NLP Models Behind RASA Stack

Ajith Shenoy, Sushma Ravindra Y, Akash Sharma, Akshay Rajan, Akshay GV

Abstract: ***This paper brings about the foundation of a platform for conversational AI the Rasa platform. This Rasa stack contains a block of open source machine learning tools exclusively used in intend to create a contextual chatbots and assistants. The services hold by this platform undergoes a major classification of powerful APIs and embedded together with Rasa stack which includes Rasa core and Rasa NLU in the form of an event stream discussed throughout this paper and also the algorithm involved in building upon this platform. Its ingredients include the Bag of words algorithm helping in simplifying representation used in the NLP, CRFs – Conditional Random Field used in statistical modelling and machine learning platforms and also advanced technology such as LTSM neural networks. This paper discusses all the algorithms involved in building up the platform and also the result produced in building up the student assistant chatbot using this platform. It also encourages the use of this RASA platform for the user required custom format as per their requirements and also promotes to contribute in developing the platform for better efficiency of the platform to function.***

Index Terms**: Bag of words, Chatbot, CRFs, NLP, Rasa stack**

I. INTRODUCTION

The RASA stack is one of the most widely acclaimed open source chat-bot building frameworks. The RASA NLU handles intent recognition and entity discovery, whereas the RASA core handles context and dialogue management. Rasa NLU internally uses Bag of word algorithm to find intent and Conditional Random Field (CRF) to find entities. The RASA core utilizes keras framework to implement a LTSM neural network for dialogue management. This paper aims at providing an overview into the above mentioned techniques which are used in RASA under the hood. This study was done to better understand the student assistant chatbot that we had built to serve the university requirements.

Revised Manuscript Received on April 24, 2019
Ajith Shenoy, School of C&IT, REVA University, India.
Sushma Ravindra Y, School of C&IT, REVA University, India.
Akash Sharma, School of C&IT, REVA University, India.
Akshay Rajan, School of C&IT, REVA University, India.
Akshay GV, School of C&IT, REVA University, India.

II. RELATED WORK

During the interest for the processing of natural language in the late 1960s the restriction parameter of "blocks worlds" associated with restricted vocabularies for the working system of natural language were SHRDLU and later the simulation of ELIZA simulated the Rogerian psychotherapist which was written by Joseph Weizenbaum and mid years of 1964 to 1966. ELIZA was defined to be a program which aided in the conversation of natural language with the computer as a possibility. Its present implementation is with the MAC time-sharing system at MIT considered to be written in MAD-SLIP for the IBM 7094.

Its name was considered to emphasize and to be improved by the user since it had its abilities which was continually imporved by a "teacher". It could be made appear even more civilized in in appearance to reality. Then as the time passed the development and curiosity lead to development of hidden Markov models during the late 1980s this development lead to processing of parts-of-speech inhibition of natural language processing. This Markov chain provides us something about the happening of sequences of states which can take any value from the set and random variables. These set includes words or tags or symbols which represents anything for instance. It makes a bold assumption that if any prediction in the future to be happening in the sequence all it matters is about the present state. The state before the present state has no impulse on the future state. As in for instance to predict the temperature of tomorrow you would examine todays temperature but not any previous measures.

Considering a system which is to be described at any time as in one of the state of set of N distinct state, S1, S2, S3, · · ·, SN,. At regular time intervals the system would undergo a change of state according to set of probability associated with the state. We represent the time associated with state change as t =1, 2 .., and we represent actual state at time t as qt . A full probabilistic description of the given system, is in general require specification of present state and previous states. For any special cases of distinct, first order, Markov chain, this probabilistic description is shortened to just present and the previous state i.e

$$P[qt = Sj|qt-1 = Si, qt-2 = Sk \ldots] = P[qt = Sj|qt-1 = Si] \quad (1)$$

Moving towards more we consider those processes in which the right hand side of above equation is independent of time , which leads to the set of state transition probabilities a_{ij} of the form

$$a_{ij} = P[qt = Si|qt-1 = Sj]\ 1 \le i, j \le N \quad (2)$$

with the state transition coefficients having the properties

$$\sum_{j=1}^{N} a_{ij} = 1 \quad (3)$$

The above process may be called as an observable Markov Model, since the output of the process is the set of states at every instant of time, where each state correspond to a observable event. In the recent years of 2010, representation learning and deep neural network-style machine learning methods became the influence in natural language processing, in advent of this natural language processing became more efficient and more in practical use in today's technology leading in the support service systems of various organization.

III. METHODOLOGY

A .LTSM in RASA Core:

In Recurrent Neural Networks (RNN) the activation outputs from the neurons propagate in a bidirectional manner, unlike the unidirectional propagation of activation outputs in the feed forward neural networks. This enables loops in the architecture of the neural network which serves as the 'memory state' of the neural networks. This enables the neural networks to retain what they've learnt. The memory retention is indeed an improvement over the traditional neural networks but are prone to the phenomenon of 'Vanishing Gradient'. This is where Long Short Term Memory (LSTM) networks come in, they enable better tuning of the parameters of the earlier layers in multilayer networks.

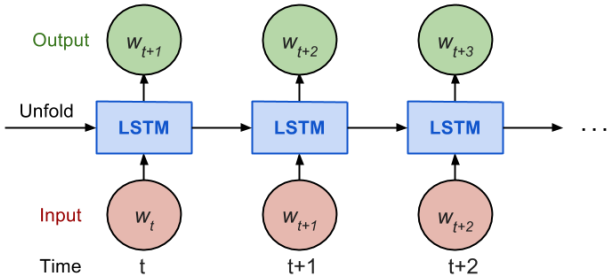

Fig1. LSTM Model

The information flow is adjusted using an additional state called the cell state, this enables the LTSM model to remember or forget its learnings discerningly [1].
A three layer model typically consists of:

1. Input Layer : Takes the sequence of words as input
2. LSTM Layer: Computes the output using LSTM units. Typically hundred units have been added. This can be tuned later.
3. Dropout Layer: A regularization layer which randomly turns-off the activations of some neurons in the LSTM layer. It helps in preventing over fitting.
4. Output Layer: Computes the probability of the best possible next word as output.

The RASA Core implements a custom LTSM using Keras to manage context and choose the next intent. This implementation can be overriden using the keras policy of the RASA core.

Bidirectional LSTM:

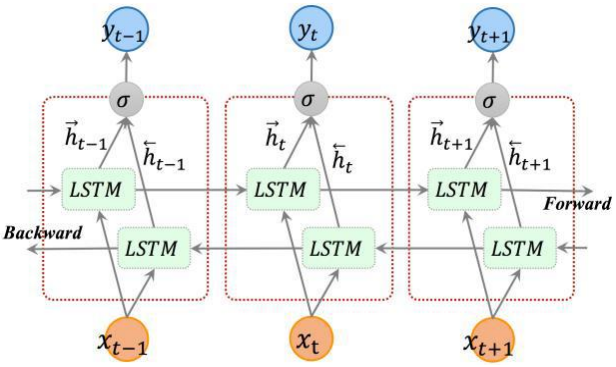

Fig2. Unfolded architecture of bidirectional LTSM with three consecutive steps

The possibility of BDLSTMs originates from bidirectional RNN, which forms grouping information in both forward and in reverse ways with two separate concealed layers. BDLSTMs associate the two shrouded layers to a similar yield layer. It has been demonstrated that the bidirectional systems are generously superior to anything unidirectional ones in numerous fields, similar to phoneme arrangement [2] and discourse acknowledgment [3]. In any case, bidirectional LSTMs have not been utilized in rush hour gridlock forecast issue, in light of our audit of the writing .The formation of an extended BDLSTM layer, consisting of a forward LSTM layer and a retrogressive LSTM layer, is presented and outlined in Fig. 2. The forward layer yield grouping, $\boldsymbol{h}^{\rightarrow}$, is iteratively determined utilizing contributions to a positive succession from time $T - n$ to time $T - 1$, while the retrogressive layer yield arrangement, $\boldsymbol{h}^{\leftrightarrow}$, is determined utilizing the switched contributions from time $T - n$ to $T - 1$. Both the forward and in reverse layer yields are determined by utilizing the standard LSTM refreshing conditions.

ft=σg(Wfxt+Ufht−1+bf) (4)
it=σg(Wixt+Uiht−1+bi) (5)
ot=σg(Woxt+Uoht−1+bo) (6)
C̃t=tanh(WCxt+UCht−1+bC) (7)

The BDLSTM layer

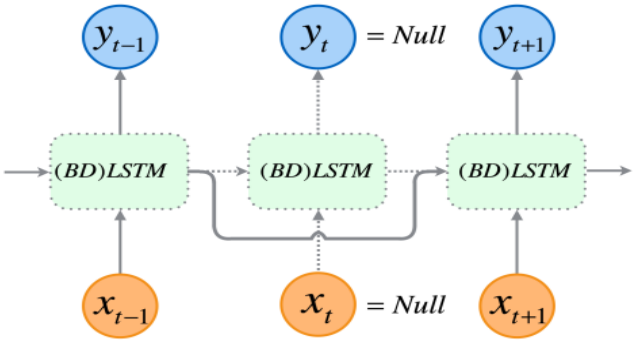

Fig 3. Masking layers for time series data with missing values.

produces a yield vector, $\boldsymbol{YT}$, in which every component is determined by utilizing the accompanying condition: $yt = \sigma(h^{\rightarrow} t , h^{\leftrightarrow} t)$ (11) where σ work is utilized to consolidate the two yield arrangements. It

very well may be a linking capacity, a summation work, a normal capacity or a duplication work. Like the LSTM layer, the last yield of a BDLSTM layer can be spoken to by a vector,

$\boldsymbol{YT} = [yT{-}n, \ldots, yT{-}1\ ],$ (8) in

which the last component, $yT{-}1$, is the anticipated speed for whenever cycle when taking rate forecast for instance.. Concealing Layer for Time Series Data with Missing Values in actuality, traffic sensors, as inductive-circle locators, may flop because of breakdown of wire protection, poor sealants, harm brought about by development exercises, or hardware unit disappointment. The sensor disappointment further causes missing qualities in gathered time arrangement information. For the LSTM-based expectation issue, if the information time arrangement information contains missing/invalid qualities, the LSTM based model will flop because of invalid qualities can't be registered amid the preparation procedure. On the off chance that the missing qualities are set as zero, or some other pre-characterized values, the preparation and testing results will be very one-sided. Consequently, we embrace a covering component to defeat the potential missing qualities issue. Fig. 3 shows the subtleties of the veiling instrument. The (BD) LSTM cell indicates a LSTM-based layer, similar to a LSTM layer or a BDLSTM layer. A cover esteem, $\emptyset$, is pre-characterized, which ordinarily is 0 or Null, and every single missing an incentive in the time arrangement information are set as $\emptyset$. For an information time arrangement information XT, if xt is the missed component, which equivalents to $\emptyset$, the preparation procedure at the t-th step will be skipped, and along these lines, the determined cell condition of the $(t-1)$-th step will be straightforwardly contribution to the $(t+1)$- th step. For this situation, the yield of t-th step likewise equivalents to $\emptyset$, which will be considered as a missing worth and, if necessary, contribution to the resulting layer. Also, we can manage input information with continuous missing qualities utilizing the concealing component. Stacked Bidirectional and Unidirectional LSTM Networks Existing examinations have appeared profound LSTM models with a few concealed layers can develop logically larger amount of portrayals of succession information, and in this manner, work progressively compelling. The profound LSTM models are systems with a few stacked LSTM concealed layers, in which the yield of a LSTM shrouded layer will be bolstered as the contribution to the ensuing LSTM concealed layer. This stacked layers component, which can improve the intensity of neural systems, is received in this investigation. As referenced in past segments, BDLSTMs can utilize both forward and Fig. 2 Unfolded design of bidirectional LSTM with three back to back advances Fig. 3 Masking layer for time arrangement information with missing qualities 5 in reverse conditions. When bolstering the spatial-worldly data of the traffic system to the BDLSTMs, both the spatial connection of the rates in various areas of the traffic arrange and the fleeting conditions of the speed esteems can be caught amid the component learning process. In such manner, the BDLSTMs are truly appropriate for being the main layer of a model to take in progressively valuable data from spatial time arrangement information. While anticipating future speed esteems, the top layer of the engineering just needs to use learned highlights, in particular the yields from lower layers, to figure iteratively along the forward heading and produce the anticipated qualities. In this manner, a LSTM layer, which is fit for catching forward reliance, is a superior decision to be the last (top) layer of the model. In this investigation, we propose a novel profound engineering named stacked bidirectional and unidirectional LSTM arrange (SBULSTM) to foresee the system wide traffic speed esteems. In the event that the information contains missing qualities, a concealing layer ought to be received by the SBU-LSTM. Each SBU-LSTM contains a BDLSTM layer as the principal include learning layer and a LSTM layer as the last layer. For purpose of utilizing the information and learning intricate and extensive highlights, the SBU-LSTMs can incorporate at least one discretionary center LSTM/BDLSTM layers. The SBU-LSTM is additionally equipped for foreseeing esteems for numerous future time steps dependent on authentic information. The point by point spatial structure of info information is depicted in the test segment.

B. Bag-of-Words Model:

The BoW model is a method of showing textual data while modelling text with ML procedures. A BoW is a way of mining features from text for use in modelling. A bag-of-words is an illustration that designates the rate of occurrence of words within a document. It involves two things:

* Vocabulary of known words.
* Quantity of the occurrence of known words.

It is called a “bag” of words, as any info about the order or assembly of words in the document is discarded. This model is only apprehensive with whether known words befall in the document, not where in the document they occur.

A very communal article abstraction processes for sentences and documents is the bag-of-words approach (BOW).We look at the histogram of the words within the text, i.e. considering each word count as an article. Additional, that from the content alone we can learn something about the significance of the document. [4]

Ex of the Bag of Words Model

Step1: Collect Data

Consider the following 3 lines

The thief ran away after stealing.

The king was very upset.

The minister was not keen on participating.

We can now consider each line as a separate “document” and the 4 lines as our entire corpus of documents.

Step2: Design the structure of vocabulary

A summation of all of the words in our model vocabulary are made, the unique words here are:

- “thief”
- “was”
- “the”
- “king”
- “not”
- “minister”
- “participating”
- “very”
- “stealing”

This is the vocabulary of 10 words from a corpus comprising 24 words.

Step3: presenting document vectors

Words in the document are scored accordingly, the main task here is to convert each document of free text into a vector that we can use as input or output for a ML model. The meekest counting technique is to mark the presence of words as a Boolean value, 0 if they are absent and 1 if they are present.

Thief=1, was=0, the =1, king=0, not=0, minister = 0,
Participating=0, very=0, stealing=1
The binary vector: (1,0,1,0,0,0,0,0,1)
Similarly other documents:
The king was very upset. – (0,1,1,1,0,0,0,1,0)
The minister was not keen on participating.
– (0,1,1,0,1,1,1,0,0)

All assembling of the words is supposedly rejected and we have a consistent way of mining features from any document in our body, ready for use [5]. New documents that overlay with the vocabulary of known words, but may contain words outside of the vocabulary, can still be prearranged, where only the occurrence of known words are scored and unidentified words are ignored.

C. Conditional random Fields:

Conditional random Fields is a popular probabilistic method for structured prediction, mostly
used for entity prediction.

Conditional Random Fields comes under the differential model, used for the prognostic corollary. To make the accuracy of prediction greater they use the information from the context of the user from the previous label.

Two of the main categories of Machine Learning models are
1. Generative and
2. Discriminative.

Conditional Random Fields come under Discriminative classifier. Decision boundary between the different classes is modelled as such. Generative models learn to model how the data was generated and make classifications based on what is learnt. Our input data is sequential in CRFs, the previous context has to be taken into account when making predictions on a data point. Feature [6] Functions is used to model this behavior, which has multiple input values, which are as follows

1. Set of input vectors, X
2. Position i of the data point that we are predicting
3. Label of data point i-1 in X
4. Label of data point i in X

We define the feature function as:

$f(X,i,l_{i-1},l_i)$ (9)

To express some kind of characteristic [7] of the sequence which the data point represents feature
function is used. For example, if CRFs is used for Parts-of-Speech tagging, then
if L{i - 1}
(10) could be a Noun,
and L{i}
(11) is a Verb then
f (X, i, L{i - 1}, L{i}) = 1
(12)
else
f (X, i, L{i - 1}, L{i}) = 0
(13)

As an outline, we use Conditional Random Fields by initially defining the feature functions needed and then initializing the weights to random values, and applying Gradient Descent repetitiously until the parameter values (lambda) combine [8]. We observe that CRFs are similar to Logistic Regression which uses a conditional probability distribution, but the main difference is algorithm can be extended by applying Feature function as our sequential inputs [9].

IV. DISCUSSIONS

Considering the student assistant chatbot which was designed by us for asisting students. Its training data is of the form :

```
{
"text": "What is my attendance r15cs019 ?",
"intent": "getAttendance",
"entities": [
{
"start": 22,
"end": 30,
"value": "r15cs019",
"entity": "SRN"
}
]
}
```

for user query : “What is my attendance r15ec211 ?”
Response:

```
{
"intent": {
"name": "getAttendance",
"confidence": 0.9636583924293518
},
"entities": [SRN],
"intent_ranking": [
{
"name": "getAttendance",
"confidence": 0.9636583924293518
},
{
"name": "greet",
"confidence": 0.03462183475494385
},
{
"name": "goodbye",
"confidence": 0
}
],
"text": “What is my attendance r15ec211 ?”
}
```

Successfully predicts the accurate intent and its entities using Bag of words

algorithm and Conditional random field.

V. CONCLUSION AND FUTURE SCOPE

The paper discussed in detail about the algorithms and methods that RASA uses under the hood. This information and understanding shall empower chat-bot builders who use RASA to tweak these inbuilt methods into something more custom, to suit their specific requirements. As RASA community is open source, the understanding of the underlying algorithms shall enable others to contribute better implementations to achieve the functionalities.

REFERENCES

1. Wenpeng Yin, Katharina Kann, Mo Yu, Hinrich Schütze ,Comparative Study of CNN and RNN for Natural Language Processing arXiv:1702.01923 , 7 Feb 2017
2. A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," Neural Networks, vol. 18, no. 5, pp. 602–610, 2005.

3. A. Graves, N. Jaitly, and A.-r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013, pp. 273–278.

4. Rong Jin,Zhi-Hua"Understanding bag-of-words model: A statistical framework" page-1, 2010. Article in International Journal of Machine Learning and Cybernetics • December 2010

5. More than Bag-of-Words:Sentence-based Document Representation for Sentiment Analysis 2013 Georgios Paltoglou Faculty of Science and TechnologyUniversity of Wolverhampton & Mike Thelwall Faculty of Science and Technology University of Wolverhampton

6. An introduction to conditional random fields Charles Sutton, Andrew McCallum
Foundations and Trends® in Machine Learning 4 (4), 267-373, 2012

7. Conditional random field based named entity recognition in geological text N Sobhana, Pabitra Mitra, SK Ghosh International Journal of Computer Applications 1 (3), 143-147, 201

8. Automatic keyword extraction from documents using conditional random fields
Chengzhi Zhang Journal of Computational Information Systems 4 (3), 1169-1180, 2008

9. Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction
Anurag Arnab, Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Måns Larsson, Alexander Kirillov, Bogdan Savchynskyy, Carsten Rother, Fredrik Kahl, Philip HS Torr

Authors Profile

Mr.Ajith Shenoy is currently pursuing Bachelor of Technology in Computer Science and Engineering from REVA University,Bangalore,India

Mrs. Sushma Ravindra Y holds M. Tech. degree in Computer Science and Engineering from VTU. She has 4 years of teaching experience. Her areas of interest include C Programming and Data Structures, Algorithm Design, Advanced Algorithms, Logic Design and Operating Systems. She is interested in pursuing research in the field of machine learning

Mr. Akash Sharma is currently pursuing Bachelor of Technology in Computer Science and Engineering from REVA University,Bangalore,India

Mr.Akshay Rajan is currently pursuing Bachelor of Technology in Computer Science and Engineering from REVA University,Bangalore,India.

Mr.Akshay GV is currently pursuing Bachelor of Technology in Computer Science and Engineering from REVA University,Bangalore,India.