**Deriving High-Quality Book Comprehension Benchmarks**

**via Book-Movie Script Alignments**

* Acquiring high-quality annotated benchmarks for book story understanding is challenging, as they usually require the annotators to read the whole books before annotations.
* On the other hand, movie scripts are semi-structured documents. There are many kinds of important story structure information that can be automatically extracted from scripts.
* By collecting movie scripts and books adapted from the movies (such as Star Wars trilogy and their book versions), we can use sentence embedding and paraphrasing techniques to align the movies with books adapted from movies.
* Then we propose to project the structures from movie scripts to annotate books, so as to automatically create book understanding benchmarks with high quality.

**Brainstormed annotations that can be projected (and more):**

* **Scene segmentation**: one fundamental question is, what is the basic content unit of a story. Previous NLP works like NarrativeQA (Kočiský et al., 2018) and BookSum (Kryściński, et al., 2021) use paragraphs or fixed-length chunks. However, actually in movie scripts, a natural unit , i.e., scenes, has been defined by the authors. Splitting the books into scenes is a unique idea and also an important problem for follow-up annotation tasks.
* **Settings**: when and where a specific subplot happens.
* **Characters**: Identifying the characters involved in a scene and speakers of dialogue utterances, if there are no explicit identifiers; this also extends to NER/coreference over characters.
* **And more to discuss** … something like functional structures of stories and character-centric plot sequence.

**Key References:**

* <https://arxiv.org/pdf/1506.06724.pdf> : a very relevant work. But they are not defining narrative tasks so the focus is very different.
* <https://github.com/booknlp/booknlp> : [**TODO**] survey the dataset papers done by the same author, and finish a write-up to summarize their supported tasks and the amount of data (in terms of both tokens and books)

**References to tools:**

* **SentenceBERT**: <https://sbert.net/> . And how it works to construct paragraph-level alignment in BookSum: <https://arxiv.org/pdf/2105.08209.pdf>
* **Script Parser**: We have an accurate script parser (<https://openreview.net/pdf?id=HK-_DteWlGq>). We have already processed the movie scripts with the parser. But we encourage you to check the details, because some of the ideas can be applied for bootstrapping the sentence alignment models, if out-of-box SentenceBERT does not provide very accurate results.

**Paper Reference**

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., & Grefenstette, E. (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, *6*, 317-328.

Kryściński, W., Rajani, N., Agarwal, D., Xiong, C., & Radev, D. (2021). Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

A Tutorial of MiniLM in SBert: <https://www.sbert.net/examples/applications/computing-embeddings/README.html>

4/6/2022 Discussion Points

1. Meaningful Unit, i.e. scene
2. Application: QA

contribution:

1. Segmentation is a task
2. Alignment is our method
3. We propose a new model to resolve the segmentation problem