Estimating Context-Specific Subjective Content Descriptions using BERT

Magnus Bender*, Felix Kuhr*, Tanya Braun† and Ralf Möller*

*University of Lübeck
Institute of Information Systems
Ratzeburger Allee 160, 23562 Lübeck
{bender, kuhr, moeller}@ifis.uni-luebeck.de

†University of Münster
Computer Science Department
Einsteinstr. 62, 48155 Münster
tanya.braun@uni-muenster.de

An agent in pursuit of a task may work with a corpus containing text documents associated with Subjective Content Descriptions (SCDs) [1]. SCDs provide additional location-specific data for documents and add value in the context of the agent's task. On the pursuit of new documents to add to the corpus, the agent may come across documents without associated SCDs or documents where *content* and SCDs are interleaved. Therefore, this paper presents approaches estimating SCDs using the well-known BERT [2] language model. Furthermore, the paper presents approaches separating SCDs and actual *content* given interleaved in a document.

An evaluation compares the performance of BERT with approaches from Kuhr et al. [1], [3], which use an SCD-word distribution represented by an SCD matrix $\delta(\mathcal{D})$. In this paper's evaluation we use text documents annotated with additional textual definitions.

We start by describing two problems from the field of SCDs to solve with BERT.

I. Inline Subjective Content Descriptions

The inline SCD (iSCD) problem [3] problem asks to separate SCDs and *content* given interleaved in a document d'. For each word of each document, the agent has then to decide whether the word is part of an SCD or belongs to the *content*. Formalized, the iSCD problem's input is a document d' and the output is the *content* $d \subseteq d'$ as sequence of words and a set of SCDs $g(d)$.

In our evaluation, the iSCD problem is a classification problem with two classes, namely SCD and *content*. We refer to the approach by Bender et al. with `Matrix iSCD`.

II. Most Probably Suited Subjective Content Descriptions

In the scenario of the Most Probably Suited Subjective Content Descriptions (MPS²CDs) problem [1], documents without SCDs shall be associated with SCDs. The MPS²CD problem asks for the M most probably suited SCDs $t_1, ..., t_M$ for a document d' given the SCD matrix $\delta(\mathcal{D})$. Solving the MPS²CD problem allows us to estimate SCDs from the set of SCDs known by $\delta(\mathcal{D})$ for each sentence in a document. We refer to the approach by Kuhr et al. with `Matrix MPS²CD`.

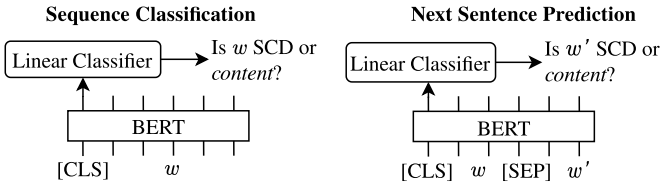

Figure 1. Use-cases of BERT solving the iSCD problem.

III. Bidirectional Encoder Representations from Transformers (BERT)

BERT, introduced by Devlin et al. [2], encodes a sequence of input tokens, i.e. words, into a sequence of vector representations. In the following, BERT is depicted as a box getting the input tokens at the bottom and yielding the encoded vector outputs at the top. The contribution of this paper is to solve the two previously described problems regarding SCD using BERT. For each of the two problems, we present how to apply two different use-cases of BERT.

A. Solving the iSCD Problem with BERT

In Figure 1, BERT's use-cases sequence classification and next sentence prediction are depicted already adjusted to SCDs.

1) `BERT Classify`*:* Using BERT's sequence classification to solve the iSCD problem is straightforward. For each sentence w as input sequence the encoded representation is calculated. From the encoded representation, only the first output vector is needed to classify the sentence as SCD or *content*.

We fine-tune the BERT on a corpus $\mathcal{D}$ and measure the model's performance on a different corpus $\mathcal{D}'$. Especially, the sets of SCDs $g(d)$ are disjoint, because it is important to prevent the model from simply memorizing all SCDs.

2) `BERT Next`*:* Applying BERT's next sentence prediction to SCDs, we assume a *next sentence* to be an SCD of the current sentence. The model is fine-tuned on tuples of two sentences, meaning the first sentence w is always a sentence from the *content* and the second sentence w' may be a related SCD or the subsequent sentence from the *content*. Thus, the

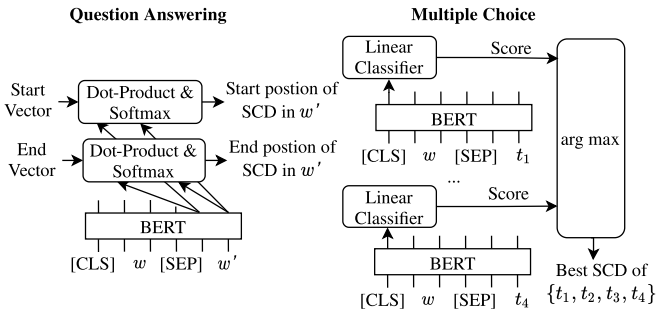

Figure 2. Use-cases of BERT solving the MPS²CD problem.

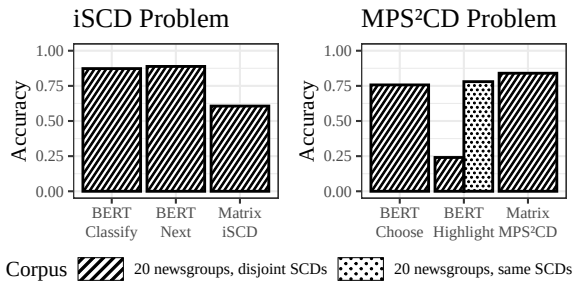

Figure 3. Accuracies gained for all scenarios.

model classifies the second sentence w' as (related) SCD or no SCD.

B. Solving the MPS²CD Problem with BERT

BERT's use-cases multiple choice and question answering can be used to choose a most probably suited SCD from set of, in our case, four SCDs $\{t_1, t_2, t_3, t_4\}$. The solution always has to be unique, i.e., one SCD may be associated with the sentence w while the other three must not be associated. In Figure 2 both use-cases are depicted.

1) `BERT Choose`: Using multiple choice is straightforward, BERT calculates a score for each pair of sentence w and possible SCD $t_i, i = 1, ..., 4$. Then, the SCD reaching the highest score is chosen.

2) `BERT Highlight`: Analogous to multiple choice, BERT's question answering use-case gets a sentence w and four SCDs to select one SCD from. The four SCDs are randomly shuffled and concatenated to a short document w'. BERT returns an interval to highlight the words in this interval of the short document w' as SCD for the sentence w.

IV. Evaluation

We use the 20 newsgroups dataset and definitions from the online dictionary Wiktionary[1]. We annotate each sentence with a definition from Wiktionary acting as SCD.

All experiments using the SCD matrix run on a virtual machine featuring 8 Intel 6248 cores at 2.50GHz and 16GB RAM, all experiments using BERT run on an NVIDIA A100 40GB graphics card. Each experiment is run five times to take the arithmetic mean of the results and thus, increase the statistical correctness.

During each experiment, we split the corpus randomly into a training set containing 80% of the documents and a test set containing the remaining 20%. If a disjoint set of SCDs is used in the current experiment, we also split the set of SCDs into 80% and 20% of the definitions from Wiktionary.

We use the pre-trained `bert-base-uncased` model of BERT and a batch size of 40 during fine-tuning (10 for `BERT Choose`). The pre-trained model uses a dropout of 0.1 and cross-entropy loss to determine the model's error.

We run the fine-tuning for 3 epochs using the following hyperparameters for AdamW: $\alpha = 5 \cdot 10^{-5}$ (also called learning rate), $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$ and $\lambda = 0.01$ (also called weight decay). The learning rate rises linear from 0 to α in the first 500 steps of the fine-tuning.

The accuracies in Figure 3 demonstrate that BERT is good at solving the iSCD problem. There is only a very small difference between `BERT Classify` and `BERT Next`. The small difference indicates that BERT does not benefit much when getting a pair of sentence and associated SCD simultaneously. The scenario using the SCD matrix reaches an accuracy of 0.61 only and thus `Matrix iSCD` is clearly worse than `BERT Classify` and `BERT Next`. In our scenario of the iSCD problem, the chance for a sentence being an SCD is 50%. Thus, an accuracy of 0.61 gained by `Matrix iSCD` comes close to random guessing.

For the MPS²CD problem, the scenarios using BERT and the SCD matrix result in similar values. Only `BERT Highlight` with a disjoint set of SCDs achieves a very low accuracy. As `BERT Highlight` asks to highlight the matching SCD out of four SCDs, the accuracy of 0.25 is as worse as randomly highlighting an SCD. Hence, we simplify the problem for `BERT Highlight` and do not split the set of SCDs. Using `BERT Highlight` with the same set of SCDs, then, shows a similar performance as the other two scenarios.

An extended version of this paper can be found at our institute's homepage[2].

Summarized, BERT performs better than the SCD matrix on the iSCD problem, while the SCD matrix performs slightly better on the MPS²CD problem. We conclude that BERT is able to grasp the concept of SCDs, in a way that BERT can be trained to solve SCD-related tasks.

Acknowledgement

The authors thank the AI Lab Lübeck for providing the hardware used in the evaluation.

References

[1] F. Kuhr, T. Braun, M. Bender, and R. Möller, "To Extend or not to Extend? Context-specific Corpus Enrichment," *Proceedings of AI 2019: Advances in Artificial Intelligence*, pp. 357–368, 2019.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[3] M. Bender, T. Braun, M. Gehrke, F. Kuhr, R. Möller, and S. Schiff, "Identifying subjective content descriptions among text," *Proceedings of the 15th IEEE International Conference on Semantic Computing*, 2021.

[1] http://qwone.com/~jason/20Newsgroups/, https://en.wiktionary.org/

[2] https://www.ifis.uni-luebeck.de/~bender/ma/poster_long.pdf