ISSN 2319 – 7595
Volume 5, No.2, March - April 2016
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs01522016.pdf

Comparing Two Novel Machine Learning Approaches For Temporal Information Extraction

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

ABSTRACT: Temporal Expression in the document is an important structures in any natural language text document. Temporal information processing is challenging research area in recent years. Temporal information processing is a fundamental task to be done for applications like question answering, temporal information retrieval, text summarization, and exploring search results in timeline manner. In this paper author have compared two novel machine learning approaches, Conditional Random Fields and Hidden Markov Model for temporal information extraction. The author have achieved average precision, recall and f-measure 92.5% , 96.19% 94.28% , 89.16%, 91.46%, 90.19%, in CRF and HMM respectively.

Keywords: Temporal Expression, CRF, HMM

1. INTRODUCTION

Temporal information recognition from English text documents has been an emerging research area since past few years. In the applications like generating timeline, it is important to interpret temporal aspects of temporal expression present in the document.
Temporal knowledge helps to filter information and flow of temporal events in the document, which is essential in applications like text summarization and time line generation. For realization of such applications, it is important to identify temporal expressions correctly. Expressions denoting time comes into a wide variety of forms ([1]):

- Fully-specified time references: 15th July 2008, the twentieth century, Monday at 5am.
- Expressions whose reference depends on the context: five days after the meeting, the next month, April the following year.
- Anaphoric expressions and expressions relative to the time when the expression is written: that day, yesterday, recently, then.
- Durations or intervals: a month, three days, some hours in the afternoon.
- Frequencies or recurring times: monthly, every day, twice a week, every first Monday of a month.
  Culturally dependent time denominations: Easter, the month of Ramadan, St. Valentine.
- Fuzzy or vaguely specified time references: the future, some day, eventually, anytime.

Processing of above temporal expression is a part of task known as temporal expression recognition and normalization(TERN), in which the aim is to recognize and identifying the meaning of temporal expressions in text and represent into intermediate standard form into XML tags with their attributes for further processing. For example:

Narendra Modi visited Gandhi Ashram on this republic day.
Narendra Modi visited Gandhi Ashram on < TIMEX3 tid="t1" type="DATE" value="2014-01-26" > this republic day < /TIMEX3 >.
TIMEX is a standard which defines guidelines for annotation of temporal expressions. Complete description of timexes can be found in [2]. TERN is very challenging task because some temporal expressions(implicit) are straight forward to recognize and normalize(e.g. 24/3/2010) whereas some are context dependent (e.g next friday, Last Christmas.) So finding appropriate value for the temporal expressions require analysis of surrounding task. Some temporal expressions are ambiguous for e. g(I may come late/ I will come in May). In 1st sentence , May is a verb which is not a temporal expression whereas in a 2nd sentence, May refers to a noun which is a temporal expressions. Traditional methods for extracting temporal information are rule based and heuristics of rules are hand tuned by experts. Such system gives good accuracy but large amount of hand crafted rules makes system context dependent not generalizable and robust. While machine learning methods are widely used in recognition task. With the advent of new linguistic corpora, supervised machine learning approaches becomes practically possible for many natural language processing. In this paper, the author have compared three novel supervised machine learning based approach, Conditional Random Field and Hidden Markov Model and Maximum Entropy Model.

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. The idea of annotating temporal expressions automatically from textual documents has appeared for the first time in MUC[3]. The importance of the proper treatment of timexes is reflected by the relatively large number of NLP evaluation efforts where they play role . Extraction of temporal expressions based on a cascaded finite state grammar is reported in [4]. Ahn D and Adafre, S.F. De Rijke has developed successful machine learning

based system as per TIMEX2 annotation guideline . They have used token by token classification of temporal expressions represented by B-I-O encoding with a set of lexical and syntactic features. e.g part of speech tag, token itself, numeric years and weekdays name etc. Their performances are shown in table 1.

Approach	Precision	Recall	F1(detection)
David Ahn. et. al.[4]	0.980	0.842	0.906
Poveda et.al.[5]	0.798	0.726	0.757
Oleksandr et.al [6]	0.85	0.84	0.845

Many systems that extract temporal expressions were developed in the scope of the ACE Temporal Expression Recognition and Normalization (TERN), in which TIMEX2 tags were associated with temporal expressions. TIMEX2 includes post-modifiers (prepositional phrases and dependent clauses). Boguraev and Ando [5] and Kolomiyets and Moens [6] reported performance on recognition of temporal expressions using TimeBank as an annotated corpus. Boguraev and Ando's[5] work is based on a cascaded finite-state grammar (500 stages and 15000 transitions) and Kolomiyets and Moens[6] first filter certain phrase types and grammatical categories as candidates for temporal expressions and then apply Maximum Entropy classifiers. Ahn et al. [7], Hachioglu et al. [8] and Poveda et al. [9] 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. [10] have developed a rule based technique for recognition and normalization of temporal expressions in TempEval-2.

From literature review it is observed that , not only rule based approach, but machine learning algorithms can also gives good results. Most of the machine learning algorithm stated above give good results on explicit and relative temporal expressions, but some implicit temporal expressions are difficult to identify like 'last ' or 'last diwali' which specifically are present in Indian news document. The author have developed a novel machine learning based approach to recognize temporal expressions based on context information around text.

3. RESEARCH METHODOLOGY

In this section the author have compared two different supervised machine learning approaches(CRF & HMM) for proper extraction of temporal expression from the text. Research Methodology includes following steps:
(1) Defining scope of Temporal expressions to be extracted
(2) Selection of training data and testing data from the available corpus
(3) Creating models by using two algorithms CRF & HMM
(4) Testing model and comparing results

3.1 Temporal Expression Extraction Definition

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

3.2 Training and Testing Data Preparation

The author have used WikiWar data set 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 [11]. The author used 17 documents from the whole collection as the training data set and 5 documents as the testing data set. Wikiwar dataset doesn't contain any Indian festivals. Our main approach is to extract temporal expressions from Indian news document where possibilities of such temporal expressions are more. So the author have used 15 manually annotated news documents along with wikiwar dataset for training. So that model can incorporate Indian festival feature during training. So first the author have performed data cleaning operation on all training documents. Data cleaning includes removing all xml tagsexcept temporal tags starting with <TIMEX> from documents. Then preprocessing step is applied on the documents. It includes sentence splitting, tokenization and POS tagging. For preprocessing the author have extensively used Stanford tools. For example, Sentence splitter is used to split all sentences of all documents. Then each sentence is split into tokens with their respective position in the sentence by using Stanford tokenizer. Stanford POS tagger is applied on each token to extract all its part of speech features of each token.

3.3 Selecting features and Creating models of CRF and HMM

3.3.1 Selecting feature set

The author 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. The author 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.

3.3.2 Introduction to Supervised machine learning algorithm

The current trend in information extraction is to use machine learning algorithms which are more adaptive and trainable than traditional rule based approach. Several classification methods have been applied in information extraction. Zhou and Su have used Hidden Markov model with large variety of features for Named entity extraction [12]. McCallum and Li[13] have used CRF for named entity recognition. The author have used two machine learning based algorithm to extract temporal expressions. (a) Conditional Random Field. (b) Hidden Markov Model. The author have used same feature set and training data for above two algorithms and tried to compare results.

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 [14], shallow parsing or noun chunking [15], and named entity recognition [16]. 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 [17].

Information extraction task depends on powerful modeling of context and current observations. Finite state machines like hidden Markov models, is a probabilistic finite state machine and tool to model sequences of observations. They have been applied in many natural language processing task such as part of speech tagging, Named entity recognition and other information extraction tasks. Hidden markov model is widely used in speech recognition. The author have used MALLET tool for using HMM.

4. RESULTS

	Precision	Recall	F-Score
Conditional Random Field			
WikiWar	91.3%	98.13%	94.59%
Aquaint	94.5%	97.3%	95.87%
TempEval	96.11%	94.21%	95.19%
15 news articles	88.09%	95.14%	91.47%
Hidden Markov Model			
Wikiwar	89.17%	92.12%	90.62%
TempEval	91.22%	90.32%	90.76%
Aquaint	92.16%	91.27%	91.47%
15 news articles	84.09%	92.14%	87.93%

From above results, the author have analyzed that both algorithms works well for same corpus but CRF is giving better result than HMM in extracting information from documents.

5. CONCLUSION

The author have used two novel approaches for temporal expression extraction. The author have compared and analyzed results of CRF and HMM. HMM is widely used in speech recognition and gives good accuracy. CRF is a good classifier for textual information extraction related activities. HMM is also giving better results. Our comparison shows that contrast to rule based approach, machine learning approaches are also giving good accuracy for information extraction.

6. REFERENCES

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

[2] G. Wilson, I. Mani, B. Sundheim, and L. Ferro. A multilingual approach to annotating and extracting temporal information. In Proc. Workshop on Temporal and Spatial Information Processing Vol. 13, pages 1–7, Morristown, NJ, USA, 2001. ACL.

[3] Mani I. and Wilson, G. Robust Temporal Processing of News. In proceedings of 38th Annual Meeting on Association for Computational Linguistics(Hong Kong 2000), 69-76.

[4] Boguraev, B. Ando, R.K: "TimeBank Driven TimeML Analysis in annotating extracting and reasoning about time and events". Dagstuhl Seminar Proceedings, Dagshtul, Germany(2005).

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] G. Zhou and J.Su . Named Entity recognition using an HMM chunk tagger. In proceeding of the 40th Annual Meeting of the Association for Computational Linguistics – 2002.

[13] A. McCallum and W.Li. Early results of named entity recognition with conditional random fields, feature induction and web enhanced lexicons. In proceeding of the 7th conference on natural language Learning, 2003.

[14] 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.

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

[16] 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.

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