MŰEGYETEM 1782

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Automation and Applied Informatics

Deep Learning Based Chatbot Models

Scientific Students' Associations Report

Author:
Richárd Krisztián Csáky

Supervised by
Gábor Recski

2017

Kivonat

A konverzációs ágens (chatbot) egy olyan program, mely természetes nyelvet használva képes emberekkel kommunikálni. A beszélgetés modellezése fontos feladat a természetes nyelvfeldolgozás és mesterséges intelligencia (MI) területén. Az MI tudományág megszületése óta egy jól működő chatbot létrehozása még mindig az egyik legnehezebb kihívás. A chatbotok sokféle feladatra használhatók, de mindegyik esetében elvárt, hogy megértsék a felhasználó mondandóját és az adott problémához releváns válaszokat generáljanak.

A múlt chatbot architektúrái kézi szabályokra és sablonokra, vagy egyszerű statisztikai módszerekre támaszkodtak. 2015 óta, a mélytanulás (deep learning) elterjedésével ezek a modellek gyorsan felcserélődtek elejétől végéig tanítható neurális hálózatokkal. Manapság a rekurrens enkóder-dekóder modell [Cho et al., 2014] dominál a konverzáció modellezésben. Ezt az architektúrát a neurális gépi fordítás területéről adaptálták, ahol rendkívül jó eredményeket ért el. Azóta sokféle változata [Serban et al., 2016] és kiegészítése született annak érdekében, hogy minél jobb minőségű legyen a chatbotok által folytatott beszélgetés.

Munkám során részletes irodalmi kutatást végeztem, melyben az elmúlt 3 évben publikált, több mint 70, a chatbotokkal kapcsolatos publikációt vizsgálok meg. Ezután amellett érvelek, hogy a konverzáció modellezés sajátosságai a jelenlegi state-of-the-art architektúráktól eltérő megközelítést igényelnek. Szakirodalmi példákon alapulva bemutatom, hogy a jelenlegi chatbot modellek miért nem vesznek figyelembe elég ún. priort a válasz generálása során, és ez hogyan befolyásolja a beszélgetés minőségét. Ezek a priorok olyan külső információt hordoznak, melyen a beszélgetés kondicionálva lehet, mint például a beszélők személye [Li et al., 2016a] vagy hangulata. Amellett, hogy bemutatom az okait, javaslatokat is teszek a probléma orvoslására.

A dolgozat következő részében egy nemrég bemutatott modellt, mely jelenleg state-of-the-art-nak számít a neurális gépi fordításban, az úgynevezett Transformer-t [Vaswani et al., 2017] adaptálom a beszélgetés-modellezés feladatára. Először az eredeti cikkben leírt modell tanításával kísérletezek, tanítóadatként a Cornell Movie-Dialog Corpus [Danescu-Niculescu-Mizil and Lee, 2011] dialógusait használva. Emellett továbbfejlesztem a modellt saját, az enkóder-dekóder architektúra hiányainak orvoslására született ötletekkel. További priorokat adok bemenetként a modellbe, mint a beszélgetők személye vagy hangulata. Végül korábbi chatbot modellekkel való összehasonlítás útján részletes elemzést végzek arról, hogy az eredeti modell mennyire teljesít jól dialógus adattal és hogyan befolyásolják a generált válaszok minőségét az általam implementált további kiegészítések.

Abstract

A conversational agent (chatbot) is a piece of software that is able to communicate with humans using natural language. Modelling conversation is an important task in natural language processing and artificial intelligence (AI). Indeed, ever since the birth of AI, creating a good chatbot remains one of the field's hardest challenges. While chatbots can be used for various tasks, in general they have to understand users' utterances and provide responses that are relevant to the problem at hand.

In the past, methods for constructing chatbot architectures have relied on hand-written rules and templates or simple statistical methods. With the rise of deep learning these models were quickly replaced by end-to-end trainable neural networks around 2015. More specifically, the recurrent encoder-decoder model [Cho et al., 2014] dominates the task of conversational modelling. This architecture was adapted from the neural machine translation domain, where it performs extremely well. Since then a multitude of variations [Serban et al., 2016] and features were presented that augment the quality of the conversation that chatbots are capable of.

In my work, I conduct an in-depth survey of recent literature, examining over 70 publications related to chatbots published in the last 3 years. Then I proceed to make the argument that the very nature of the general conversation domain demands approaches that are different from current state-of-the-art architectures. Based on several examples from the literature I show why current chatbot models fail to take into account enough priors when generating responses and how this affects the quality of the conversation. In the case of chatbots these priors can be outside sources of information that the conversation is conditioned on like the persona [Li et al., 2016a] or mood of the conversers. In addition to presenting the reasons behind this problem, I propose several ideas on how it could be remedied.

The next section of my paper focuses on adapting the very recent Tranformer [Vaswani et al., 2017] model to the chatbot domain, which is currently the state-of-the-art in neural machine translation. I first present my experiments with the vanilla model, using conversations extracted from the Cornell Movie-Dialog Corpus [Danescu-Niculescu-Mizil and Lee, 2011]. Secondly, I augment the model with some of my ideas regarding the issues of encoder-decoder architectures. More specifically, I feed additional features into the model like mood or persona together with the raw conversation data. Finally, I conduct a detailed analysis of how the vanilla model performs on conversational data by comparing it to previous chatbot models and how the additional features, affect the quality of the generated responses.

Contents

1 Introduction

A conversational agent (chatbot) is a piece of software that is able to communicate with humans using natural language. Ever since the birth of AI, modelling conversations remains one of the field's toughest challenges. Even though they are far from perfect, chatbots are now used in a plethora of applications like Apple's Siri [Apple, 2017], Google's Google Assistant [Google, 2017] or Microsoft's Cortana [Microsoft, 2017a]. In order to fully understand the capabilities and limitations of current chatbot architectures and techniques I conduct an in-depth survey, where I examine related literature published over the past 3 years and I implement my own chatbot based on a novel neural network model.

The paper begins with a brief overview of the history of chatbots in Section 2, where I discuss the properties and objectives of conversational modelling. I present early approaches and the current dominating model based on neural networks for building conversational agents.

In the following section I present key architectures and techniques that were developed over the past 3 years relating to chatbots. I group publications into 11 groups based on the specific techniques or approaches that are discussed by the authors. After this I present criticism regarding some of the properties of current chatbot models and I show how several of the techniques used are inappropriate for the task of modelling conversations.

In the next section I conduct experiments by training a novel neural network model, the Transformer [Vaswani et al., 2017] using dialog datasets [Danescu-Niculescu-Mizil and Lee, 2011, Tiedemann, 2009]. I run several trainings using these datasets detailed in Section 4. I present the results of the different training setups by qualitatively comparing them to previous chatbot models and by using standard evaluation metrics.

In the final section before concluding I offer possible directions for future work. More specifically, I propose several ideas in order to remedy the problems discussed in Section 3.2.

2 History of Chatbots

2.1 Modelling Conversations

Chatbot models usually take in as input natural language sentences uttered by the user, and output a response. There are two main approaches for generating responses. The traditional approach is to use hard-coded templates and rules to make chatbots, which I present in Section 2.2. The more novel approach, which I discuss in detail in Section 2.3 was made possible by the rise of deep learning. Neural network models are trained on large amounts of data to learn the process of generating relevant and grammatically correct responses to input utterances. Models have also been developed to accommodate for spoken or visual inputs. They oftentimes make use of a speech recognition component to transform speech into text [Serban et al., 2017] or convolutional neural networks that transform the input pictures into useful representations for the chatbot [Havrylov and Titov, 2017]. The latter models are also called visual dialog agents, where the conversation is grounded on both textual and visual input [Das et al., 2017].

Conversational agents exist in two main forms. The first one is the more traditional task-oriented dialog system, which is limited in its conversational capabilities, however it is very robust at executing task specific commands and requirements. Task-oriented models are built to accomplish a specific task like making restaurant reservations [Joshi et al., 2017, Bordes et al., 2016] or promoting movies [Yu et al., 2017], just to name a few. These systems often don't have the ability to respond to arbitrary utterances since they are limited to a specific domain, thus users have to be guided by the dialog system towards the task at hand. Usually they are deployed to tasks where some information has to be retrieved from a knowledge base. They are mainly used to replace the process of navigating through menus and user interfaces like making the process of booking flight tickets or finding a public transportation route between two locations conversational [Zhao et al., 2017].

The second type of dialog agents are the non-task or open-domain chatbots. These conversation systems try to imitate human dialog in all its facets. This means that one should hardly be able to distinguish such a chatbot from a real human, but current models are still far away from such claims. These models are usually trained with conversation examples extracted from movie scripts or from Twitter-like post-reply pairs [Vinyals and Le, 2015, Shang et al., 2015, Serban et al., 2016, Li et al., 2016a]. For these models there isn't a well defined goal, but they are required to have a certain amount of world knowledge and commonsense in order to hold conversations about basically anything.

Recently an emphasis has been put on integrating the two types of conversational agents. The main idea is to combine the positive aspects of both types, like the robust abilities of goal-oriented dialog systems to perform tasks and the human-like chattyness of open-domain chatbots [Zhao et al., 2017, Yu et al., 2017, Serban et al., 2017]. This is beneficial because the user is more likely to engage with a task-oriented dialog agent if its more human-like, and handles out of domain responses well.

2.2 Early Approaches

ELIZA is one of the first ever chatbot programs written [Weizenbaum, 1966]. It uses clever hand-written templates to generate a reply that resembles the user's input utterance. Since then countless hand-coded rule-based chatbots have been written [Wallace, 2009, Carpenter, 2017, Worswick, 2017]. Furthermore, a number of programming frameworks specifically designed to facilitate building dialog agents have been developed [Marietto et al., 2013, Microsoft, 2017b].

These chatbot programs are very similar in their core, namely that they all use hand-written rules to generate replies. Usually simple pattern matching or keyword retrieval techniques are employed to handle the user's input utterance. Then rules are used to transform a matching pattern or a keyword into a predefined reply.

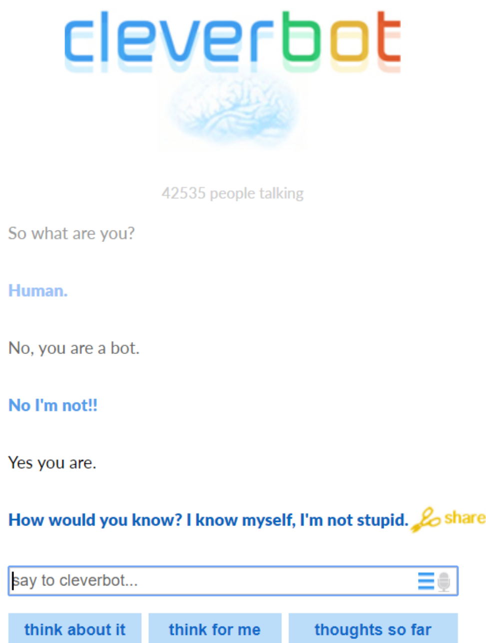

Figure 1: Sample conversation with Cleverbot [Carpenter, 2017]

2.3 The Encoder-Decoder Model

The main concept that differentiates rule-based and neural network based approaches is the presence of a learning algorithm in the latter case. Instead of using hand-written rules deep learning models transform input sentences into replies directly by using matrix multiplications and non-linearities over millions of parameters. We can further divide neural network based models into two categories, retrieval-based and generative models. The former simply returns a reply from the dataset by computing the most likely response to the current input utterance based on a scoring function, which can be implemented as a neural network [Cho et al., 2014] or by simply computing the cosine similarity between the input utterance and candidate replies [Li et al., 2016b]. Generative models on the other hand synthesize the reply one word at a time by computing probabilities over the whole vocabulary [Sutskever et al., 2014, Vinyals and Le, 2015]. There have also been approaches that integrate the two types of dialog systems by comparing a generated reply with a retrieved reply and determining which one is more likely [Song et al., 2016].

As with many other applications the field of conversational modelling has been transformed by the rise of deep learning. More specifically the encoder-decoder recurrent neural network (RNN) model (also called seq2seq [Sutskever et al., 2014]) introduced by [Cho et al., 2014] and its variations have been dominating the field. This model was originally developed for neural machine translation, but it was found to be suitable to *translate* source utterances into responses within a conversational setting [Shang et al., 2015, Vinyals and Le, 2015].

3 Background

3.1 Recent Chatbot Architectures and Augmentations

3.1.1 Attention

3.1.2 Context

3.1.3 Objective Functions

3.1.4 Additional Features

3.1.5 Knowledge Bases, Copying and Information Retrieval

3.1.6 Task-Oriented Approaches

3.1.7 Decoding and Beam Search

3.1.8 Reinforcement Learning

3.1.9 Pretraining

3.1.10 Additional Encoder-Decoder Models

3.1.11 Evaluation Methods

3.2 Criticism

3.2.1 Datasets

3.2.2 Loss Function

3.2.3 Memory

3.2.4 Evaluation Metrics

3.3 Summary

4 Experiments

4.1 The Tranformer Model

4.2 Datasets

5 Results

6 Future Work

7 Conclusion

References

[Apple, 2017] Apple (2017). Siri. `https://www.apple.com/ios/siri/`. Accessed: 2017-10-04.

[Bordes et al., 2016] Bordes, A., Boureau, Y.-L., and Weston, J. (2016). Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

[Carpenter, 2017] Carpenter, R. (2017). Cleverbot. `http://www.cleverbot.com/`. Accessed: 2017-10-04.

[Cho et al., 2014] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[Danescu-Niculescu-Mizil and Lee, 2011] Danescu-Niculescu-Mizil, C. and Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

[Das et al., 2017] Das, A., Kottur, S., Moura, J. M., Lee, S., and Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*.

[Google, 2017] Google (2017). Google assistant. `https://assistant.google.com/`. Accessed: 2017-10-04.

[Havrylov and Titov, 2017] Havrylov, S. and Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *arXiv preprint arXiv:1705.11192*.

[Joshi et al., 2017] Joshi, C. K., Mi, F., and Faltings, B. (2017). Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

[Li et al., 2016a] Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., and Dolan, B. (2016a). A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

[Li et al., 2016b] Li, X., Mou, L., Yan, R., and Zhang, M. (2016b). Stalematebreaker: A proactive content-introducing approach to automatic human-computer conversation. *arXiv preprint arXiv:1604.04358*.

[Marietto et al., 2013] Marietto, M. d. G. B., de Aguiar, R. V., Barbosa, G. d. O., Botelho, W. T., Pimentel, E., França, R. d. S., and da Silva, V. L. (2013). Artificial intelligence markup language: A brief tutorial. *arXiv preprint arXiv:1307.3091*.

[Microsoft, 2017a] Microsoft (2017a). Cortana. `https://www.microsoft.com/en-us/windows/cortana`. Accessed: 2017-10-04.

[Microsoft, 2017b] Microsoft (2017b). Microsoft bot framework. `https://dev.botframework.com/`. Accessed: 2017-10-04.

[Serban et al., 2017] Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., Kim, T., Pieper, M., Chandar, S., Ke, N. R., et al. (2017). A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

[Serban et al., 2016] Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

[Shang et al., 2015] Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

[Song et al., 2016] Song, Y., Yan, R., Li, X., Zhao, D., and Zhang, M. (2016). Two are better than one: An ensemble of retrieval-and generation-based dialog systems. *arXiv preprint arXiv:1610.07149*.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

[Tiedemann, 2009] Tiedemann, J. (2009). News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[Vinyals and Le, 2015] Vinyals, O. and Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

[Wallace, 2009] Wallace, R. S. (2009). The anatomy of alice. *Parsing the Turing Test*, pages 181–210.

[Weizenbaum, 1966] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

[Worswick, 2017] Worswick, S. (2017). Mitsuku. `http://www.mitsuku.com/`. Accessed: 2017-10-04.

[Yu et al., 2017] Yu, Z., Black, A. W., and Rudnicky, A. I. (2017). Learning conversational systems that interleave task and non-task content. *arXiv preprint arXiv:1703.00099*.

[Zhao et al., 2017] Zhao, T., Lu, A., Lee, K., and Eskenazi, M. (2017). Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. *arXiv preprint arXiv:1706.08476*.