JU_CSE_TEMP: A First Step towards Evaluating Events, Time Expressions and Temporal Relations

Anup Kumar Kolya[1], Asif Ekbal[2] and Sivaji Bandyopadhyay[3]

[1,3]Department of Computer Science and Engineering, Jadavpur University,
Kolkata-700032, India
[2]Department of Computational Linguistics, Heidelberg University,
Heidelberg-69120, Germany
Email: anup.kolya@gmail.com[1], asif.ekbal@gmail.com[2]
and sivaji_cse_ju@yahoo.com[3]

Abstract

Temporal information extraction is a popular and interesting research field in the area of Natural Language Processing (NLP). In this paper, we report our works on TempEval-2 shared task. This is our first participation and we participated in all the tasks, i.e., A, B, C, D, E and F. We develop rule-based systems for Tasks A and B, whereas the remaining tasks are based on a machine learning approach, namely Conditional Random Field (CRF). All our systems are still in their development stages, and we report the very initial results. Evaluation results on the shared task English datasets yield the precision, recall and F-measure values of 55%, 17% and 26%, respectively for Task A and 48%, 56% and 52%, respectively for Task B (event recognition). The rest of tasks, namely C, D, E and F were evaluated with a relatively simpler metric: the number of correct answers divided by the number of answers. Experiments on the English datasets yield the accuracies of 63%, 80%, 56% and 56% for tasks C, D, E and F, respectively.

1 Introduction

Temporal information extraction is, nowadays, a popular and interesting research area of Natural Language Processing (NLP). Generally, events are described in different newspaper texts, stories and other important documents where events happen in time and the temporal location and ordering of these events are specified. One of the important tasks of text analysis clearly requires identifying events described in a text and locating these in time. This is also important in a wide range of NLP applications that include temporal question answering, machine translation and document summarization.

In the literature, temporal relation identification based on machine learning approaches can be found in Boguraev et el. (2005), Mani et al. (2006), Chambers et al. (2007) and some of the TempEval 2007 participants (Verhagen et al., 2007). Most of these works tried to improve classification accuracies through feature engineering. The performance of any machine learning based system is often limited by the amount of available training data. Mani et al. (2006) introduced a temporal reasoning component that greatly expands the available training data. The training set was increased by a factor of 10 by computing the closure of the various temporal relations that exist in the training data. They reported significant improvement of the classification accuracies on event-event and event-time relations. Their experimental result showed the accuracies of 62.5%-94.95% and 73.68%-90.16% for event-event and event-time relations, respectively. However, this has two shortcomings, namely feature vector duplication caused by the data normalization process and the unrealistic evaluation scheme. The solutions to these issues are briefly described in Mani et al. (2007). In TempEval 2007 task, a common standard dataset was introduced that involves three temporal relations. The participants reported F-measure scores for event-event relations ranging from 42% to 55% and for event-time relations from 73% to 80%. Unlike (Mani et al., 2007; 2006), event-event temporal relations were not discourse-wide (i.e., *any* pair of events can be temporally linked) in TempEval 2007. Here, the event-event relations were restricted to events within two consecutive sentences. Thus, these two frameworks produced highly dissimilar re-

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 345–350,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

sults for solving the problem of temporal relation classification.

In order to apply various machine learning algorithms, most of the authors formulated temporal relation as an event paired with a time or another event and translated these into a set of feature values. Some of the popularly used machine learning techniques were Naive-Bayes, Decision Tree (C5.0), Maximum Entropy (ME) and Support Vector Machine (SVM). Machine learning techniques alone cannot always yield good accuracies. To achieve reasonable accuracy, some researchers (Mao et al., 2006) used hybrid approach. The basic principle of hybrid approach is to combine the rule-based component with machine learning. It has been shown in (Mao et al., 2006) that classifiers make most mistakes near the decision plane in feature space. The authors carried out a series of experiments for each of the three tasks on four models, namely naive-Bayes, decision tree (C5.0), maximum entropy and support vector machine. The system was designed in such a way that they can take the advantage of rule-based as well as machine learning during final decision making. But, they did not explain exactly in what situations machine learning or rule based system should be used given a particular instance. They had the option to call either component on the fly in different situations so that they can take advantage of the two empirical approaches in an integrated way.

The rest of the paper is structured as follows. We present very brief descriptions of the different tasks in Section 2. Section 3 describes our approach in details with rule-based techniques for tasks A and B in Subsection 3.1, CRF based techniques in Subsection 3.2 for tasks C, D, E and F, and features in Subsection 3.3. Detailed evaluation results are reported in Section 4. Finally, Section 5 concludes the paper with a direction to future works.

2 Task Description

The main research in this area involves identification of all temporal referring expressions, events and temporal relations within a text. The main challenges involved in this task were first addressed during TempEval-1 in 2007 (Verhagen et al., 2007). This was an initial evaluation exercise based on three limited tasks that were considered realistic both from the perspective of assembling resources for development and testing and from the perspective of developing systems capable of addressing the tasks. In TempEval 2007, following types of event-time temporal relations were considered: **Task A (**relation between the events and times within the same sentence), **Task B** (relation between events and document creation time) and **Task C (**relation between verb events in adjacent sentences). The data sets were based on TimeBank, a hand-built gold standard of annotated texts using the TimeML markup scheme[1]. The data sets included sentence boundaries, timex3 tags (including the special document creation time tag), and event tags. For tasks A and B, a restricted set of events was used, namely those events that occur more than 5 times in TimeBank. For all three tasks, the relation labels used were before, after, overlap, before-or-overlap, overlap-or-after and vague. Six teams participated in the TempEval tasks. Three of the teams used statistics exclusively, one used a rule-based system and the other two employed a hybrid approach. For task A, the range of F-measure scores were from 0.34 to 0.62 for the *strict scheme* and from 0.41 to 0.63 for the *relaxed scheme*. For task B, the scores were from 0.66 to 0.80 (*strict*) and 0.71 to 0.81 (*relaxed*). Finally, task C scores range from 0.42 to 0.55 (*strict*) and from 0.56 to 0.66 (*relaxed*).

In TempEval-2, the following six tasks were proposed:

A: The main task was to determine the *extent of the time expressions* in a text as defined by the TimeML timex3 tag. In addition, values of the features *type* and *val* had to be determined. The possible values of *type* are time, date, duration, and set; the value of *val* is a normalized value as defined by the timex2 and timex3 standards.

B. Task was to determine the *extent of the events* in a text as defined by the TimeML event tag. In addition, the values of the features tense, aspect, polarity, and modality had to be determined.

C. Task was to determine the *temporal relation* between an *event* and a *time expression* in the same sentence.

D. *Temporal* relation between an *event* and the *document creation* time had to be determined.

E. *Temporal* relation between two *main events* in consecutive sentences had to be determined.

F. *Temporal relation* between two *events,* where one event syntactically dominates the other event.

In our present work, use handcrafted rules for Task A and Task B. All the other tasks, i.e., C, D, E and F are developed based on the well known statistical algorithm, Conditional Random

[1] www.timeml.org for details on TimeML

Field (CRF). For CRF, we use only those features that are available in the training data. All the systems are evaluated on the TempEval-2 shared task English datasets. Evaluation results yield the precision, recall and F-measure values of 55%, 17% and 26%, respectively for Task A and 48%, 56% and 52%, respectively for Task B. Experiments on the other tasks demonstrate the accuracies of 63%, 80%, 56% and 56% for C, D, E and F, respectively.

3 Our Approach

In this section, we present our systematic approach for *evaluating events*, *time expressions* and *temporal relations* as part of our first participation in the TempEval shared task. We participated in all the six tasks of TempEval-2. Rule-based systems are developed using a preliminary handcrafted set of rules for tasks A and B. We use machine learning approach, namely CRF for solving the remaining tasks, i.e., C, D, E and F.

3.1 Rules for Task A and Task B

We manually identify a set of rules studying the various features available in the training data. There were some exceptions to these rules. However, a rule is used if it is found to be correct most of the time throughout the training data. It is to be noted that these are the very preliminary rules, and we are still working on finding out more robust rules. Below, we present the rules for tasks A and B.

Task A. The time expression is identified by defining appropriate regular expression. The regular expressions are based on several entities that denote month names, year, weekdays and the various digit expressions. We also use a list of keywords (e.g., day, time, AM, PM etc.) that denote the various time expressions. The values of various attributes (e.g., *type* and *value*) of time expressions are computed by some simple template matching algorithms.

Task B. In case of Task B, the training data is initially passed through the Stanford PoS tagger[2]. We consider the tokens as the events that are tagged with POS tags such as *VB*, *VBG*, *VBN*, *VBP*, *VBZ* and *VBD*, denoting the various verb expressions. Values of different attributes are computed as follows.

a. Tense: A manually augmented suffix list such as: "*ed*","*d*","*t*" etc. is used to capture the proper tense of any event verb from surface level orthographic variations.

b. Aspect: The Tense-Aspect-Modality (TAM) for English verbs is generally associated with auxiliaries. A list is manually prepared. Any occurrence of main verb with continuous aspect leads to search for the adjacent previous auxiliary and rules are formulated to extract TAM relation using the manually generated checklist. A separate list of auxiliaries is prepared and successfully used for detection of progressive verbs.

c. **Polarity**: Verb-wise polarity is assigned by the occurrence of previous negation words. If any negation word appears before any event verb then the resultant polarity is negative; otherwise, the verb considered as positive by default.

d. **Modality**: We prepare a manual list that contains the words such as: *may*, *could*, *would* etc. The presence of these modal auxiliaries gives modal tag to the targeted verb in a sentence otherwise it is considered a non-modal.

e. **Class**: We select '*occurrence*' to be class value by default.

3.2 Machine Learning Approach for Tasks C, D, E and F

For tasks C-F, we use a supervised machine learning approach that is based on CRF. We consider the temporal relation identification task as a pair-wise classification problem in which the target pairs–a TIMEX3 tag and an EVENT–are modelled using CRF, which can include arbitrary set of features, and still can avoid overfitting in a principled manner.

Introduction to CRF. CRF (Lafferty et al., 2001), is used to calculate the conditional probability of values on designated output nodes given values on other designated input nodes. The conditional probability of a state sequence $S =< s_1, s_2, ..., s_T >$ given an observation sequence $O =< o_1, o_2, ..., o_T)$ is calculated as:

$$P_{\wedge}(s \mid o) = \frac{1}{Z_o}\exp\left(\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(s_{t-1}, s_t, o, t)\right)$$

where, $f_k(s_{t-1}, s_t, o, t)$ is a feature function whose weight λ_k is to be learned via training. The values of the feature functions may range between $-\propto .....+\propto$, but typically they are

[2] http://nlp.stanford.edu/software/tagger.shtml

binary. To make all conditional probabilities sum up to 1, we must calculate the normalization factor,

$$Z_0 = \sum_s \exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t)),$$

which, as in HMMs, can be obtained efficiently by dynamic programming.

To train a CRF, the objective function to be maximized is the penalized log-likelihood of the state sequences given the observation sequence:

$$L_\wedge = \sum_{i=1}^{N} \log(P_\wedge(s^{(i)} \mid o^{(i)})) - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma^2},$$

where, $\{<o^{(i)}, s^{(i)}>\}$ is the labeled training data. The second sum corresponds to a zero-mean, σ^2 -variance Gaussian prior over parameters, which facilitates optimization by making the likelihood surface strictly convex.

CRFs generally can use real-valued functions but it is often required to incorporate the binary valued features. A feature function $f_k(s_{t-1}, s_t, o, t)$ has a value of 0 for most cases and is only set to 1, when s_{t-1}, s_t are certain states and the observation has certain properties. Here, we set parameters λ to maximize the penalized log-likelihood using Limited-memory BFGS (Sha and Pereira, 2003) a quasi-Newton method that is significantly more efficient, and which results in only minor changes in accuracy due to changes in σ.

We use the OpenNLP C^{++} based CRF++ package[3], a simple, customizable, and open source implementation of CRF for segmenting /labeling sequential data.

3.3 Features of Tasks C, D, E and F

We extract the gold-standard TimeBank features for events and times in order to train/test the CRF. In the present work, we mainly use the various combinations of the following features:

(i). **Part of Speech (POS)** of event terms: It denotes the POS information of the event. The features values may be either of ADJECTIVE, NOUN, VERB, and PREP.

(ii). **Event Tense**: This feature is useful to capture the standard distinctions among the grammatical categories of verbal phrases. The tense attribute can have values, PRESENT, PAST, FUTURE, INFINITIVE, PRESPART, PASTPART, or NONE.

(iii). **Event Aspect**: It denotes the aspect of the events. The aspect attribute may take values**,** PROGRESSIVE, PERFECTIVE and PERFECTIVE PROGRESSIVE or NONE.

(iv). **Event Polarity**: The polarity of an event instance is a required attribute represented by the boolean attribute, polarity. If it is set to 'NEG', the event instance is negated. If it is set to 'POS' or not present in the annotation, the event instance is not negated.

(v). **Event Modality**: The modality attribute is only present if there is a modal word that modifies the instance.

(vi). **Event Class**: This is denoted by the 'EVENT' tag and used to annotate those elements in a text that mark the semantic events described by it. Typically, events are verbs but can be nominal also. It may belong to one of the following classes:

REPORTING: Describes the action of a person or an organization declaring something, narrating an event, informing about an event, etc. For example, *say*, *report*, *tell*, *explain*, *state* etc.

PERCEPTION: Includes events involving the physical perception of another event. Such events are typically expressed by verbs like: *see, watch*, *glimpse*, *behold*, *view*, *hear*, *listen*, *overhear* etc.

ASPECTUAL: Focuses on different facets of event history. For example, *initiation*, *reinitiation*, *termination*, *culmination*, *continuation* etc.

I_ACTION: An intentional action. It introduces an event argument which must be in the text explicitly describing an action or situation from which we can infer something given its relation with the I_ ACTION.

I_STATE: Similar to the I_ACTION class. This class includes states that refer to alternative or possible words, which can be introduced by subordinated clauses, nominalizations, or untensed verb phrases (VPs).

STATE: Describes circumstances in which something obtains or holds true.

Occurrence: Includes all of the many other kinds of events that describe something that happens or occurs in the world.

(vii). **Type of temporal expression:** It represents the temporal relationship holding between events, times, or between an event and a time of the event.

(viii). **Event Stem**: It denotes the stem of the head event.

[3] http://crfpp.sourceforge.net

(ix). **Document Creation Time**: The document creation time of the event.

4 Evaluation Results

Each of the tasks is evaluated with the TempEval-2 shared task datasets.

4.1 Evaluation Scheme

For the extents of events and time expressions (tasks A and B), precision, recall and the F-measure are used as evaluation metrics, using the following formulas:

Precision (P) = tp/ (tp + fp)
Recall (R) = tp/ (tp + fn)
F-measure = 2 *(P * R)/ (P + R)

Where, tp is the number of tokens that are part of an extent in both keys and response,
fp is the number of tokens that are part of an extent in the response but not in the key, and
fn is the number of tokens that are part of an extent in the key but not in the response.

An even simpler evaluation metric similar to the definition of 'accuracy' is used to evaluate the attributes of events and time expressions (the second part of tasks, A and B) and for relation types (tasks C through F). The metric, henceforth referred to as 'accuracy', is defined as below:

Number of correct answers/ Number of answers present in the test data

4.2 Results

For tasks A and B, we identify a set of rules from the training set and apply them on the respective test sets.

The tasks C, D, E and F are based on CRF. We develop a number of models based on CRF using the different features included into it. A feature vector consisting of the subset of the available features as described in Section 2.3 is extracted for each of <event, timex>, <event, DCT>, <event, event> and <event, event> pairs in tasks C, D, E and F, respectively. Now, we have a training data in the form (W_i, T_i), where, W_i is the i^{th} pair along with its feature vector and T_i is it's corresponding TempEval relation class. Models are built based on the training data and the feature template. The procedure of training is summarized below:

1. Define the training corpus, C.
2. Extract the corresponding relation from the training corpus.
3. Create a file of candidate features, including lexical features derived from the training corpus.
4. Define a feature template.
5. Compute the CRF weights λ_k for every f_K using the CRF toolkit with the training file and feature template as input.

During evaluation, we consider the following feature templates for the respective tasks:

(i) **Task C**: Feature vector consisting of current token, polarity, POS, tense, class and value; combination of token and type, combination of tense and value of the current token, combination of aspect and type of current token, combination of aspect, value and type of the current token.

(ii) **Task D**: Feature vector consisting of current token and POS; combination of POS and tense of the current token, combination of polarity and POS of the current token, combination of POS and aspect of current token, combination of polarity and POS of current token, combination of POS, tense and aspect of the current token.

(iii). **Task E**: Current token, combination of event-class and event-id of the current token, combination of POS tags of the pair of events, combination of (tense, aspect) values of the event pairs.

(iv). **Task F**: Current token, combination of POS tags of the pair of events, combination of tense values of the event pairs, combination of the aspect values of the event pairs, combination of the event classes of the event pairs.

Experimental results of tasks A and B are reported in Table 1 for English datasets. The results for task A, i.e., recognition and normalization of time expressions, yield the precision, recall and F-measure values of 55%, 17% and 26%, respectively. For task B, i.e., event recognition, the system yields precision, recall and F-measure values of 48%, 56% and 52%, respectively. Event attribute identification shows the accuracies of 98%, 98%, 30%, 95% and 53% for *polarity*, *mood*, *modality*, *tense*, *aspect* and *class*, respectively. These systems are the *baseline* models, and the performance can further be improved with a more carefully handcrafted set of robust rules. In further experiments, we would also like to apply machine learning methods to these problems.

Task	precision (in %)	recall (in %)	F-measure (in %)
A	55%	17%	26%
B	48%	56%	52%

Table 1. Experimental results on tasks A and B

Evaluation results on the English datasets for tasks C, D, E and F are presented in Table 2. Experiments show the accuracies of 63%, 80%, 56% and 56% for tasks C, D, E and F, respectively. Results show that our system performs best for task D, i.e., relationships between *event* and *document creation time*. The system achieves an accuracy of 63% for task C that finds the temporal relation between an *event* and a *time expression* in the same sentence. The system performs quite similarly for tasks E and F. It is to be noted that there is still the room for performance improvement. In the present work, we did not carry out sufficient experiments to identify the most suitable feature templates for each of the tasks. In future, we would experiment after selecting a development set for each task; and find out appropriate feature template depending upon the performance on the development set.

Task	Accuracy (in %)
C	63%
D	80%
E	56%
F	56%

Table 2. Experimental results on tasks C, D, E and F

5 Conclusion and Future Works

In this paper, we report very preliminary results of our first participation in the TempEval shared task. We participated in all the tasks of TempEval-2, i.e., A, B, C, D, E and F for English. We develop the rule-based systems for tasks A and B, whereas the remaining tasks are based on a machine learning approach, namely CRF. All our systems are still in their development stages. Evaluation results on the shared task English datasets yield the precision, recall and F-measure values of 55%, 17% and 26%, respectively for Task A and 48%, 56% and 52%, respectively for Task B (event recognition). Experiments on the English datasets yield the accuracies of 63%, 80%, 56% and 56% for tasks C, D, E and F, respectively.

Future works include identification of more precise rules for tasks A and B. We would also like to experiment with CRF for these two tasks. We would experiment with the various feature templates for tasks C, D, E and F. Future works also include experimentations with other machine learning techniques like maximum entropy and support vector machine.

References

Boguraev, B. and R. K. Ando. 2005. TimeML Compliant Text Analysis for Temporal Reasoning. In *Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, Scotland, August, pages 997–1003.

Chambers, N., S., Wang, and D., Jurafsky. , 2007. Classifying Temporal Relations between Events. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic, June, pages 173–176.

Lafferty, J., McCallum, A., and Pereira, F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of 18th International Conference on Machine Learning*, 2001.

Mani, I., B., Wellner, M., Verhagen, and J. Pustejovsky. 2007. Three Approaches to Learning TLINKs in TimeML. *Technical Report CS-07-268*, Computer Science Department, Brandeis University, Waltham, USA.

Mani, I., Wellner, B., Verhagen, M., Lee C.M., Pustejovsky, J. 2006. Machine Learning of Temporal Relation. In *Proceedings of the COLING/ACL*, Sydney, Australia, ACL.

Mao, T., Li., T., Huang, D., Yang, Y. 2006. Hybrid Models for Chinese Named Entity Recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*.

Sha, F., Pereira, F. 2003. Shallow Parsing with Conditional Random Fields. In *Proceedings of HLT-NAACL*, 2003.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, and J.: SemEval-2007 Task 15: TempEval Temporal Relation Identification. 2007. In *Proceedings of the SemEval-2007*, Prague, June 2007, pages 75-80.