# Extracting Temporal Information from Electronic Patient Records

**Min Li, MIT[1], Jon Patrick, PhD[1]**
**[1] School of IT, the University of Sydney, Sydney, NSW, Australia**

**Abstract**

*A method for automatic extraction of clinical temporal information would be of significant practical importance for deep medical language understanding, and a key to creating many successful applications, such as medical decision making, medical question and answering, etc. This paper proposes a rich statistical model for extracting temporal information from an extremely noisy clinical corpus. Besides the common linguistic, contextual and semantic features, the highly restricted training sample expansion and the structure distance between the temporal expression & related event expressions are also integrated into a supervised machine-learning approach. The learning method produces almost 80% F- score in the extraction of five temporal classes, and nearly 75% F-score in identifying temporally related events. This process has been integrated into the document-processing component of an implemented clinical question answering system that focuses on answering patient-specific questions (See demonstration at http://hitrl.cs.usyd.edu.au/ICNS/).*

**Introduction**

Temporal information is an important and pervasive component of electronic patient records (EPR) and is needed for addressing the problems of answering time-oriented clinical questions. For instance common questions are: "*Did the patient's temperature exceed 38C in the last 48 hours?*", "*What was the trend of his blood sugar after he got up?*", etc. These sorts of questions could not be answered with simple keyword matching approaches used by most search engines and question-answering systems. For example: 1. the temporal expression in the questions like "*in the last 48 hours*", "*after he got up*", and "*the last time*" have a low chance of being coded in EPRs. 2. as the temporal events change over time, even if this temporal event information was coded in the patient record, its information can be only correct at the time of reporting.

In order to answer these sorts of questions, the first step is to identify and extract temporal expressions and corresponding clinical events from EPRs.

**Background**

Temporal study is an active subject of research in the newswire domain. After several revisions of the TIMEX schema in the MUC-6[1] and MUC-7[2], TIMEX2 schema[3] in TIDES, etc., TIMEML[4] has become one of the most widely used annotation schema in the natural language processing (NLP). The TIMEML standard was developed to provide an annotation scheme for identifying events mentioned in a text document and orienting them on a timeline. Besides the TIME3 tag which represents a variation of the TIMEX2 scheme, TIMEML focuses on among other things, ways of systematically anchoring event predicates to a broad range of temporally defining expressions, and on ordering such event expressions. The main tags within the TIMEML framework are TIMEX3, SIGNAL, EVENT, TLINK, SLINK and ALINK. By their definition, the TIMEX3 tag is primarily used to mark up explicit temporal expressions, such as times, dates, durations, etc. The temporal signal is represented by the SIGNAL tag. EVENTs are situations that happen or occur. The LINK tag is used to encode a variety of relations that exist between the temporal elements in a document, as well as to establish an explicit ordering of events. Three subtypes to the LINK tag are used to represent strict temporal relationships between events or between an event and a time (TLINK), subordination between two events or an event and a signal (SLINK), and aspectual relationship between an aspectual event and its argument (ALINK). With the aim of automatic temporal annotation, a modular system – TARSQI[5, 6] was developed to detect and resolve temporal relations using a combination of hand-crafted rules and machine learning. In TARSQI, the temporal expressions were tagged by the GUTime which achieved 82% F-score, while the event expressions were labelled by Evita with 80% F-score. This task was continued in the TempEval2 challenge[7]. In the TempEval2 challenge, by using a rule-based approach, the best system – HeidelTime[8] performed at 86% on temporal span extraction and 96% on temporal type assignment. As a result of the loss in the temporal type classification, the performance of the best system is quite close to the GUTime. However, for the event extraction, the best system – TIPSem[9] which is also a rule-based approach outperforms Evita by approximately 3%.

While there is widespread use of the TIMEML scheme within the NLP community, temporal information in EPRs exhibits complex and unique characteristics[29]. In the first place, some temporal expressions are written by using

shorthand, such as "3/7" means three days, "bd" means twice daily, etc. In addition, besides the four types (date, time, duration, and set) of TIMEML temporal expression, other classes also exist in EPRs, like the event dependent temporal expressions, e.g., "post-op", "at time of intubation", etc., as well as the fuzzy time, e.g., "at shift start", "for most of day", etc. Moreover, the clinical events can refer to any medical-related phenomena, and most of them are expressed by nouns in EPRs. Apart from this, the semi-structured written manner, the possible temporal relationships between two events, etc., also prevent the unmitigated adoption of a temporal standard from the news domain in the clinical domain. Temporal information in medical texts has not been widely studied. Some first steps were taken by Zhou and colleagues[10-12]. Through the temporal structure analysis in discharge summaries, a robust annotation schema was proposed which contains six major categories of temporal expressions ("date and time", "relative date and time", "duration", "event-dependent temporal expression", "fuzzy time" and recurring times) and medical event. These studies became the guiding principle of the following studies. For instance, a revised temporal annotation schema was proposed by Mowery and colleagues[13] for labelling six types of clinical reports. Subsequently, they built a statistical model to classify whether a clinical condition is historical or recent by exploring the temporal features, such as the temporal expression, tense and aspect, and trigger terms, etc.[14] Furthermore, a system called TN-TIES[15] was developed to process clinical temporal information, which focuses on temporal information extraction and classification. The temporal segments were identified by a rule-based chunker which successfully chunked 91% of the temporal segments. Subsequently, these segments were classified into five temporal classes by a Naive Bayes classifier with an F-score in the 60s. On the other hand, the Semantic Web and the Web Ontology Language (OWL) was adopted to represent temporal information in clinical records most recently, e.g., the CNTRO ontology[26, 27], because of the advantages in the OWL, such as a standard mechanism with semantic knowledge representation, powerful reasoning and inference capabilities offered by Description Logic (DL) forms, etc. However, due to the fact of that the OWL is initially designed for general domain and focuses on absolute time, some of the clinical temporal information cannot be covered, e.g., relative time, domain event dependent temporal, etc. Therefore, a certain amount of clinical records need to be analysed and evaluated for the robustness of the ontology, while the capabilities of the CNTRO ontology has been only evaluated on five clinical records. In contrast, a temporal information model which was designed by Zhou and colleagues based on an analysis of 200 clinical records and an evaluation of 100 clinical records[11], which has been adopted as the basis of the present work. Similar to the task in TN-TIES, in this study a fine-grained level of temporal information extraction is the main objective, that is to target the exact temporal expression span of each temporal class and its corresponding event span.

**Methodology for Temporal information Extraction[1]**

There are four main steps in our methodology: ①defining the information to be extracted. ②preparing data ③using natural language processing technologies to build a temporal information extraction system.

***Extraction Definition***

The annotation schema is an extension of the work that Zhou and colleagues completed on hospital discharge summaries, since we are working on a different corpus. For the temporal expression, five subclasses are considered.

1. Date and time (both specific and relative): e.g., *"@ 0730hrs", "3yrs ago", "at 0145 (14/08)", "0600"*, etc.
2. Frequency: e.g., *"1-2hrly", "mane", "q2/12", "3 days/week",* etc.
3. Duration: e.g., *"1-2/24", "for >2hours", "from previous day", "until 1330hrs"*, etc.
4. Key event dependent temporal: *"post op day 1", "prior to assessment", "since OT", "after d/c",* etc.
5. Fuzzy time: e.g., *"since then", "at start of shift", "at later date", "at the moment", "until early evening",* etc.

The definition of medical event is also revised to meet the ICU dataset which includes:

1. Medical related state/condition: e.g., *"fell well", "stable", "mild agitation", "SaO2 95%",* etc.
2. Medical related process: e.g., *"use propofol", "given filling", "Central line removed",* etc.
3. Medical related changes: e.g., *"increasing difficulty with mobility", "no significant change", "Headaches improved",* etc.
4. Medical related behaviour: e.g., *"sitting out of bed", "appropriate to commands", "slept well",* etc.

[1] In the following sections, all the examples are the raw data in the ICU corpus, except de-identification. Therefore, examples may be in bad English.

5. Medical plan: *"With an aim to extubation", "For CT brain", "aiming for neutral balance",* etc.
6. Other events: e.g., *"social visit by sister", "no phone enquiry", "wash pt",* etc.

***Dataset***

The data used in this study consists of clinical notes collected from the Intensive Care Unit at the Royal Prince Alfred Hospital, Sydney, Australia. The corpus consists of progress notes both nursing and medical spanning 5 years from 2002 to 2006 with 60 million tokens in total. A random sample of 200 patient notes was selected from the pool, which contains approximately 30,000 words in total. An iterative process was conducted to develop the gold standard annotation. In the beginning, the 200 notes were distributed to two annotators, who independently marked the five classes of temporal expressions and corresponding event expressions. An F-score of 86% was computed to establish the inter-annotator agreement. Then the inconsistent instances were presented to a clinical language research group which comprises six computational linguists and one doctor for further discussion and modification. Finally, the outcome of this iterative process was reviewed and finalised to achieve 100% annotation agreement. The final frequencies for each class are presented in the Table 3.

***System design***

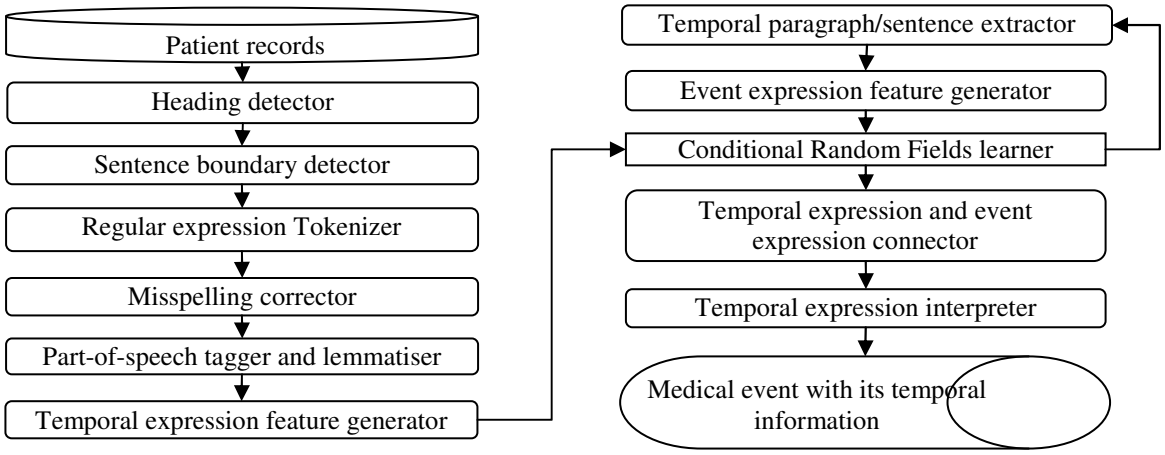


**Figure 1**. Temporal event processor system architecture

A high level overview of the proposed system architecture is presented in Figure 1. In order to provide a good understanding, this diagram starts from the beginning of the clinical text processing. Initially the EPRs were segmented into section headings and sentences by heading detector and sentence boundary detector. Both of them are designed by statistical methods, which produce an F-score of 93% and 94% respectively. Then headings and sentences were tokenised by a regular expression tokeniser. After that, an automated misspelling correction process was performed on the tokenised records. Once the corrected version was obtained, a part-of-speech (POS) tagger[16] was applied to each sentence which was the last step of the pre-processing to compute the POS, chunk, and the lemma for each token. Subsequently, learning features were prepared by the temporal expression feature generator for the conditional random field learner to extract five types of temporal expressions. Next, only the sentences with tagged temporal expressions were passed to the event expression feature generator. After that, the temporally related events were labelled by using the event learning features, and stored with temporal expressions for further studies. Learning feature selection is one of the most crucial issues to impact the performance of a machine learner. The features should be general enough to support the variation of different instances that belong to one class, and strict enough to capture the differences between instances from different classes. Consequently, with regard to the characteristics of temporal expressions and event expressions, two feature groups are considered for these two extraction tasks.

***Temporal expression extraction features***

1. Bag of words: provides the content information, since words surrounding the target word are useful for predicting the instance. A five word window is used which includes the current token, two tokens before it, and two tokens after it.
2. Bag of proofed words: The content information is smoothed by removing some rare cases. It includes the current token and a five word window of proofed tokens.

3. Bag of lemmas with proofs: By applying lemmatization, the bag of proofed words is generalized. It includes the current token and a five word window of lemmatized proofed tokens.
4. POS: This low level grammatical information is helpful for determining the boundaries of instances.
5. Chunk: It's a phrase chunker based on 'POS' which assists in determining expression boundaries.
6. Gazetteer: We constructed several different features to capture the existence of an instance in a closed dictionary, e.g. a week list, month list, etc. They are constructed by extracting all instances from the training data in each fold.
7. Orthography: is to capture the rendition of words, such as capitalization, digitalization, uppercase, etc.
8. Symbol: is to capture the structure of temporal expressions, e.g. *24/07/01 -> /_/_/, (1-2/7) -> (-_/),* etc.
9. Prefix and Suffix: is used to distinguish whether a non-word is a temporal expression or a score, e.g. GCS10 vs. q2/12, 2000ml vs. 2000hrs, etc.
10. The length of a digit: it was theorized that an integer with an even numbered set of digits has a higher probability to be a time expression than an odd number of digits.
11. Scores and measures: A trainable finite state automata[17] is performed to identify scores and measures. Thus, the digit based temporal items and scores can be distinguished.
12. Unit of measurement: The trainable finite state automata also used to identify units of measurement.
13. Highly restricted training sample expansion mapping (rule-based method): The mapping results are encoded as features, that has been proven as an efficient way to assist the learning process[18]. In the first place, the temporal instances in each fold of the training data are extracted in a proofed form, except the digit only temporal instances. Then, these instances are normalized by using digit normalization, e.g. *for 3 days -> for (/d+) days*. After that, the gazetteers (month, week, the preposition of duration and key-event, etc.) are applied for a second level of normalization which is category based. E.g. *for 3 days -> (for|since|during|through|until|till|…) (\d+|one|two|…) (hours|days|weeks|months|years)*. Finally, these normalized samples are used to capture the instance with the same canonical form in the test data.

***Event expression features***

Besides the common linguistic, contextual and semantic features which are designed for the temporal learning model listed above (except "Gazetteer", "Symbol", "The length of a digit", and "Highly restricted training sample expansion mapping"), several other features were designed for the event learning model:

1. SNOMED CT (SCT) concept: By applying a "Text to SNOMED CT" (TTSCT)[19] conversion process, the medical terms in the temporal sentences are labelled to distinguish whether it is a medical event or non-medical event.
2. SCT category: The SCT top category is also generated by using TTSCT which provides the semantic information to indicate the medical category of each event.
3. Temporal class: The five temporal classes are used to inform the Conditional Random Fields (CRF) learner[28] which temporal class exists in each sentence.
4. Temporal expression: The span of the temporal expressions is also considered to avoid mis-classification between temporal expressions and event expressions.
5. Number of temporal expressions: The total number of temporal expressions in each sentence is used to target multiple event expressions.
6. The position of the temporal expression: Three values are considered, e.g., beginning of a sentence, middle of a sentence, and end of a sentence. It might helpful to locate the event expression in the sentence.
7. Punctuation and the Distance between a temporal expression and an event expression. This feature is designed as a substitution for a parser result which has been a proven success in the TIMEML event extraction, e.g., the dependency tree[20] and the semantic roles[21]. Due to the extremely noisy nature of the clinical notes they are unable to be successfully parsed despite experiments with several widely used parsers in the biomedical domain, e.g., CCG parser[22], Charniak parser[23], and Enju parser[24]. Hence, a

structural distance is used to find the relation between a temporal expression and an event expression. At first, the temporal expressions and punctuation marks are extracted from each sentence. Then, the calculation starts in two directions with a value of zero at the beginning of a temporal expression (incrementally) and from the end of a temporal expression (decrementally). An example is presented in Table 1.

| Tokens | Punctuation | Distance | Gold Standard |
|---|---|---|---|
| NEURO | : | 2\|: | O |
| : | O | 2\|: | O |
| Patient | O | 1\|, | B-event |
| very | O | 1\|, | I-event |
| lethargic | O | 1\|, | I-event |
| , | , | 1\|, | O |
| awake | O | 0\|O | B-event |
| most | B-TE\|O | B-TE\|O | O |
| of | I-TE\|O | I-TE\|O | O |
| the | I-TE\|O | I-TE\|O | O |
| night | I-TE\|O | I-TE\|O | O |
| however | O | 0\|O | O |
| rarely | O | 0\|O | B-event |
| moving | O | 0\|O | I-event |
| . | . | -1\|. | O |

**Table 1**. Example of calculating punctuation distance[2]

| Model # | Features |
|---|---|
| **1** | Bag-of-words |
| **2** | Bag-of-words with correction |
| **3** | **Bag-of-lemma with correction** |
| **4** | Model #3 + part-of-speech |
| **5** | **Model #3 + part-of-speech (2)** |
| **6** | Model #3 + part-of-speech (4) |
| **7** | **Model #4 + Chunk** |
| **8** | **Model #4 + Chunk (2)** |
| **9** | **Model #8 + Gazetteer** |
| **10** | Model #9 + Orthography |
| **11** | **Model #9 + Orthography (2)** |
| **12** | Model #9+ Orthography (4) |
| **13** | Model #11 + Symbol |
| **14** | Model #11 + Prefix |
| **15** | Model #11 + Prefix (2) |
| **16** | **Model #11 + Prefix + Suffix** |
| **17** | Model #11 + Prefix + Suffix (2) |
| **18** | Model #11 + Prefix (2) +Suffix (2) |
| **19** | Model #16 + Length of a digit |
| **20** | **Model #16 + Scores & Measures** |
| **21** | **Model #20 + Unit of measurement** |

**Table 2.** Features applied in each temporal extraction experiment[3]

### Extraction model and evaluation method

In order to discover the best feature sets, a selective incremental method was used. If the performance benefited by one feature set, then this feature set was retained, otherwise, it was dropped. The evaluation mechanism used 10-fold cross-validation and calculated the precision, recall and the F-score.

### Results

***Temporal Expression Extraction Results (the feature models were presented in Table 2 while the performance of each feature model were illustrated in Figure 2)***

The main purpose of the temporal expression extraction is to extract five types of temporal expressions, e.g., "date and time", "event dependent temporal", "fuzzy time", "duration", and "frequency". From Figure 2, it can be seen that the overall F-score of these five classes is improved effectively by various feature sets (except the "highly restricted training sample expansion mapping"). The features which were applied in these experiments are presented in Table 2 with the detailed scores of the baseline model and the best model in Table 3. A micro-average F-score of 71.1% can be achieved by the best model which is more than 20% higher than the baseline model. The performance of the rule-based method (Temporal expression extraction feature 13. Highly restricted training sample expansion mapping) is presented in Table 4, as well as the performance of integrating the rule-based method into the previous best feature model (model #22). From table 4, it is clear that a similar micro-average F-score was obtained by the rule-based method. Meanwhile, by integrating the rule results as a feature set in the previous best model, the final performance of the combined model increased to almost 80%.

[2] "O" in the "Punctuation" and "Distance "column means no punctuation for the current token. "B-TE"/"B-event" means the first token of the temporal/event expression, while the "I-TE"/"I-event" means the remaining tokens. No-event tokens are labelled as "O" in the "Gold standard" "column.

[3] "(2)" means a window of one label before/after the current label. "(4)" means a window of two labels before/after the current label.

***Temporally Defined Medical Event Expression Results (the feature models were presented in Table 6 while the performance of each feature model were illustrated in Figure 3)***

The models that benefit the temporal extraction task are in bold in Table 6. A set of similar experiments were carried out on the temporally related medical events. The results of those experiments are illustrated in Figure 3 and numerical values of the baseline model and the best model are presented in Table 5, while the corresponding features are listed in Table 6 with the best feature models in bold. By exploring a rich feature model, the micro-average of the best model is more than 25% higher than the baseline mode.

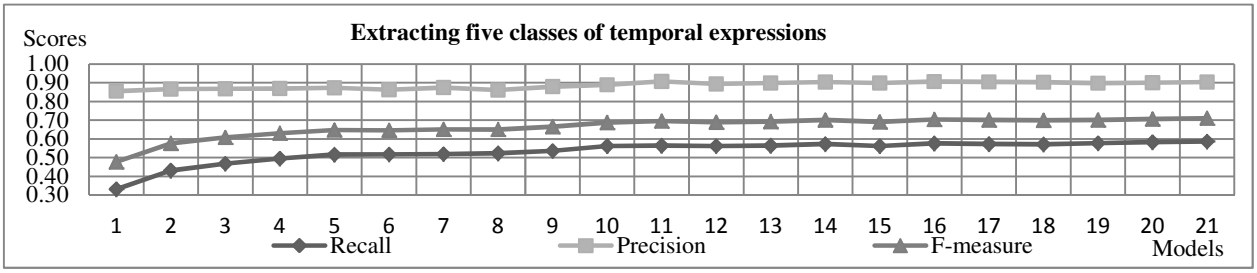


**Figure 2**. Overall performance of five classes of temporal expressions.

| Temporal Class | Frequency | Baseline (model #1) [R/P/F] | Best model (model #22) [R/P/F] |
|---|---|---|---|
| **Date and time** | 362 | 0.377/0.870/0.526 | 0.674/0.915/0.776 |
| **Event dependent temporal** | 72 | 0.000/0.000/0.000 | 0.042/0.500/0.077 |
| **Fuzzy time** | 243 | 0.432/0.846/0.572 | 0.697/0.884/0.780 |
| **Duration** | 116 | 0.273/0.833/0.412 | 0.543/0.909/0.680 |
| **Frequency** | 103 | 0.035/0.750/0.066 | 0.200/0.955/0.331 |
| **Micro Average** | 896 | 0.331/0.855/0.477 | 0.586/0.904/0.711 |

**Table 3**. The temporal expressions performance for each class (with frequency) by applying the baseline model and the best model[4].

| Temporal Class | **Rule method** (highly restricted training sample expansion mapping) **[R/P/F]** | **Combined model [R/P/F]** |
|---|---|---|
| **Date and time** | 0.687/0.777/0.729 | 0.749/0.935/0.831 |
| **Key event** | 0.214/0.882/0.345 | 0.236/0.944/0.378 |
| **Fuzzy time** | 0.729/0.854/0.787 | 0.798/0.898/0.845 |
| **Duration** | 0.438/0.875/0.583 | 0.655/0.883/0.753 |
| **Frequency** | 0.732/0.740/0.736 | 0.699/0.960/0.809 |
| **Micro Average** | 0.634/0.805/0.710 | 0.703/0.920/0.797 |

**Table 4**. The scores of the highly restricted training sample expansion mapping and the combined learning model for temporal events[4].

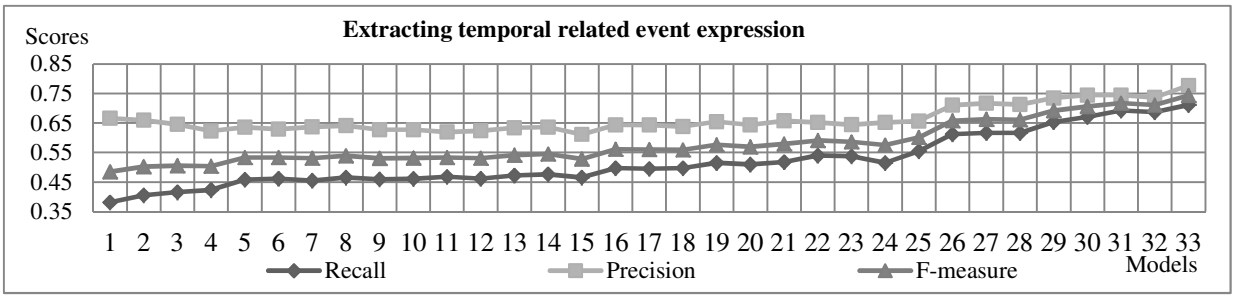


**Figure 3.** Performance of medical event extractions.

| Temporal Class | Baseline model (model #1) [R/P/F] | Best model (model #33) [R/P/F] |
|---|---|---|
| **Temporal related medical event** | 0.381/ 0.666/0.485 | 0.712/0.778/0.743 |

**Table 5.** The scores of baseline model and the best model in the temporal related medical event extraction[4]

[4] R: Recall, P: Precision, F: F-score

| Model # | Features | Model # | Features |
|---|---|---|---|
| **1** | Bag-of-words | **18** | Model 16 + Prefix |
| **2** | Bag-of-words with correction | **19** | **Model 16 + Prefix (2)** |
| **3** | **Bag-of-lemma with correction** | **20** | Model 16 + Prefix (4) |
| **4** | Model 3+ part-of-speech | **21** | **Model 19 + Suffix** |
| **5** | **Model 3 + part-of-speech (2)** | **22** | Model 19 + Suffix (2) |
| **6** | Model 3 + part-of-speech (4) | **23** | **Model 19 + Suffix (4)** |
| **7** | Model 5 + Chunk | **24** | Model 23 + Scores & Measures |
| **8** | **Model 5+ Chunk (2)** | **25** | **Model 23 + Unit of measurement** |
| **9** | Model 5+ Chunk (4) | **26** | **Model 23 + Temporal expression** |
| **10** | Model 8+ SCT concept | **27** | **Model 26 + Temporal class** |
| **11** | **Model 8 + SCT concept (2)** | **28** | **Model 27 + Number of temporal expression** |
| **12** | Model **8** + SCT concept (4) | **29** | **Model 28+ The position of the temporal expression** |
| **13** | Model 11 + SCT category | **30** | Model 29 + Punctuation |
| **14** | **Model 11 + SCT category (2)** | **31** | **Model 29 + Punctuation (2)** |
| **15** | Model 11 + SCT category (4) | **32** | Model 29 + Punctuation (4) |
| **16** | **Model 14 + Orthography** | **33** | **Model 31 + Distance between the temporal expression & event expression** |
| **17** | Model 14 + Orthography (2) | | |

**Table 6**. Features applied in each event extraction experiment[5]

## Discussion

### *Temporal Expression Extraction Discussion*

Since temporal expressions are commonly formed as a multi-word expression, the context needs to be represented not only by the tokens to the left and right, but also the features of the token in the context window. We have established that the feature window does not necessarily have to be the same as the token window. Models 5 and 6 demonstrate that the features of two POS windows (#5: current token POS, the preceding POS, and the succeeding POS) provides better results than using a larger window of four (#6: current POS, two preceding POS, and two succeeding POS) where an incremental method was used to discover the best window size for context. It should be noted that the "highly restricted training sample expansion mapping" was not applied in the initial experiments, since the reliability of a rule-based approach needs to be proven first. Even though, a satisfactory improvement is produced by the best model, which is approximately 20% higher than the baseline model. The detailed scores of these two learning models are demonstrated in Table 3. However, a considerable gap exists between the precision and recall, about 50% in the baseline while around 30% in the best model. This phenomenon is caused by the fact that:

1. The contexts which surround the temporal expressions are not quite helpful enough, since the same temporal expressions can be mentioned in various contexts. Consequently, the context knowledge from the training set can only capture limited distinctiveness in the test set.

2. The diverse distribution of temporal expressions also creates difficulties in using known instances to predict unseen instances. The scores for "Key event" is a good example which is hardly recognised in the test set, since it is constructed by plain words and has limited samples.

The "highly restricted training sample expansion mapping" method is able to pinpoint this problem by applying a knowledge based expansion on the training instances. The performance of this rule-based approach and the combined model (adding the mapping result into the previous learning model) is promising (see Table 4). Through Table 3 and 4, it is clear that almost the same F-score (micro-average) can be obtained by the best model and the rule-based method. In contrast to the initial learning model, the rule method can produce a higher recall and lower precision than the automatic expansion. The major advantage of this rule method is that it is effortless work. No regular expression rules need to be manually compiled for each temporal expression in the training set, even when new training data becomes available in the future. Finally, the rule result is adopted as a learning feature and added into the previous best learning model to evaluate whether the rule method can make any contribution for the CRF learner. As a result, by studying the positive and negative rule result, both precision and recall are improved in the combined learning model with a significant jump in recall. The best result in the temporal extraction model produces an F-score of 79.7%. The prediction results reveal that the false positives come from:

[5] "(2)" means a window of one label before/after the current label. "(4)" means a window of two labels before/after the current label.

1. Temporal adjectives and adverbs which describe medication events are classified as a temporal expression: e.g., *"Denies pain on hourly assessment.", "Check gent trough with mane bloods please", "Daily ECG attended at 1300.",* etc. The underlined expression should not be labelled.

2. Non medical events related to temporal expressions, which commonly exist in the "Family history" and "Social" section: e.g., *"Her younger brother (30) is now engaged", "She was unemployed since Dec 03", "found by flatmate yesterday after having taken an overdose of venlafaxine",* etc. The underlined expression should not be labelled.

3. Temporal expression with multiple meanings: e.g., *"Patient has woken once prior to take back to ot.", "Bair Hugger in place more warm than earlier",* etc. The underlined expressions should not be labelled, since the meaning of these expressions in these examples are not temporal related. For example, the *"once"* in the first example means *"as soon as"* rather than *"one time"* (e.g., "once a day"), while the *"earlier"* in the second example means*" a previous condition"* rather than *"near to the beginning of a period of time"* (e.g., "early morning").

4. Plural has more weight for the duration: e.g., *"crying most days", "Late to settle but appeared to have slept well for a good few hours."* The underlined expression should be labelled as "fuzzy time" rather than "duration", since the plural like *"days"* and *"hours"* have more change to be part of a "duration", like *"three days"* and *"2 hours"*.

5. Partially identified: e.g., *in "EtOH +++, ex-smoker quit Dec 2004"* only 2004 is labelled as a "date and time" *while the gold annotation is "Dec 2004" in "Pt oliguric most of the shift",* only "the shift" is marked as a "fuzzy time" while the correct annotation should be "most of the shift", etc.

It is important to stress that the "Partially identified" issues produces not only the false positives by also false negatives. The remaining false negatives are caused by the diverse content surrounding the temporal expressions, and unseen temporal structures.

***Temporally Defined Medical Event Expression Extraction Discussion***

In this task, only temporally defined medical events are considered rather than all medical events. However, it is common for the same medical events to be temporally related or non-temporally related depending on the context. Thus, a word level smoothing and generation strategy may increase the similarity between them and create false positives, which leads to a decrease in the precision through the first three feature model. In the previous temporal expression extraction task, due to temporal expressions commonly being formed by prepositions and non-word tokens, a consistent improvement of precision and recall is achieved by the effect of smoothing and generalisation.

Furthermore, another different facet of this task is that the medical event is generally recorded by multiple words. This indicates that a larger window size (two) for features is beneficial in this task as is demonstrated in Table 6. However, significant drop in scores happens when the window size was increased to four, like "SCT category" or the "Scores and measures" were embedded in the feature model. It seems that the probability distribution is broken by the multiple semantic categories near the event instance and the poor performance of "Scores and measures" on the unseen data. The performance is dramatically improved when the knowledge of temporal expressions were appended in model #26, #27 and #28. This information can not only filter the non-relevant content for the task but also indicate the approximate location of the temporally related events. In addition, the "Punctuation and the Distance between the temporal expression & event expression" is very beneficial for dealing with the relation between temporal expressions and event expressions, especially for the phenomena of multiple corresponding events on a single temporal expression. Overall, the best model (#33) produced an F-score of 74.3% for this task (see Table 5 for the scores of the baseline model and the best model). The closeness of precision and recall was supported by the punctuation study which separates the sentence into chunks and assigns values to each chunk. In this way, the CRF learner is fed by not only the text level features, but also the structure features. The text level features cannot recognise new events in a different grammatical structure and different context. Even worse, the similarity within the training examples is very weak that leads to the high precision and low recall. Converting the sentence into weighted chunks can provide more general structure and context for CRF learner (see Event expression feature 7. Punctuation and the Distance between a temporal expression and an event expression).

Through analysis of the output of the best model, the errors (both false positive and false negative) in most cases occurred in long sentences rather than short sentences. The short sentences, like "Remains on PSV tomorrow." , "No midazolam given this morning." "BP 119/38(61) at time of report.", "Restart Warfarin tonight." etc., have a very

simple structure (a medical event followed by a temporal expression). Major systematic errors in long sentences are listed below (true positive events are underlined, false positive events are in wave underline, false negative events have a dotted underline, and temporal expressions are in double underline):

1. Some instances in the unrecognised event list: e.g., *"Placed onto SIMV onc sedation took full effect, currently on SIMV, RR 14, set TV 550, TVe 521, PIP 16, MV 7.6, Fi02 40%, Sp02 95-97%, alst ABg unsatisfactroy, nurse taking over and nurse in charge informed.", "CVS: Monitored in SB initially, now SR with MAP75-95, CVP 12-14, Cool peripheries, pulses present in all 3 limbs.",* etc. In these two examples, the false negative events (*"MV 7.6", "Fi02 40%","Sp02 95-97%","alst ABg unsatisfactroy", etc.*) are too far away from the temporal expressions (in double underline).

2. The conjunctive events can happen at two different times : e.g., *"Pt was extubated and she passed away at 1645.", "Pt is on Vancomycin and level has not been taken this morning",* etc. The first example indicates that the patient was *"extubated before 1645"* and then *"passed away at 1645"*, while in the second example, both two events (*"Pt is on Vancomycin"* and *"level has not been taken"*) happened in the morning.

3. The various usages of conjunctions: e.g., *"In late afternoon, patient woke inappropriately, grabbing at ETT, thrashing around on the bed despite reassurance, adn reorientation, became tachycardic 999-105bpm), hypertensive with MAP 98, and tachypneic with RR 33 bpm, and droppingperipheral saturations to 90%."*. The long sentences with multiple conjunctions (*"and"*) brought trouble for the CRF learner to decide which medical event is temporal related. As a consentience, *"tachypneic with RR 33 bpm",* and *"droppingperipheral saturations to 90%"* were missed in the event targeting.

4. Multiple events separated by punctuation and temporal expressions. *e.g., "Pt was sitting out in bed this morning eating breakfast.", "BNO this shift, BS present.",* etc. In these examples, the false negative events (*"eating breakfast"* and *"BS present"*) were caused by the lack of dependency to the temporal expressions.

5. Non-medical events: e.g., the *"lives in Port Macquarie with husband"* in the sentence *"Current: Assist with all ADLs, lives in Port Macquarie with husband.",* the *"visitors state this is usual"* in the sentence *"Pt sweaty and clammy, visitors state this is usual at night.",* etc.

**Conclusion**

In this paper, we have described a novel information extraction method for clinical temporal information extractions and corresponding event expressions by using the 'highly restricted training sample expansion' and the 'structure distance between the temporal expressions & related event expression' as effective learning features. The performance of temporal expression extraction is quite encouraging as it produces a score close to the state of the art systems in the news domain[5, 6, and 8] achieved, for example, in TARSQI[5,6], the temporal expressions tagger can achieved 82% F-score, as well the best system in TempEval2 challenge (HeidelTime[8] ) performed at 86% on temporal span extraction and 96% on temporal type assignment. In the fact of domain dependent (the newswire domain), a dramatic deterioration happened when applying them to the clinical domain. A recent study shows an F-score of 15% can be achieved by using the Tarsqi Toolkit on clinical notes[25]. However, a significant gap (approximately 10%) in the performance on temporally related event extraction exists between our system and the best systems in news domain, but it does outperform the TN-TIES[15] which can chunk 91% of the temporal segments and classified them into five temporal with an F-score in the 60s. Due to the poor grammatical structure in EPRs, a more robust method needs to be designed, for instance, a sophisticated clinical parser with grammar correction working in concert with domain specific knowledge.

**References**


1. Grishman R, and Sundheim B. Message Understanding Conference-6: a brief history. Proceedings of the 16th International Conference on Computational Linguistics (COLING). 1996: 466-471.
2. Chinchor NA. MUC-7 Named Entity Task Definition (Version 3.5). Proceedings of the Seventh Message Understanding Conference. 1997.
3. Ferro L, Mani I, Sundheim B, and Wilson G. "TIDES Temporal Annotation Guidelines - Version 1.0.2". MITRE Technical Report MTR 01W0000041. McLean, Virginia: The MITRE Corporation. June 2001.
4. Pustejovsky J, Castaño J, Ingria R, Saurí R, Gaizauskas R, Setzer A and Katz G. TimeML: Robust Specification of Event and Temporal Expressions in Text. IWCS-5, Fifth International Workshop on Computational Semantics. 2003.

5. Verhagen M, Mani I, Saurí R, Knippen R, Littman J and Pustejovsky J. Automating Temporal Annotation with TARSQI. Demo Session. Proceedings of the ACL 2005.
6. Verhagen M, Mani I, Saurí R, Knippen R, Littman J and Pustejovsky J. Temporal Processing with the TARSQI Toolkit. In proceedings Coling 2008: Companion volume - Posters and Demonstrations. 2008:189-192.
7. Verhagen M, Sauri R, Caselli T, Pustejovsky J. SemEval-2010 task 13: TempEval-2. In Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010). 2010:57-62.
8. Strotgen J and Gertz M. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. Proceedings of the 5th International Workshop on Semantic Evaluation. 2010: 321-324,
9. UzZaman N and Allen JF. TRIPS and TRIOS System for TempEval-2: Extracting Temporal Information from Text. Proceedings of the 5th International Workshop on Semantic Evaluation. 2010:276-283.
10. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. J Biomed Inform. 2006:424-39.
11. Zhou L, Friedman C, Parsons S, Hripcsak G. System architecture for temporal information extraction, resentation and reasoning in clinical narrative reports. AMIA Annu Symp Proc. 2005:869-73.
12. Zhou L, Parsons S, Hripcsak G. The Evaluation of a Temporal Reasoning System in Processing Clinical Discharge Summaries. J Am Med Inform Assoc. 2007.
13. Mowery DL, Harkema H and Chapman WW. BioNLP '08 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. 2008.
14. Mowery DL, Harkema H, Dowling JN, Lustgarten JL, Chapman WW. Distinguishing historical from current problems in clinical reports: which textual features help? Proceeding BioNLP '09 Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. 2009:10-18.
15. Irvine AK, Haas SW and Sullivan T. TN-TIES: A System for Extracting Temporal Information from Emergency Department Triage Notes . AMIA Annu Symp Proc. 2008: 328-332.
16. Tsuruoka Y, Tateishi Y, Kim JD, Ohta T, McNaught J, Ananiadou S and Tsujii J. Developing a Robust Part-of-Speech Tagger for Biomedical Text, Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, 2005:382-392
17. Patrick J and Sabbagh M. An Active Learning Process for Extraction and Standardisation of Medical Measurements by a Trainable FSA. In Proceedings of CICLing. 2011:151~162.
18. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. Journal of the American Medical Informatics Association. 2010;17:524-527.
19. Patrick J, Wang Y, Budd P. An automated system for conversion of clinical notes into SNOMED clinical terminology, In Proc. 5rd Australasian symposium on ACSW frontiers. 2007; 68: 219-226.
20. Steven B and Martin JH. Identification of event mentions and their semantic class. In EMNLP: Proceedings of the Conference on Empirical Methods in NLP. 2006:146-154.
21. Llorens H, Saquete E, and Navarro-Colorado B. TimeML events recognition and classification: learning CRF models with semantic roles. Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010). 2010.
22. Curran JR, Clark S, and Bos J. Linguistically Motivated Large-Scale NLP with C&C and Boxer. Proceedings of the ACL 2007 Demonstrations Session. 2007: 33-36.
23. Lease M, and Charniak E. Parsing Biomedical Literature. In Second International Joint Conference on Natural Language Processing (IJCNLP'05). 2005:58-69.
24. Miyao Y, Tsujii J. Probabilistic Disambiguation Models for Wide-Coverage HPSG Parsing. In Proceedings of ACL. 2005:83-90.
25. Ong FR. The Tarsqi Toolkit's Recognition of Temporal Expressions within Medical Documents. Master's Thesis of Vanderbilt University. 2009. http://etd.library.vanderbilt.edu/available/etd-05242010-101636/ (accessed 14 March 2012).
26. Tao C, Wei WQ, Solbrig HR, Savova G, and Chute CG. CNTRO: A Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives. AMIA Annu Symp Proc. 2010: 787-791.
27. Tao C, Solbrig HR, Sharma DK, Wei WQ, Savova GK and Chute CG. Lect Notes Comput Sci. 2010; 6497: 241-256.
28. Lafferty J, McCallum A, and Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In Proc. of ICML, 2001:282-289.
29. Galescu L and Blaylock N. A corpus of clinical narratives annotated with temporal information. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012: 715-720.