abstract:** Clinical narratives contain useful information that can complement the patient progress records which are obtained throughout the patient's medical and treatment duration. In order to understand the clinical narratives content, medical concepts that include events and temporal information should be performed. This study addresses this issue based on a linguistic rule-based approach which combines domain knowledge, extraction modules and temporal linker component. This is in contrast to the fundamentals adopted by the major works based on machine learning. The proposed work's performance is therefore evaluated against a machine learning based approach and a knowledge intensive approach. Results have shown its strength regardless of its different nature.

**Key words:** Narratives, extraction modules, temporal linker component, learning, strength

## INTRODUCTION

The quality of clinical care in recent years are increased due to the advancement of the Electronic Health Records (EHRs) that also support improved clinical decision support system and medical research (Uzuner and Stubbs, 2015). Clinical narratives are produced to complement the structured format of data. Implementation of EHRs are however still majorly limited to the structured format of data which requires manual information recording (Roberts *et al*., 2013). There is limited means of EHR systems to extract and update the structured data because the clinical narratives are in textual format. In fact, the clinical narratives often contain information that are useful to record the complete medical status and treatment history of the patients for applications like progression tracking. Therefore, neglecting the extraction of the clinical narratives may lead to missing information which would benefit the clinical decision support system and medical research.

Existing works for medical concepts extraction in clinical narratives vary from the identification and extraction of medical events, temporal expression and temporal relations (Gobbel *et al*., 2014; Reeves *et al*., 2013; Roberts *et al*., 2013; Sharef and Madzin, 2012, 2013; Stubbs and Uzuner, 2015). Methods for the clinical narratives understanding are divided into natural language processing and machine learning. The Natural Language Processing (NLP) often produces linguistic rules to identify the segments of the clinical narratives (Jahiruddin *et al*., 2010; Serban *et al*., 2007) while machine learning is used to identify medical concept boundaries and types (Chang *et al*., 2015; Delen *et al*., 2010; Jindal and Roth, 2013; McCallum, 2002; Meyfroidt *et al*., 2009). However, the advantage of the NLP methods is the human understandable representation of linguistic patterns through application such as text extraction and text annotation. In contrast, the machine learning provides more compressed version of information representation. The scope of the paper is on NLP in order to extract the meaningful phrases and mentions of medical concepts in the clinical narratives. These will then be used to populate the EHRs.

The existing NLP-based approaches range from lexicon-based tools such as Meta-Map, UMLS, MeSH and SNOMED-CT which are knowledge intensive, developed by experts and contain predefined medical conceptual representation and which benefits from heuristic rules. The purpose of this study is to improve the performance of SNOMED-CT approach with a framework based on Linguistic Rules for Clinical Narratives (LRCN) extraction in identifying medical events and time expressions towards enabling patients' progression tracking and clinical decision support.

## CONCEPTUAL REPRESENTATION OF CLINICAL NARRATIVES

A conceptual representation of medical events is essential for medical concepts extraction and patient

**Corresponding Author:** Nurfadhlina Mohd Sharef, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Seri Kembangan, Malaysia

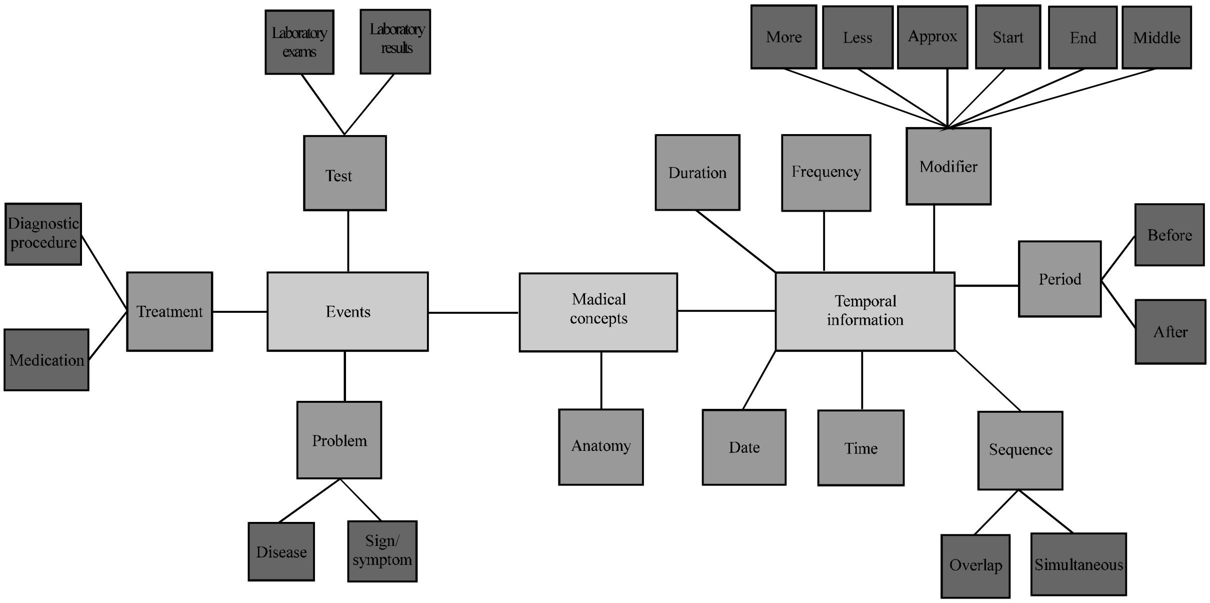

Fig. 1: Medical concepts representation

timeline discovery. Figure 1 shows the general representation of medical concepts which consists of medical problems, treatment, test, anatomy and temporal information.

Clinical narratives are rich in temporal information about events such as: admission date, discharge date, frequency of a medication, duration of a hospital course and so on. For example: 'Mr. Young needs Cardiac catherization at the end of October' contains the following information:

Diagnostic procedure: Cardiac catherization
Date: end of October

Each of these concepts are further represented with events, for example signs, symptoms and disease to denote medical problems, the laboratory exams and results to indicate medical test and diagnostic procedure and medication for the treatment. Linguistically, these concepts are expressed by combination of words such as in 'This is a 73 years old man with squamous cell carcinoma of the lung' which contains the following information: Age: 73; Gender: Male; Disease: Squamous Cell Carcinoma; Anatomy: Lung.

The temporal information is further shown with TIMEX3 to represent the time, date, duration or frequency. The medical concepts can further be extracted based on the events and temporal description in the clinical narratives as explained in the following.

**Events extraction:** For the purpose of extraction, the EVENT tag is used to annotate the events in the clinical discharge summary. In this research the patients symptoms, syndromes, disorders, problems, procedures, medications, laboratory exams, laboratory results, tests are considered as events. Every EVENT has three attributes: type, polarity and modality.

Every EVENT has a type attribute associated to it. These have been categorized into three main categories namely: problem (symptoms, syndromes, diseases and other complaints), treatment (medication, surgical procedure) and test (laboratory exams, laboratory results and tests) which belong to the EVENT type field. The EVENT tag is not allowed to have overlapped offsets and offset across multiple sentences. The EVENT tag has also a Polarity attribute which states whether the event is positive or negative. The Modality attribute specified whether the event has happened or not which can have three different values: factual, conditional and possible. The factual value is assigned to facts to specify whether an event has really occurred (in case of positive polarity), or if the event has not occurred (in case of negative polarity). The conditional value is allocated to EVENTS which are assumed to occur under certain situation. The possible value is allocated to EVENTS that are assumed to have occurred.

**Temporal extraction:** Extracting temporal information is important to understand the occurrence time, duration, sequence and relation between medical events. The TEMPORAL expressions include all references to the point in time, time periods, durations, and frequencies. Every TEMPORAL expression is assigned a TIMEX3 tag

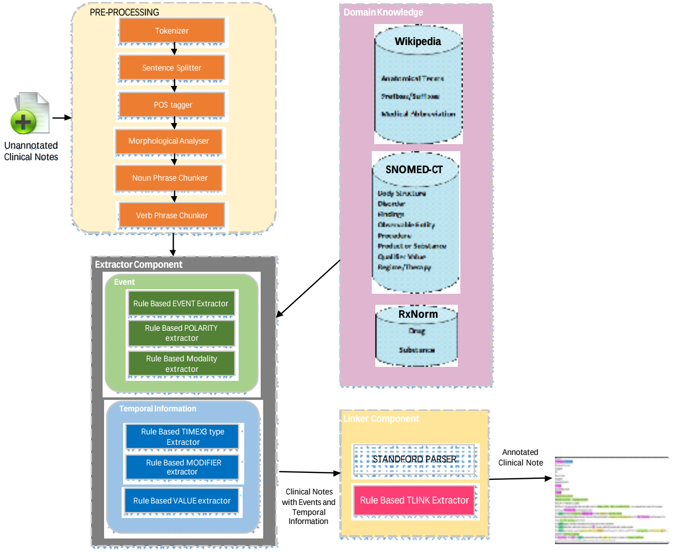

Fig. 2: Linguistic rule-based framework for patient progression tracking pre-processing

and has three attributes namely Type, Val and Mod. The temporal expression Type is marked as date, time, duration and frequency. Since we can use many different formats to express temporal information, the Val attribute normalizes the value of the temporal expression into standard ISO 8601, so the machine can recognize and process them. The format for Date value in ISO 8601 is: [YYYY]-[MM]-[DD]. If the month and day values cannot be inferred from the text it is simply omitted as: [YYYY]-[MM] and [YYYY], respectively. The format for TIME value in ISO 8601 is: [hh]:[mm]:[ss], where [hh] ranges from 00-24; [mm] and [ss] both range from 00-59. Similarly, [mm] or [ss] can be omitted as: [hh] and [hh]:[mm]. Durations are temporal expressions that describe a period of time, e.g. for eleven days, for half a year. The syntax of duration representation is P[n][Y/M/W/D]. So, "for 11 days" will be represented as "P11D", meaning a period of 11 days.

## LINGUISTIC RULE-BASED FRAMEWORK FOR PATIENT PROGRESSION TRACKING

The medical system proposed in this research, called LRCN is a rule-based pipeline which runs within GATE (Konchady, 2008) framework. In order to craft the rules the Java Annotation Patterns Engine (JAPE) language was used. JAPE allows pattern matching and evaluation of text annotations using a regular expression-like syntax. An annotation represents a marked range in the text, corresponding to some entity or mention, with start and end nodes, a document-unique identifier and a set of features (attributes on the annotation). Each node points to a character offset in the document. One of the benefits of JAPE is that annotations not specified in the input are ignored for pattern matching purposes which enables patterns to be generalized when, for example, intervening punctuations and prepositions are not significant. Our system combines GATE ANNIE text segmentation components with custom names-entity annotators in order to annotate the EVENTS and TEMPORAL Expressions in clinical text. The framework of our system has been shown in the following Fig. 2. It is composed of four components: Pre-processing, Extractor, Domain Knowledge and the Linker.

**Pre-processing:** The framework starts with pre-processing which assembles a processing pipeline following standard GATE-General Architecture for Text Engineering ANNIE components for tokenization, sentence splitting, Part Of Speech (POS) tagging, Noun Phrase (NP) chunking and Verb Group (VG) chunking. The English Tokeniser splits a text into annotations of type Token and allows for the application to apply rules for annotating the concepts. The sentence splitter was used to split the boundaries in text, in order to calculate the line offset. The POS tagger assigns POS tags to Token annotations. This Penn Treebank tag set is used. VG identifies verb groups and NP identifies noun phrases in text. In clinical notes there are paragraph headings like: physical examination, laboratory data, history of present illness, Infectious disease, fluids, electrolytes and nutrition which divide the text into sections and should not be annotated as medical events. Thus, we composed rules which split the clinical note into subsections and excluded the paragraph headings from medical event classification.

**Domain knowledge:** The domain knowledge is the set of dictionaries incorporated in the system as gazetteer. In the GATE framework, a gazetteer comprises of one or more plain text files that function as lookup lists, each of which is described in an index file that classifies each list according to major and minor types.

We used a drug lexicon derived from the RxNorm dataset in the UMLS (Segura-Bedmar *et al*., 2008) which includes drugs with brand names, generic names and common abbreviations to identify drugs. We created a dictionary comprising of 269474 terms which included the following attributes: drug name, RxCUI code and term types to assist in the look up process using RxNorm. The term types have four different type: Semantic Clinical Drug Form (SCDFs), Semantic Clinical Drugs (SDCs), Semantic Branded Drugs (SBDs) and ingredients. A python program was written to separate all the drugs with respect to their type and have them stored in a separate file. These separate files were used as dictionaries for drug names.

The SNOMED CT concepts were extracted from the 'Concept' table. The 'Concept' table was pre-processed for the extraction of concepts with their respective categories. The concept table contained the following attributes: CONCEPT ID, CONCEPT STATUS, FULLY SPECIFIED NAMES, CTV3 ID, SNOMED ID, IS PRIMITIVE. In the first step a Python program was written to remove all attributes from the SNOMED CT concept file except "FULLY SPECIFIED NAMES". The attribute "FULLY SPECIFIED NAMES" contained names of concepts along with their semantic categories. In the second step, another code was written to separate all the concepts with respect to their semantic categories from the attribute "FULLY SPECIFIED NAMES" and have them to be stored in a separate file. These separate files were used as dictionaries of each semantic category listing concepts.

Measurement information indicates the amount and unit in a medication administration. It generally appears in the form of a numeric with a unit of measure, for example, 'one tablet', '0.5 mg'. A list of all units of measure has been stored in several gazetteer lists.

**Extractor component:** The extractor component is comprised of two sub parts: EVENT and TEMPORAL information. To help identify medical events and temporal information we extended the existing ANNIE named entity recognition and wrote a pattern-based recognizer for general named entities such as measurement, date, time, frequency, duration and number using gazetteer lists of primitives and JAPE expressions. In addition we used rule based recognizer to match medical events: PROBLEM, TREATMENT and TEST.

We also use information from dictionaries and additional features to overcome the deficiencies of the dictionary-only approach. Since the lexicon provided in the dictionaries are not sufficient to extract medical concepts, the outputs of the lookups generated from the Gazetteer lists constructed for measurement, frequency, route, their abbreviations and modifiers were combined with JAPE rules to extract important concepts including frequency, dose and route.

Due to the lack of annotation using lexical resources, some prefixes and suffixes can provide good clues for classifying named entities. For example words which have the suffix "pathy" indicate a disease or condition (e.g., neuropathy means disease of the nervous system) or the suffix "oma" means tumour (e.g., retinoblastoma is tumour of the eye) and so on. Similarly, prefixes like "dys" indicate a difficulty or pain (e.g., dysfunction), "hyper" above normal, "hyp(o)" below normal or "para" abnormal and so on.

Due to the incompleteness of the dictionaries, context clues are further used to aid the EVENT extraction. Context clues are pieces of text that gives us clues about an EVENT. Words preceding or following a target word may be useful for modelling the local context. It is clear that the more context words analyzed, the better and more precise the results become. The selection of the context clues should be performed wisely as it results in some unnecessary EVENTs to be extracted. The noun phrase which includes the context clue is annotated as the EVENT.

A possible medication is defined as any non-drug-name text (drugs not identified using the lexicon) surrounded by drug information such as DOSE, MODE, FREQUENCY to indicate that it refers to a misspelled drug name or a drug name not in the RxNorm dataset. For example:

> The patient was treated with Atenolol 100 mg [DOSE] p.o [MODE]. Daily [FREQUENCY]....

Assume that Atenolol has not been annotated as treatment using the lexicon. However, using the drug information such as DOSE, MODE and FREQUENCY, we can infer that Atenolol is a medication name. It should be noted that the drug information may come in any order. There are some generic medical terms for symptoms and syndromes in a Noun Phrase which mention a PROBLEM such as: disease, abscess, disorder, injury, disability, exposure, dysfunction, lack of, inability, failure, symptom and so on which can be identified using the following rule:

Table 1: Performance of EVENT Extraction between LRCN, MEDTime and SNOMED-CT

| | Precision | | | Recall | | | F-measure | | |
|---|---|---|---|---|---|---|---|---|---|
| Approaches | Test | Treatment | Problem | Test | Treatment | Problem | Test | Treatment | Problem |
| MEDTime | 0.96 | 0.93 | 0.91 | 0.87 | 0.86 | 0.88 | 0.89 | 0.91 | 0.90 |
| LRCN | 0.77 | 0.82 | 0.90 | 0.73 | 0.70 | 0.71 | 0.74 | 0.72 | 0.78 |
| SNOMED-CT | 0.50 | 0.50 | 0.52 | 0.18 | 0.18 | 0.18 | 0.22 | 0.22 | 0.25 |

Table 2: Performance of TEMPORAL expression extraction between LRCN and MEDTime

| Approaches | Precision | Recall | F-Measure |
|---|---|---|---|
| MEDTime | 0.8296 | 0.9055 | 0.8659 |
| LRCN | 0.83 | 0.77 | 0.80 |

{SyndromeContextClues | SymptomContextClues}
({Token.string ==~ "(?i)in | with"})?{NP}

For example,

> Mr. Anders presented with an injury most consistent with meniscal tear [PROBLEM]

There are some generic medical terms for medical treatments in a Noun Phrase which mention a TREATMENT such as: surgery, catheter, procedure, operation, therapy, and so on. There are some generic medical terms for medical tests in a Noun Phrase which mention a TEST such as: exam, culture, test, rate, ratio, pulse, and so on. There are some keywords that when combined with an anatomical term state a PROBLEM, such as failure, deficit, enlarged, hyperactive, broken, block, injury, problem, pain and so on. For example:

> Mr Smith is a 40 years old man with a history of GERD [PROBLEM], who presents with abdominal [ANATOMICAL TERM] pain [DISEASE-KEYWORD] of 7 days duration.
> The chest x-ray shows lung [ANATOMICAL TERM] infiltration [DISEASE-KEYWORD].

In this 2 week interval he denied having CP/chest discomfort or difficulty breathing, no cough or hemoptysis Based on this, the following two rules were constructed:

- {AnatomicalTerm}[1,2] {Symptom|Syndrome}
- {Symptom|Syndrome}{(?i)of|on|from|to|in|under|over|below|above}({Token.category=~"^DT|PRP})?{AnatomicalTerm}

Similar to that, there are some keywords that require an anatomical term preceding or following in order to be considered a PROCEDURE mention (Table 1 and 2).

## EVALUATION ON MEDICAL EVENTS AND TEMPORAL EXTRACTION

The i2b2 (Informatics for Integrating Biology and the Bedside) 2010 challenge dataset is used to evaluate the performance of the proposed framework, LRCN against the benchmark approaches namely SNOMED-CT and several other benchmark studies. Table 1 shows the comparison of EVENT extraction performance between LRNC, MEDTime (Lin *et al*., 2013) and SNOMED-CT. MedTime comprises a cascade of rule-based and machine-learning pattern recognition procedures. MEDTime's result is included to denote the superior performance of machine learning based methods but as explained in the introduction, the machine learning method does not focus on the information extraction and EHR population. Therefore, focus should be given to the performance of LRNC and SNOMED-CT. The results indicate that the combination of linguistic rules and domain knowledge has better performance compared to SNOMED-CT which is solely based on domain knowledge approach.

Meanwhile, the performance of LRCN in TEMPORAL Expression extraction is very close with the performance of MEDTime. SNOMED-CT is not included for this evaluation because SNOMED-CT does not have ability for temporal representation.

## CONCLUSION

Understanding of clinical narratives can improve the quality of clinical decision support system by integrating its contents with the database of medical treatments which is in structured format. Existing works are divided into NLP and machine learning. Many works have focused on the machine learning methods which also utilize lexicon based approaches and basic NLP approach as features representation to build classification models. However, these models are not able to annotate the medical concepts in the clinical narratives. Therefore, this research explores the development of linguistic rules for clinical narratives understanding and its performance is compared against SNOMED-CT, a lexicon based approach and MEDTime, a machine learning approach. Results have indicate that the LRCN approach provides promising achievement. The next stage of the work is the implementation of the linguistic-based framework to support patient health analytics.

## REFERENCES

Chang, N.W., H.J. Dai, J. Jonnagaddala, C.W. Chen, R.T.H. Tsai and W.L. Hsu, 2015. A context-aware approach for progression tracking of medical concepts in electronic medical records. J. Biomed. Inform., 58: S150-S157.

Delen, D., A. Oztekin and Z.J. Kong, 2010. A machine learning-based approach to prognostic analysis of thoracic transplantations. Artificial Intell. Med., 49: 33-42.

Gobbel, G.T., R. Reeves, S. Jayaramaraja, D. Giuse and T. Speroff *et al*., 2014. Development and evaluation of RapTAT: A machine learning system for concept mapping of phrases from medical narratives. J. Biomed. Inform., 48: 54-65.

Jahiruddin, M. Abulaish and L. Dey, 2010. A concept-driven biomedical knowledge extraction and visualization framework for conceptualization of text corpora. J. Biomed. Inform., 43: 1020-1035.

Jindal, P. and D. Roth, 2013. Extraction of events and temporal expressions from clinical narratives. J. Biomed. Inform., 46: S13-S19.

Konchady, M., 2008. Building Search Applications: Lucene, LingPipe and Gate. Mustru Publishing, Oakton, VA., USA., ISBN-13: 978-0615204253, Pages: 448.

Lin, Y.K., H. Chen and R.A. Brown, 2013. MedTime: A temporal information extraction system for clinical narratives. J. Biomed. Inform., 46: S20-S28.

McCallum, A.K., 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu/.

Meyfroidt, G., F. Guiza, J. Ramon and M. Bruynooghe, 2009. Machine learning techniques to examine large patient databases. Best Pract. Res. Clin. Anaesthesiol., 23: 127-143.

Reeves, R.M., F.R. Ong, M.E. Matheny, J.C. Denny and D. Aronsky*et al*., 2013. Detecting temporal expressions in medical narratives. Int. J. Med. Inform., 82: 118-127.

Roberts, K., B. Rink and S.M. Harabagiu, 2013. A flexible framework for recognizing events, temporal expressions and temporal relations in clinical text. J. Am. Med. Inform. Assoc., 20: 867-875.

Segura-Bedmar, I., P. Martinez and M. Segura-Bedmar, 2008. Drug name recognition and classification in biomedical texts: A case study outlining approaches underpinning automated systems. Drug Discovery Today, 13: 816-823.

Serban, R., A. ten Teije, F. van Harmelen, M. Marcos and C. Polo-Conde, 2007. Extraction and use of linguistic patterns for modelling medical guidelines. Artif. Intell. Med., 39: 137-149.

Sharef, N.M. and H. Madzin, 2012. IMS: An improved medical retrieval model via medical-context aware query expansion and comprehensive ranking. Proceedings of the International Conference on Information Retrieval and Knowledge Management, March 13-15, 2012, Kuala Lumpur, Malaysia, pp: 214-218.

Sharef, N.M. and H. Madzin, 2013. Semantic-based medical records retrieval via medical-context aware query expansion and ranking. J. Theoret. Applied Inform. Technol., 58: 697-706.

Stubbs, A. and O. Uzuner, 2015. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J. Biomed. Inform., 58: S78-S91.

Uzuner, O. and A. Stubbs, 2015. Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. J. Biomed. Inform., 58: S1-S5