Summary for “Recent Advances in Chatbot”

The Summary is divided into four major parts that are:

1. Basic Models Used
2. Dataset for Language Training
3. Evaluation of Chatbot
4. Suggestion (You can write any suggestion if you have w.r.t anything in paper.)

The Application of Chatbots section in paper is not summarized, it majorly just contains the part where it tells the difference of application of machine learning according to purpose of use of chatbot. Like most chatbot usage at a smaller scale is of information retrieval form and hence rarely some chatbots may contain application of deep learning and where it is used.

Models used:

* Rule Based Model – It is the most primitive model dependent on pattern matching
* AI (Artificial Intelligence) Models -
  + **Information Retrieval Model** - It is the model based on information retrieval that searches the information from predefined set of ans for a given question. Mainly based on ques-ans matching and displaying the predefined ans.
  + Model based on **local textual co-occurrence and map hierarchical information** across domain for more distant terms. It is like categorizing the input based on set of words that occur frequently and are occurring together. This leads to better search retrieval. Like an input contains ‘dog,’ ‘cat’ and ‘pet’ co-occur and are frequently used then it refers to the hierarchical information to match it to the pet term which is better way to categorize the following input rather than the matching it based on frequently occurring term which is ‘dog.’
  + **Sequence to Sequence Model** – based on generating sentences by itself. It is that the model is trained on large no. Of conversational phrases to make it learn the syntax as well as the punctuation. It is best for making social bots that aim to converse with the user in a natural manner. Based on Encoder-Decoder Neural Network model with Long-Short-Term-Memory Mechanisms to counterbalance the vanishing gradient in vanilla RNNs
  + **Transformer** – First presented in the paper “Attention is all you need.” Based on Differentially weighing relevance of each portion. Provided with training parallelization enabling to train the model on larger dataset than was earlier possible. BERT (Bidirectional Encoder Representations from transformers) and GPT (Generative Pre-trained Transformer), which were trained on huge language datasets.
  + **Transformer-XL** – They are aimed at breaking the bound of fixed size of context in the setting of language model. It enables learning dependencies beyond a given length without breaking the temporal coherence. The authors of Transformer-XL propose a solution called "sentence-level recurrence." This means that the model can remember information from previous sentences and use it to better understand the current sentence. By doing this, the model can capture longer-term dependencies in the text.

Encoder-Decoder Mechanism in Sequence-to-Sequence model is different from the vanilla RNNs as the latter processes the input and output together thus is not able to capture long range dependencies. Thus, seq-to-seq models decouple the input and output processing and are better at handling long range dependencies.

Here is an example of how the encoder-decoder mechanism would work for the English sentence "I am a student." To French.

1. The encoder would read the English sentence one word at a time. For each word, the encoder would calculate a hidden state. The hidden state would represent the information about the word that the encoder has learned so far.
2. After reading the entire English sentence, the encoder would calculate the context vector. The context vector would represent the overall information about the English sentence that the encoder has learned.
3. The decoder would then start generating the French sentence one word at a time. For each word, the decoder would calculate a hidden state. The hidden state would represent the information about the French sentence that the decoder has generated so far.
4. The decoder would also use the context vector to generate the French sentence. The context vector would provide the decoder with information about the English sentence, which could help the decoder to generate a more accurate French sentence.

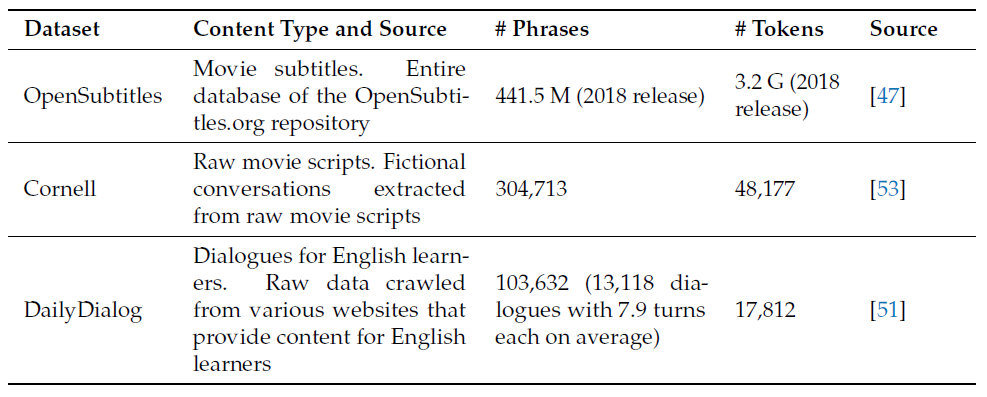
Major difference between Transformer and Encoder-Decoder mechanism/RNNs:

* Transformers use self-attention instead of recurrent connections: Recurrent connections are a type of neural network connection that allows information to flow through the network in a sequential manner. This can be problematic for sequence-to-sequence tasks, as the recurrent connections can make it difficult for the model to capture long-range dependencies. Transformers, on the other hand, use self-attention, which allows the model to attend to any part of the input sequence, regardless of its position. This makes transformers much better at capturing long-range dependencies.

Datasets Used:

Open Domain Datasets:

* OpenSubtitles – It is a dataset of various movie subtitles in different languages. It is a large dataset but is filled with screen description, closed captioning, and segmented sentences. Thus, the quality of the text is not good. These extra things can affect the cohesiveness of the dialogues (cohesiveness is lost) when the model is being trained.
* Cornell – Cornell Movie Dialogue presents a meta-data rich collection of fictional conversation taken from raw movie screenplays, consisting of more than 300,000 total utterances. So, it is of decent quality, but the dataset size is small if we want to train more advanced language models.
* DailyDialog – It is a human-made dataset that consists of several multi turn conversations. It contains about 13000 conversations with each speaker having on average about 8 turns. The data is annotated, annotation contains information about the intent and emotions. It is small and unsuited for large and complex language models.



Evaluation:

A Chatbot is evaluated based on its purpose. A home assistant chatbot should be evaluated based on the ability to complete tasks the user gives and in difficulty of communicating the task by the user. Whereas a social chatbot should be evaluated on its ability to keep the conversation going and ease of conversation for the User.

* **Human Evaluation**: Consist of asking a group of people to test and evaluate the chatbot and rate them according to their experience. This method cannot be generalized as it varies from person to person and can be very time consuming and inefficient, but Human Evaluation can consider various aspects that can be missed out.
* **Automatic Evaluation**: They are much more efficient in terms of resource and time needed to evaluate the chatbot. They cannot collectively assess the quality, effectiveness, and efficiency of the conversation. Evaluation Metrics are standard evaluation metrics used for Machine Translation and other Natural Language Processing tasks such as BLEU, METEOR and TER.

In Automatic Evaluation F1 score is used frequently (There are others such as F0.5 or F2). This score quantifies the relation between the precision of the model v/s the no. Of recalls to the system.

Many methods used in calculation of chatbot are (Their respective formulae are not included you can refer to the original paper for the specifics):

* **Perplexity** – It is the test sets inverse probability normalized by number of words. It is not applicable to unnormalized language model.
* **BLEU** – This metrics was majorly used in evaluation of machine translation outputs. The metrics is between 0 to 1, the closer the translations is to 1, the more closely resembles human translation. But BLEU metrics does present some issues.
* **Metric for Evaluation of Translation with Explicit Ordering (METEOR)** - It was created to overcome the problems with the BLEU metrics. It scores translations based on explicit word-for-word matches between the translation and a reference translation. If many reference translations are available, the given translation is evaluated independently of each reference and the best score is reported. METEOR produces an alignment between two strings when given a pair of translations to compare (a system translation and a reference translation).
* **Translation Error Rate (TER)** - It has been used less compared to other methods for evaluating chatbots performance, but it is widely used to evaluate textual entailment. It is determined by the edit distance. It calculates the mistake rate by calculating the number of revisions necessary to convert a machine translated output sentence to a human0translated reference sentence. Thus, the complement of this error rate is considered when computing the similarity score.

In recent times rather than n-grams base evaluation models some recent studies have been conducted to study the usage of adversarial evaluation methods to evaluate dialogue models.

“Discussion” part of the paper can be read at it discusses about the future of the chatbot and what all challenges are there that are needed to be faced for us to produce a chatbot capable of human like conversations.

Future of Chatbot:

To bridge these gaps, smaller, flexible, less domain dependent models would be beneficial. Improved, scalable, and flexible language models for industry specific applications, more human-like model architectures, and improved evaluation frameworks would surely represent great steps forward in the field.

Suggestions:

* As written in the article, the process of creating and adding all the knowledge base as question and ans base is highly time taking and inefficient. We use the database method in information that usually changes regularly or is subjected to change occasionally. This will help train the model on data that usually does not change often (like once in 2 to 3 years) and train the chatbot on that information as the source for that topic. Now what else can we do is if the database is becoming too big and difficult to handle as the knowledge base, we sacrifice the data or information with least frequency as well as data that is related to minor topic so that even if the data is changed regularly, we can retrain the model on the said topic with methods like partial retraining. This will help manage data sources and maintain the balance between data sources of the chatbot and its knowledge base.