Review of Thesis Proposal

Reviewer: doc. RNDr. Ondřej Bojar, Ph.D.; bojar@ufal.mff.cuni.cz
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, 181 00

Date: 12. 9. 2022

Thesis Title: Multimodal summarization
Candidate: Mgr. Mateusz Krubiński
Supervisor: doc. RNDr. Pavel Pecina, Ph.D.
ÚFAL MFF UK

The thesis proposal by Mateusz Krubiński focuses on a very up-to-date and challenging problem, the task of automatic summarization of documents available concurrently in multiple modalities (text, images, video) into a similarly multi-modal summary.

The proposal is well structured and clearly written, with only occasional writing errors, see below. The only change I would do in this regard is to clearly distinguish the MLASK dataset and the model and experiments with it, which all appear in Section 3.1.

At times, it would be useful to make the thesis proposal text more self-contained, i.e. not just citing related work but also providing its gist, esp. for topics upon which Mateusz is building. One example is Li et al. (2020d) in Section 2.4.2 where the concatenation across modalities vs. time is unclear, esp. given that the input modalities can be not directly synchronized in time. Other examples are in Section 2.5.2 where we would like to know if the approach of Zhu et al. (2018) convincingly worked and whether the questions "generated from the reference summary" in Li et al. (2020e) were automatic or manual, which can affect the evaluation considerably. Another example is Hessel et al. (2021) where image captioning evaluation did not need references.

Questions to the proposal:

- In Section 2.4.1, you mention that no specific order can be imposed on the images and thus dedicated encoders are rare. I do not see the logical relation here; a dedicated encoder can be designed to ignore the order.
- Looking at Figure 7, multi-modal summary evaluation, I do not like any of the images offered. Comparably, the middle picture can be seen as a better one, although the depicted area of the city is a totally different one; I suspect none of the pictures is actually from Prague at all. Can the annotator indicate that there could be a much better visualization for the given text?
- I am very delighted to see that you did the contrastive experiment with random noise, as mentioned in Section 3.2. Such sanity checks are critical and often neglected, which results in wrong assumptions about the system performance, and consequently prohibits successfully building upon the approach. What is your working explanation? I suspect that there is much more of the signal coming from the visual part but it is not very informative in comparison to the text. The "information density" is much lower, so the network decides to ignore the visual input.
- Related to the previous point, I would be very careful in interpreting results on MLASK in Section 3.1. ROUGE-L of 12.93 is with visual data only, text information

moves us to 13.26 and more text to 14.32, but replacing visual input with random data here works equally well? What is the performance of just random visual data and no text? Around 12 points ROUGE-L? This would be surely possible but it would tell us something important about ROUGE or your data or its domain.

- I am very happy to see that you measured the kappa statistic on your data (although 0.217 is rather low in practice). Would it make sense to, e.g. automatically extract object labels from the images, list them under each image and ask for each "Does the 'tram' element in the picture work well for the summarization purpose? Yes / No / 'tram' is actually not in the picture". This fine-grained (but grounded with an exact list) evaluation could be easier to agree upon and would tell us more about the pictures.
- Could you elaborate more on your future plans in Challenge 3, task-specific pre-training? What else in addition to text summarization from other domains are you considering? Or are you simply waiting for an opportunity such as a novel shared task that may appear?
- I really do think that Challenge 4, evaluation of multi-modal summarization, is the most critical area and also likely the area where your work could be most influential. Machine translation is seeing a revival of manual error annotation and these error annotations are not quite in line with (manual or automatic) "overall scores" but they are a good guidance as to what aspect of the system to focus on. Proposing (and polishing) a more principled evaluation methodology could become the steering element for the whole field of multi-modal summarization. Do you have such an ambition (and does your supervisor have the annotation resources) for this goal?

Outside of the main thesis topic, I highly value Mateusz's work on MT and textual summary evaluation. Both the MTEQA metric (best scoring metric in Chinese-to-English translation in 2021) and COMES metric for summaries based on COMET are great achievements and I am hopeful that Mateusz will be able to directly build also upon this work for this thesis.

Other comments:

- Popel et al. (2020) cited in the first paragraph actually evaluated MT *in the context of surrounding sentences*, not on isolated sentences.
- In challenge 4 in Section 3, the "thus" between human annotation and domain/language specifics does not really connect a matching premise and consequent.
- In challenge 4, you refer to "recent guidelines" but they were not mentioned in the previous text.
- Please be always very clear about your contribution. In the last paragraph of Section 3.1, I suggest you say "we created and used", not just "we used" (the annotation tool).
- I suggest avoiding the abbreviation CNN. In your context, it can refer to convolutional NNs or the CNN + Daily Mail summarization corpus.
- Text flow in Section 3.2 is not ideal, you start by "One negative result". This word order implies that the negative result has already been mentioned, which is not the case.
- I am happy to see that you are avoiding excessive decimal places (beyond what the actual experiment's precision provides) but in Table 1, using one decimal place for averages would be useful to indicate that these are counts any more.

English and wording issues, some of them a little embarrassing:

- “methods have became” in the first sentence is incorrect; the correct form is “become”.
- the word “data” is ambiguous with respect to plurality; my personal preference is to treat this noun as plural, i.e. “data were”, not “data was” (in the first paragraph of Section 2.2).
- The correct spelling is “web scraping”, not “scrapping” (two occurrences in the text).
- In 2.4.2, you probably mean “cross-modal", not “cross-model” modules.
- “Annotators were give a collection” should say “given” in Section 2.5.2.
- The correct spelling is “assess”, not “asses” (last paragraph of Section 3.1).
- “On pair” should be “on par” in Section 4.1.

In sum, the thesis proposal by Mateusz Krubiński is of sufficient quality and documents Mateusz’s expertise in the area. The proposed plan is well structured and I support it. I especially highlight Mateusz’s carefulness in evaluation (measuring inter-annotator agreement, using contrastive 'dummy' baselines) and I am looking forward to seeing further good results from good models but also bad results from intentionally bad models. I fully recommend accepting the proposal.

In Prague, September 12, 2022.

Ondřej Bojar