International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-3, January 2020

Text Based Restaurant Recommendation System Using End-To-End Memory Network

Shrikanth Subramanian, Shanmukha Surapuraju, C.N.Subalalitha

Abstract*: With growing use of online content streaming websites, online shopping, and other exclusively online services, it becomes more and more imperative for technology companies to invest a lot of funds into a system to gauge user needs and requirements. To bridge this gap, there has been an influx of recommendation systems in the markets. From advertisements, to movies, and products we buy, recommendation engines are feeding on new data everyday to learn user trends. This paper tries to focus on improving the text based recommendation systems that can be implemented to leverage the vast review data that can be found on websites. We suggest using a novel memory based end-to-end network mechanism to reduce the need for long term dependencies and to reduce the need for memory intensive systems. As we generate more and more reviews and textual data on the web everyday, we need to be able to use this data to make meaningful analytical and business predictions. With the ability to perform multiple lookups, implement attention mechanism and back-propogation, this system was found to perform much better when compared to CNN, RNN and LSTM alternatives in our testing.

Keywords: end-to-end memory network, CNN, RNN, attention, LDA, LSTM

I. INTRODUCTION

Recommendation systems are programs and algorithms that help users in making a choice based on a set of predefined criteria. With the advent of content streaming websites and online retail, recommendation systems play a pivotal role in pushing new content to users. It is, therefore, very critical for companies to invest in building state of the art recommending engines to make sure all users are directed to products of their liking. There are predominantly two paradigms of recommendation systems, namely:

1) Collaborative filtering techniques
2) Content based filtering techniques

Collaborative methods are recommendations based on the past interactions recorded between users and items, and between different users.

Revised Manuscript Received on January 30, 2020.
***** Correspondence Author
Dr. C.N. Subalalitha*, Associate Professor, SRMIST, Department of Computer Science and Engineering, SRM Institute of Science and Technology. Kattankulathur, Tamilnadu, India.
Shrikanth Subramanian, B.Tech, Computer Science, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.
Shanmukha Surapuraju, B.Tech, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

These interactions are stored in a "user-item interactions matrix". Content based approaches use additional information about users and/or items. These generally include personal information about the user to tailor-make suggestions.

Natural Language Processing is a study of linguistics using machine intelligence, generally used to parse and analyze textual and speech data. Generally, for NLP tasks, Recursive Neural Network (RNN) are preferred for temporal structures and to parse complex lexical grammar due to their ability to preserve sequential order and model long-distance contextual information, and Convolution Neural Nets (CNN) are chosen for dealing with spatial structure due to their ability to mine semantic clues in contextual windows. While these are very useful models, they have certain limitations, namely:

1) out-of-order access
2) long-term dependency
3) unordered sets

Overcoming these limitations is a model called end to end memory networks (MemN2N), which has the following salient features:

1) Reads from memory with soft attention
2) Performs multiple lookups (hops) on memory
3) End-to-end training with backpropagation
4) Only requires explicit supervision of attention during output validation.

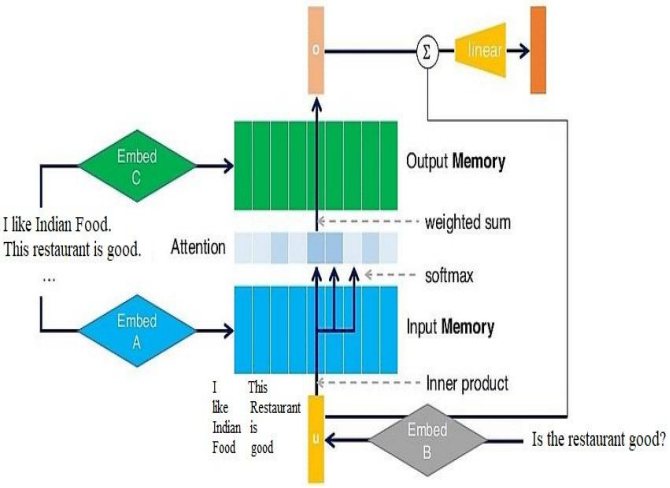

Figure 1: MemN2N workflow diagram

Figure (1) is a representation of and end-to-end memory network.

Topic modeling is a statistical method to determine the different abstract "models" that occur in a text, this helps in identifying hidden semantic structures in a text.

A popular topic modeling technique is called Latent Dirichlet Allocation (LDA). LDA is a generative probabilistic model,

1. Children love cheese.
2. Cheese and butter are dairy products.
3. Cheese in made in Italy, not butter.

After LDA:

Sentence 1 contains 100% of Topic A,
Sentence 2 contains 50% of Topic A and 50% of Topic B
Sentence 3 contains 66% of Topic A and 33% of Topic B

From analysis, we find that Topic A is about cheese and Topic B is about Butter.

Figure 2: Example of LDA

specifically, a hierarchical Bayesian model. Using LDA, every element or an object in a sentence is treated as a mixture of the extracted topics. This technique can also be used in Document structuring and collaborative filtering. An example of LDA topic modeling is illustrated below in Figure(2):

Attention is a mechanism that was originally invented to improve the performance of the Encoder-Decoder type RNN on machine translation. An attention mechanism takes into account the input from several time steps to make a single prediction and can be defined as components of memory networks, which focus their attention on external memory storage rather than a sequence of hidden states in an RNN. There are 2 types of attentions in Neural Networks, namely:

1) Hard Attention, which is non-deterministic and uses probability density function.

2) Soft Attention, which is Deterministic and differential.

A diagrammatic representation of attention mechanism is given below in figure(3):

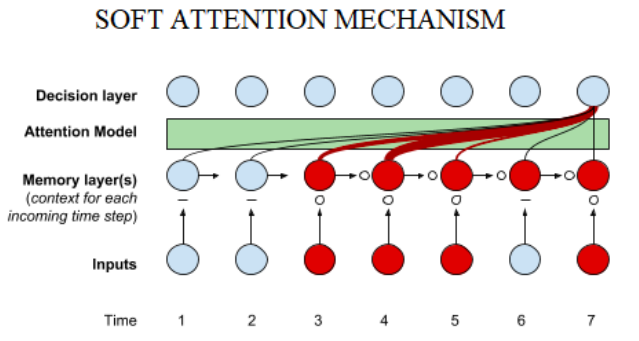

Figure 3: A diagrammatic representation of the soft attention mechanism

I. RELATED WORK

In view of increasing dependency on outdated methods for text analysis and information retrieval, a lot of research has been undertaken to devise a more efficient method of text parsing.

Various works have shown the efficiency of LDA systems to topic model and to tag important words, including LDA ad-hoc information retrieval **(Xing Wei , 2006)** and sentiment analysis as discussed in **(Xianghua, 2013)**. The latter in particular, discusses the cutting edge Latent Dirichlet Allocation mechanisms for sentiment analysis of user social sentiment by leveraging and mining chinese social reviews. This therefore, also sets precedence in terms of LDA modelling for reviews.

In **(Xing Wei, 2006)**, the authors discuss the effective information retrieval systems that are ad-hoc specialized and very narrow and concentrated in their applications. It proves that in ad-hoc conditions, an LDA is able to return a 95% confidence based on the wilcoxon test.

The earliest reference to MemN2N models was made in **(Sukhbataar, 2015)** which introduced a novel mechanism of end to end memory nets which over multiple hops easily outperformed other neural network models such as LSTM, RNN and CNNs. Furthermore, we find that end to end memory networks suggested in this paper are more suitable for multiple simultaneous computational hops. Due to its apparent efficiency in dealing with heavy duty data and large data sets, this mechanism is chosen over RNNs and LSTMs.

In **(Huang, 2016),** the hashtag recommendation system based on end to end memory, a model similar to recommendation system for restaurants is suggested. Similar to the hashtag recommendation system that prescribes a novel approach to tag and create relations between topics (bag of words), we want the restaurant recommender system to have a low recall and high efficiency.

(Xu Chen, 2018) defines an alternative memory mechanism for sequential recommendation, we draw inspiration from the novel approach in finding that though user's previous preferences and behavior records are not all equally important, in the sense that some behavioral aspects are much more prevalent and useful for future predictions as compared to others. A memory mechanism is able to overcome this. Despite the fact that the paper defines its mechanism based on a memory augmented neural network, it can be found that the latent memory matrix storage and manipulation similar to that of the end to end system. According to the paper, memory storage mechanism has again been recorded to be consistently more efficient as compared to RNNs and markov chains.

(Zhang, 2016) paper on phrase-level textual sentiment analysis across multiple categories or LRPPM, is helpful in breaking down the reviews into phrases, each with an individual context and finding correlation between the words in the individual lines.

(Chen Cheng, 2013) links the LRPPM model in the previous paper by discussing sequential correlation in his paper on successive points-of-interest recommender systems. These recommender systems offer higher accuracy and reliability as compared to other systems.

II. PROBLEM FORMULATION

This paper aims to develop an optimal recommendation algorithm which can accurately identify the preferences of the user based on historical data, user reviews, and the personal information obtained from the user. The aim is to build such a recommendation system that uses both content based and collaborative filtering methods to reach an optimal result. In this work, we aim to utilize the end-to-end memory network setup to parse the input reviews and to successfully identify and tag the key words that will be useful in predicting user behavior. Using state of the art NLP techniques such as attention mechanism,

LDA topic modeling and bag-of-words implementation, this paper aims to make significant improvements in the performance of text based recommendation systems. An example of end to end memory system is shown below in figure(4) to demonstrate the use of temporal data and soft attention.

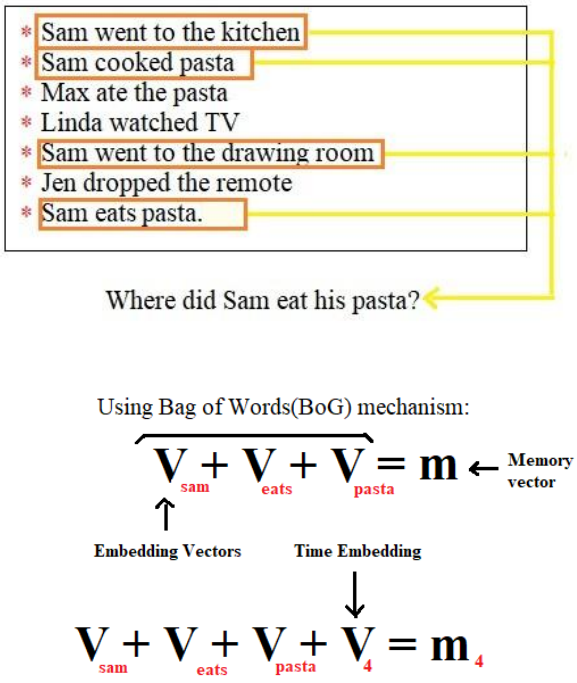

Figure 4: Working of a MemN2N system with time embeddings

As shown in the diagram above, the key difference in MemN2N architecture is the addition of memory and time embeddings separately while externalizing the memory.

III. WORKING MODEL

(a) Getting user's personal information from the application.
(b) Extracting User-item interaction matrix for user history.
(c) Reading the generated restaurant reviews for the chosen restaurant.

1) Cleaning extracted review words:

Prior to proceeding with text based analysis and modeling, there often is a need for cleaning and parsing the text. This is owing to the fact that almost all the text is created and stored in human-readable form, and it is challenging for a computer to process that text accurately. Most of the cleaning and parsing of text involves increasing the regularity and adding structure to the text. This involves:

1. Removing stop words
2. Fixing typing errors
3. Tagging some words as important, such as name, title, and etc
4. Lemmatization, grouping words with common roots

2) Latent Dirichlet Allocation (Topic tagging)

Every element or object is treated as a mixture of the extracted topics in a certain way. With respect to the text classification capabilities, the topic probabilities provide an almost accurate representation of the document. After LDA is completed, the topics of the review are all individually tagged.

3) TF-IDF bag of words:

TF-IDF, or term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection of words, or a sentence. It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The tf–idf value increases proportionally to the number of times a word appears in the document.

4) Implementation of attention mechanisms:

While implementing attention mechanism, we find an unfortunate side-effect of using attentions in computational models. To successfully use the attention mechanism, we need to calculate an attention value for each combination of input and output word. Take for instance, a 100-word input sequence and generate a 100-word output sequence, that would be 10000 attention values. That doesn't sound too bad for the smaller denomination of word sequences, but if you do character-level computations and deal with sequences consisting of more than a hundred tokens the above attention mechanisms can become prohibitively expensive. Actually, that's quite counterproductive. Human attention is something that's supposed to save computational resources. By focusing on one thing, we are able to neglect many other things. But that does not seem to be possible with computational models. We're essentially looking at everything in detail before deciding what to focus on.

5) Word Embedding are created:

We use MemN2N as a language model. For instance, we parse any random review as the example: "The movie was great, it couldn't have been better. The first half was better than the second." Instead of 1 sentence per memory entry, we store only one word per entry as shown in figure(5):

WORD EMBEDDINGS

Memory slot	Word
1	The
2	movie
3	was
4	great
5	it
6	could
7	not
8	....

Fig 5: Word embedding

Algorithm for parsing reviews to meaningful information:

I. READ_TEXT
II. USE PRE-TRAINED LIBRARY MODULES TO IGNORE INCORRECT WORDS
III. REMOVE STOP WORDS
IV. CREATE WORD EMBEDDINGS,

so now we have memory vector and embedding vector.

V. FIND TOPICS FROM EMBEDDING VECTOR

VI. if(TOPIC_ELEMENTS = Restaurant_tags) then add TOPIC_ELEMENTS to User-item Interaction matrix.

VII. END

Algorithm for collaborative recommendation implementation:

I. READ HISTORY DATA

II. LOAD DATASET CLASS

III. IMPLEMENT MATRIX FACTORIZATION, for eg. non-negative matrix factorization (NMF)

IV. CALCULATE VALID Root Mean Square Error (RMSE) values

V. RETURN SUGGESTION WITH THE LOWEST RMSE value.

VI. END

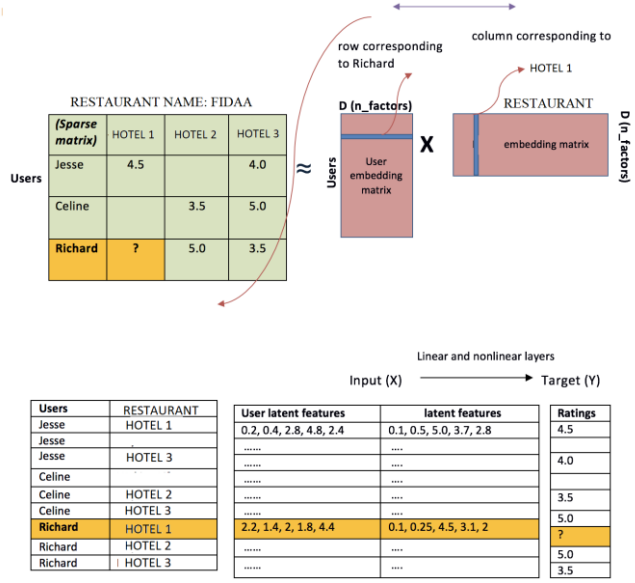

Figure 6: Figure depicting recommendation system

IV. COMMON MISTAKES

- Inaccurate model architecture.
- Inefficient train-test splits
- Not eliminating stop words
- Lack of lemmatization
- Implementing attention mechanism without rein-forcement learning approach, whereby increasingcomputational burden
- Overfitting the model on the database
- Tokenization and node optimization.

V. RESULT AND DISCUSSION

The results showed an interesting trend:
1) The best MemN2N models are reasonably close to the supervised models, although the supervised models are still superior.
2) All variants of our proposed model comfortably beat the weakly supervised baseline methods.
3) Joint training on all tasks helps.
4) Increased computational hops gives improved performance.

The recommendation system returned a validation accuracy of 95% in our implementation. The MemN2N model outperformed LSTM, RNNs and CNNs in our attempt. Given below is the MSE of the implementation Figure (7).

Table : Performance Metric, as we can see increasing hops increases performance of the system. (10K training examples).

Model	Hidden	No. of Hop	Mem Size	Validation	**Test Perf.**	**Mean Error**	**Failed Task (Error>5%)**
RNN	300	-	-	133	**129**	**26%**	**15**
LSTM	100	-	-	120	**115**	**36%**	**18**
CNN	100	-	-	120	**115**	**39%**	**20**
MemN 2N	150	2	100	128	**121**	**6%**	**4**
MemN 2N	150	4	100	127	**120**	**4.5%**	**3**
MemN 2N	150	6	75	122	**114**	**4%**	**2**

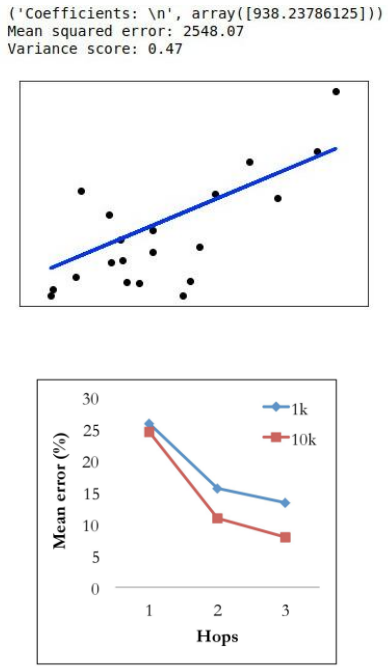

Figure8. MSE, Variance and 1K vs 10K training data comparison

VI. CONCLUSION AND SCOPE

In this work we showed that a neural network with an explicit memory and a recurrent attention mechanism for reading the memory can be used in recommendation tasks. It can be successfully trained to perform tasks in the NLP domain such as language modeling. Our model outperforms RNNs and LSTMs of comparable complexity. On both tasks we can see that increasing the number of memory hops improves performance. Using MemN2N reduced the need for training supervision and reduces memory load.

Compared to the Memory Network implementation of [9] there is no overall supervision required, because of this our model can be used in a plethora of use cases. Our model approaches the same performance of that model, and is significantly better than other systems with comparable supervision measures. On language modeling tasks, our model outperforms tuned RNNs and LSTMs of comparable complexity by a small margin.

In the future,

- more work needs to be done in particular to focus on making highly specialized specific recommendation systems such as this one.
- Work must be done to further reduce the memory burden.
- There is scope for addition of sentiment analysis in this system to further gauge the user sentiment.
- Currently the system only accepts reviews that follow a stringent grammar, with additional training, the system can be trained to handle human errors.

REFERENCES

1. Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to- end memory networks. In Advances in neural information processing systems, pages 2440–2448.
2. Dzmitry Bahdanau ; Jan Chorowski ; Dmitriy Serdyuk ; Philémon Brakel ; Yoshua Bengio, End-to-end attention-based large vocab-ulary speech recognition, Published in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)ISSN: 2379-190X
3. Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In Proceedings of the conference on empirical methods in natural language processing, pages 583–593. Association for Computational Linguistics.
4. Jatin Ganhotra, Lazaros Polymenakos: Knowledge-based end-to-end memory networks, arXiv:1804.08204 [cs.CL]
5. Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In AAAI, volume 16, pages 3776–3784.
6. J. Weston, S. Chopra, and A. Bordes. Memory networks. In International Conference on Learning Representations (ICLR), 2015
7. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR), 2015.
8. A. Graves, G. Wayne, and I. Danihelka. Neural turing machines.arXiv preprint: 1410.5401, 2014
9. C. G. Atkeson and S. Schaal. Memory-based neural networks for robot learning. Neurocomputing, 9:243–269, 1995
10. J. Chung, C¸ . Gulc¸ehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural¨networks on sequence modeling. ArXiv preprint: 1412.3555, 2014.
11. S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735– 1780, 1997
12. Tom Young; Devamanyu Hazarika; Soujanya Poria; Erik Cambria,Recent Trends in Deep Learning Based Natural Language Process-ing, arXiv preprint: 1503.08895, 2018.

AUTHORS PROFILE

Shanmukha Surapuraju, B.Tech SRM Institute of Science and Technology, currently pursuing masters in Computer Science from Indiana University, Bloomington.

Shrikanth Subramanian,
B.Tech in Computer Science, SRM Institute of Science and Technology, ACM (SIGCHI- chapter) member.
Areas of Interest: NLP, Computer Vision, Deep Learning

Dr.C.N.Subalalitha,
Associate Professor, SRMIST
Department of Computer Science and Engineering, Kattankulathur, SRM Institute of Science and Technology