Survey on question answering mechanism in real world call center applications

[1]Hemant P
[1]Research Scholar
[1]Shri JJT University

***Abstract* - Machine reading comprehension has created a tremendous change in today's Call-Centre Bots and question answering system. In real-time call centre environment aid to agents is very important as it helps to solve complex problems and in turn improve customer satisfaction. Machine Reading Comprehension can help in understanding customer issues after running analysis on the speech to text output and provide right set of questions to be asked to the customer so that the conversation can be channeled properly. Due to this the average handling time would reduce by a substantial number. We have made a comparison of different system build on MRC and parameters like the dataset used for training and validation, performance on various benchmarks and features. So according to the research done it has been observed that the question answering system has been shifted from query matching to read the exact passage to generate the answers. The datasets which were used for training and how the performance differs according to the dataset and the training mechanism were also discussed. Then the evaluation and performance of all the tasks are compared and best algorithm was chosen. The use case on which the specific algorithm was working is also important. In this way the final algorithm was concluded by observing all these factors.**

***keywords* - Generative Chatbot, Machine Reading Comprehension, Question Answering**

I. INTRODUCTION

By having latest improvement in the domain of deep learning techniques in these last few years, there is tremendous growth in speech recognition systems. The main task in speech recognition is to make human to machine communication more interactive and useful for use in various domains. Suppose if we take an example that in a certain company there is a forum for solving customer queries and the agents which are provided to solve the customers are new and don't have the exact questions on what to ask the customer for a specific product and query and how to solve their issues. So here comes the major role of a chatbot that will be used to make a conversation with the help of text or speech to text done by a speech to text model for the question asked by th7/78/e user to get the responses. So in this chatbot, there will be many mechanisms to solve the queries of customers. There will be features like it should be able to do speech to text conversion, the converted text will be supplied to the question and answer mechanism to get correct answers for that question. This system will help the organization to provide a unique outcome for the queries of the customer and this way the customer will be satisfied by the solution. Speech recognition will help the system to identify the words in human spoken language to text and this conversion is very important to understand the words said by the user and to identify the correct text from the speech. The corrected text obtained after the speech to text conversion will be supplied to the QA mechanisms so that it can identify the correct response. MRC will be useful to get a brief of the products and services which are provided by the organization in an unstructured text format. It will use this paragraph to read and understand the context of the text provided. After all this process it will be skillful to provide relative answer to the questions which is related to the products and services which a customer can ask the agent. A chatbot should also understand the emotions of the user and give a response according to the emotions. It will result in better human interaction and as humans speak with emotions so understanding only words and grammar won't help. So the result will be more effective if the emotion analysis of chatbot is better. It uses natural language understanding to generate the response using a large amount of labeled training data in this training helps in building a good architecture of the conversational systems. This architecture is graded to users in a long way. Open Ai GPT models used for the pre-training of transformer models. GPT is a multilayer transformer decoder and it consists an attention model with training on a large corpus. The dataset is used so as to have high coverage of text. These chatbots are also domain-specific related to the task they are used as if it is related to products or services or feedbacks, etc. As they have to work on the regeneration of text, a bidirectional LSTM model is used to carry out the embedding of words and to encode and decode the text. So sequence-2-sequence model techniques are used basically in generative chatbots to generate the text which is not answerable or doesn't even exist in the system. The focus is that chatbots should function like machine translation and transform the input sequence to the output sequence. Neural Networks have contributed in applying the encoder-decoder architecture so as the data and pre-trained architecture can be studied in better way. Conversion of words into vectors and using vectorized input helped the deep learning models to understand the text in a more sophisticated and good manner. And this helped in building fast regenerative QA systems that make the answers more feasible according to the conversation happening between the humans.

II. LITERATURE REVIEW

Zhenzhong Lan et al 2020 observed that if we increase the model size it will result in better performance. But there are memory limitations as the model size increases and to solve these problems there are 2 point parameter techniques introduced so that training speed will improve and memory usage will be managed. Using self-supervised loss helps modeling of sentences

it helps in training the downstream task to perform better and this results in better results of the models on benchmarks like SQuAD, RACE, and GLUE with small number of parameters.

Jacob Devlin et al 2019 proposed BERT which pre-trains the deep bidirectional networks on very large unlabelled data. The model is fine-tuned to have an additional output layer to achieve good results of the models on large scale tasks, for example, question answering task, language inference task. The model is simple and powerful because it has achieved a tremendous good result for eleven natural language processing tasks. Transfer learning has resulted in significant improvements in unsupervised training of language models. The use of less resources is activated by these results from unidirectional architecture. These results are generalized for bidirectional architecture which would have the pre-trained model to overcome the NLP tasks.

Aniruddha Kembhavi et al 2018 represented that Machine Reading Comprehension is highly skillful to answer the questions from the specified paragraph but it requires modeling interaction between questions and paragraphs. Attention mechanism with Machine Comprehension has resulted in good results. So basically these models use attention mechanisms for small sentences or parts of paragraph for summarization using a fixed vector and this attention will form unidirectional attention. Here we have introduced the Bi-directional Attention Flow network, which helps to represent the context of the paragraph at various measures and it uses a bidirectional attention mechanism for representation on relevant queries context avoiding premature summarization. The model has resulted in good results in SQuAD and CNN/Daily Mail cloze tests. The above analysis helped to determine the importance of every component in our model. The model is effective enough to answer difficult questions by extracting the right sentences from a given paragraph. It has further scope to improve the attention layer to deploy multiple hops.

Siva Reddy et al 2019, researched that humans gather information by reading passages and then answer questions according to paragraph. To build a system that to do the same requires to train in such a way to answer the informal questions. So here we have presented a CoQA, a dataset used for conversational datasets. The dataset consists of total 127000 questions and answers acquired from 8000 conversations passages extracted from various areas. The dataset has some challenges which do not exist in current comprehension datasets. Then model training has been done on these datasets to achieve a conversational and reading comprehension model. The system achieved an F1 score 65.4 which is only some points lagging form human performance.

III. COMPARATIVE ANALYSIS

As we have trained our systems on various datasets for different natural language tasks we have done a comparison of all the algorithms and achieved good accuracy in LAMBADA Winograd schema challenge technique. This language model task is trained on web text. The task has a significant-good result on the GPT-2 model and has procured a state-of-art on various language modeling datasets.

Use Case	Pros	Cons	Dataset Used	Accuracy
Text Entailment Generalization Language Representation[5]	It is 1.7x faster than bert and transition from layer to layer is much smoother.	Data throughput of ALBERT-xxlarge is 3.17 times lesser than BERT large.	Training: BOOKCORPUS , Wikipedia	RACE accuracy is 89.4%, GLUE benchmark result is 89.4%, and the F1 score of SQuAD 2.0 is 92.2.
Relevant text generation based on keywords[6]	Can be used to improve other NLP applications.	There still exists a trade-off between the parameters depending on the generation intended.	*Wikipedia * submissions from 45 subreddits * OpenWebText * a large collection of news data * Amazon Reviews	Very Accurate
Text Similarity and Question answering [7]	Transformer benefits up to 9% for MultiNLI	1.Large unlabeled data are abundant	*SNLI *MultiNLI *Question NLI *RTE * SciTail *Question Answering RACE *Sentence similarity MSR Paraphrase Corpus	
Multiple language understanding[8]	CLM and MLM approaches supplystrong cross-lingual features used for pretraining models	monolingual, and largely focused around English benchmarks	Unsupervised machine translation using monolingual corpora only. In ICLR	XNLI by 4.9% accuracy on average than the previous.

Text Entailment	DistilBERT is greater than 60% faster and smaller than BERT and 120% faster and smaller than ELMo+BiLSTM	Does not support token ids	Evaluation: GLUE.	92.82% on IMDB and 78.7% on SQuad
Language Modelling	It is 1,800+ times faster than a vanilla Transformer while evaluating on language modeling tasks.	The model can only recognize the seed context and provide on limited data	enwik8, WikiText-103, text8, Penn Treebank, and One Billion Word.	Transformer-XL improves the SoTA BPC/perplexity from 1.06 to 0.99 on enwik8, from 1.13 to 1.08 on text8, from 20.5 to 18.3 on WikiText-103
Foster research and downstream applications for French NLP:	Model evaluation is on four downstream tasks and achieving state-of-the-art results in most tasks	performance lags behind models trained on the original English training set in the TRANSLATE-TEST setting	Wikipedia, Tatoeba, and SETimes. NER-annotated French Treebank dataset. French part of the XNLI dataset.	81.2%
Language models are unsupervised multitask learners	Vocabulary was expanded to 50,257 and Context size was increased from 512 to 1024 tokens	All models still underfits on WebText.	WebText	*LAMBADA:59.23 *Children's book test:85.7 *93.3% on common nouns *89.1% on named entities
Multiple language understanding:	The pretraining multilingual language models have significant performance gains for a wide range of cross lingual transfer tasks.	Uncovering the high-resource versus low-resource trade-off, the curse of multilinguality and the importance of key hyperparameters	github.com/attardi ,wikiextractor ,opus.nlpl.eu , phontron.com/kytea	80.0% average accuracy
Question answering	Outperforms strong baselines and measure multimodal performance	limitations of time	quora,wikitext,github	70%
Converting any language problem into a Text-to-Text format.	pre trained T5 model can be used as a starting point for building systems	Model size is large	An unlabelled dataset developed by Google called C4.	88.9 on SuperGLUE language benchmark

So depending on the performance and accuracy parameter we choose the LAMBADA Winograd schema challenge technique for MRC activities.

IV. CONCLUSION

So in this paper we have approached various methods of question answering and reading comprehension techniques. Use of speech to text for conversion of human words resulted in getting exact format of words that can be supplied to the reading mechanisms. The main aim was to build a system that can help new agents to get suggestion of answers for the question asked by the user. This system has multiple features from extracting correct words from customer to search for the answers for that question. The model is trained on various dataset to understand the patterns of the questions and answers. This also helped us to understand the domain of the passage and if the model will be compatible for that particular domain. By comparing all the systems performing various experiments on various dataset we can say that this system will be able to perform better for information extraction and extract answers with the passage context.

REFERENCES

[1] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In ICLR.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[3] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603.

[4] Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. CoQA: A conversational question answering challenge. TACL.

[5] Zhu, Chen, et al. "Freelb: Enhanced adversarial training for natural language understanding." International Conference on Learning Representations. 2019.

[6] Faure, David, and Claire Nédellec. "A corpus-based conceptual clustering method for verb frames and ontology acquisition." LREC workshop on adapting lexical and corpus resources to sublanguages and applications. Vol. 707. No. 728. 1998.

[7] Nicosia, Massimo, et al. "QCRI: Answer selection for community question answering-Experiment for Arabic and English." Association for Computational Linguistics, 2015.

[8] Wang, Alex, et al. "Glue: A multi-task benchmark and analysis platform for natural language understanding." arXiv preprint arXiv:1804.07461 (2018).