Legalbot: a Deep Learning-Based Conversational Agent in the Legal Domain

Adebayo Kolawole John, Luigi Di Caro, Livio Robaldo, and Guido Boella

Dipartimento di Informatica, Universita Di Torino
Corso Svizzera 185, Torino, 10149, Italy
{kolawolejohn.adebayo}@unibo.it, {dicaro,guido.boella}@di.unito.it,

Abstract. This paper present a deep learning-based dialogue system which has been trained to answer user queries posed as questions during conversation. The proposed system, though generative, takes advantage of domain specific knowledge for generating valid answers. The evaluation analysis shows that the proposed system obtained a promising result.

Keywords: Recurrent Neural Networks, Long Short-Term Memory, Chatbot, Conversational agent

1 Introduction

Dialogue Systems (DS), a.k.a. Conversational Systems (CS), have been a subject of research since mid-60's (cf. [17]). In the domain of DS, since the work of [11], exciting results have been reported by systems which model human conversation and response with Neural Networks (NN) systems [16, 14, 13].

Modeling human conversation is quite challenging since it involves generating plausible and intelligible response to a message giving some contexts. Conversational systems can be either of 1) retrieval-based [7, 6] and 2) machine translation-inspired generative systems [4, 1, 16, 14]. Retrieval-based systems use a repository of predefined responses and some kind of heuristic to pick an appropriate response based on the input and context. These type of systems benefit from some clearly defined templates which could be used to limit the potential responses from which to search from. Thus, they tend to give coherent responses since they are less prone to making linguistic or grammatical mistakes [13].

Our work is more challenging since the text have longer sequences, compared to the short text employed in [13, 16]. Also, our work is close to IR-based response ranking systems, except that ours is generative. Furthermore, with the exception of the work of [16], most CS systems are trained on open-end data. Because of this, the conversation tends to be off point most of the time. We introduce a more challenging corpus curated from law textbooks, providing prose answers to some basic legal questions specifically within the US jurisdiction. Thus, we deal with messages with longer sequences while also being domain specific. We picture a machine legal advisor that is able to inform, advice and educate its users on some basic every-day legal issues. Giving our limited data, we show that our system works and that it's capable of scaling better with more training data. Our system, henceforth *legalbot*, benefits from the sequence-to-sequence (Seq2Seq) model [15] integrated with special attention scheme in order to focus on some important information between the context and response pair.

2 Conversational Systems

A conversation is made up of the context, message and response. The context is an aggregation of previous query and responses, while the message is the query which leads to the response. Conversational systems (CS) could lean towards the short text conversation [13] or the long text conversation. Our work is inspired by the recurrent neural networks (RNNs) machine translation approach in [1, 4]. The authors used an encoder to transform the input into a high dimensional vector representation and then use a decoder to generate a translation from the input representation. The work described in [16] is consistent with this approach, also benefiting from *Seq2Seq* [15]. [14] incorporate word embeddings in order to capture long range dependency. Word embeddings in particular have shown to capture more semantic information with excellent result on many NLP tasks [9]. A copy-based attention RNNs was employed in [5]. The model was trained to focus on important information to transfer from the message to the response. Li et. al., [8] used the Long Short-Term Memory (LSTM) to automatically mine user information about entities in the dialogue. An intent-based attention RNN was introduced in [18], using three RNNs, each for the message, response and intention, they keep track of the structural knowledge of the conversation process. Specifically, our work is close to [18] since we also introduce an attention scheme for modeling an *intense-focus* between important words appearing in the message as well as the response. We use a variant of RNNs, -the LSTM, which is more robust to vanishing gradient problem while having the ability to retain information over longer time steps.

At each time step t, let an LSTM unit be a collection of vectors in $\mathbb{R}^d$ where d is the memory dimension: an *input gate* i_t, a *forget gate* f_t, an *output gate* o_t, a *memory cell* c_t and a *hidden state* h_t. The LSTM transition can be represented with the following equations (x_t is the an input vector at t, σ represents sigmoid activation function and $\odot$ the elementwise multiplication. The u_t is a tanh layer which creates a vector of new candidate values that could be added to the state):

$$
\begin{aligned}
i_t &= \sigma\Big(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}\Big),\\
f_t &= \sigma\Big(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}\Big),\\
o_t &= \sigma\Big(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}\Big),\\
u_t &= \tanh\Big(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}\Big),\\
c_t &= i_t \odot u_t + f_t \odot c_{t-1},\\
h_t &= o_t \odot \tanh c_t
\end{aligned}
\tag{1}
$$

Giving an input symbol in a sequence, the LSTM learns a probability distribution by being trained to predict the next symbol. Normally, the input X is paired with some outputs $Y = \mathrm{y}_1, \mathrm{y}_2,...,\mathrm{yN}_y$. The softmax function is used to distribute the probability over the outputs as below:

$$p(Y|X) = \prod_{k=1}^{N_y} p(y_k|x_1, x_2, ..., x_t, y_1, y_2, ..., y_{k-1})$$

$$= \prod_{k=1}^{N_y} \frac{exp((h_{t-1}, e_{yt})}{\sum_{y'} exp(f(h_{t-1}, e_{y'}))} \quad (2)$$

where f(h_{t-1}, e_{yt}) is the activation function between h_{t-1} and e_{yt}). h_{t-1} is the hidden state at the time step $t-1$.

Enquirer	Am I legally obligated to answer a police officer's questions?
Legal Advisor	No. Refusing to answer a police officer's questions is not a crime. The Fifth Amendment to the U.S. Constitution guarantees the right to silence.
Enquirer	So in that case, I can easily walk away from being questioned
Legal Advisor	Yes you can. Unless a police officer has probable cause to make an arrest, or a reasonable suspicion to conduct a stop and frisk, you have the legal right to walk away from a police officer.

Fig. 1. Sample Conversation From our Dataset

Original Enquiry	Am I legally obligated to answer a police officer's questions?
Test sample formatted	Am I legally obliged to allow a police officer interrogate me
Original Enquiry	Given this kind of situation, how do you advise that I present my case?
Test sample formatted	Given this kind of situation, what advise do you give on how to talk to the policeman?

Fig. 2. Context conversion for test samples.

2.1 Encoder-Attention-Decoder

Given a set of conversation between two or more persons, we denote all the conversation as $D = C_1, C_2....C_m$ where each C_i is a sequence of some tokens. At each time step, the encoder reads the embedding vector $\overrightarrow{\boldsymbol{x_t}}$ forward and converts each C_i into a fixed high dimensional representation. Equations (3) and (4) shows the forward and backward context being computed by the non-linearity function as a recurrence. The representation is taken as the final hidden state h_T (see equation (5)) value computed by merging the forward and backward context.

$$\overrightarrow{\boldsymbol{h_t}}^{forward} = f(x_t, h_{t-1}) \quad (3)$$

$$\overleftarrow{\boldsymbol{h_t}}^{backward} = f(x_{t-1}, h_{t-1}) \quad (4)$$

$$h_T = concat(\overrightarrow{\boldsymbol{h_t}}^{forward}, \overleftarrow{\boldsymbol{h_t}}^{backward}) \quad (5)$$

where the function f(,) in this work is a LSTM block. To instatiate our network, we used *Glove* vectors [10]. This is consistent with the approach employed by [14]. Out-of-Vocabulary (OOV) Words were assigned embedding weights initialized from random Gaussian distribution. Our choice of *Glove* is because it was trained on a huge data and it is semantically rich. Throughout the training, the weights of the embeddings remain fixed. Next, we feed the ensuing representation into another LSTM for the *intense-focus*. For clarity, the *intense focus* is achieved with a bi-directional LSTM which searches the input sequence and create attention for the important words. In our data, the message is usually of short sequences when compared to the response which can be of arbitrarily longer sequences, our idea is to identify the most semantically important words and by looking through the sequence both left and right, we capture again the relationship of this important words with their context. The decoder is also another LSTM block with the hidden state computed similarly to equation (5) except that it is is non-bidirectional. The representation from the hidden state h_T of the *intense focus* is passed to the decoder which predicts each response/target symbol. The conditional probability is distributed by a non-linear function, in this case *softmax*. Our model maximizes the conditional log likelihood where θ is the parameter to be learned. Instead of translating the target language to the source as usually done in machine translation [1] tasks, we instead predict the likelihood of a response word, given a query word.

$$\frac{\max}{\theta} \frac{1}{N} \sum_{i=1}^{N} logp\theta(y_n|x_n) \tag{6}$$

3 Dataset

Our intention is to develop a machine advisor that is able to give simple legal advice to users. A user may want to know whether (s)he has the rights to ignore a police interrogation, for instance, when no crime has been committed. We develop a dataset which has been curated from some online law textbooks about criminal law, property rights, family, divorce and company rights. Even though the data follow the question-answer pattern in our source books, the sequences of progression in the conversation makes it suitable for a conversational system. This is more so since it exhibits a kind of phenomenon which we called *context loop*, i.e., a question leads to an answer, the answer given also leads to another question which leads to another answer and vice versa. Where necessary, we have manually adjust the tone of the conversation in order to reflect true dialog while still preserving the information it serves to pass across. We have a total of 1200 question-answer pairs formatted into dialogues,yielding a total of 2400 message and response samples. Figure 1 shows a sample conversation from the dataset.

4 Training

We trained our model with the log-likelihood function. We used Keras[1] deep learning library. To avoid any imprecision in evaluation, we randomly select

[1] `https://github.com/fchollet/keras`

50 context from the dataset and for each, we created identical sentences as queries. We optimized the log likelihood with ADAM optimizer. Our best model was achieved with the following configuration: batch size=32, epochs=700, and hidden state size=400. We stick to our surface model since we observed no real improvement while stacking LSTMs in order to have an increased depth.

5 Evaluation

There are different metrics often used in evaluating dialogue systems. A prominent example is the perplexity score as used in [12]. However, this metric doesn't suffice in our case. Following the work of Vinyals et al., in [16], we employed human judgment as our choice evaluation approach. The goal is to present a question, the gold-standard response as well as the machine's response to 3 human judges. Each judge has to score a response using three scales of measure, i.e., acceptable (assigned score=1), borderline (assigned score =0.5) and unacceptable (assigned score=0.0). The total maximum score per judge equals the total number of test samples and the maximum score obtainable is a product of the number of judges and samples. Our accuracy is obtained by a simple formula given in equation 7. Our assessment aims at observing the level of acceptance of the response from our model by human users.

$$Acceptance = \frac{TotalScoreObtained}{MaximumSCoreObainable} * 100 \tag{7}$$

The three judges were each assigned the 50 questions to score using our acceptance metric. Using the formula in 7, our system achieved a score of 0.48 which implies that roughly 24 out of 50 questions were within the range borderline or acceptable. Few things might have contributed to the poor performance of our model. As earlier highlighted, the sentences in the response are rather too long (66 tokens on the average) which is not so for the question which averages 15 tokens per question. Also, we have limited amount of data since it requires significant manual effort in formatting information from our sources in order to look quite conversational. Nevertheless, the result is promising and with potential for improvement with huge amount of data. Sadly, these kind of data is quite scarce in the legal domain especially with the stringent copyright issues.

6 Conclusion

We have presented legalbot, a *Seq2Seq* conversational agent. Our model was trained on a micro dataset curated from question-answer information on some civil legal issues. Our neural network model achieved an acceptance score of 48% as evaluated by 3 human judges. The result is promising considering the nature and size of the data. Going forward, we would like to increase the samples in our data. Also, we are thinking of streamlining the number of sentences per response. For empirical evaluation, we would like to implement some existing systems and test their performance on our data while also testing our approach on available short-text conversation datasets. We plan in our future works to apply the approach presented here to our systems in legal informatics [3], [2].

Acknowledgments. Kolawole J. Adebayo has received funding from the Erasmus Mundus Joint International Doctoral (Ph.D.) programme in Law, Science and Technology. Luigi Di Caro have received funding from the European Union's H2020 research and innovation programme under the grant agreement No 690974 for the project "MIREL: MIning and REasoning with Legal texts".

References

1. D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
2. G. Boella, L. Di Caro, M. Graziadei, L. Cupi, C. Salaroglio, L. Humphreys, H. Konstantinov, K. Marko, L. Robaldo, C. Ruffini, K. Simov, A. Violato, and V. Stroetmann. Linking legal open data: Breaking the accessibility and language barrier in european legislation and case law. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 171–175. ACM, 2015.
3. G. Boella, L. Di Caro, and L. Robaldo. *Semantic Relation Extraction from Legislative Text Using Generalized Syntactic Dependencies and Support Vector Machines.* Springer Berlin Heidelberg, 2013.
4. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
5. J. Gu, Z. Lu, H. Li, and V. Li. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*, 2016.
6. Z. Ji, Z. Lu, and H. Li. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988*, 2014.
7. R. Kadlec, M. Schmid, and J. Kleindienst. Improved deep learning baselines for ubuntu corpus dialogs. *arXiv preprint arXiv:1510.03753*, 2015.
8. J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*, 2016.
9. T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
10. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
11. A. Ritter, C. Cherry, and W. Dolan. Data-driven response generation in social media. In *Proc. of Empirical Methods in Natural Language Processing*, 2011.
12. I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. *arXiv preprint arXiv:1507.04808*, 2015.
13. L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*, 2015.
14. A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*, 2015.
15. I. Sutskever, O. Vinyals, and Q. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 2014.
16. O. Vinyals and Q. Le. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.
17. J. Weizenbaum. Eliza: a computer program for the study of nl communication between man and machine. *Communications of the ACM*, 9(1), 1966.
18. K. Yao, G. Zweig, and B. Peng. Attention with intention for a neural network conversation model. *arXiv preprint arXiv:1510.08565*, 2015.