Topic modelling for a virtual librarian assistant tool

Linnea Strand[1]
gusstrlip@student.gu.se

Robert Rhys Thomas[1]
gusroberth@student.gu.se

Simon Dobnik[2]
simon.dobnik@gu.se

Mehdi Ghanimifard[2]
mehdi.ghanimifard@gu.se

[1]FLoV
University of Gothenburg, Sweden

[2]CLASP and FLoV
University of Gothenburg, Sweden

Abstract

We outline a topic classifier for a virtual librarian assistant trained on a corpus of conversations between patrons and librarians. Because this conversation data is limited, we explore the possibility of training the system on a domain of artificial data generated from distilled versions of real chats and applying transfer learning techniques to augment learning on real data. By training the model on artificial dialogues and fine-tuning it on real data, we hope to achieve better overall performance for the classifier even when little data is available. This abstract describes the data generation (augmentation) process, the model architecture and presents some preliminary results.

1 Introduction

In the interaction between librarians and patrons there is an increasing need for language technology tools that would provide assistance to librarians, especially when the interaction is split between different modalities that include on-site personal conversations, conversations on the phone, email and chat as in Göteborgs stadsbibliotek (SB, Gothenburg City Library). One of the tools that was identified useful in this domain is a topic modelling system that detects a topic of the conversation in written chats and can then support a rule-based dialogue system or routes the conversation to a human librarian if the topic is better handled by them. A similar idea has been applied in dialogue systems as in (Smith et al., 2011) where topic modelling has be achieved through application of different data-driven techniques (Pulman et al., 2010) that supported a rule-based system.

The main aim of this project is to train a topic classification system that detects the topic of a patron conversation with a librarian. However, a major drawback of applying data-driven learning techniques is that they require a large number of domain-specific conversations. Within the project with SB a dataset of 1,600 conversations have been collected to date which were annotated with topics by a domain expert. While this is not a negligible amount of dialogue data for a novel dialogue domain, in terms of the data-driven techniques the size of the corpus is still limited. To address this limitation we introduce a novel way of expanding the dataset through artificial dialogue generation. To achieve this, we distil rules from the real data and generate additional samples from it. Effectively, this is an integration of a rule-based approach with a data-driven approach. The rules are translated into data on which the system is pretrained. This provides biases encoded in the rule to the system in the forms of distributions in the data upon which the system can then fine-tune real examples. If our experiment(s) are successful, this approach could increase the chances of companies and other organisations to bootstrap robust dialogue systems without having access to large amounts of domain-specific data.

2 Data Collection

The application which the current library chat function is built with is LibraryH3lp.[1] The software works as an instant messaging platform and contains functionality that allows librarians to tag each conversation. The tags were introduced and refined incrementally by a domain expert and senior librarian who identified 34 topics. However, further data collection and annotation may introduce new topics or lead to a refinement of the existing topics. It is important to note that each conversation may be assigned several tags, as a conversation may be about more than one topic, e.g. *app Libby* or *suggestions, opinions and questions*. These tags are what the topic classifier algorithm will be trained to predict in a multi-class classification task. So far, a total 1,600 chat transcripts have been collected over the

[1] https://libraryh3lp.com/

Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue, July 18-19, 2020, Online, hosted from Massachusetts, USA.

course of 6 months. Each chat is labelled with one or more of the 34 available topics that have been chosen by the domain expert.

3 Distilling and Generating

The process consists of two steps: (i) the real dialogues are first distilled into a set of templates, and then (ii) the templates are expanded back into several examples that represent our artificial dataset. (Larsson et al., 2000) proposes guidelines for applying distilling techniques for dialogue to templates. We ensure that the artificial transcripts remain true to the real samples but differ enough to provide features that may be important when the data is fed through the machine learning model. Distillation is achieved by splitting transcripts into utterances and rephrasing them. From the distilled templates we can generate multiple dialogues that follow the same basic structure as the real examples by varying certain slots in the templates (examples in Appendix A). The generations are limited to how many phrases are added to the templates and to the variations of the slots. We also apply simple data augmentation techniques such as Random Deletion and Synonym Substitution (Wei and Zou, 2019), (Zhang et al., 2015), (Wang and Yang, 2015). The augmentation techniques allow for up to around 10,000 unique examples per tag.

4 The Model

We employ a simple, two-step neural architecture to classify the topic(s). First, each dialogue – represented as a list of utterances – is encoded using the pre-trained Swedish BERT model[2] which is based on BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019) and is trained over 200 million sentences over various sources. Each embedded utterance is then passed through a max pooling layer before being concatenated with its speaker vector (1 for patron, 0 for librarian) and fed into an LSTM cell. Finally, each hidden state is fed into a linear classification layer, thus predicting the topic(s) at every time step[3] (see Figure 1).

5 Evaluation

We compare the performance of the model trained on real dialogues (169), generated dialogues (2908 dialogues, 3000 with duplicate generations removed) and mixed dialogues containing a combined bag of the two on three categories: *app Libby*, *app PressReader*, and *suggestions*. The generated dialogues have one target topic whereas real data may have several target topics (multi-class prediction). The held-out test set contains of 23 real dialogues. We train the model for 20 epochs. While the system can predict the topic incrementally for parts of dialogue, we evaluate the system for the entire dialogue by comparing *loss* on the test data for different models. We will report the figures in the final version of this paper.

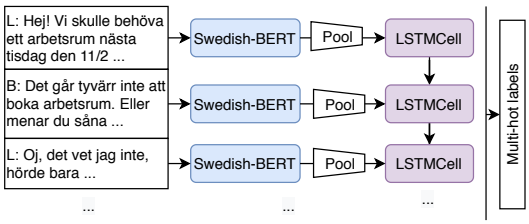

Figure 1: The model diagram.

6 Discussion and Conclusions

In this paper, a method for creating of training data has been introduced that is inspired by dialogue distillation techniques from dialogue system design coupled with data augmentation techniques. Additionally, we also investigate how we could benefit from transfer learning in contextual word embeddings such as BERT. We train a model of dialogue topic classification using this data and pre-trained knowledge and test it on real data. So far the experiments have given mixed results. We hope that the dialogue generation approach can be applicable in practically any domain, not just for a library chat bot. Furthermore experiments in future could explore different curriculum learning techniques on augmented data to be able to maximise the benefits of transferring knowledge from other domains. Future work includes improving our data generation method, evaluating the naturalness of the generated dialogues and including other forms of pre-trained knowledge.

Acknowledgements

The research of Dobnik and Ghanimifard reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

[2] https://github.com/Kungbib/swedish-bert-models

[3] One variation of the system is designed for incremental topic classification, but evaluations are done on dialogue classification.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv*, arXiv:1909.11942 [cs.CL].

Staffan Larsson, Lena Santamarta, and Arne Jönsson. 2000. Using the process of distilling dialogues to understand dialogue systems. In *Proceedings of ICSLP 2000*, pages 374–377.

Stephen G Pulman, Nigel Crook, Ramon Granell, Simon Dobnik, and Manjari Chandran-Ramesh. 2010. Deliverable D4.2.3: Report on topic detection with Dirichlet Process Priors. Technical report, Department of Computer Science, Oxford University, Oxford, United Kingdom.

Cameron Smith, Nigel Crook, Simon Dobnik, Daniel Charlton, Johan Boye, Stephen Pulman, Raul Santos de la Camara, Markku Turunen, David Benyon, Jay Bradley, Björn Gambäck, Preben Hansen, Oli Mival, Nick Webb, and Marc Cavazza. 2011. Interaction strategies for an affective conversational agent. *Presence: Teleoperators and Virtual Environments*, 20(5):395–411.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal. Association for Computational Linguistics.

Jason W Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv*, arXiv:1901.11196 [cs.CL].

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

A Appendices

A real transcript
L: hej! vad behöver du hjälp med? **U:** hej är studieplatserna fortfarande öppna så man kan sitta och plugga några timmar? **L:** nej tyvärr **U:** gäller det alla bibliotek i Göteborg eller bara stadsbiblioteket? **L:** det kan jag inte svara på jag tror att detsamma gäller alla bibliotek **U:** Tack
Artificially generated transcript
L: hej! vad behöver du hjälp med? **U:** hej! är studieplatserna fortfarande öppna så man kan sitta och plugga några timmar? **L:** det är dom tyvärr inte **U:** är det bara stadsbiblioteket eller alla bibliotek? **L:** det kan jag inte svara på jag tror att det är begränsat på de flesta bibliotek **U:** tack så mycket
A real transcript (translated to English)
L: hello! what do you need help with? **U:** hey, the study sites are still open so you can sit and study for a few hours? **L:** no unfortunately **U:** applies to all libraries in Gothenburg or just the city library? **L:** I can't answer I believe that the same applies to all libraries **U:** Thanks
An artificially generated transcript translated to English
L: hello! what do you need help with? **U:** hello! are the study sites still open so you can sit and study for a few hours? **L:** unfortunately they are not **U:** is it just the city library or all libraries? **L:** I can't answer that, I think it's limited in most libraries **U:** Thank you so much

Table 1: An example of real and artificial transcripts.

Sample 1		
Grammar	Template	Utterance
what_help	<librarian> hello! what do you need help with?	- <librarian> hello! what do you need help with?
user	- <user>	
hello	- hello - hi - hi! - hello!	
study_places	- are the study places still open to sit and study for a few hours? - is it possible to sit and work at the study places? - are the study places open? - do you still have study places?	<user> hello! are the study sites still open so you can sit and study for a few hours?
librarian	- <librarian>	
no	- no i'm afraid - unfortunately not - they are not unfortunately - i'm sorry but no - unfortunately they are not	<librarian> Unfortunately they are not
user	- <user>	
all_lib	- is it all libraries in Gothenburg or just the city library? - is this the same for all libraries? - is it just the city library or all libraries?	<user> is it just the city library or all libraries?
librarian	- <librarian>	
cant_say	- i'm afraid I can't say - I can not answer - I am not sure	
probably	- i think it is limited in most libraries - i believe it may be the same in all libraries - i think it is the same in all libraries	<librarian> I can't answer that, I think it's limited in most libraries
user	<user>	
thanks	- thank you so much - thank you - thanks a lot - thanks	<user> thank you so much

Table 2: Template generation