A RECURRENT NEURAL MODEL FOR TEMPORAL INFORMATION EXTRACTION

Parul Patel

Department of ICT, Veer Narmad South Gujarat University,
Surat, Gujarat, India

ABSTRACT

Temporal information extraction is an emerging area of information retrieval. Understanding temporal nature of a document is very important in application like answering time sensitive queries, doing temporal analysis of a document, document clustering etc. Lot of research is done in temporal reasoning using rule based or machine learning based approaches. In this paper, deep learning is used to extract temporal expressions from the text documents. Bi-directional Long Short term Memory Recurrent Neural Network (Bi-LSTM RNN) is used to extract temporal expression from the text. Gold standard datasets are used for training and evaluation. Performance of the proposed system is compared with existing system.

KEYWORDS

Long Short Term Memory, temporal expression, TIMEX

1. INTRODUCTION

As the amount of digitized information increases, time plays important role in getting relevant information, documents or organizing document on a timeline. Time embedded in a document can be useful in answering temporal questions like "When did last olympic held?" . Information extraction is the process of extracting specific information like entities, relations, events, objects from a large volume of unstructured data. Temporal information extraction is very promising research area nowadays as it plays very fundamental role in applications like timeline generations, document classifications, temporal query processing, temporal information retrieval etc. Nowadays machine learning algorithms are more popular in performing such task. Some deep learning algorithms such as recurrent neural network, convolutional neural network, bi directional LSTM are used for temporal information extraction. In this paper, a deep learning model is proposed based on Long Short Term Memory (LSTM) recurrent Neural network (RNN) for extracting temporal expressions.

Rest of the paper is organized as follows: Literature review of the temporal information extraction task is listed in Section 2, Section 3 describes proposed model, evaluation and results are included in section 4 and section 5 contains conclusion.

2. LITERATURE REVIEW

Over the last few years, 'temporality' has drawn a significant attention to the community of Natural Language Processing (NLP) and Information Retrieval (IR). Time is an intrinsic property that aids in ordering events in a sequential order from the past to present to future. This ordering of events is very crucial in analyzing a document. Recently, Recurrent Neural Network have shown impressive result in the area of NLP. Recurrent Neural Network are very widely used for

computer vision, Natural Language Processing, Text mining etc. It is suitable to text classification because it is able to capture local features and different combinations of such features. Conventionally, for temporal expression extraction, features are extracted from dataset and then fed to a learning algorithm for classification of a text. For automatic feature extraction, Neural Network process text as much as possible and pass it to different layers of network that process the inputs. Recently, there have been many attempts to extract temporal expressions with neural network using convolutional neural network and recurrent neural network has been done. Every document contain temporal expression, lot of research has been done in the area of temporal information extraction. A system for recognizing temporal expression and identifying relation between events have been proposed in [1, 2] . [3] proposed a method in which finite state automata is used for extracting temporal expression. Some expressions were filtered using existing rules.[4] has used a CRF classifier to extract temporal expression. HeidelTime [5] is a temporal tagger with high accuracy designed for temporal information processing. SUTime [6] is a rule based system developed by Stanford University for temporal information extraction and interpretation. INDTime [7] is a rule based temporal tagger that focus on relative temporal expression like diwali, independence day, republic day etc. Hachioglu et al. [8] adopted a left to-right, token-by-token, discriminating, deterministic classification scheme to determine the tags (for each token. Poveda et al. [9] have compared statistical (support vector machines) and one of rule induction (FOIL). Their analysis shows that SVM are superior than FOIL. Yuanliang Meng [10] et.al has proposed context aware neural model for temporal information extraction by using convolutional neural network.

3. RESEARCH METHODOLOGY

Temporal information extraction problem is formalized as a binary classification problem. For each token in a sentence, it is classified as a temporal or a temporal expressions. It is a very challenging task because of the variety of way in which time can be expressed. Moreover, sequence of a token play very important role in deciding whether a given expression is temporal or not. For example, expression ‘today’ or ‘tomorrow’ do not need require any extra information regarding surrounding text to get recognized as temporal expression. It can be easily classified as temporal expression, but expression like ‘this’ requires details of surrounding text for classification as next token can decide whether it is a temporal expression or not. For eg. ‘this’ ’year’ and ‘this’ ’table’. As it is sequence labelling task, bi directional long short term memory Recurrent Neural Network (BILSTM-RNN) is used for this classification problem. Various TIMEX annotated corpus are available like TempEval, Wikiwar, Aquaint, IndiaTimes, TempEval-2, TempEval-3 etc. In this research, 3 datasets Wikiwar, India Times, and TempEval-2 are combined to get variety of temporal expression available in them. Wikiwar contains all historical events, India Times covers news articles covering 1l different categories like sports, culture, biographies etc. One another reason to select this corpus it that it covers not only national holidays but all festivals that are celebrated in various parts of India like Diwali, Holi, Onam etc.

Linguistic Processing: In this phase, the documents are passed through linguistic processing pipeline that involves tokenization, sentence splitting, Part of Speech (POS) tagger and a syntactic parser. The tokenizer divides input text into tokens like words, numbers, punctuation marks etc. These tokens are useful in identifying locations of sentence limits which can be helpful in POS tagging and syntactic analysis. All tokens with their POS tagging and are they Timex or not are stored in a file prepared for training.

Original Text

Document will explain marketing schemes carried out by Indian companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for

agrochemicals, pesticide, herbicide, fungicide, insecticide fertilizer, predicted sales, market share, encourage demand, price out, volume of sales.

Output of POS tagger

Document/NN will/MD explain/VB marketing/NN schemes/NNS carried/VBD out/IN by/NN Indian/NNP companies/NNS for/IN their/PRP agricultural/JJ chemicals/NNS, report/NN predictions/NNS chemicals/NNS,/, agrochemicals/NNS, or/CC for/IN report/Nn pesticide/NN, market/NNS share/NN market/NN herbicide/NN, of/IN such/JJ statistics/NNS for/IN fungicide/NN, insecticide/NN fertilizer/NN, predicted/VBN sales/NNS,/ ,market/NN share/NN,/, encourage /VB demand/NN,/, price/NN out/NN ,/, volume/NN of/IN sales/NNS ./.`

All documents are pre-processed and all tokens along with their POS tagging and a label indicating whether they are time x or not are assigned to it. Entire dataset is split into training data (80 %) and testing data (20%). In training data, 20% data is selected for validation testing during training of model.

Word embedding is used to represent same words by similar value distribution. It gives dense representation of words represents word in to a vector.

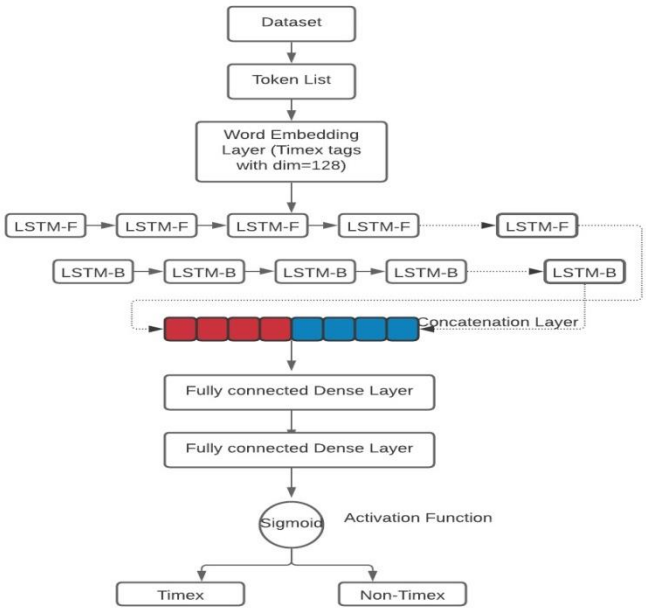

Figure 1. Proposed System Architecture

It takes all words as input and convert them into a vector. In this model, 128 dimensional vector is used for embedding layer where each word is represented into a 128 dimensional space. Initially weights for embedding layer are randomly initialized and they are gradually adjusted by using back propagation technique. Next layer is LSTM RNN layer. Regular LSTM model read sentences only in forward direction. Regular machine learning models allows to keep the window size (-m ,+m) in both direction of word. To implement the same, bi directional LSRM is used in which one LSTM can read in forward direction where as another LSTM works in backward direction from the current token. In proposed architecture, LSTM-B and LSTM-F are used for backward and forward directions respectively. Output of embedding layer as vector is applied to

Bi-LSTM layer that has used tanh as activation function for mapping words with its corresponding label. Finally output of backward and forward LSTM is concatenated at concatenation layer. Next layer is a dense layer, that uses relu as an activation function .The last layer is a dense layer that uses soft max activation function to decrease the dimension of the vector and generates a final output. Finally, the sigmoid activation function is applied at final dense layer as it is binary classification problem. The final output is a value between 0 & 1. where 0 indicates non time x and 1 indicates time x expressions.

4. EVALUATION & RESULTS

The proposed system is designed for temporal information extraction. We have achieved the best result during training by selecting parameters (Optimizer = Adam, learning rate = 0.0001, loss function= binary cross entropy, batch size=32, No. Of epochs=14). Table 1 shows results of the proposed system on different standard data sets. Table 1 shows comparison of proposed system with existing system.

Table 1. Results on different data set

Dataset	Precision	Recall	F-measure
TempEval-2	0.94	0.91	0.92
Aquaint	0.88	0.90	0.88
IndiaTimes	0.89	0.85	0.86
TempEval-3	0.91	0.90	0.90
Wikiwar	0.92	0.91	0.91

Table 2. Comparison of proposed system with existing system

System	Precision	Recall	F-Score
SUTime	0.88	0.96	0.94
HidelTime	0.90	0.82	0.85
KUL-1	0.78	0.82	0.80
KUL-2	0.75	0.85	0.79
KUL-3	0.85	0.84	0.84
HeidelTime-1	0.90	0.82	0.86
HeidelTime-2	0.82	0.91	0.86
TRIPS/TRIOS	0.85	0.85	0.85
Proposed System	0.96	0.94	0.89

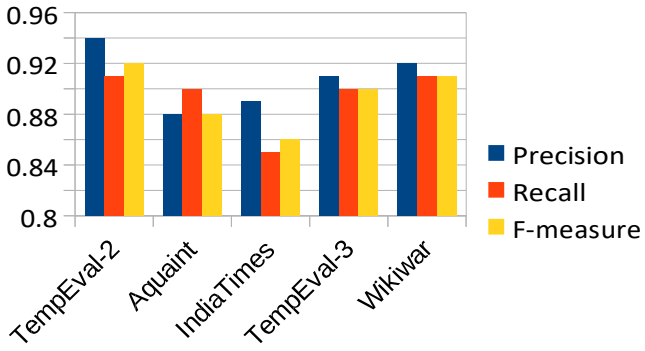

Figure 2. Comparison of result of proposed system on different gold standard dataset

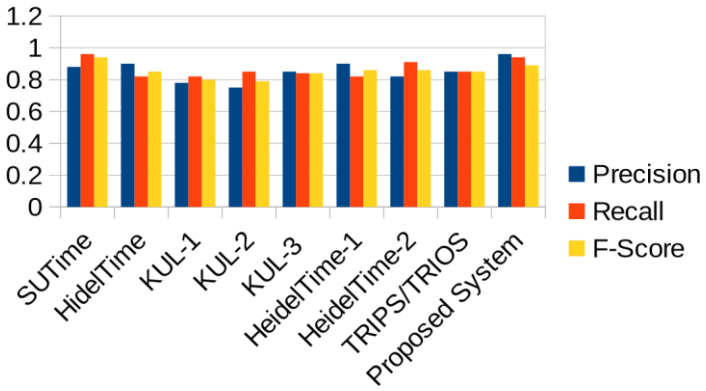

Figure 3. Comparison of proposed system with existing system

5. CONCLUSION & FUTURE WORK

In this paper, a deep learning neural network model is used to extract temporal information present in the English text document. The system has been trained by using existing corpora and performance of the model is compared with existing system and on other available corpus. In this paper, primary focus is on English text document only, but using this model for other languages may introduce new challenges in this domain. In future, I would like to do work for other languages like Hindi, Gujarati etc. Moreover, temporal expression interpretation using deep learning can also be a future research direction.

REFERENCES

[1] I. Mani, B. Schiffman, and J. Zhang, "Inferring temporal ordering of events in news," in Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, Canada, 2003, pp. 55-57.

[2] I. Mani and G. Wilson, "Robust temporal processing of news," In Proceedings of the 38th Annual Meetingof the Association for Computational Linguistics, Hong Kong, China, 2000, pp. 69-76.

[3] P. Kim and S. H. Myaeng, "Usefulness of temporal information automatically extracted from news articles for topic tracking," ACM Transactions on Asian Language Information Processing (TALIP), vol. 3, no. 4, pp. 227-242, 2004.

[4] Parul Patel and S V Patel. Article: CRF based Approach for Temporal Information Recognition from English Text Documents. *IJCA Proceedings on International Conference and Workshop on Emerging Trends in Technology* ICWET 2015(1):1-4, May 2015.

[5] Strötgen, J., & Gertz, M. (2010). HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010* (pp. 321–324). Uppsala, Sweden, 15–16 July 2010.

[6] Chang, A. X., & Manning, C. D.: SUTime: A library for recognizing and normalizing time expressions. In *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*.

[7] Patel P., Patel S.V. (2016) INDTime: Temporal Tagger—First Step Toward Temporal Information Retrieval. In: Satapathy S., Joshi A., Modi N., Pathak N. (eds) Proceedings of International Conference on ICT for Sustainable Development. Advances in Intelligent Systems and Computing, vol 408. Springer, Singapore.

[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] Yuanliang Meng, Anna Rumshisky "Context aware Neural Model for Temporal Information Extraction", Proceeding of the 56th Annual Meeting of the Association for Computational Linguistics Pages 527-536, Melbourne, Australia, July 15-20, 2018.

AUTHOR

Dr. Parul Patel is an Assistant Professor in Department of ICT in Veer Narmad South Gujarat University, Surat. She has 15 years of experience in teaching. She has published more than 10 research paper in the information retrieval area.