University of Regensburg
Institute for Language, Literature and Cultural Studies
Chair of Media Informatics

Exposé for a bachelor thesis on the topic:

CookBERT – Cooking with BERT

Domain-adaptive pre-training of BERT on cooking conversation data

Submitted by:

Pascal Strobel
Gluckstraße 3
93053 Regensburg
E-Mail: pascal.strobel@stud.uni-regensburg.de

Matriculation number: 2106133
Bachelor thesis in Media Informatics
7. Fachsemester
Supervisor: Mr. Alexander Frummet (M. Sc.)
First reviewer: Prof. Dr. Udo Kruschwitz
Second reviewer: PD Dr. David Elsweiler

Datum: 26.11.2021

Table of contents

Problem definition and state of research

The introduction of large-scale pre-trained transformers, such as BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019) and ELMo (Peters et al., 2018) has led to a small revolution in the field of NLP. Such language models (LMs) are pre-trained on a massive amount of unlabeled text data with self-supervised objectives, e. g. masked language modelling and next sentence prediction (Devlin et al., 2018), and achieve state-of-the-art performance on a wide variety of tasks, including text classification (Sun et al., 2019), named entity recognition (Symeonidou et al., 2019) and machine translation (Zhu et al., 2020). However, the pre-training data mostly originates from general topics and domains, e.g. Wikipedia articles and book corpus (Devlin et al., 2018), leading to limited knowledge of the LM about domain-specific vocabulary or text structures (like dialogues).

To counteract this lack of domain knowledge, different approaches can be found in the literature, which basically all try to do so by feeding domain- or task-specific data into the model, be it by pre-training a model from scratch (Beltagy et al., 2019), or by further pre-training an already existing model (Araci, 2019; Lee et al., 2020; Wang et al., 2021). Not surprisingly, such approaches generally outperform the standard pre-trained models when applied to tasks in their domain.

While meanwhile an extensive number of domain-specific language models exists, there is not yet a model that is geared towards cooking conversation data and thus is a candidate for a conversational agent for the kitchen.

Objective and planned procedure

The main goal of this work is to answer the following research question:

> *„How does domain-adaptive pre-training of BERT on cooking conversation data affect the performance of downstream tasks relevant to conversational agents in the kitchen?“*

To answer this question, an already pre-trained LM (BERT) is enriched with domain-specific knowledge about cooking conversations. This is done via domain-adaptive pre-training, which is nothing but further pre-training the model on domain-specific data on the two target objectives: masked language modeling and/ or next sentence prediction. Since no matching data is publicly available, a textual data set of unlabeled cooking conversations is first to be created.

The model is then finetuned and evaluated on (several) downstream tasks that are relevent for conversational agents in the kitchen, e. g. intent classification and named entity recognition.

➔ It is still unclear, how many (and what) tasks the model will be finetuned and tested on, since the data sets have to be suitable, i.e. of the cooking conversation domain

By comparing the performance on those tasks with the performance of other (base) models (that are directly finetuned and thus do not have the additional domain knowledge), the research question can then be answered.

➔ Comment: Pellegrini et al. (2021) trained on the Recipe1M dataset via domain-adaptive pre-training (i.e., on food data in general, and not on cooking conversation data as I will do). However, since this domain is quite similar to my target domain, a performance comparison of the two models would also be very interesting!

Summary of the objectives of the work:

- Creation of an unlabeled, textual cooking conversation data set
- Domain-adaptive pre-training of an existing LM (BERT)

- Finetuning and evaluating the model on relevant downstream tasks
- Model comparison and answering research question

It is expected that the proposed model with domain-specific knowledge will outperform existing, general models on all conversational agent relevant tasks.

Current state of progress

In addition to the literature review, the cooking conversation data for the domain-adaptive pre-training has already been gathered[1]. The data generally comes from two sources:

1. Subtitles of live cooking TV broadcasts (approx. 16 mb; however, some mb will be omitted during data cleanup)
2. Transcripts from cooking podcasts (approx. 14 mb; however, some mb will be omitted during data cleanup)

The subtitles are all taken from cooking broadcasts that are part of the ARD, a joint organization of Germany's regional public-service broadcasters and have almost all the same structure (see fig. 1): A chef and a host are in the kitchen together. The chef cooks a particular recipe, explains his procedure and gives tips, and the host asks questions about the recipe preparation, ingredients, etc.

Figure 1: Typical structure of the cooking broadcasts used, in this case taken from the show "ARD Buffet". (from https://www.ardmediathek.de/)

It should be noted that the subtitles do not correspond to an exact transcription of the wording used, which in a way contributes to the loss of the "naturalness" of the conversations. Instead, the subtitles follow a scheme established by ARD, for example:

[1] Data can be found in the "conversational_data_raw" folder at: https://github.com/paschistrobel/Cook-BERT

“Die Untertitel bestehen idealer Weise aus Hauptsätzen oder aus einem kurzen Haupt- und Nebensatz. Die Untertitel sind immer in direkter Rede verfasst. Es gibt aber dennoch keine Anführungszeichen für direkte Rede. […] Der Untertitel fasst einen knappen Gedanken auf zwei Zeilen zusammen, Sätze erstrecken sich nur in Ausnahmefällen über mehrere Untertitel.” (Erstes Deutsches Fernsehen [ARD], n.d.)

While the TV-show subtitles originate from conversations **during** the cooking process, this is not the case for the gathered podcast transcripts: they come from conversations **about** cooking and cooking-related topics, but they offer more naturalistic data, as the transcriptions were not simplified as much as those of the subtitles. Although the podcast data does not completely match the desired target data and may include some conversations on irrelevant topics, it is still considered useful because the language model learns both the characteristics of conversations, but also the use of cooking-specific vocabulary in the context of conversations.

The next step would be to prepare and clean the collected data (extract the plain text from PDFs, translate it into English, etc.).

Preliminary outline

1. Introduction (incl. research question)
2. Objectives
3. Related Work
4. Creation of a cooking conversation dataset
   - 4.1 Origin and properties of the gathered data
     - ➔ Where does the data come from (cooking tv shows, podcasts, …) and what kind of data is it?
   - 4.2 Preparation of the data
     - ➔ How was data prepared for the domain-adaptive pre-training?
   - 4.3 Advantages and limitations of the data
5. Explanation of the base model used (BERT)
   - ➔ Explain the general structure of BERT
6. Domain-adaptive pre-training
   - 6.1 Pre-training objectives
     - ➔ Masked language modelling and next sentence prediction
   - 6.2 Data preprocessing
     - ➔ Tokenizing etc.
   - 6.3 Pre-training specifications
     - ➔ What optimizer was used? How many training epochs? Handling of over-/underfitting?
7. Finetuning
   - 7.1 Task 1
     - ➔ Brief explanation of task (e. g. intent classification) and corresponding data set
   - 7.2 Task 2
   - 7.3 …
8. Evaluation
   - 8.1 Models used for comparison
   - 8.2 Comparison of model performances on downstream tasks
9. Conclusion
10. Summary

Rough time schedule

Available time: approx. 13 weeks (26.11.2021–01.03.2022), official registration of the bachelor thesis: beginning-mid January

- Already done: Literature research
- Already done: Data collection (for domain-adaptive pre-training)
- Until 03.12.: Data preparation and cleaning
- Until 10.12.: Familiarizing with huggingface/ FARM library
- Until 24.12.: Domain-adaptive pre-training of existing BERT model (via huggingface or FARM by deepset-ai)
- Until 14.01.: Finetuning on different downstream tasks
- Until 28.01.: Evaluation
- Until 18.02.: Writing (also already during the practical parts)
- Until 25.02: Test reading and finetuning
- Until 01.03.: Submission

Bibliography

Araci, D. (2019, August 27). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. http://arxiv.org/pdf/1908.10063v1

Beltagy, I., Lo Kyle, & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. http://arxiv.org/pdf/1903.10676v3

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/pdf/1810.04805v2

Erstes Deutsches Fernsehen. *Gestaltung von Untertiteln - ARD | Das Erste.* ARD. https://www.daserste.de/specials/service/gestaltung-untertitel100.html

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Pellegrini, C., Özsoy, E., Wintergerst, M., & Groh, G. (2021, February 11–13). Exploiting Food Embeddings for Ingredient Substitution. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies* (pp. 67–77). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0010202000670077

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018, February 15). *Deep contextualized word representations*. http://arxiv.org/pdf/1802.05365v2

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*(1(8)), 9.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019, May 14). *How to Fine-Tune BERT for Text Classification?* http://arxiv.org/pdf/1905.05583v3

Symeonidou, A., Sazonau, V., & Groth, P. (2019). Transfer Learning for Biomedical Named Entity Recognition with BioBERT. *SEMANTICS Posters&Demos*.

Wang, P., Fang, J., & Reinspach, J. (2021). CS-BERT: A pretrained model for customer service dialogues. *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, 130–142.

Zhu, J., Xia, Y., Wu, L., Di He, Qin, T., Zhou, W., Li, H., & Liu, T.-Y. (2020, February 17). *Incorporating BERT into Neural Machine Translation*. http://arxiv.org/pdf/2002.06823v1