Dialogue Systems using Sequence-to-sequence models for Spanish

Joel Oswaldo Gallegos Guillen

Universidad Catolica San Pablo, Arequipa

Abstract. Nowadays the natural language study is indispensable for building conversational models. These models allow realizing several interesting applications in different study fields with the Recurrent Neural Networks help. In this way, there are different approaches to made conversational models, the traditional approaches require a rather complicated pipeline of many stages and they have a specific domain knowledge, thus they are restricted. Under this background, this paper proposes a conversational model based on the sequence to sequence approach. This approach is able to cope the disadvantage of conventional approaches because it requires fewer stages and it is not limited or restricted a specific domain knowledge.

Our model has the ability to answer to customer queries in Spanish. This model was trained end-to-end of neural network with a dataset based on questions and answers from a bank sent by chat. The results have indicated a good performance to get proper knowledge for simple questions both with a specific domain dataset and general domain dataset.t.

Keywords: *Conversational model, recurrent neural networks, sequence to sequence approach.*

1 Introduction

Currently, computer techniques perform an important role for studying natural language processing. The natural language requires computer vision techniques for the purpose of learning, understanding and producing human language content according to a given previous knowledge.

Recently research advances indicate neural networks have a good performance in many fields such as pattern recognition, speech recognition, and natural language processing. In this way, end-to-end training neural networks have the ability to serve in more complex works than the classification process. They can be used to map complicated structures to other complicated structures with fewer features than other conventional methods, one example of this is the natural language understanding [1].

Conversational models have been studied in the last years in order to fit them according to a previous knowledge based on answers and questions between people. Therefore, it requires a mapping between queries and responses. Due to the complexity of this mapping, conversational modeling has previously been designed to have a very restricted domain. This paper, purpose the use an end-to-end training neural network using the sequence to sequence approach in order to develop a chat session in the Spanish language. The tests have been-made using a dataset in Spanish of customer inquiries made by chat to a bank. The results show good performance at simple and basic conversations.

2 Related Works

Several studies about the natural language processing have been proposed over time, so early computational approaches to language research focused on automating the analysis of the linguistic structure of language and developing basic technologies such as machine translation, speech recognition, and speech synthesis [2] [3]. However, natural language conversation is one of the most challenging artificial intelligence problems, which involves language understanding, reasoning, and the using of common sense knowledge [4].

In the last years, Deep Learning have been refined and they have been used for real-world applications, creating a conversational language or better known as chatbots. In this way, [5]presents a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure using a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. The result is that on an English to French translation task from the WMT-14 dataset, the translations produced by the LSTM achieve a BLEU score of 34.8.

For building a chatbot according to [6] the recurrent neural networks have good performance and consistency in a general domain providing useful answers to the user. Also, this research indicates recurrent neural networks obtain better perplexity compared to the n-gram model and capture important long-range correlations. Other research of great impact proposes Neural Responding Machine (NRM), a neural network-based response generator for Short-Text Conversation. It formalizes the generation of response as a decoding process based on the latent representation of the input text, while both encoding and decoding are realized with recurrent neural networks (RNN). Empirical study shows that NRM can generate grammatically correct and content-wise appropriate responses to over 75% of the input text [4].

Another interesting research is based on videos, [7] proposes to translate videos directly to sentences using a unified deep neural network with both convolutional and recurrent structure getting good performance in their results. Finally, one of the most recent works is presented by [8]. It presents MILABOT: a deep reinforcement learning chatbot developed by the Montreal Institute for Learning Algorithms (MILA) for the Amazon Alexa Prize competition. MILABOT is capable of conversing with humans on popular small talk topics through both speech and text. The system consists of an ensemble of natural language generation and retrieval models, including neural network and template-based models.

3 Concepts

3.1 Recurrent Neural Network

A recurrent neural network (RNN) models an input sequence of tokens $w_1, ..., w_N$ using the recurrence:

$$h_n = f(h_{n-1}, w_n) \tag{1}$$

Where $h_n \in \Re^{d_h}$ is called a recurrent, or *hidden*, state and acts as a vector representation of the tokens seen up to position n. In particular, the last state h_N may be viewed as an order-sensitive compact summary of all the tokens. In language modeling tasks, the context information encoded in h_n is used to predict the next token in the sentence:

$$P_\theta(W_{n+1} = v | w \leq n) = \frac{exp(g(h_n, v))}{\sum_{v'} exp(g(h_n, v'))} \tag{2}$$

The functions f and g are typically defined as:

$$f(h_{n-1}, w_n) = tanh(Hh_{n-1} + I_{w_n}) \tag{3}$$

$$g(h_n, v) = O_{w_n}^T h_n \tag{4}$$

The matrix $I \in \Re^{d_h x|V|}$ — contains the input word embeddings, i.e. each column I_j is a vector corresponding to token j in the vocabulary V . Due to the size of the model vocabulary V , it is common to approximate the I matrix with a low-rank decomposition, i.e. $I = XE$, where $X \in \Re^{d_n x d_e}$ and $E \in \Re^{d_e x|V|}$, and $d_e < d_h$. This approach has also the advantage that the embedding matrix E may separately be bootstrapped (e.g. learned) from larger corpora. Analogously, the matrix $O \in \Re^{d_h x|V|}$ represents the output word embeddings, where each possible next token is projected into another dense vector and compared to the hidden state h_n. The probability of seeing token v at position $n + 1$ increases if its corresponding embedding vector O_v is near the context vector h_n. The parameter H is called a *recurrent* parameter, because it links h_{n-1} to h_n. All parameters are learned by maximizing the log-likelihood of the parameters on a training set using stochastic gradient descent.

H is usually an elementwise application of a sigmoid function. However we have found that the Long Short-Term Memory (LSTM) architecture, which uses purpose-built memory cells to store information, is better at finding and exploiting long range context. For the version of LSTM, H is implemented by the following composite function:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{5}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{6}$$

$$c_t = f_t c_{t-1} + i_t tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{7}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{8}$$

$$h_t = o_t tanh(c_t) \tag{9}$$

where σ is the logistic sigmoid function, and i, f, o and c are respectively the *input gate, forget gate, output gate* and *cell* activation vectors, all of which are the same size as the hidden vector h. The weight matrix subscripts have the obvious meaning, for example W_{hi} is the hidden-input gate matrix, W_{xo} is the input-output gate matrix etc. The weight matrices from the cell to gate vectors (e.g. Wci) are diagonal, so element m in each gate vector only receives input from element m of the cell vector. The bias terms (which are added to i, f, c and o) have been omitted for clarity [9] [10] [11].

4 Proposal

This research proposes to building a conversational model using recurrent neural networks based on end-to-end training. The conversational model better knows as chatbot has been trained using a dataset from the Spanish language. For this purpose, we have used a sequence-to-sequence model from Tensorflow take into account the encoder-decoder idea.

The dataset was obtained from chats, this dataset is the queries, claims, among others of customers of a bank. In total, 16225 sets of questions and answers were obtained, giving us 34426 words total of vocabulary, of which 32161 are only words and 2265 are numbers..

The Figure 1 shows a brief flowchart of the research:

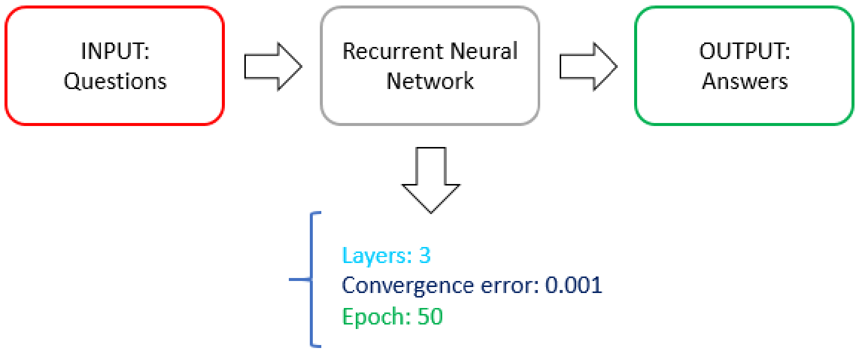

Fig. 1. Research flowchart

5 Experimental Results

In this section, we can found the results and some examples of the conventional conversations realized with the chatbot.

In the experiments, several comparisons were made, where depending on the number of words in the phrase, the questions and answers of the dataset are accepted or not. With small sentences there is a more concrete answer, but the amount of the data set decreases a lot; and with quite long sentences, we have a higher percentage of the dataset, but the answers have no meaning with the questions. The following table shows the amount of the dataset reduced depending on the number of words in the phrase:

Number of Words	Reduced Data
25	80%
35	64%
50	45%
200	3%

Table 1. Reduced Data

The figure 2, 3, 4 shows an example of one of the conversations with the chatbot.

		With Numbers		Without Numbers	
Size accepted	Size Stored	With meaning	Incomplete answer	With meaning	Incomplete answer
25	25	yes	no	yes	no
35	50	yes	yes/no	yes	yes/no
50	50	yes	yes/no	yes	yes/no
50	50	yes	no	yes	no
200	200	yes/no	no	yes/no	no

Table 2. Reduced Data

```
 Query > amigos de ibk les he dejado un mensaje por inbox para que me
ayuden
 Answers> que unk gracias saber hola nos comisión vez xq la manera

 Query > pueden responder es urgente
 Answers> que unk por con banca el respuesta de para dni tu

 Query > Me han robado, que puedo hacer
 Answers> que unk nos señores de

 Query > la banca movil no funciona
Answers > que unk atención encargada hola teléfono lo por con banca
para

 Query > mi tarjeta esta bloqueado
 Answers> que unk a señores gracias encargada ha amplíanos para pe
saludos
```

Fig. 2. Conversation example.

```
Query > amigos de ibk les he dejado un mensaje por inbox para que me
ayuden
> hola unk continuaremos con tu atención por dicho medio quedamos a
> hola unk continuaremos con tu atención por dicho medio saludos

Query > pueden responder es urgente
> hola unk te comentamos que estamos atendiendo tu consulta vía inbox
> hola unk te comentamos que hemos enviado tu consulta al área

Query > Me han robado, que puedo hacer
> hola unk por favor amplíanos tu consulta a través de nuestro
> hola unk por favor envíanos el número de tu dni y

Query > la banca movil no funciona
> hola unk sentimos el malestar generado asimismo te informamos que
nuestra
> hola unk te comentamos que nuestra página web está funcionando con

Query > mi tarjeta esta bloqueado
> hola jose por favor envíanos un inbox con tu dni y
> hola unk por favor envíanos el detalle de tu consulta vía
```

Fig. 3. Conversation example.

```
Query > amigos de ibk les he dejado un mensaje por inbox para que me
ayuden
> hola julio continuaremos tu atención por dicho medio quedamos a tu
> hola julio continuaremos con tu atención por dicho medio quedamos a

Query > pueden responder es urgente
> hola jose te hemos respondido por inbox aguardamos tu respuesta
martin
> hola jose estamos atendiendo tu consulta vía inbox gracias

Query > Me han robado, que puedo hacer
> hola marco por favor envíanos el número de tu dni vía
> hola marco por favor amplíanos tu comentario para asesorarte
estaremos atentos

Query > la banca movil no funciona
> hola jose sentimos lo sucedido estamos trabajando para solucionar el
inconveniente
> hola jose sentimos lo que comentas por favor envíanos el datos

Query > mi tarjeta esta bloqueado
> hola luis por favor envíanos tu comentario vía el dni para
> hola jose por favor amplíanos tu comentario para asesorarte saludos
angie
```

Fig. 4. Conversation example.

6 Conclusion

According to our research, we can conclude that recurrent neural networks based on training end-to-end have good performance in simple and basic conversations in the Spanish language. However, for better results is indispensable have a greater amount data of different contexts. Therefore, we purpose like future research the creation of a dataset from different contexts and improve our results by the application of a better robust deep learning method that allows a fluent conversation using speech recognition.

References

1. Sak, H., Senior, A., Rao, K., Beaufays, F.: Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947 (2015)
2. Hirschberg, J., Manning, C.D.: Advances in natural language processing. Science **349**(6245) (2015) 261–266
3. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (icassp), 2013 ieee international conference on, IEEE (2013) 6645–6649
4. Shang, L., Lu, Z., Li, H.: Neural responding machine for short-text conversation. arXiv preprint arXiv:1503.02364 (2015)
5. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. (2014) 3104–3112

6. Vinyals, O., Le, Q.: A neural conversational model. arXiv preprint arXiv:1506.05869 (2015)
7. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)
8. Burgess, A.: Ai in action. In: The Executive Guide to Artificial Intelligence. Springer (2018) 73–89
9. Graves, A., Jaitly, N.: Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning. (2014) 1764–1772
10. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A.C., Pineau, J.: Building end-to-end dialogue systems using generative hierarchical neural network models. In: AAAI. Volume 16. (2016) 3776–3784
11. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015)