Master Thesis Review

Faculty of Arts, Charles University

Thesis Author	Štěpán Lars Laichter		
Thesis Title	Silence in Dialogue		
Submission Year	2023		
Study Program	Logic	**Branch of Study**	Logic

Review Author	Ondřej Dušek	**Role**	Opponent
Department	Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics		

Review Text:

Contents Summary The topic of Štěpán Lars Laichter's master's thesis is silence in dialogue (and in dialogue systems), which is an important but frequently overlooked issue. The length of silence between turns is one of important turn-taking cues in dialogues, it helps facilitate successful dialogues and is dependent on many variables, including the dialogue topic/content, relation of the interlocutors, or cultural background. The present thesis aims to bring contributions to research of silence in dialogue on two levels: theoretical, with an overview of how various theoretical descriptions of dialogue approach silence, and practical, evaluating a chosen large language model (LLM) on the task of predicting the length of silence between two dialogue turns based on movie subtitle data.

The author shows that most of the six theories of dialogue he analyzed do not devote a lot of attention to how silence is handled. However, he is able to place or interpret silence in terms of most of them. In his LLM experiment, he shows that the chosen GPT-3 LLM can be, to some extent, finetuned to predict the length of silence in dialogues using a simple textual encoding (based on inserting a a number of "*" characters). The accuracy gets better with larger amounts of training data and surpases the mean and uniform baselines in terms of KL divergence from the true silence distribution, but variance in predictions is high and GPT-3 mostly does worse than the baselines in terms of mean average error.

The text of the thesis is split into five numbered chapters:

- Chapter 1 gives a quick description of the thesis's main aims and the underlying motivation for the research.
- Chapter 2 provides some basic background on the concept of dialogue and on human-computer dialogue, reviewing main architectures of dialogue systems.
- Chapter 3 analyzes six different theoretical descriptions of dialogue, giving a brief overview of each of them, adding general comments or criticisms from the point of view of the thesis author, and pointing out where silence comes into the equation for each of them.
- Chapter 4 gives a detailed account of the experiments with the GPT-3 language model. The author chose 8 movies by 2 directors (all split 50:50 into training and test sets) and examined performance on a gradually enlarged subset of his corpus, starting from a single movie. The chapter includes an analysis and discussion of the results, and adds some notes on potential future work on this problem.
- Chapter 5 summarizes the conclusions and lists some limitations and potential ethical issues of this work.

Overall Evaluation I believe the text generally fulfills the requirements for a master's thesis in terms of both the extent and the execution. However, I do have a lot of reservations towards the latter. A few minor ones relate to the theoretical analysis part, but quite substantial ones are related to the experiments with GPT-3. These limit the overall impact of the results of this thesis in my view. In effect, the experiments give a notion of a single model's performance in a single, not very realistic setting, but would be hard to generalize to LLMs' performance on silence length prediction in general.

The writing is excellent and only involves very few language errors (and these are very minor, such as punctuation). The overall structuring of the text is very well done, the text is mostly quite clear and easy to read.

I am not very well qualified to comment on the choice or completeness of dialogue theories analyzed in this thesis, but I believe the analysis to be quite extensive and reasonably performed given the space and format of the thesis. On the other hand, I believe the author may be a little overtly critical in some places. I do not think any theory can be truly complete, and it is obvious that the merits and focus of the theories just do not align very well with the thesis author's own focus. In addition, some of the theories could be better explained to a reader who is not very familiar with them (see detailed comments below).

The experiments with GPT-3 are reasonably well-defined and the overall experimental setup is quite standard – it uses separate training and test sets, a standard approach to LLM finetuning, appropriate evaluation metrics, and trivial but reasonable baselines. On the other hand, the experiments have a lot of limitations or caveats. A few are just technical and rather minor: No statistical significance tests have been used, mean squared error could have been an additional helpful metric, and prompting a vanilla GPT-3 model is an obvious but omitted baseline. Other issues are more design-related and seem more important to me. Of these, a few are admitted by the author in the discussion section (limited input context, inaccurate subtitle timing, question of human performance on the task, model exploiting data statistics), some of them are rather glossed over (problems of turn/scene boundary detection and textual silence representation), and some of them are left out from the text entirely while having a potentially major influence on the outcomes (question of movie realism, appropriateness of model choice and training approach). Details on the individual issues are given below.

Detailed Comments These comments give a more detailed explanation of my evaluation above. They are ordered around related topics and attempt to follow the thesis's text. As noted above, comments regarding Chapter 4 are most important.

First, a few minor comments on the descriptions in the background Chapter 2:

- *Statistical dialogue systems*: Note that "statistical data-driven machine learning" is basically the same thing, three times. Most of these systems actually do not use any sophisticated grammars, entities and intents are the usual representation. The large number of dialogue states and potential replies is generally the case and the systems are trained to cope with it.
- *End-to-end dialogue systems*: These systems generally do use an explicit representation (dialogue states) if trained for task-oriented dialogue. They also do not outperform previous architectures in the task-oriented setting (their main advantage is easier training).
- *Remarks on dialogue systems not prioritizing/neglecting silence*: It is true that mimicking humans on silence length is generally not a priority and that no theories are applied here, but silence is not completely ignored. Voice-based bots do need to work with silence. It is important to catch the beginning and end of user turns, and in practice it is hard to get a right balance in practice between appearing slow and barging in on the user.

Notes on the theoretical analysis in Chapter 3:

- The concepts/terms/keywords of the theories presented in Sects. 3.1, 3.4, 3.5, and 3.6 are not very well explained and would need more detail.
- It feels like the author criticizes a lot the lack of focus on silence, but he does not discuss turn-taking analysis, which would be spot-on in terms of this focus, for more than a few sentences in Sect. 3.2.
- The discussion in Sect. 3.4 seems contradictory (is Clark arguing that language is or is not an individual and social process?) and the end of the first paragraph on p. 23 seems unfinished.

The problems with experiments in Chapter 4 revolve around three thematic areas. First, detailed comments regarding the data used:

- *Dialogue depiction in movies*: This is an important question which is completely neglected in the thesis – how realistic are the dialogue depictions in movies, from the point of view of silence? I believe they do not necessarily have to be very realistic, as directors may want to work with suspense and dramatic effects. This means that movies might not have been the best choice, if we aim at reproducing real dialogues.

- *Subtitle timing accuracy:* This is an important issue and definitely limits the realism of the results, so I am happy that the author is aware of it. However, the author does not consider that the issue might be even more severe: the same dialogue can have different timing when played at different frame rates – all of 23.96/24/25/29.97 FPS is very common and may have been used in different rips of the same movie. The extent of the timing problem could have been measured by analyzing the same dialogues in multiple different subtitle files for the same movie. Overall, using speech recognition output and mapping it onto transcriptions/subtitles would likely yield a more accurate timing.

- *Turn segmentation:* This is another important issue and limitation, and it seems rather glossed over in the thesis. I believe that raw subtitles are not the best choice, and they could have been complemented by movie scripts, which are also available online. This would allow to distinguish scenes and characters more accurately. See (Lison & Meena, 2016) for a more detailed description of this problem and a dataset which aligns scripts and subtitles.[1]

- *Input Context:* This is again an important point, and I agree with the author that using a single preceding turn is a huge oversimplification. However, I do not understand why there would be just a few turns with longer preceding context – dialogues in movies are typically composed of multiple turns, so I do not really see the issue. In addition, the lost visual context is also a problem (cf. Lison & Meena). Ultimately it might have been better to choose a form of communication without a visual context, such as phone calls. There are corpora of phone conversations with annotation (Switchboard, Callfriend), though gaining access might be problematic as they are not freely available.

Regarding the experiment architecture:

- *Model choice:* The choice of GPT-3 is not discussed beyond stating that it is one of the largest available. I am not sure if using the largest possible model brings so many benefits; on the other hand, using a model only available through an external API brings a lot of limitations. The first and obvious one is cost – without paying for every API call, the number of experiments and/or the training data size might have been bigger. There is, however, also the issue of architectural limitations – if the model's internal representations were available, they could have been used as features for a real regressor or some other kind of architecture, without the need to rely on text generation as the output. I believe that either one of the newer open LLMs, or even a smaller model such as GPT-2 or BERT would have been a more interesting choice.

- *Training approach:* As far as I understand from the thesis, the GPT-3 model was simply provided pairs of dialogue turns and the corresponding outputs with "*"'s representing silence. However, GPT-3 is known for prompting and use of few-shot examples, so this seems like the training approach was not best suited for this model.

- *Silence representation:* I believe using the number of generated "*" characters to represent the silence length is a good choice, likely better than attempting to generate a numeric representation. However, I am not sure about the granularity, and the thesis is not clear here either – it mentions 10ms per "*" on pgs. 29 and 36 and 100ms on pg. 33. I believe optimizing the granularity would need some specifically targeted experiments. In addition, I am not convinced by the idea of distinguishing different types of silence (Sect. 4.5.4) – these types would need to be manually annotated, which would make the task quite expensive.

And regarding results and their interpretation:

- *Human performance:* This is a very important issue, and I believe this should have been a part of the thesis, at least on a very small scale. The task is likely very hard even for humans, especially with the limited context. Note that Lison & Meena report that dialogue turn segmentation is hard for humans too, without the visual context. Also note that the term "human evaluation" used in the thesis to refer to this task is misleading, as the default reading would be a human evaluation of the LLM's outputs, not establishing a human baseline.

[1] P. Lison and R. Meena, "Automatic turn segmentation for Movie & TV subtitles", *2016 IEEE Spoken Language Technology Workshop (SLT)*, San Diego, CA, USA, 2016, pp. 245-252, doi: 10.1109/SLT.2016.7846272.

- *Model exploiting data statistics:* I agree that the model most likely exploits relatively low-level statistics (Sect. 4.5.2). Some more analysis on this would be a good idea. However, this means that the model just may not be able to discern differences between movies and directors, not that there are no major differences (as the author implies in Chapter 5).
- A minor note: There seems to be something missing between pages 47 and 48, the sentence does not hold well together.

Finally, one small remark on the concluding Chapter 5:

- I appreciate an ethics statement, but I am not convinced that most of what is included here is very relevant for the work at hand.

Questions I have a few questions for the author, some of which could be discussed during the defense:

- What was the reason for not choosing other, more easily accessible language models (be it smaller models such as BERT or GPT-2, or larger open models)?
- Have you tried vanilla GPT-3 without finetuning? Have you used any prompts giving the model instructions, or any few-shot examples given to the model at runtime?
- Why are works that you cite in Chapter 4 introduction as focusing on silence (Jaworski 1997; Ibrahim & Ambu Mohammad, 2021) omitted from the theory discussion in Chapter 3?
- How come your result for *Inglourious Basterds* is exactly the same in the single-movie and four-movie experiments (in Tables 4.3 and 4.4)? I would expect a slight deviation, given that the models were finetuned on different amounts of data.
- There are a lot of spikes in the predicted outputs (most visible in Fig. 4.5) – could it be that the model has a preference for round values?
- It appears that the theoretical analysis was not really used in the experiment (there are no references to Chapter 3 from Chapter 4). Could it be used somehow? How could the theories, or their potential improvements, inform the choice of models and representations for automatic silence length prediction?

I recommend that the thesis be defended, with Grade 2, "very good (velmi dobře)".

I do not nominate the thesis for a special award.

Prague, 10 June 2023

Signature: