Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Anna Kuznetsova
Název práce	Multilingual Multimodal Detection of Humour in Stand-Up Comedy
Rok odevzdání	2023
Studijní program	Computer Science
Studijní obor	Language Technologies and Computational Linguistics
Autor posudku	Mateusz Krubiński
Role	Oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Review text

In her Master thesis, Anna Kuznetsova challenges the problem of multimodal humor detection in stand-up comedy videos. The main contribution of this work is the collection of a novel dataset based on stand-up routines in Russian and the corresponding experiments focusing on data preprocessing (e.g., word-level audio-to-text alignment or sentence segmentation) and methods of automatic laughter detection and humor labeling. By extending the dataset with videos of stand-up comedy shows performed in English, the student approaches the novel problem of multi-lingual, multimodal humor detection. In the latter part of the thesis, uni-modal modeling approaches (e.g., text-only model operating on subtitles) are compared to multi-modal approaches that consider various combinations of textual and visual features.

The thesis is structured into five Chapters. Chapter 1 presents the background and relevant work, Chapter 2 describes the process of dataset collection and automatic labeling. Several models are introduced and trained in Chapter 3, with Chapter 4 presenting the quantitative results and Chapter 5 dedicated to more qualitative findings and broader implications. As an electronic attachment, both the collected dataset and relevant scripts are provided.

Overall, the work is well-structured, and the quality of the writing makes it rather easy to follow, with only several non-crucial flaws that I was able to identify.

I highly value the second chapter, which compares several laughter detection methods that are crucial for automatic labeling and the corresponding work to assert the quality of the automatic process, including rounds of manual annotations and sanity checks. In my opinion, the biggest limitation has to do with the multi-modal approaches presented in Chapter 3. While the experiments on the textual modality make use of recent advances, with models such as ColBERT (Annamoradnejad and Zoghi, 2022) considered, the ones on the multi-modal data are based on early fusion (feature concatenation) and modeled only with the Support Vector Machine (SVM) learning algorithm. Adapting the very recent general-purpose multimodal foundation models for the humor detection task is probably out of the scope of this work. However, I would expect a comparison with a hierarchical, cross-modal attention-based approach, such as the one presented in FunnyNet (Liu et al., 2022), or even a simpler early fusion-based approach but modeled end-to-end with neural networks. Based on the fact that the student managed to implement the ColBERT architecture herself and that she acknowledges the simple nature of multi-modal and multi-lingual experiments conducted (Section 5.3), I would still consider this as a promising start for future work. Below, I include detailed comments and questions.

Major comments:

1. **Section 1.3** – This section introduces the datasets and models used for multimodal humor detection. I am missing a part that would introduce training signal formulation and evaluation metrics. Based on the following sections, it becomes clear that humor detection is modeled as a binary classification task, thus Precision/Recall/F-score and Binary Cross Entropy are typically used. The following sentence from Page 9 gives a clue, but is not sufficient: „*To assess the humour content in the clips, the authors [Mittal et al., 2021] used a different scoring system compared to binary classification.*".
2. **Section 2.1** – It is unclear from the text whether the collected subtitles are of human origin, an outcome of an ASR system, or a mix of both. Based on my understanding of the code base provided, only manually created transcripts were considered.
3. **Section 4.3, Table 4.5** – The results on the multilingual text are not very informative. Language-specific results, similar to Table 4.6, should be included.

Minor comments:

1. The UK/US English spelling is used inconsistently, with „humor" occurring 65 times and „humour" 117 times.
2. **Section 2.5.1, Figure 2.10** – Instead of the word `True`, I believe that the correct labels on the sub-plots in the right-most column should be `kmeans` and `clustering_avg`. The sub-plots should also be column-aligned, i.e., the outcome of `kmeans` for both videos should be placed one below the other.
3. **Section 3.2** – „*To process the facial features, I averaged the context frames into one vector and the utterance frames into another vector.*" - I believe that *features* extracted from the frames were averaged, not the frames themselves.

Questions for the defense:

1. Would it be feasible to compare the different laughter detection methods (peak detection approach, machine learning-based ones) and the study on hyper-parameter choices on one of the multimodal datasets introduced in Section 1.3.1, instead of using the small sample manually annotated by the author (Section 2.2)?

2. One of the characteristics of stand-up comedy is the active interaction of the speaker with the audience. While some of the routines involve the monologue of one person, others are interrupted by constant banter with the crowd. On the other hand, the monologue may consist of a single, long story but may also consist of several short, unrelated stories. In the author's opinion – could this dynamic be an influential factor, that could require different modeling methods, e.g., a more dynamic context window formulation?

3. How exactly was the manual humor detection, reported in Table 4.1 and Table 4.2, conducted? Was the annotator looking only at the utterances or context+utterance pairs? I have the same question regarding the multimodal annotations – was the annotator looking at the muted videos? Were the subtitles included? In the author's opinion, would looking at the whole transcript and annotating the humorous sentences/phrases sequentially lead to better results?

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

V Praze dne 28. srpna 2023

Podpis: