CRF based Approach for Temporal Information Recognition from English Text Documents

Parul Patel
M.SC(I.T) Programme
VNSGU, Surat,
India

S. V. Patel, Ph.D
Department of Computer Science
VNSGU, Surat
India

ABSTRACT

Temporal expressions are very important structure in a natural language. In order to use it in information retrieval, it needs to be extracted and normalized into its absolute value. In this paper we have presented a novel approach for temporal information extraction system from English documents. We have used CRF based classifier for extraction of temporal expression. System is evaluated it on Wiki war corpus and manually annotated documents. It achieves Precision of 91.3% , recall of 98.13% on data set 1 and 88.09% and 95.14% on data set 2 for detection of temporal expression.

General Terms

Information Extraction, Natural Language Processing

Keywords

temporal expression, temporal information extraction, Conditional random fields

1. INTRODUCTION

The problem of temporal information extraction refers to identification of the temporal expression from free format English document. Temporal expression can be very useful in the applications like question answering, search exploration, text summarization etc. Temporal expressions can be classified into following categories according to Schilder and Habel[1].

Explicit: Expressions denoting date such as '13/08/2013','15th August' refer explicitly to entries of a calendar system and can be mapped directly to temporal Chronons in a timeline.

Implicit: All temporal expressions that can be evaluated via a given time ontology and capability of the named entity extraction approach such as name of holiday(last chritmas), next diwali etc.

Relative : Some temporal expressions express vague temporal information and it is rather difficult to precisely place the information expressed on a time line. Such temporal expressions can be only anchored in a timeline in reference to another explicit or implicit already anchored temporal expressions. For example, 'on Monday','Before June and After March' etc. If the document has creation date, then they can be easily anchored. This date then can be used as a reference for that expression which can be then mapped to a chronon.

Various temporal taggers are developed to extract and normalize temporal expressions from the document. They have focused on Explicit and some implicit temporal expression, but in Indian documents, we may have implicit temporal expression like last diwali, next holi. We have developed novel approach to temporal tagging by using conditional random field(CRF) model which extract implicit, explicit and relative temporal expressions including Indian holidays from text documents .

2. RELATED WORK

Over past few years, several NLP tools have been used for temporal expression extraction. A lot of research in the area of temporal information extraction has been conducted on multiple languages, including English and several European languages. A large number of systems that extract temporal expressions were developed in the scope of the ACE Temporal Expression Recognition and Normalizationg (TERN), in which TIMEX2 tags are associated with temporal expressions. There are few differences between TimeML TIMEX3 and TERN TIMEX2, notably TIMEX2 includes post-modifiers (prepositional phrases and dependent clauses) but TIMEX3 doesn't. But to a large extent TIMEX3 is based on TIMEX2. Boguraev and Ando [8] and Kolomiyets and Moens [9] reported performance on recognition of temporal expressions using TimeBank as an annotated corpus. Boguraev and Ando's work is based on a cascaded finite-state grammar (500 stages and 15000 transitions) and Kolomiyets and Moens first filter certain phrase types and grammatical categories as candidates for temporal expressions and then apply Maximum Entropy classifiers. Ahn et al. [10], Hachioglu et al. [11] and Poveda et al. [12] used approaches with a token-by-token classification for temporal expressions represented by a B-I-O encoding with lexical and syntactic features and tested on the TERN dataset. Recently, Strotgen et al. [13] used a rule based technique for recognition and normalization of temporal expressions in TempEval-2. We have developed a novel machine learning based approach to recognize temporal expression.

3. METHODOLOGY

Following steps are included into our methodology.
1. Defining temporal expressions to be extracted.
2. Training and Testing Data Preparation
3. Using Standard natural language processing techniques to develop temporal expression extraction system.
4. Testing model with test data to the system & evaluation of the final result.

3.1 Temporal Expression Extraction definition

Our goal is to develop accurate and comprehensive temporal expression system which not only extract explicit temporal expression(like date, month , year etc.), but it extracts all implicit temporal expressions like(last Christmas, last valentines day etc.) and relative temporal expressions like (on Monday, Before June etc.)

3.2 Data Preparation

To prepare training data, WikiWar data set has been uased which contains 22 Xml historical documents containing different type of temporal expressions. Wikiwar data set contains total of almost 1,20,000 tokens and 2671 temporal expressions annotated in TIMEX2 format [7]. We used 17 documents from the whole collection as the training data set and 5 documents as the testing data set. Then we have performed data cleaning on all 22 wikiwar documents. All the Xml tags are removed except the temporal tags.

3.3 Temporal Expression Extraction System Architecture

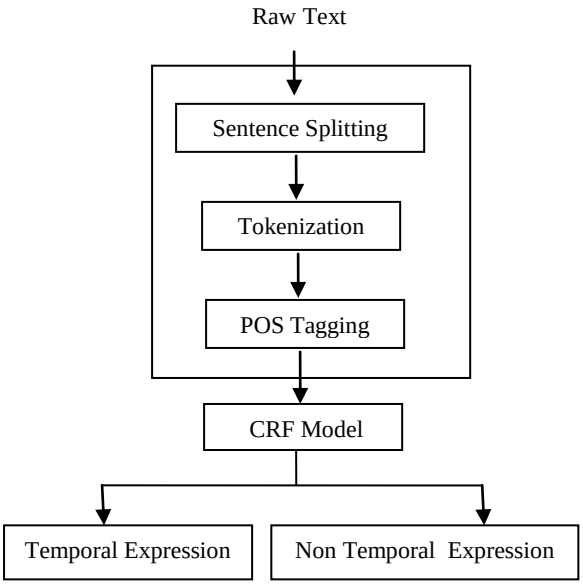

Figure 1: System Architecture

3.3.1 Sentence Splitting

We have used Stanford sentence splitter to split all sentence of all document.

3.3.2 Tokenization

Each sentence is then split into tokens with their respective position in the sentence by using Stanford tokenizer.

3.3.3 Pos Tagging

We have applied Stanford POS tagger to extract all part of speech (POS) features of each token .

3.3.4 Selection of Suitable machine learning algorithm

A machine learning technique that has recently been introduced to tackle the problem of labeling and segmenting sequence data is Conditional Random Fields . Unlike Hidden Markov Models [3], CRFs are based on exponential models in which probabilities are computed based on the values of a set of features induced from both the observation and label sequences. This enables the incorporation of overlapping and interacting features into the model. CRFs have been shown to perform well in a number of natural language processing applications, such as POS tagging [3], shallow parsing or noun chunking [5], and named entity recognition [4]. Their characteristics make CRFs ideally suited for the specific task of recognizing timexes as they provide us with a framework for combining evidence from different sources to maximize performance. W. Cohen used the implementation of CRFs from the minor- Third toolkit for extracting timexes from text [2].

3.3.5 CRF Feature builder

We have extracted Months, days and year features from the tokens. These tokens are helpful in identifying the temporal expressions. This way each token will have two features extracted namely a) Calendar features that involves months, days and year and b) POS features. We have considered following temporal expressions :

1. A List of periodic temporal set: Hourly, Daily, Weekly, Monthly, Yearly

2. A List of Seasons: Spring, Winter, Monsoon, Summer etc.

3. A list of relative days: Yesterday, Today, Tomorrow etc

4. A list of all Indian festivals occurring on fixed days: Independence Day, Republic Day, Teacher's Day, Gandhi Jayanti etc.

5. A list of all Indian Festivals occurring on variable days: Diwali, Holi, Navratri, Women's Day, Rakhi, Durgashtami etc.

6. A list of months: January, February...December.

7. A list of temporal expression modifier: Last, This, Mid, Recent, Earlier, Beginning, Late

8. A list of decades: twenties, thirties etc.

9. A list of Week Days: Monday, Tuesday……...Sunday etc.

3.3.6 CRF model building and classification

The CRF feature builder has generated features for the CRF machine learner. The context window for the CRF was set to be five words. Each pair of temporal expressions at the unigram sentence level is used for training. We have used crf++ 0.58, an open source implementation of the conditional Random Field (CRF) machine learning classifier for our experiments. CRF++ templates have been used to capture the relation between the different features in a sequence to identify temporal expressions. Template file containts unigram template which defined number of tokens to be considered for prediction of token as temporal expression. The training set prepared is given as the input to the CRF algorithm. The CRF algorithm learns from the training samples and gives a model. The training set contains a) token, b) its calendar feature, c) POS tag and the d) label which is either temporal expression or non-temporal expression.

Input		
He	OTH	PRP
had	OTH	VBD
handled	OTH	VBN
distribution	OTH	NN
of	OTH	IN
cheques	OTH	NNS
worth	OTH	JJ
crores	OTH	NNS
during	OTH	IN
the	OTH	DT
Khoraj	OTH	NNP
land	OTH	NN
acquisition	OTH	NN
drive	OTH	NN
in	OTH	IN
Sanand	OTH	NNP
in	OTH	IN
December	CAL	NNP
last	CAL	JJ
year	OTH	NN
.	OTH	.

Output			
He	OTH	PRP	nonTemporalExp
had	OTH	VBD	nonTemporalExp
handled	OTH	VBN	nonTemporalExp
distribution	OTH	NN	nonTemporalExp
of	OTH	IN	nonTemporalExp
cheques	OTH	NNS	nonTemporalExp
worth	OTH	JJ	nonTemporalExp
crores	OTH	NNS	nonTemporalExp
during	OTH	IN	nonTemporalExp
the	OTH	DT	nonTemporalExp
Khoraj	OTH	NNP	nonTemporalExp
land	OTH	NN	nonTemporalExp
acquisition	OTH	NN	nonTemporalExp
drive	OTH	NN	nonTemporalExp
in	OTH	IN	nonTemporalExp
Sanand	OTH	NNP	nonTemporalExp
in	OTH	IN	nonTemporalExp
December	CAL	NNP	**temporalExp**
last	CAL	JJ	**temporalExp**
year	OTH	NN	**temporalExp**
.	OTH	.	nonTemporalExp

4. RESULTS & EVALUATION

The token level features (calendar feature and POS tags) are obtained from the testing set. When the model is run on this set we obtain temporal or non-temporal labels for each of the token. The tokens corresponding to temporal labels are the temporal expressions. In this we have considered recognition problem as classification problem of sentence constituents.

For the recognition of time expressions, precision, recall and F1-score have been used as evaluation metrics, using the following formulae:

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 keys and response, fp is the number of token 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.

We have tested model on two set of documents. DataSet1 is Wikiwar corpus which is a collection of historical data and contains rich set of temporal expressions. Dataset 2 is a is a collection of 15 news articles. Dataset 2 is chosen for testing because it contains Indian festival as temporal expression which was not there in dataset 1.

1. From wikiwar corpus, 7 documents has been taken as test data.

2. We have taken 15 manually annotated 15 english news document which are rich of temporal expressions. This documents contains maximum relative and Implicit Temporal Expressions. DataSet 1 is rich of implicit and explicit temporal expressions. So for rigorous testing of all type of temporal expressions, we have selected two different data set. After Testing we have got following result shown in Table 1. Figure 2 shows comparison of results of two datasets. After analyzing result of two datasets, we came to know that model gives good accuracy for both iexplicit and relative temporal expression .collection of 15 news articles. Dataset 2 is chosen for testing because it contains Indian festival as temporal expression which was not there in dataset 1.

Table 1. Results

Data	Precision	Recall	F-Measure
DataSet 1	91.3%	98.13%	94.59%
DataSet 2	88.09%	95.14%	91.47%

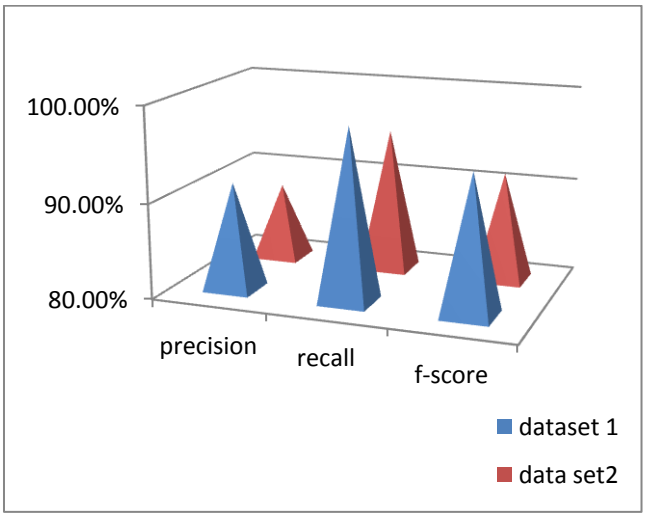

Figure 2: Comparing Result of two datasets

5. CONCLUSION & FUTURE WORK

Rule based system is good enough in recognizing temporal expressions but it demands lots of time in developing large amount of handcrafted rules. Our Experiments suggests that even machine learning approach can lead to a better result if we have good annotated corpus. For retrieving temporal

expressions like Indian festivals, special days celebrated in india (e.g. birthday's of freedom fighters) we still need good annotated corpus for Indian news documents which includes above mentioned temporal expression. In future, such corpus can be developed and normalization of such implicit temporal expression into absolute value is also challenging work.

6. REFERENCES

[1] Frank Schilder and Christopher Habel: 'From Temporal Expression to Temporal Information :Semantic tagging of News Messages' : In Proceeding of the ACL2001 Workshopon Temporal and Spatial Information Processing, 2001.

[2] W. Cohen. Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data, 2004.

[3] J. Lafferty, F. Pereira, and A. McCallum. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the International Conference on Machine Learning, 2001.

[4] A. McCallum and W. Li. Early results for Named Entity Recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the 7th CoNLL, 2003.

[5] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In Proceedings of Human Language Technology-NAACL, 2003.

[6] D. Jurafsky and J. Martin. Speech and Language Processing. Prentice-Hall, 2000.

[7] WikiWars: A New Corpus for Research on Temporal Expressions: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing MIT, Massachusetts, USA, 9-11 October 2010. c 2010 Association for Computational Linguistics

[8] B. Boguraev and R. K. Ando, "TimeBank-Driven TimeML analysis," presented at the Annotating, Extracting and Reasoning about Time and Events, Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2005.

[9] O. Kolomiyets and M.-F. Moens, "Meeting TempEval-2: Shallow Approach for Temporal Tagger," presented at the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, 2009.

[10] D. Ahn, *et al.*, "Extracting Temporal Information from Open Domain Text: A Comparative Exploration," *Digital Information Management*, 2005.

[11] K. Hachioglu, *et al.*, "Automatic Time Expression Labeling for English and Chinese Text," presented at the CICLing, 2005.

[12] J. Poveda, *et al.*, "A Comparison of Statistical and Rule-Induction Learners for Automatic Tagging of Time Expressions in English," presented at the International Symposium on Temporal Representation and Reasoning, 2007.

[13] J. Strotgen and M. Gertz, "HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions," presented at the International Workshop on Semantic Evaluations (SemEval-2010), Association for Computational Linguistics (ACL), 2010.