Dave the debater: a retrieval-based and generative argumentative dialogue agent

Dieu Thu Le
Institute for Natural Language Processing (IMS)
University of Stuttgart
thu@ims.uni-stuttgart.de

Cam-Tu Nguyen
National Key Laboratory for Novel Software Technology
Nanjing University
ncamtu@nju.edu.cn

Kim Anh Nguyen
FPT Technology Research Institute
FPT University
anhnk14@fpt.com.vn

Abstract

In this paper, we explore the problem of developing an argumentative dialogue agent that can be able to discuss with human users on controversial topics. We describe two systems that use retrieval-based and generative models to make argumentative responses to the users. The experiments show promising results although they have been trained on a small dataset.

1 Introduction

Research in argument mining has mainly focused on the problem of identifying claims, premises (Boltužić and Šnajder, 2014, 2015; Levy et al., 2014), assessing arguments, classifying stances, detecting political beliefs (Hasan and Ng, 2013; Iyyer et al., 2014; Bamman and Smith, 2015) or finding connection between claims (Stab and Gurevych, 2014). Very few research has addressed the problem of generating arguments directly in a conversational form.

To study and analyse debates, it is important to understand how to formulate claims, how arguments develop and relate to each other, what factors influence the next argument. In this work, we explore the question whether we can teach computers to make or generate arguments and follow the ideas/stances/sides of actors in a debate. To start inspecting this challenging problem, we develop two debate dialogue systems, a retrieval-based and a generative model. The aim of the system is to mimic a debater, make arguments and give relevant responses to users on given topics.

Such argumentative dialogue systems could be useful in a lot of future applications, such as in information campaigns, where the users can get objective answers for controversial topics to make evidence-based decisions; in an interactive argumentative dialogue system, where the users can practice making arguments, learning to persuade people.

2 Related work

Analyzing public debates about controversial issues is a well-studied area in social and political science. Natural language processing and machine learning could help building a scalable and data-driven predictive modelling for public debates. In this rapidly growing field, most of the work has focused on the identification of claims and justifications in text (Boltužić and Šnajder, 2014, 2015; Levy et al., 2014), connecting claims (Stab and Gurevych, 2014), actors with discourse analysis (Peldszus and Stede, 2015), stance detection (Hasan and Ng, 2013), or the categorization of political beliefs (Iyyer et al., 2014; Bamman and Smith, 2015).

Most of these studies have focused on public debates which can be found in newspaper articles, written essay (Stab and Gurevych, 2017) or parliament debates (Koehn, 2005). Another line of research works on Internet dialogues such as those in social networks, online forum debates (Walker et al., 2012a). The dialogic language used in these forms is usually different from that found in newspapers. While it also contains stances, arguments, opinions, this language is usually more informal, can contain typos and subjective acts such as sarcasm (Justo et al., 2014; Swanson et al., 2017). There are a number of studies focusing on this kind of data, working on sarcasm and nastiness detection (Justo et al., 2014; Swanson et al., 2017) as well as topic stance classification (Walker et al., 2012b).

Little research has been done on using machine learning to generate arguments in a conversation. The most relevant idea is the one reported in (Rakshit et al., 2017; Rach et al., 2018). In (Rakshit et al., 2017), the authors describe Debbie, a debate bot of the future. It is an initial working prototype, in which the system retrieves the most appropriate counter-arguments using a similarity al-

Proceedings of the 5th Workshop on Argument Mining, pages 121–130
Brussels, Belgium, November 1, 2018.

gorithm. In particular, they used Latent Semantic Similarity for word similarity and Wordnet, together with hierarchical agglomerative clustering to retrieve the most similar responses. Evaluation has been done based on the time the system took to retrieve the results. While this work is the most similar to our ideas, it is currently an initial prototype and fully-retrieved based. Our work on the other hand explores several options focusing on the currently challenging direction, generative model in argumentative dialogue systems.

Research on conversational systems has two main directions: task-oriented dialogue systems (Williams et al., 2017; Bordes and Weston, 2016; Eric and Manning, 2017) and general/open-domain chatting systems such as those described in (Vinyals and Le, 2015; Zhou et al., 2017; Li et al., 2016a; Serban et al., 2016). For the task-oriented dialogue systems, it is usually important to have an intent classifier together with a dialogue state tracker that keeps track of which information is needed to be requested and finally a language generation module (Williams et al., 2017; Bordes and Weston, 2016). Our work is more related to the second direction, a general/open-domain chatting system. Besides the common retrieval-based approach, a growing interest in the research area focuses on a generative, end-to-end system. One of the first study using sequence to sequence model for building conversational models is described in (Vinyals and Le, 2015). These systems can generate new responses in daily conversational topics, but are still quite limited in making sense of these responses. The main problem often lies in the decoder and objective parts, where usually the most generic and *safe* responses such as "I don't know" are selected. To deal with this problem, (Li et al., 2016a) proposed using another objective function to promote diversity in responses. Some other work investigates the problem of integrating emotion and persona into the conversational agents such as (Zhou et al., 2017; Li et al., 2016b).

3 The debater system

In this section, we describe our argumentative conversational system, which can give responses in two different modes: using a retrieval-based approach and using a generative model.

3.1 Format of a debate

The aim of the chatbot is to be able to carry a conversation with humans to debate about a given topic. At the initial step, the system suggests a topic (Table 1) and the user can decide to debate on this topic or move on to another one. When a topic is selected, the user can give her opinions and the system should generate coherent responses to the user's message. Ideally, the system's response should be meaningful, relevant to previous messages and present opinions/arguments about the given topic.

3.2 Dataset

In the demo, we use the Internet Argument Corpus (Abbott et al., 2016), which is a collection of 65K posts in 5.4K debate topics (Table 1) retrieved from Convinceme website[1]. While debates from medias such as those in newspapers, broadcast news are more officially and formally written, online debate posts are often more colorful, personal and may be rational, contain emotional languages. Such kinds of debates tend to be more subjective and naturally present how humans debate with each other. Topics of discussion in this online forum are various, ranging from political debates (e.g., *should guns be controlled?*) to everyday life topics (e.g., *How much should I tip the pizza man for my 20$ lunch order?*).

Star Wars vs. Lord of the Rings
Pepsi vs. Coke, the true taste test
A billboard saying "There is no God and life is still great" is offensive?
Is atheism a taboo in the USA?
Should .50 Cals be allowed in warfare?
Pencils vs. Pens
Should the Government allow NAZI rallies in neighborhoods where Holocaust survivors live?
Pronunciation: The letter Z, 'Zed' or 'Zee"?
Would you be more disappointed to find out that your child cheated on a test or smoked a cigarette?
How much should I tip the pizza man for my 20$ lunch order?
Cellphones While Driving
Smoking should be banned?
Should we judge motives or actions?

Table 1: Examples of debate topics

A debate can contain multiple posts from several users. We use each debate as a training sample of a dialogue for the argument system, where two consecutive posts are served as a quote and response pair.

To build the conversational argument system,

[1] http://www.convinceme.net/

we employ two typical approaches: a retrieval and a generative method.

3.3 The retrieval-based system

In the retrieval-based system, the task is to select the most relevant response given a user's message and the context of the conversation. While limited in the sense that it cannot generate new responses, retrieval-based systems are still often selected as a base method for many applications including summarisation, tasked-oriented dialogue systems. The aim of the system is to learn how to select the best argument from a pre-defined topic that matches the current user's response and the history of the conversation. The architecture of the system is depicted in Figure 1.

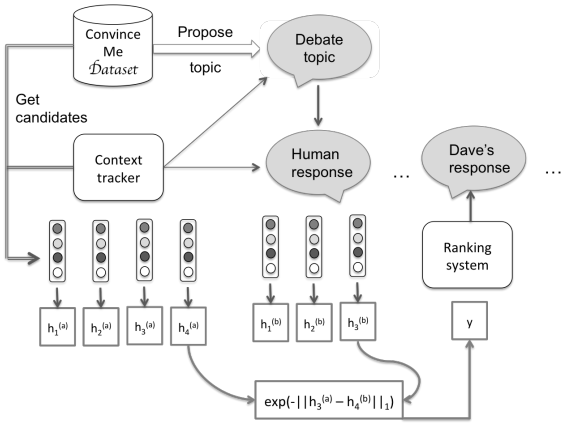

Figure 1: The debater retrieval system architecture

We use a siamese adaption of the LSTM network, which is called the Manhattan LSTM (MaLSTM) model (Mueller and Thyagarajan, 2016) to learn the similarity representation of two given messages. Common approaches usually use neural networks to represent sentences whose word vectors are trained on large corpora (Mikolov et al., 2013b; Le and Mikolov, 2014). The MaLSTM model on the other hand learns the semantic similarity directly with its representation learning objective function. It is reported to achieve state-of-the-art results on the task of assessing semantic similarity between sentences (Mueller and Thyagarajan, 2016).

The model is composed of two networks LSTM_a and LSTM_b, where the weights of these two networks are shared. The first LSTM_a represents candidate responses taken from the dataset, while the second LSTM_b represents the current human's response. A context tracker helps keeping track of which responses have been retrieved before to avoid repetition. The current user's message is compared to all candidates from the given debate to find the most relevant one r^k_{top}, where k is the index of the response in the dataset. Finally, r^{k+1}_{top}, the next response of r_{top} is selected to become Dave's response. Based on this approach, the first response r^1 will never be selected. To avoid irrelevant responses (when the user's message is not similar to any of the posts in the debate), we set a similarity threshold τ. For cases when the system cannot find a response that is similar (i.e., similarity value $\mathcal{S} < \tau$), the system will select the first post to return since the first post is usually the most general one that describes the topic of the debate. After a new response is selected, the context tracker will add the response to the context and only reset it when all responses have been achieved to promote diversity in the whole conversation.

For similarity metric, we use the simple function $\mathcal{S}(h^{(a)}_{T_a}, h^{(b)}_{T_b}) = \exp(-||h^{(a)}_{T_a} - h^{(b)}_{T_b}||_1)$ where $h^{(a)}_{T_a}$ and $h^{(b)}_{T_b}$ are representation of posts and user's messages respectively. The similarity value $\mathcal{S} \in [0, 1]$. l_1 norm is used in the similarity function instead of l_2 in order to avoid the problem of correcting errors in early stages due to vanishing gradients of the Euclidean distance (Chopra et al., 2005). It has also been reported to perform slightly better than other metrics such as cosine similarity (Mueller and Thyagarajan, 2016; Yih et al., 2011).

To train the MaLSTM, one needs to have a parallel corpus with similarity annotation between pairs of sentences. Unfortunately, there is no such corpus that is directly representing posts' similarities in debates and is large enough for training. We therefore use the Quora question pairs Kaggle competition dataset[2] which contains 404,302 question pairs annotated with similarity information (i.e., whether they are having the same meaning or not). Examples of questions in the training set is given in Table 2. This dataset has an open domain with questions covering many topics, which are suitable to be applied to our online post similarity assessment task. As can be seen from Table 2, computing similarity between sentences requires more than just word/word meaning matching. A similarity classifier should be able to do reasoning and take into account the struc-

[2] https://www.kaggle.com/c/quora-question-pairs

ture of the sentences. For the embedding layer, we use the pre-trained word2vec of Google News dataset[3](Mikolov et al., 2013a).

Questions that are equal
How can I be a good geologist? What should I do to be a great geologist?
How do I read and find my YouTube comments? How can I see all my Youtube comments?
What can make Physics easy to learn? How can you make physics easy to learn?
Questions that are not equal
What are the types of immunity? What are the different types of immunity in our body?
What is abstract expressionism in painting? What are major influences of abstract expressionism?
Why do girls want to be friends with the guy they reject? How do guys feel after rejecting a girl?

Table 2: Examples of Quora questions used for training the MaLSTM for the retrieval-based approach

3.4 The generative system

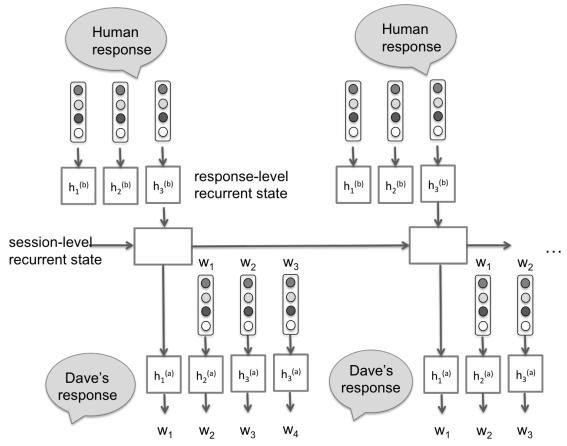

Figure 2: The debater generative system architecture

Although retrieval-based method is straightforward and guarantee to produce high quality messages, it is limited to only arguments that are available in the dataset and cannot adapt or tailor to every new responses from the users. A common trend in dialogue system community is to push towards generative models, where they are able to generate new messages based on the context and/or current states of the conversation (Vinyals and Le, 2015; Li et al., 2016a; Zhou et al., 2017).

Debating is different from normal open-domain conversation: argumentative responses may present attributes such as emotion, agreement, disagreement, sarcasm and stance.

[3]https://code.google.com/archive/p/word2vec/

To study if an end-to-end model could generate such responses, we use a hierarchical recurrent (RNN) encoder-decoder architecture as depicted in Figure 2. The original hierarchical RNN was introduced in (Sordoni et al., 2015) for the task of generating context-aware query suggestion for search engines. Its model attempts to capture the context of user queries based on sessions and sample suggestion one word at a time.

Applying to the task of generating debater responses, this architecture could take into account previous users' responses and is context sensitive. The order of messages in history is captured and encoded in a session-level recurrent state and the current response is represented in a response level recurrent state.

A given topic is treated as the first message starting a conversation. When the user submits the first response, it is fed into a bidirectional RNN (Jain and Medsker, 1999), in our case using GRU cells (Cho et al., 2014). Each word in the response is embedded using the pre-trained word embeddings. The encoder RNN then updates its internal vector, the *response-level* recurrent state. To capture the context of the previous messages in the dialogue and condition the next response generation based on the context, the *session-level* recurrent state is updated using another RNN on top of the previously computed current response-level encoder. This therefore forms a hierarchical architecture that could be able to capture the deep dialog context together with the current response encoder.

Given a set of responses $\mathcal{R} = \{r^1, r^2, ..., r^M\}$ where M is the number of responses in the given session and the responses are submitted in a chronological order. Each response is represented by a set of words $r^m = \{w_1^m, w_2^m, ..., w_{N_m}^m\}$, where N_m is the total number of words in that response.

Response-level encoder. For each word w_n, the response-level recurrent encoder state $h_{(enc),n}^m$ is computed based on the previous state and the current word:

$$h_{(enc),n}^m = g_{(enc)}(h_{(enc),n-1}^m, w_n^m) \quad (1)$$

The first initial state h_0 is set to 0. $h_{(enc),n}^m$ stores information about the current response r_m and word w_n^m.

Session-level encoder. In the session-level encoder, we encode the context of the previous re-

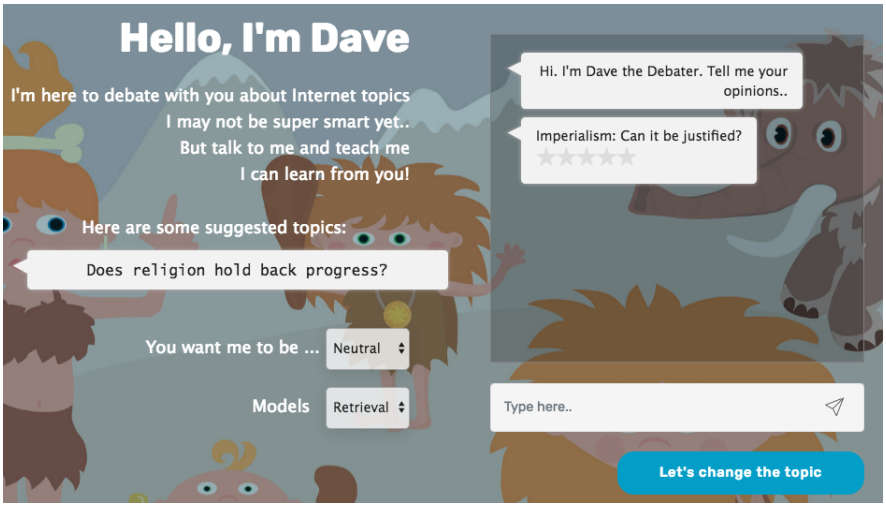

Figure 3: Dave the debater web demo

sponses in state c

$$c^m_{(enc)} = g_{(enc)}(c^{m-1}_{(enc)}, r^m) \quad (2)$$

The session level encoder sums up the context of all responses that the network has seen so far in a chronological order. It additionally builds up on the response vector r^m.

Response decoder. In this model, the response is sampled one word at a time. In particular, the prediction of the next response r^m based on the context $r^{1:m-1}$ is based on the estimation of the probability:

$$P(r^m|r^{1:m-1}) = \prod_{n=1}^{N_m}(w_n|w_{1:n-1}, r^{1:m-1}) \quad (3)$$

The current state d^m_n of the decoder is computed using another GRU:

$$d^m_{(dec),n} = g_{(dec)}(d^m_{(dec),n-1}, w^m_n) \quad (4)$$

To embed the context information into the decoder space d^m, we initialize the first recurrent state d^m_0 using a tanh function:

$$d^m_{(dec),0} = \tanh(D_0 c^{m-1} + b_0) \quad (5)$$

where D_0 projects the context summary vector c^{m-1} into the decoder space and b_0 is a bias vector.

Finally, the probability of a word w^m_n takes the form u is computed based on the previous words and given context as:

$$P(w^m_n = u|w^m_{1:n-1}, r^{1:m-1}) = \frac{\exp(e^T_u f(d^m_{n-1}, w^m_{n-1}))}{\sum_k \exp(e^T_k f(d^m_{n-1}, w^m_{n-1}))}$$

where e_u and e_k are the word embeddings of word u and k; f is the function that is computed based on both response-level and session-level states, similar to those used in (Sordoni et al., 2015; Cho et al., 2014).

$$f(d^m_{n-1}, w^m_{n-1}) = H_0 d^m_{n-1} + E_0 w^m_{n-1} + b_0 \quad (6)$$

Objective function. In this framework, we use the maximum mutual information (MMI) as proposed in (Li et al., 2016a) instead of the tradition likelihood function. As reported in (Li et al., 2016a), MMI objective function helps produce more diverse and interesting responses.

The likelihood objective function is computed as:

$$r^* = \arg\max_r \{\log P(r|r^{1:m})\} \quad (7)$$

while the MMI objective function is defined as:

$$r^* = \arg\max_r \{\log P(r|r^{1:m}) - \log P(r)\} \quad (8)$$

Generation and reranking. We use sampling method, where each word is sampled based on the output distribution. The results are finally reranked based on the log likelihood or the MMI score.

4 Web Demo for Dave the debater

A demo of the system described in the previous sections is available at http://114.212.80.16:8000/debate/. The interface is illustrated in Figure 3. A topic is randomly selected for a user, he or she can start talking about this topic by

giving own arguments. The user can also select to switch to a new debate topic with the button "Let's change the topic".

There are two different modes depending on which debater system the user wants to talk to. In the retrieval mode, each topic has a pre-defined list of responses. Note that these responses are ordered chronically. The selected responses are chosen based on the comparison of the user's current message and the history of the debate. In particular, the message from the user is being compared to every responses in the pre-defined set and the response coming after the most similar one is going to be selected. We keep track of all responses that have been retrieved to not get repeated responses again. More details about the approach is explained in section 3.3.

For the generative mode, the user can talk about the given topic or any other topics. The responses are generated based on the current message (encoded in the response-level recurrent state) and the context, which is the history of the whole conversation from the beginning (encoded in the session-level recurrent state) (section 3.4).

To collect feedbacks for each response to retrain the models, users are able to rate each response with rating from 1 (very bad) to 5 (very good). These ratings are stored in the database together with the responses and their contexts to use in the future so the models can learn from them.

5 Experiments and results

In this section, we report the first evaluation results we have so far for both retrieval-based and generative argumentative dialogue system.

5.1 Retrieval-based model

We evaluate the MaLSTM as describe in section 3.3 for the similarity assessment task on the Quora dataset. As we do not have directly gold standard data for similarity evaluation performing on the argumentation dataset, we use the Quora as a measurement of the similarity assessment for the MaLSTM. We used 50 hidden layers, word embedding length of 300, batch size of 2048, dropout rate of 30%. For the loss function, we used the mean square error and Adam Optimizer (Kingma and Ba, 2014). After several experiments, we observe that the model accuracy and loss converge at around 50 epochs. The model accuracy and loss on the train and validation dataset are plotted in Figure 4. The final result obtained on the test set is 81.92%.

Topic: *Is anything in this world free?* **Human:** Nothing in this world is completely free, there must always be a cost somewhere **Dave:** Nothing in this world is free.. Everybody has to pay a price for anything they want or recieve.. Life is not free.. Death is not free.. Happiness is not free.. No matter what it is, there will always be a toll.. Nothing in this world is free.. **Human:** I totally agree with you **Dave:** Close, there is no energy cost to giving up life, entropy takes care of that.. But there is an energy cost to having life to give up.. I was running under the assumption that "free" was referring to currency, not energy.. If we're going to argue that energy is a currency for the purposes of this debate, then there's really no debate to be had at all.. **Human:** Happiness is free **Dave:** I'm sorry to ruin your fun, but "picking up a penny on the street" is MOST definitely not free, the effort to bend down and grab that penny costs you energy loss etc.. of course on minuscule level but still causes some sort of work.. In general for one thing to be free another has to lose true, BUT there is still never anything free for the one that is receiving the "gift", even the portion of receiving is some sort of exchange! **Human:** how about dead? **Dave:** Death comes with many expense's there is a energy cost, you have to give up life to gain death, Death can come at the price of pain..

Table 3: Examples of a conversation in the retrieval mode

5.2 Generative model

Evaluation metric. To evaluate the responses from the generative model, following (Vinyals and Le, 2015; Li et al., 2016a; Zhou et al., 2017), we use the perplexity metric PP evaluated on the test data:

$$PP = 2^{-\sum_y \log P(y)} \quad (9)$$

Usually the lower the perplexity the better the model is, and more specialized corpus also often obtains lower perplexity.

To measure the degree of diversity of responses, we use the *distinct-1* and *distinct-2* metrics following (Li et al., 2016a). They are the number of distinct unigrams and bigrams in generated responses, scaled by total number of tokens.

Settings. In our experiments, for both encoder and decoder, the number of recurrent layers is set to two, the number of dimensions for the recurrent layer is 512 and the drop-out ratio is 0.2. We use the batch size of 192, the Adadelta method for adapting learning rate (Zeiler, 2012).

For the decoder, we examine two methods: the Sampling method, in which responses are sampled from output distribution token by token. For each

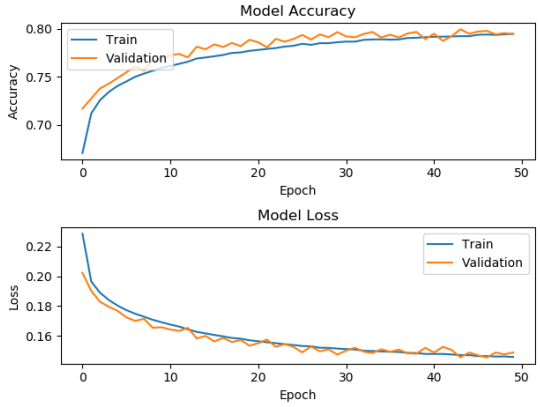

Figure 4: Retrieval-based model: MaLSTM results

token, the temperature transform is performed prior to sampling. Generally, the low temperature sample is more grammatically correct while the high temperature sample represents greater linguistic variety. In our experiments, we set the default temperature to 0.5. The second model is Sampling with MMI re-ranking. In this model, the sampling procedure is done the same as the other one, but candidates are ranked according to their MMI-score (see section 3.4).

Human: I think guns should be banned **Dave:** So, if you want to ban guns, then they're not completely legitimate. I don't think that the government should start a legal system..
Topic: There are no good arguments for the non-existence of God. **Human:** sure, I believe in God. **Dave:** I am not a christian but there are so many of the bible.. The bible has written to it, and if a dog do have the same in the existence of god, then that would be the only true gods..
Human: it does not make much sense what you are saying **Dave:** I think they are..
Topic: evolution: scientific or unscientific? **Human:** scientific should be better **Dave:** Well i'm not to say he was the first person who would have been on the same - life **Human:** alright.. **Dave:** You are a good debater!
Human: which one would you pick? **Dave:** I am a fan of both..

Table 4: Random context-sensitive examples in the generative mode

Finally, to prevent the model from generating the same tokens over and over again, we use a penalized coefficient method, in which tokens that have been generated before will have their probabilities lower down by a factor of 10.0. Tokens that are not penalized including modifiers (such as *a*, *an*, *the*) and punctuations.

Testing data. For testing the perplexity of the model and the distinctness metrics, we prepare two datasets: a context free dataset that contains 200 argumentative sentences and a context sensitive dataset containing 200 consecutive argumentative responses from an online debate.

The results are reported in table 5. The context sensitive perplexity is higher than the context free perplexity. In both cases, the reranking method helps reducing the perplexity of the model by 13% (in context sensitive case) and 3.1% (in context free case). Reranking also helps increasing the diversity of the responses: the *distinct-1* increased by 3.8% and the *distinct-2* increased by 13.16%.

While these metrics may not be the most useful in evaluating conversational systems, they are most widely used metrics and could somehow give a reflection on how different models perform.

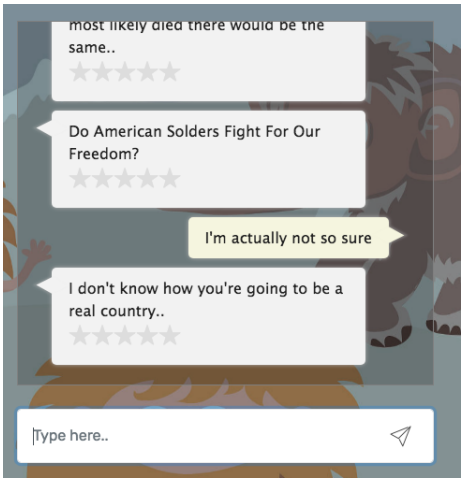

Figure 5: Generative model on random army topic: white bubbles (left) are responses from Dave

6 Conclusion and future work

In this paper, we have described an argumentative dialogue agent, whose aim is to be able to debate with human on a given topic. We explored two approaches, using a retrieval-based and a generative system. The systems have been trained on a limited open-domain dataset, but have shown interesting and promising results. Still there is a lot of work that can be done to improve the system, including training on a much larger dataset, combining both retrieving and generating methods alternatively to give interesting responses to the users based on different scenarios. For the retrieval-

Model	Sampling	Sampling & MMI re-ranking
Context-sensitive Perplexity	88.51	76.97 (-13.0%)
Context-free Perplexity	75.53	73.17 (-3.1%)
Distinctness *distinct-1*	0.708 %	0.736 (+3.8%)
Distinctness *distinct-2*	5.94%	6.84% (+13.16%)

Table 5: Perplexity and distinctness for the sampling method and sampling with MMI re-ranking method

Figure 6: Generative model on random gay/society topic: white bubbles (left) are responses from Dave

based system, one can try the un-tied version of the Manhattan LSTM, since responses could vary in length and may not be symmetric. In the generative system, different decoding methods could be applied such as a traditional beam search, sampling output based on topics, increasing the depth and power of the model. One can also integrate argument strategies as those described in (Rosenfeld and Kraus, 2016) to the generative system to have a more structural and persuasive conversation. Such system can put the first milestones in developing a machine that can someday fully engage in a debate and discussion with human on controversial topics.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

David Bamman and Noah A. Smith. 2015. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85, Lisbon, Portugal. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the First Workshop on Argumentation Mining*, Denver, Colorado, USA. Association for Computational Linguistics.

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *ArXiv e-prints*, abs/1406.1078.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 539–546, Washington, DC, USA. IEEE Computer Society.

Mihail Eric and Christopher D. Manning. 2017. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. *CoRR*, abs/1701.04024.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122. Association for Computational Linguistics.

L. C. Jain and L. R. Medsker. 1999. *Recurrent Neural Networks: Design and Applications*, 1st edition. CRC Press, Inc., Boca Raton, FL, USA.

Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Know.-Based Syst.*, 69(1):124–133.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. Cite arxiv:1603.06155Comment: Accepted for publication at ACL 2016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2786–2792. AAAI Press.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.

Niklas Rach, Saskia Langhammer, Wolfgang Minker, and Stefan Ultes. 2018. Utilizing argument mining techniques for argumentative dialogue systems. In *Proceedings of the 9th International Workshop On Spoken Dialogue Systems (IWSDS)*.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. *CoRR*, abs/1709.03167.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *ECAI*, volume 285 of *Frontiers in Artificial Intelligence and Applications*, pages 320–328. IOS Press.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3776–3783. AAAI Press.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 553–562, New York, NY, USA. ACM.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Comput. Linguist.*, 43(3):619–659.

Reid Swanson, Stephanie M. Lukin, Luke Eisenberg, Thomas Chase Corcoran, and Marilyn A. Walker. 2017. Getting reliable annotations for sarcasm in online dialogues. *CoRR*, abs/1709.01042.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012a. A corpus for research on deliberation and debate. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig H. Martell, and Joseph King. 2012b. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53:719–729.

Jason D. Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *CoRR*, abs/1702.03274.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 247–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *CoRR*, abs/1704.01074.