Giving Sentence Encoders Poor Input for Comparison Works Well

Emma Kleiner
emmakleiner@me.com
29.05.2021

Abstract
A simple chatbot to provide information to people with addiction issues performs its classification of the user query on three differently formatted question and answer datasets. Not surprisingly, providing many question patterns in the dataset turned out to help the bot provide more accurate predictions of what answer a query is requiring of. But an even bigger performance increase resulted from simply evening out terminology throughout the dataset and proving different but constant syntactical patterns for each question answer pair. The smaller performance increase of adding many patterns comes at the expense of upsizing the dataset, while the considerably larger performance increase requires only slightly more data than simply providing one sample pattern per question.

1. Introduction
The main purpose of the chatbot is to create a contact point for users somewhere on the spectrum of alcohol use disorder to reach out to. The users may be doubting whether their relationship to alcohol is healthy and want to find out more about boundaries, they may be dependent on alcohol and are looking for guidance on how to get sober, or they may have ventured on a journey of sobriety and simply want to check in. The bot is essentially a sobriety coach that helps with addiction prevention and recovery by providing information and referring the user to relevant further contacts.

A chatbot is a good tool to perform this task, because it can deliver this service to a large population, without the constraints that come with human contact or group memberships. It is available on demand anytime and can thus help when someone is dealing with immediate cravings or impulses. It can also help streamline the care work of humans, as the bot can assess the situation the user is facing and then refer the relevant cases to the relevant services, taking over that administrative work.

With this bot as an example, I wanted to find out which method of retrieving answers from a dataset of corresponding questions and answers yields most fitting results. I worked with a small dataset of frequently asked questions from prevention websites, and their answers. In the answer retrieval step, the most similar question to the user query is determined, and the bot replies with the corresponding answer. I am comparing whether collecting several different patterns for how one question is posed is conducive or detrimental to the bot making an accurate prediction about which question was asked.

Our interest was sparked because both results could easily be justified in hypothesis and would arguably be revealing about the ins and outs of the embedding model used, namely distilBERT's sentence embeddings. A larger set of question sample patterns for each tag gives more chances for the user query to be very close to one of the patterns. But that also may be a reason for performance to be inhibited by a larger set of question sample patterns, since a way of posing the question similar to the user query is bound to be among the question patterns of an unfitting tag. If storing a larger variety of question patterns in the dataset for each tag does inhibit performance in our example, I think that could potentially reveal that sentence embeddings privilege the semantics of sentence structure over the semantics of word content.

2. Related Work
Conneau et al. have observed that while sentence embeddings have achieved impressive results, such as sentence classification relying on semantic subtleties, "I still have poor understanding of what they are capturing" and it is "difficult to pinpoint the information a model is relying upon" (Conneau et al. 2018). In order to compare different models in terms of how they arrive at classifications, they opt for probing tasks checking a variety of linguistic properties. A task is built specifically to reveal the performance in just one aspect of a model's linguistics awareness. Interestingly, Conneau et al. distinguish between the encoder model's "ad hoc" performance on the task that they propose to assess the model's ability to capture a linguistic property and that information being encoded itself (Connau et al. 2018).

Belinkov et al. similarly aimed to uncover the properties that are encoded in

sentence representation and what language information they actually capture or overlap with (Belinkov et al. 2017).

Pre-trained, transformer-based language models, such as BERT and its daughters, generally don't perform worse if pre-processing steps, such as stemming and stop word removal, are not performed before encoding, as BERT was trained on unformatted text and learned to filter out which stop words are not useful, while preserving the relational information stored in suffixes and stop words, that are useful and informative to the embedding (Qiao et al. 2019). Thus, the same word may receive different embeddings depending on its context, which can be very useful for tasks that require a degree of disambiguation and semantic awareness.

I use distilBERT as the language model to encode the questions and queries into embeddings, because it 40% smaller and thereby 60% faster than its larger counterparts (Sahn et al. 2020). It is great that it was so cheap to train, also in terms of the environmental costs pointed out by Bender et al., and it works much more swiftly on our hardware.

3. Data collection and dataset description

The data collection process for a dataset of corresponding questions and answers was very time consuming and tedious, because corpora of addiction therapy or reliable real life question answer pairs are hard to come by. Few corpora on this topic even exist, presumably because researchers refrain from risking the privacy of people with addiction, and those that do exist are located behind paywalls. Question answer pairs that can be found in forums proved to be too conversational, the answers often consisting of sharing own experiences rather than information, and if they did include factual information, it wasn't necessarily reliable enough to be appropriated by a bot into a matter-of-fact response.

Instead, I collected the question answer pairs that were listed in the 'frequently asked questions' (FAQ) section (or similar) on websites of four more reliable prevention programs. Different answers to similar questions from different FAQ pages were synthesized manually into one coherent answer, and the sample questions that would be to be matched were also devised manually.

The dataset that forms the foundation of the different methods to obtain the fitting answer remains in the same structure for each method. The questions and corresponding answers each form one intent in a json file. An intent consists of a descriptive tag labelling the type of question asked, one or several sample patterns of the question type, and the corresponding answer, accessible by the keys 'tag', 'questions', and 'answers' respectively.

4. Methods

I am trying to peak into what properties a sentence embedding actually pays attention to for the encoder model distilBERT. Rather than trying the distilBERT encoder on some of these probing tasks developed previously, I look at the difference in performance of the distilBERT sentence encoder model by comparing its output accuracy in a downstream application, as judged by humans. The downstream application is the sobriety coach chatbot. Its main algorithm is simple. It compares a user query to each of the question patterns in the dataset, as accessed by a for loop, using cosine similarity between their respective embeddings. The tag of the question pattern that was determined most similar then leads to the answer with the same tag.

This process is repeated for three slightly altered datasets in a json file. The first only provides one sample question pattern for each answer and tag. The second provides up to eight different sample question patterns for a single answer and tag, each pattern using different grammar and vocabulary to prompt the same answer. And the third provides even sample question patterns for each answer. By even, I mean, that

a. there are exactly three different phrasings for each answer and tag,
b. the patterns consistently use the same terminology throughout the dataset,
c. and the three different phrasings exhibit three different semantic and syntactic structures that stay as constant as possible for each pattern triple

Constant terminology ensures that no keyword that shows up in only one pattern (although it would also fit for other patterns) skews the bot to judge this sample question pattern as most similar. All question patterns about alcoholism or alcohol dependence, for example, are always phrased in terms of 'alcohol use disorder' in the even dataset, and questions about quitting

alcohol or stopping cold turkey for example are always phrased in terms of 'stopping drinking'. Constant structures ensure that neither word order, nor opting for passive vs. active voice, nor a perspective as indicated by personal pronouns, etc. skews the bot to judge a sample question pattern as most similar. I made sure that, where it makes sense, one of the three patterns is a question posed from a first-person singular perspective. In the majority of cases, the second and third patterns are posed in passive and active voice respectively but are otherwise kept as close to each other as makes sense.

This way I aim to ensure that only the features of a question that delineate it from the others in terms of content and nothing beyond content differ in the question patterns provided. I then compare the bot's performance per dataset format by prompting the bot's predictions for a set of devised question and seeing what percentage of queries it classified correctly.

5. Findings

Feeding the algorithm the dataset with only sample question pattern fitting an answer was not very successful. The bot found the fitting answer for 76.19% of the test queries. Feeding it the dataset with several different sample question patterns per answer was more successful. The bot found the fitting answer for 90.48% of the test queries. But the dataset with a quantity of sample patterns between the other two, namely three that were evened out for vocabulary and structure, performed the best. The bot found the fitting answer for 100.0% of the test queries. This is somewhat surprising as the bot had much less variability in the question patterns to closely match the query with than when relying on the dataset with several different patterns. But as all sample patterns were more or less equally poor, if you will, the relevant information stood out beyond irrelevant accuracy, as would be found in a random match of the same phrasing, for example.

One inference about what distilBERT pays attention to that can arguably be derived from the bot's performance quite safely, is that distilBERT sentence embeddings privilege exact terminology over synonyms, as in a small dataset like this, a query that included the same term generally had the closest cosine similarity. This is what the success of the even dataset can largely be attributed to, since as soon as an exact overlap in terminology doesn't point to one answer anymore, the surrounding of the terminology and its semantic context receive attention as well.

6. Limitations

There are considerable limitations to assessing the properties an encoder is aware of by comparing its performance in slightly different downstream applications. Each judgment made about the which properties the embedding can reproduce is an inference based on a highly contingent output.

Additionally, our dataset of questions and answers is much too small to trust that the difference in performance stems from the added question patterns.

The size of the dataset also is a limitation to the utility of the bot, at the current size, the bot is not agile enough to deal with user's questions that I did not foresee, even though each answer is detailed enough to cover some peripheral problems to a single question asked. But for that reason, a dataset or collection of information that the bot automatically searches in would have been better, rather than only searching in answers to manually selected problems that users, who would turn to a bot like this, were predicted to experience.

Plus, it would be easy to add a filter to the bot, that only lets the bot respond if the similarity between query and question pattern exceeds a certain limit, since as of right now, the bot still gives advice on alcohol addiction, if one asks it about the latest soccer scores.

7. Discussion

By creating even datasets as the basis for a bot to search for matching questions or answers in, the margin for error seems to be decreased significantly. I suspect this can be attributed to sentence embeddings placing emphasis on linguistic properties that lie beyond the core factual content of a sentence in some cases, this may be the choice for passive vs. active voice, sentence length, pronoun use, etc. while in other cases, these properties stand out rightfully, as they carry core content.

Therefore, I think it can be of use to perform targeted paraphrasing on datasets that serve as examples for comparison and classification tasks, as the multiplication of ways to express something is revealing for an encoder model which linguistic or semantic aspect of a sentence is retained and which aspects are dropped in other phrasings of the

same information and can thus be judged nonessential semantically.

Perhaps a process of multiplying a sentence's content as wrapped in different but recurrent structures and terminology can be thought of as targeted stemming and stop word removal. I've come to this conceptualization, since not an entire language's vocabulary and syntax is responsible for deciding which stop words and suffixes are superfluous for a sentence's meaning, but rather the dataset at hand sets the standard for which contexts and structural particularities carry meaning for this topic.

As only adding two evened out phrasings to each the sample question pattern set improved the bot much more radically than adding many random phrasings, this may show the path to a really data-efficient way to provide basis for classification and comparison to encoder models.

Perhaps, further research could go into formatting datasets as a means to making comparison or classification tasks more efficient, as opposed to making encoder models even more potent and aware of even more properties by training them on even larger datasets, for example. This would be in line with finding ways to keep up the benefits of large language models but downsizing to keep unjustly distributed environmental costs at bay, but starting from the opposite end than what Bender et al. outlined (2021).

References

- Belinkov, Yonatan et al. "Analysis of sentence embedding models using prediction tasks in natural language processing." IBM Journal of Research and Development, vol. 61, no. 4-5, 2017.
- Bender, Emily M. et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" FAcct, Virtual Event, 2021, pp. 1-14.
- Conneau, Alexis et al. "What you can cram into a single vector: Probing sentence embeddings for linguistic properties." 2018, pp. 1-14, arXiv:1805.01070 [cs.CL].
- Sanh, Victor et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." 2020, pp. 1-5, arXiv:1910.01108v4 [cs.CL].
- Qiao, Yifan et al. "Understanding the Behaviors of BERT in Ranking." 2019, pp. 1-4, arXiv:1904.07531v4 [cs.IR].

Sources for the Q&A dataset:

- https://www.niaaa.nih.gov/alcohols-effects-health/alcohols-effects-body
- https://www.webmd.com/connect-to-care/addiction-treatment-recovery/alcoholism-vs-alcohol-dependence
- https://lifeprocessprogram.com/alcohol-addiction/faqs/
- https://www.alcohol.org/faq/

Links embedded in the bot's answers for queries about facilities, contact points, resources:

- https://www.brijder.nl/onze-locaties/verslavingszorg-alkmaar-kliniek
- https://www.brijder.nl/onze-locaties/verslavingszorg-den-haag-kliniek-detox-diagnostiek-bopz
- https://www.brijder.nl/onze-locaties/verslavingszorg-den-haag-verslavingskliniek
- https://www.brijder.nl/onze-locaties/verslavingszorg-hoofddorp-kliniek
- https://www.brijderjeugd.nl/kliniek%2Dmistral.html
- https://www.brijderjeugd.nl/voor%2Djongeren/behandeling.html
- https://www.brijderjeugd.nl/wat%2Ddoen%2Dwij.html
- https://ggzinterventie.nl/
- https://www.changesggz.nl/verslavingen/
- https://www.yeswecanclinics.com/addictions/alcohol-addiction
- https://www.gezondheidsplein.nl/dossiers/alcoholverslaving/item42719
- https://www.alcoholinfo.nl/
- https://www.trimbos.nl/
- https://www.stap.nl/nl
- https://minderdrinken.nl/
- https://www.jellinek.nl/
- https://www.alcoholondercontrole.nl/
- https://www.alcoholenik.nl/
- www.113.nl