International Journal of u- and e- Service, Science and Technology
Vol.10, No.9 (2017), pp.101-108
http://dx.doi.org/10.14257/ijunesst.2017.10.9.10

Sequence Labeling using Conditional Random Fields

Romansha Chopra, Nivedita Singh, Yang Zhenning and N.Ch.S.N.Iyengar

School of Computer Engineering, VIT University, Vellore-632014, India
romansha31@gmail.com,nivedita.singh2016@vitstudent.ac.in

Abstract

The aim of this paper is to get some experience with sequence labeling, specifically, assigning tags or labels to each member in the sequences of utterances in conversations from a corpus. Since nowadays predicting single class label or tag is not adequate. Predicting large number of variables that depends on each other is required. In sequence labeling it is often beneficial to optimize the tags assigned to the sequence as a whole rather than treating each tag decision separately. A machine learning technique termed as Conditional Random Fields, which is designed for sequence labeling will be used in order to take advantage of the surrounding context. Conditional random fields (CRFs), is a scheme for building probabilistic models to divide and tag sequence data. With a given a labeled set of data, baseline set of features will be created and the accuracy of the CRF suite model created using those features will be measured.

Keywords*: Sequence labeling, Conditional Random Fields, Markov model*

1. Introduction

There are wide scientific fields where we need to label and divide the text. To target this problem, Hidden Markov and other stochastic models are globally used. These models are generative in nature and use joint probability to increase the joint likelihood of training examples .To define a joint probability on observations and label sequences, a generative model list for every possible sequences, specifically requiring observations such as words. The conditional probability is used for labeling the sequences which can be dependent on arbitrary, non-independent features of the observation sequences. The features which are observed contain attributes are under the different level of details of the same observation. The generative model considers this independent feature, but in real life data all the features are not independent they are related to past and future sequences, if available. Therefore Maximum entropy Markov models (MEMMs) is conditional probabilistic model which has all the above stated advantages. MEMMs have an exponential structure which considers observation as an input and gives output distributed for next possible states. An appropriate iterative scaling method in the maximum entropy framework is used to train these exponential models. But all these models have label problem, which states that the transition compete only against each other, rather than the other transition states in the model. So, to solve all the above problems this paper focuses on Conditional Random Fields (CRF) model. This is one of the most important and well used model in machine learning for sequence labeling. It is a modeling approach which has all the benefits of Maximum entropy Markov model (MEMMs). The major distinction between Conditional Random Fields and Maximum entropy Markov model is that CRF has a single exponential model.

Received (May 15, 2017), Review Result (August 21, 2017), Accepted (September 5, 2017)

ISSN: 2005-4246 IJUNESST

1.1 Sequence Labeling

In machine learning sequence labeling is interpreted as placing discrete tags or labels to all the specified features of the given observation. Under this problem, we give input 'a' and tag each component of 'a' with its class to get the output 'b'. There are various applications of sequence labeling to name the few- entity recognition problem, handwriting recognition problems, plagiarism, pattern recognition, speech recognition and most importantly is being widely used in bio medical field for deciding DNA sequence in the given DNA base. DNA is a part of gene and therefore it is a part in the gene production problem. Sequence labeling is a key aspect for analyzing human behavior. It is most important in the field of human sciences. Anyone can print a sequence of aural features with oral words (speech recognition), or a sequence of video frames with hand gestures (gesture recognition). Yet such tasks arise when visualizing time series, they are spotted in domains with non-temporal sequences, such as protein secondary structure prediction. The sequence labeling is also termed as sequence classification. It has also broad level of importance in information retrieval, health informatics and abnormal detection and many more. In genomic research, protein classification in already existing categories to get functions of new proteins. Even after using highly sophisticated features there are various potential features which are not classified properly, therefore sequence classification is a challenging task. Sequence labeling is one of the pattern recognition task used to give the categorical tag to all the possible different features passed in the model. Sequence labeling is a set of independent classification tasks, one tag per member of the sequence. However, accuracy is generally improved by making the optimal label for a given element depending on the choices of nearby elements, using special algorithms to choose the globally best set of labels for the entire sequence at once.

1.2 Conditional Random Fields

Conditional Random Fields or popularly known as CRF, it is a probabilistic model which contains all the components of stochastic model such as Maximum entropy Markov Model(MEMMs) used for labeling and augmenting data structure like sequences, trees. The basic thought behind Conditional Random Fields is to define conditional probability distribution over label sequences in a given observations. The main advantage of Conditional Random Fields over Hidden Markov Model is that its flexibility that results in the relaxation of the independence assumptions required by Hidden Markov Model in order to ensure traceable inference. And above all CRF avoids the label bias problem which was the main concern in the previous models. Another advantage of Conditional Random Field is its convexity of the loss function. All the convexity properties of general maximum entropy models are fulfilled by CRFs. But all these benefits come with the cost like class of Conditional Random Fields is very suggestive, as it permits arbitrary dependencies on the observation sequence and the features need not specify the total state or observation, so that the model can be estimated from less training data. (CRFs) are often applied in pattern recognition and machine learning for a statistical pattern, implemented for structured prediction.

1.3 CRFsuite

For CRF suite pycrfsuite is installed. Pycrfsuite is a Python interface to CRF suite. This toolkit expects training data to be in the following format. A corpus is represented as two lists: a list of features (each element is a list of features), and a list of labels. These features are binary. The presence of a feature indicates that it is true for an item. Absence indicates that the feature would be false. Here are the features for a training example using features for whether a particular token is present or not in an utterance. ['TOKEN_i', 'TOKEN_certainly', 'TOKEN_do', 'TOKEN_.']

2. Literature Survey

Sequence labeling is classified under machine learning pattern recognition technique. It is implemented through various models. Each model has its own advantages and disadvantages. The most used algorithm is Hidden Markov Model (HMMs) **[1]**. The dataset is tagged on the basis of different features taken like part of speech. This model is based on statistical approach .It has various application area like speech recognition**[2]** .Under this, we can find whether the particular speech or utterance is phoneme, or a word, or a sentence. One of the ways to capture the structure in the sequence of symbol is the use of Hidden Markov Model. The other important application is Information Extraction **[3]** tasks. These tasks give us deep knowledge about the model and help us in understanding of the labeled data. Under this, we first understand and analyze the behavior of the structure of the data and then we evaluate the importance of labeled data and distant labeled data which means it is labeled but under different domain. Here the header of research paper is used for extracting the fields which are useful for creating the database of computer science research field. So this is done by labeling each word as author, keyword, title and more. It is implemented through Hidden Markov Model (HMMs). But with this model, there arises different problems associated with it. One of the major problems is in speech recognition **[4]** is that it considers all the utterances as an independent, which is not a practical approach to any problem and especially in speech recognition, dependencies are extended to multiple stages and this type of problem is not considered in Hidden Markov Model. In a typical model, there is a need of huge data and for training also, so it's a drawback if we want to model for small dataset. So to overcome this problem, we have another model known as Conditional Random Fields (CRF) which has various advantages over Hidden Markov Model. It is a probabilistic model for sequence labeling and tagging the unlabelled data **[5]**. The main motivation to use Conditional Random Fields is dependent on arbitrary non independent features of the sequences. The features which are chosen represent attributes at different level of granularity of same observation. In sequence labeling shallow parsing is one of the most important aspects **[6].** It is used to identify the structure of the sentence by specifying noun, verb, and adjective in it but does not tell us anything about its internal structure. Conditional Random Field is known for its flexibility to include all the features and non-arbitrary input. There is one induction method which is used for increasing likelihood by iteratively constructing feature conjunctions. In any data analysis table extraction is one of the major tasks.

These tables contain lot of information in a densely packed form, so to extract information from them is important and should be done in an effective way[7].One of the technique used is Conditional Random Fields, this is done by labeling each line of the document by giving tag which describes its relation with table. This is line tagging and then comes non extraction line tagging, this contains the lines which are either outside the table or consists of punctuation mark or special characters. Then comes the header label, it contains the metadata i.e. information about the heading of the table and tells us what all information we can expect from this table. In this way table extraction can be done through sequence labeling with the help of CRF. Till now we have seen that tagging is done on the basis of one feature but this can be done through multiple ways like part of speech tagging, noun phrase segmentation simultaneously and many more features can be added with the help of dynamic markov network [8]. It represents the large literature body and it also studies the particular class of Hidden Markov Model

It is an undirected graphical model which is conditionally trained, repeated over sequence. Summarizing in DCRFs model is done using approximate method. In view of their factorize state we can utilize DCRFs to do labeling task and sharing data between them. And this model has higher joint accuracy than linear chain CRF model.

3. Research Methodology

The raw data for each utterance in the conversation consists of the speaker name, the tokens and their part of speech tags. Given a labeled set of data, firstly baseline set of features is created, and accuracy of the CRF suite model created using those features is measured. Experiment with additional features is done in an attempt to improve the performance. The best set of features developed will be called the advanced feature set. Then the accuracy of the CRFsuite model created using those features is measured. The last task is assigning dialogue act tags to a set of unlabelled data. This was done with two models. The first model uses the baseline feature set and is trained on all the labeled data. The second model uses the advanced feature set and is trained on all the labeled data.

Finally, set of testing data evaluates classification ability of the model through various evaluation measures (such as F-score, precision, and recall). Methodology can be defined in three phases namely:

Phase 1: Pre-processing of Data and Feature Selection

The labeled data picks the best features for this task. The labeled data is segmented by randomly putting roughly 25% of the data in the development set and using the rest to train the classifier. In the baseline feature set, each utterance includes:

- A feature for whether or not the speaker has changed in comparison with the previous utterance.
- A feature marking the first utterance of the dialogue.
- A feature for every token in the utterance.
- A feature for every part of speech tag in the utterance (e.g., POS_PRP POS_RB POS_VBP POS_.)
- Other set of features that we'll call advanced. The advanced feature set includes more information than the baseline feature set. It improves performance.
- In the advance feature set, each utterance includes:

1. A feature for if the utterance is a question or not which implies there is surely going to be a next utterance.

2. A feature for exclamatory sentence i.e. the utterance contains exclamation or not.

3. A feature to check if the words in an utterance is less than 5 or not.

Phase 2: Predicting tags

In this phase, the python CRFsuite (pycrfsuite) is used to predict tags for utterances. The first step of this phase deals with the training Data. The features selected are passed through CRFsuite. This toolkit expects training data to be in the following format. A corpus is represented as two lists: a list of features (each element is a list of features for an example), and a list of labels. The features are binary. The presence of a feature indicates that it is true for this item. Absence indicates that the feature would be false. Once this is done, our model is produces a set of dialogue act tags for the unlabelled data.

Phase 3: Model Evaluation

After the predicting of tags phase the testing and evaluation of the dataset is done. Evaluation measures (precision, recall and F1 score) are used to measure system performance. These evaluation measures can be mathematically defined as

$$\text{Recall} = \frac{\text{No of relevant documents and retrieved documents}}{\text{Total no relevant documents}}$$

$$\text{Precision}=\frac{\text{No of relevant documents and retrieved documents}}{\text{Total no of retrieved documents}}$$

$$\text{F1 Score} = \frac{2* \text{Precision} *\text{Recall}}{\text{Precision} +\text{Recalls}}$$

3.1 Precision

It is defined by a statistical measure called standard deviation which defines the exactness. Low exactness is indicated by high standard deviations and high exactness is indicated by low standard.

3.2 Recall

It is that the fraction of the documents that are relevant to the question that are with success retrieved in retrieving the information. In general recall is also termed as sensitivity.

3.3 F1 Score

In applied mathematics analysis of binary classification, the F1 score (also F-score or F-measure) could be a live of a test's accuracy. It considers each the exactness ‘p’ and therefore the recall r of the check to compute the score. ‘p’ is that the variety of correct positive results divided by the amount of all positive results, and ‘r’ is that the variety of correct positive results divided by the amount of positive results. The F1 score may be taken as a weighted average of the preciseness and recall, wherever associated degree F1 score reaches its best worth at one and worst at zero.

4. Implementation Details and Discussions

4.1 Experimental Setup

This experiment is performed on i7 Intel processor with 8 GB RAM size. The algorithm is implemented in python, so we have used Ubuntu 14.04 for compiling the python language code and python CRFsuite *i.e.*, pycrfsuite to predict tags for each utterance in a sequence or a dialog.

4.2 Data Description

The Switchboard (SWBD) corpus was collected from volunteers and consists of two person telephone conversations about predetermined topics such as child care. SWBD DAMSL refers to a set of dialogue act annotations made to this data. This (lengthy) annotation manual defines what these dialogue acts mean. Corpus data is divided into labeled and unlabelled (test) data sets. In all data, individual conversations are stored as individual CSV files. These CSV files have four columns and each row represents a single utterance in the conversation. The order of the utterances is the same order in which they were spoken. The columns are:

1) **act_tag** - the dialogue act associated with this utterance. This is blank for the unlabelled data.
2) **Speaker** - the speaker of the utterance (A or B).
3) **POS** - a whitespace-separated list where each item is a token, "/”.

4) **Text** - The transcript of the utterance with some clean-up but mostly unprocessed and untokenized. This column may or may not be a useful source of features when the utterance solely consists of some kind of noise.

4.3 Experimental Results

4.3.1 Output: Tags Predicted

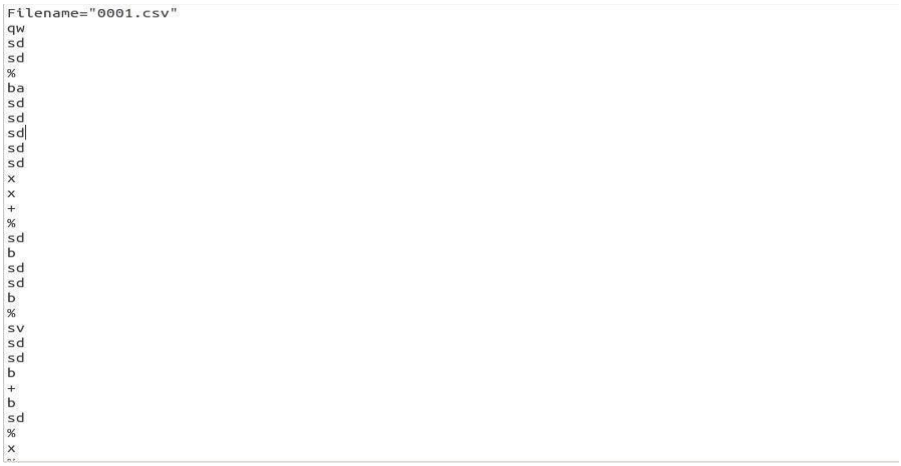

Figure 2. Tags Predicted

4.3.2 Evaluation Measures

Table 1. Evaluation Measures for Baseline Features

Model	Recall	Precision	F-score
	%	%	%
HMM	63.7	60.2	61.9
CRF	61.9	72.4	68.7

Table 2. Evaluation Measure for Baseline Features

CRF Model	Accuracy
Advance Features	72.7

4.3. Comparison of Conditional Random Fields with Hidden Markov Model

Here we compare our method Conditional Random Fields with the traditional Hidden Markov Model (HMM) which is defined as a supervised method. In the implementation we have carried out, we compare HMM-based models with Conditional Random Fields. CRF works efficiently as it combines the merits of HMM. In Table 1 evaluation measures precision, recall and F-score are calculated for both models. From this Table 1 we can infer that predicting tags through Conditional Random Fields provides better results than the traditional Hidden Markov Model. In figure 3 a graphical representation of comparison of both the models has been depicted which shows significant difference in performance of predicting tags through both models. Therefore, an elaborative comparison among these methods can make it descriptive that whether CRF possess these merits in our proposed problem.

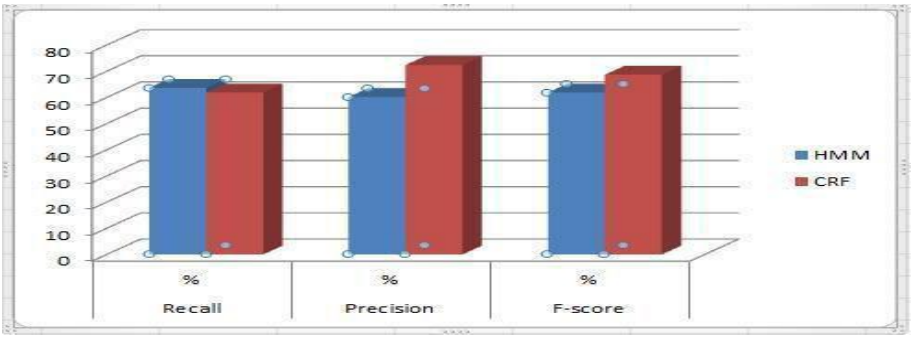

Figure 3. Comparison of Evaluation Measures of Hidden Markov Method and Conditional Random field

5. Conclusion

In this paper we have implemented the sequence labeling through Conditional Random Fields, which is the most suitable for predicting tag for an utterance. There are two major steps for improving classification accuracy: pre-processing a feature selection. We have predicted tags for an utterance in the dialog conversation. There are several criteria's to be taken in consideration for getting better performance and one of them is quality of dataset. Conditional Random Fields is one of the finest techniques we have used but we also have to see the time taken and space complexity for the accurate results. Our result is 72 per cent accurate, calculated with the help of F1 score method. This paper defines the utility of linear-chain Conditional Random Fields (CRFs) to perform robust and accurate sequence labelling by providing a principled framework that helps in the integration of domain knowledge. A probabilistic prediction of tags or labels is presented, which further improves performance.

6. Future Work

In future we plan to apply Semi-Markov method to implement sequence labelling with Conditional Random Fields. This methodology will be a combination of both Conditional Random Fields and Hidden Markov Model.

References

[1] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data", Proceedings of the Eighteenth International Conference on Machine Learning (ICML), **(2001)**.
[2] H. Wallach, "Efficient Training of Conditional Random Fields", M.Sc. thesis, Division of Informatics, University of Edinburgh, **(2002)**.
[3] F. Sha and F. Pereira, "Shallow Parsing with Conditional RandomFields", Proceedings of the Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT/NAACL), **(2003)**.
[4] A. McCallum, "Efficiently Fields", Proceedings of the 19th Inducing Features Conference in of Conditional Uncertainty in Random Artificial Intelligence (UAI), **(2003)**
[5] D. Pinto, A. McCallum, X. Wei and W. Bruce Croft, "Table Extraction UsingConditional Random Fields", Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), **(2003)**.
[6] A. McCallum, K. Rohanimanesh and C. Sutton, "Dynamic ConditionalRandom Fields for Jointly Labelling Multiple Sequences", Workshop on Syntax, Semantics, Statistics; 16th Annual Conference on Neural Information Processing Systems (NIPS), **(2004)**.
[7] A. Paz, Introduction to Probabilistic Automata, Academic Press, **(1971)**.
[8] V. Punyakanok and D. Roth, "The Use of Classifiers in Sequential Inference", NIPS Forthcoming, **(2001)**.
[9] A. Ratnaparkhi, "A Maximum Entropy Model for Part-of-speech Tagging", Proc. EMNLP, New Brunswick, New Jersey: Association for Computational Linguistics, **(1996)**.

[10] R. Rosenfeld, "A Whole Sentence Maximum Entropy Language Model", Proceedings of the IEEE Workshop on Speech Recognition and Understanding, **(1997)**; Santa Barbara, California.
[11] D. Roth, "Learning to Resolve Natural Language Ambiguities: A Unified Approach", Proc. 15th AAAI, Menlo Park, California: AAAI Press, **(1998)**, pp. 806-813.
[12] L. Saul and M. Jordan, "Boltzmann Chains and Hidden Markov Models", Advances in Neural Information Processing Systems 7, MIT Press, **(1996)**.
[13] R. Schwartz and S. Austin, "A Comparison of Several Approximate Algorithms for Finding Multiple (N-BEST) Sentence Hypotheses, Proc. ICASSP, **(1993)**; Minneapolis, MN.
[14] S. Abney, R. E. Schapire and Y. Singer, "Boosting Applied to Tagging and PP Attachment", Proc. EMNLPVLC, New Brunswick, New Jersey: Association for Computational Linguistics, **(1999)**.
[15] A. L. Berger, S. A. Della Pietra and V. J. Della Pietra, "A Maximum Entropy Approach to Natural Language Processing", Computational Linguistics, vol. 22, **(1996)**.

Authors

Romansha Chopra, She is currently pursuing her M.Tech degree in the branch of Computer Science specialized in Big data Analytics at VIT UNIVERSITY , Vellore,Tamil Nadu, India. Her area of interest is Data Mining and Analytics.

Nivedita Singh, She is currently pursuing her M.Tech degree in the branch of Computer Science specialized in Big data Analytics at VIT UNIVERSITY , Vellore,Tamil Nadu, India. Her area of interest is Data Mining and Analytics.

Yang Zhenning, he is pursuing M.ScComputer Science at School of Computing Science and Engineering, VIT University, Vellore. His area of interests are Algorithm design an Pattern Recognition, operating Systems and cloud computing

N. Ch. S. N. Iyengar (b 1961), he currently Senior Professor at the School of Computer Science and Engineering at VIT University, Vellore-632014, Tamil Nadu, India. His research interests include Agent-Based Distributed Computing, Intelligent Computing, Network Security, Secured Cloud Computing and Fluid Mechanics. He had 30+ years of experience in teaching and research, guided many scholars, has authored several textbooks and had nearly 200+ research publications in reputed peer reviewed international journals. He served as PCM/reviewer/keynotespeaker/Invited speaker for many conferences. He serves as editorial board member for many international journals, reviews papers for many conferences with an interest of serving to the education community

