Automatic Transcript Generation from Presentation Slides

Nguyen Xuan Vu Nguyen
Aimesoft JSC
Hanoi, Vietnam
vunx@aimesoft.com

Ngo Quang Huy
Aimesoft JSC
Hanoi, Vietnam
huynq@aimesoft.com

Pham Quang Nhat Minh
Aimesoft JSC
Hanoi, Vietnam
minhpham@aimesoft.com

Abstract

With the advancement of generative AI models, there is an increasing interest in research and product development for the automation of presentations. However, one significant subtask that has been overlooked is preparing transcripts for speakers. In this paper, we introduce a new task of automatically generating transcripts from contents of presentation slides and a novel dataset to build and assess the quality of slide-to-transcript generation models. We also conduct experiments with large language models including LLaMA and Alpaca for the task and propose a method of creating prompts to guide these models to generate transcripts from slides. The experimental results are promising, but they also reveal challenges that need to be overcome to achieve good results for slide-to-transcript generation.

1 Introduction

Presenting a public speech is a time-consuming and laborious task. Whether it is an academic seminar or an investor meeting, the entire process, including idea generation, slide preparation, and delivering the talk to the audience, has conventionally relied on human involvement. While humans continue to play a vital role in this domain, we now acknowledge the potential for machines to automate certain aspects of the process. In this paper, we introduce a novel task of automatically generating speech transcripts for a given presentation, exploring the challenges and potential of this task.

Existing literature and software have already addressed some challenges in AI-assisted public speech generation, such as automated slide generation (Fu et al., 2022; Sefid and Wu, 2019; Wang et al., 2017; Hu and Wan, 2014; Utiyama and Hasida, 1999), slide-transcript alignment (Chen and Heng, 2003), and virtual speaker[1]. Virtual speaker software automates the slide presentation task; it automatically displays slides and uses speech synthesis technology to generate the speech given human written transcripts. We have developed and released one of the first virtual slide presentation software packages called AimeTalk. However, to the best of our knowledge, there has been no research focusing on the generation of transcripts for a given slide deck. This process offers several benefits. Firstly, it provides a solid starting point for speakers to plan their speeches effectively. Secondly, it produces a written record of the spoken content, serving as a valuable resource for speakers and audiences. Finally, the generated transcript can be seamlessly integrated with a text-to-speech module, enabling entirely autonomous speech delivery.

We recognize that the task of generating transcripts for presentations presents several challenges compared to other text generation tasks. First, generated transcripts need to convey a natural spoken language style to mimic human speech. In addition, the transcripts must align accurately with the corresponding slides, maintaining coherence and relevance. Furthermore, presentation slides themselves are often intricate, lacking a fixed format, and frequently contain various components beyond textual information, including images and diagrams, posing challenges to effectively represent these compositions for computers. Crucially, the absence of public datasets addressing this specific problem hinders the application of AI models in this domain.

This paper presents our initial effort in tackling the task of automatically generating presentation speech transcripts. Given the absence of an established benchmark, we curated a dataset comprising 1,634 pairs of presentation slide decks and their corresponding transcripts, resulting in approximately 23,000 slide-transcript pairs. To ensure the dataset's quality, we conducted data preprocessing and filtered out any irrelevant or erroneous data.

In order to leverage the dataset effectively, we

[1] Aimetalk: https://www.aimesoft.com/aimetalk.html

conducted experiments with some of state-of-the-art large language models, namely LLaMA (Touvron et al., 2023) and Alpaca (Taori et al., 2023). In order to address the effectiveness of our dataset, we finetuned these models on our dataset and compared them against the original Alpaca model. It should be noted that these language models are limited to processing only textual input; thus, our experiments solely focused on textual inputs extracted from presentation slides; other modalities such as images or diagrams were ignored. Additionally, due to the models' length restrictions, we could only input one slide per inference. In order to ensure that the models receive adequate context, we designed a comprehensive prompt that includes not only the content of the current slide but also adjacent slides and the overall context of the presentation.

To summarize, our contributions are two folds:

- We constructed a novel dataset of public speaking transcripts, named Slide2Transcript. The dataset contains pairs of a slide deck and a transcript, whose segments are aligned with each of the slides. To the best of our knowledge, ours is the first dataset for the task of generating transcripts from slides. To facilitate research in this domain, we plan to make the dataset openly accessible as an open-source resource for further experimentation and expansion.

- We conducted experiments of generating transcripts for a given slide deck using large language models. Furthermore, since the models are limited to being able to process one slide at a time, we devised a prompt format that captures both the content of the slide of interest and the context of the presentation. This approach ensures that the language models receive the necessary context to generate accurate and cohesive transcripts.

2 Related work

2.1 Natural language processing for public speaking assistance

Public speaking presents a challenging domain for natural language processing due to its incorporation of multiple modalities, including text, images, and audio. Existing literature primarily focuses on automated slide generation, where algorithms generate slide decks from source materials such as scientific papers or reports, with notable works in this area including (Fu et al., 2022; Sefid and Wu, 2019; Wang et al., 2017; Hu and Wan, 2014; Utiyama and Hasida, 1999). The primary strategy for this task involves paraphrasing sentences from the source materials into bullet points and incorporating necessary images into corresponding slides. Furthermore, text-to-speech techniques (Dutoit, 1997) have been extensively employed for automated speech delivery, assuming that transcripts are already available. However, to the best of our knowledge, no prior work has addressed the task of generating transcripts based on a given slide deck. This task represents the missing link in the complete pipeline of automated public speech delivery, as illustrated in Figure 1.

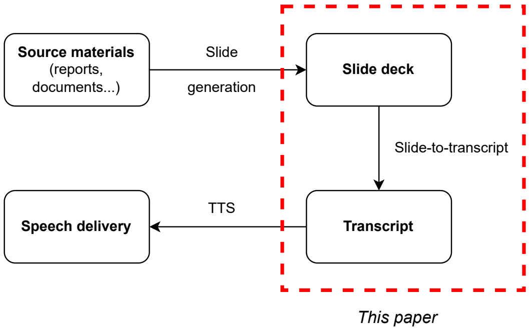


Figure 1: Fully automated public speech delivery pipeline.

2.2 Large language models

Language models refer to a family of algorithms designed to understand and generate natural language (Zhao et al., 2023). Recently, generative models, particularly autoregressive models, have been receiving a lot of attention. They have transformed from small pre-trained models requiring finetuning for specific tasks (Radford et al., 2019) to large models with billions of parameters, trained on extensive internet text, such as GPT-3 175B (Brown et al., 2020) and PaLM 540B (Chowdhery et al., 2022), allowing the understanding of complex structures of natural language.

By scaling up both model size and data size, LLMs are capable of zero and few-shot inference via prompting technique (Wei et al., 2022). When combined with instruction tuning, LLMs demonstrate the ability to follow human instructions with relatively high precision and helpfulness. Notably, models like ChatGPT and GPT-4 (OpenAI, 2023) have been reported to achieve human-like performance on a wide range of complex tasks.

Following the emergence of the GPT family, which is proprietary by OpenAI, the research community has become actively involved in developing open-source alternatives to foster community engagement in LLMs. Noteworthy mentions in this context include LLaMA (Touvron et al., 2023), a series of foundation models ranging from 7B to 65B parameters, and Alpaca (Taori et al., 2023), which was instruction tuned from LLaMA. The availability of these open-source alternatives has significantly accelerated research in LLMs.

LLMs have already demonstrated the capability of zero-shot inference across a wide range of tasks, but through additional finetuning on domain-specific data, their performance can be further enhanced for narrower domains. Nonetheless, the necessity of such a fine-tuning process for presentation transcript generation warrants additional empirical investigation.

3 Dataset

To the best of our knowledge, there is no dataset that can be used to directly address the slide-to-transcript generation task. In this work, we constructed our own dataset by crawling presentation files and their corresponding transcripts from the web. We collected data from data sources as follows.

- *videolectures.net*[2]: This platform hosts lectures covering a wide range of topics. Each lecture typically includes a video recording, a slide deck in PDF format, and a segmented transcript that aligns with the lecturer's slide transitions. The transcripts may be written by humans or generated by an automatic speech recognitions (ASR) engine. Nevertheless, due to factors such as audio quality, inaccuracies in ASR, or speakers exhibiting unclear enunciation, a certain degree of noise in the data is expected.

- Another valuable source of data is the presentation archives of various corporations, such as GSK[3], Imperial Brand[4], Nestlé[5], and Toshiba[6]. Both the slide decks and transcripts on these sites are available in PDF format. Based on our observations, the majority of the transcripts are human-generated, resulting in better overall transcript quality compared to those from *videolectures.net*. Those business-related slides occupy about 10% of our constructed dataset.

In order to improve the quality of the transcripts, we implemented the following pipeline:

- **Basic preprocessing.** We performed some basic processing steps, such as capital letter restoration, removing special characters, grammar correction using LanguageTool [7].

- **Back translation to improve fluency.** We translated the transcripts back and forth using an intermediate language. We observe that back translation could help transcripts become more coherent in some of the cases.

- **Transcript-slide alignment.** Initially, when we crawled the dataset from the web, the transcripts are already segmented to align with the slides. Each transcript segment contains a header that is expected to resemble the corresponding slide's header. However, upon observation, we notice that the alignment is often inadequate in numerous cases. Consequently, we made the decision to re-align the dataset by minimizing the disparity between the transcripts and slides.

  Since the slides are in PDF format, extracting their header is challenging due to the unstructured nature of the format. In order to address that problem, we minimized the partial edit distance between the headers of the transcripts and the extracted text from the slides. In other words, if the header of a transcript is found within a slide, we consider their similarity to be 1. This approach allows us to achieve a more accurate alignment between the transcript segments and the corresponding slides.

- **Filtering transcripts with short utterances**. We define an utterance as a segment of text bounded by punctuations. For each transcript, we calculated the average number of tokens per utterance by dividing the total number

[2] http://videolectures.net

[3] https://www.gsk.com/en-gb/investors/speeches-and-presentations/

[4] https://www.imperialbrandsplc.com/creating-shareholder-value/results-reports-and-presentations

[5] https://www.nestle.com/investors/presentations

[6] https://www.global.toshiba/ww/ir/corporate/pr.html

[7] https://github.com/jxmorris12/language_tool_python

Table 1: Dataset statistics

	# slide decks	*# slides*	*# transcript tokens*	*# avg. transcript tokens*	
				per slide deck	*per slide*
Train set	1,564	21,447	4,370,064	2,794	204
Test set	70	1849	292,125	4,713	158
Total	1634	23,296	4,662,189	2,853	200

of tokens in a transcript by the number of utterances it contains. We then filtered out transcripts whose average utterance length falls below a specific threshold. This filtering approach aims to address situations where speakers tend to stutter frequently during their presentation, resulting in shorter utterances within the recorded transcript.

- **Filtering out transcript segments with low relevance to their corresponding slides.** In order to assess the relevance between each transcript segment and the textual content of its respective slide, we computed the ROUGE-1 recall score. Segments that obtained scores below a designated threshold were filtered out. A low recall score indicates that the segment exhibits limited relevancy to the slide. This lack of relevancy could arise from various factors, including characteristics of the speech itself or errors introduced during the transcript-slide alignment process.

Ultimately, we obtained a total of 1,634 presentations accompanied by their respective transcripts, resulting in 23,296 slide-transcript segment pair. To ensure comprehensive evaluation, we partitioned the dataset into a train-set and a test-set. The detailed statistics of the dataset can be found in Table 1.

4 Model

In this work, we used Meta's LLaMA (Touvron et al., 2023) and Alpaca (Taori et al., 2023) to conduct experiments on our dataset. LLaMA is a family of foundation models, varying from 7B to 65B parameters, and pre-trained on trillions of tokens from publicly available datasets. On the other hand, Alpaca is a model finetuned from LLaMA 7B using a self-instructed learning approach with 52K instruction-following examples. LLaMA serves as the base model for further finetuning in various downstream tasks, while Alpaca has demonstrated the capability to perform a diverse range of tasks due to its ability to follow human instructions.

Within the scope of this study, given our primary objective of evaluating the dataset's effectiveness with an arbitrary language model, along with hardware limitation, we decided to conduct finetuning LLaMA 7B and Alpaca models on our constructed slide-to-transcript datasets. To assess the impact of finetuning on the models' performance in this task, we compared the finetuned models with the base Alpaca model.

5 Prompt

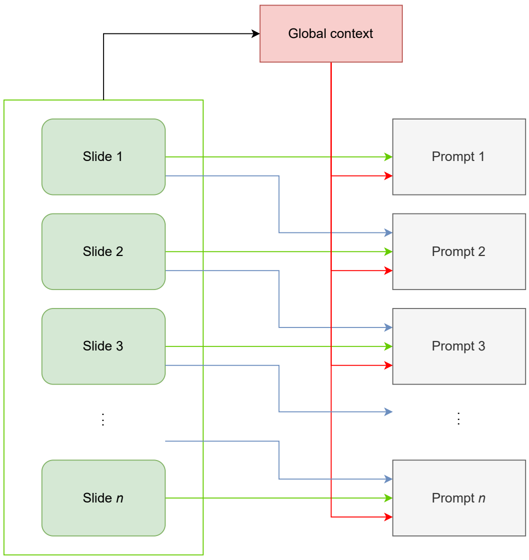
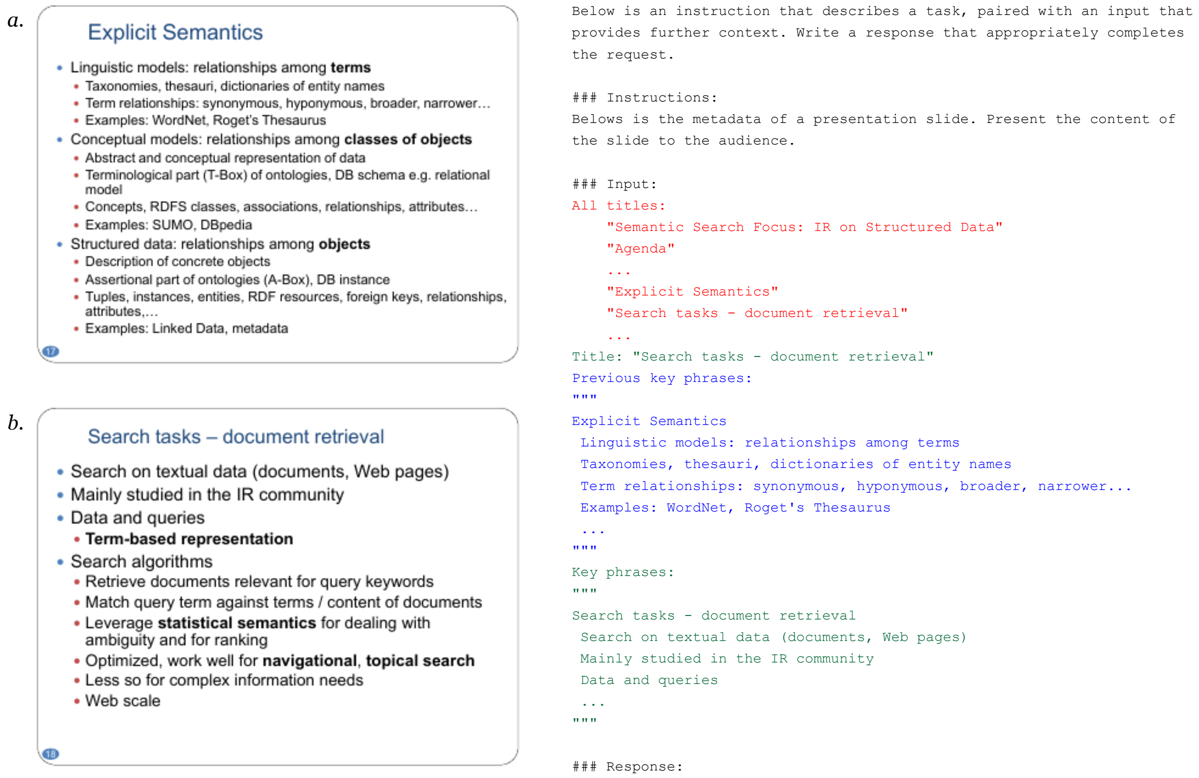


Figure 2: For each slide, the prompt contains its textual content and title (green arrow), textual content of the preceding slide (blue arrow), and the global context (red arrow).

In many cases, it is impractical to input an entire presentation consisting of multiple slides into a model and generate all the transcripts simultaneously due to limitations in model capacity. Hence, we used models to generate transcripts slide-by-slide using textual content of each slide. To ensure coherence in the generated transcripts, we supplemented the textual content of each slide with additional global information provided in the prompt (Figure 2). For each slide, the prompt encompasses components as follows.

- *Global context*: This component provides



a.

Explicit Semantics

- Linguistic models: relationships among **terms**
  - Taxonomies, thesauri, dictionaries of entity names
  - Term relationships: synonymous, hyponymous, broader, narrower…
  - Examples: WordNet, Roget's Thesaurus
- Conceptual models: relationships among **classes of objects**
  - Abstract and conceptual representation of data
  - Terminological part (T-Box) of ontologies, DB schema e.g. relational model
  - Concepts, RDFS classes, associations, relationships, attributes…
  - Examples: SUMO, DBpedia
- Structured data: relationships among **objects**
  - Description of concrete objects
  - Assertional part of ontologies (A-Box), DB instance
  - Tuples, instances, entities, RDF resources, foreign keys, relationships, attributes,…
  - Examples: Linked Data, metadata

b.

Search tasks – document retrieval

- Search on textual data (documents, Web pages)
- Mainly studied in the IR community
- Data and queries
  - **Term-based representation**
- Search algorithms
  - Retrieve documents relevant for query keywords
  - Match query term against terms / content of documents
  - Leverage **statistical semantics** for dealing with ambiguity and for ranking
  - Optimized, work well for **navigational**, **topical search**
  - Less so for complex information needs
  - Web scale

```
Below is an instruction that describes a task, paired with an input that
provides further context. Write a response that appropriately completes
the request.

### Instructions:
Belows is the metadata of a presentation slide. Present the content of
the slide to the audience.

### Input:
All titles:
    "Semantic Search Focus: IR on Structured Data"
    "Agenda"
    ...
    "Explicit Semantics"
    "Search tasks - document retrieval"
    ...
Title: "Search tasks - document retrieval"
Previous key phrases:
"""
Explicit Semantics
 Linguistic models: relationships among terms
 Taxonomies, thesauri, dictionaries of entity names
 Term relationships: synonymous, hyponymous, broader, narrower...
 Examples: WordNet, Roget's Thesaurus
 ...
"""
Key phrases:
"""
Search tasks - document retrieval
 Search on textual data (documents, Web pages)
 Mainly studied in the IR community
 Data and queries
 ...
"""

### Response:
```

Figure 3: Example of a prompt describing a slide. Left: a) The slide immediately preceding the slide of interest; b) The slide of interest. Right: The corresponding prompt for the slide of interest in b).

global context of a whole presentation, aiding the models in understanding the overall presentation structure. In our experiments, we used a list of all slide titles as global context.

- *Current slide's content*: This component includes the title and the textual content extracted from a slide. During training, in cases where no textual content can be extracted (e.g., slides containing only images), we used RAKE algorithm (Rose et al., 2010) to extract essential keywords from the golden transcript. These extracted keywords were then utilized as the content for the corresponding slides.

- *Previous slide's content*: This component contains the textual content or transcript keywords corresponding to the slide immediately preceding the current slide. This component can be left out if there is no preceding slide.

An example prompt containing all three components is shown in Figure 3 for the purpose of demonstration.

6 Experiments

6.1 Metrics

We assessed the quality of a generated transcript using two key criteria. Firstly, the transcript should exhibit fluency, enabling it to be directly read to the audience smoothly. Secondly, it should effectively convey all the content depicted on its corresponding slide, ensuring a coherent and focused discussion without deviating from the main topic.

Based on this intuition, generated transcripts are evaluated using the following metrics:

- *Perplexity*: We measured the perplexity of the models on the golden transcript in the test set.

- *ROUGE-F2 score vs. slide content*: We report ROUGE-1, ROUGE-2, and ROUGE-L between the generated transcripts and the textual content of the corresponding slides as the indicators of the relevance of the transcript to the source materials. We set *beta* to 2 when calculating the *F* score.

- *ROUGE-F1 score vs. golden transcripts:* We also report ROUGE-1, ROUGE-2 and

Table 2: Finetuning hyperparameters

Hyperparameter	Value
learning rate	3e^{-4}
warm-up steps	100
batch size	1
grad accumulation steps	128
max sequence length	1024
lora r	8
lora alpha	16
lora dropout	0.05

ROUGE-L between the generated transcripts and the golden transcripts, providing an indication of their similarity to the ground truth.

- *GPT-4 based evaluation*: We randomly selected a subset of presentations in the test set along with their corresponding models' responses and asked GPT-4 to compare and rank the outputs of models.

6.2 Baseline

To the best of our knowledge, slides-to-transcript is a novel task with no established baseline. Therefore, we adopted the original Alpaca as our baseline because Alpaca is an instruction-tuned model, which can follow human instructions. Alpaca is a quite strong baseline model in the task.

6.3 Hyperparameters

We employed several low-resource training strategies during the finetuning process. Notably, we utilized LoRA (Hu et al., 2021), an adaptation strategy that significantly reduces the number of trainable parameters, often by up to 10,000 times. In all of our experiments, we kept the parameters of LoRA fixed at: $r = 8$, $alpha = 16$, and $dropout = 0.05$. By combining this approach with the utilization of the *bfloat16* datatype, we could successfully finetune LLaMA 7B and Alpaca models on a single NVIDIA GeForce RTX 3090 24GB GPU. The full set of hyperparameters is shown in Table 2.

During the inference phase, we employed a sampling strategy with a top-k parameter set to 10 and a temperature value of 0.8.

7 Results

7.1 Conventional evaluation

Table 3 shows the performance of LLaMA 7B and Alpaca finetuned on our dataset, compared with the base Alpaca served as baseline. We compared the ROUGE scores between the model-generated transcripts and the textual content of the corresponding slides, as well as the ROUGE scores between the generated transcripts and the golden transcripts. Additionally, we calculated perplexity scores of the two finetuned models and the baseline on the golden transcripts.

The ROUGE scores with respect to the textual contents of source slides indicated that the transcripts generated by the original Alpaca are more closely related to the source slides compared the two finetuned models. However, when we analyzed outputs of models, we found that the higher ROUGE scores of the original Alpaca are primarily attributed to its tendency to directly copy the slide content into the transcript. In contrast, through finetuning on human speech transcripts, both the finetuned LLaMA and Alpaca models demonstrated the ability to paraphrase slide contents, which is more desirable for effective public speaking. Based on this observation, we conclude that while the ROUGE scores between generated transcripts and slide content can provide a sense of relevance, they should not be considered the sole decisive metric for evaluating the quality of generated transcripts. Figure 4 shows sample outputs of the three models, which exhibit that mentioned tendency.

On the other hand, by finetuning on a dataset originating from human speech, LLaMA and Alpaca models can effectively capture the speech patterns exhibited by humans in that dataset. Finetuned models achieved higher ROUGE scores with respect to the gold-standard transcripts and lower perplexity scores in comparison to the base Alpaca model.

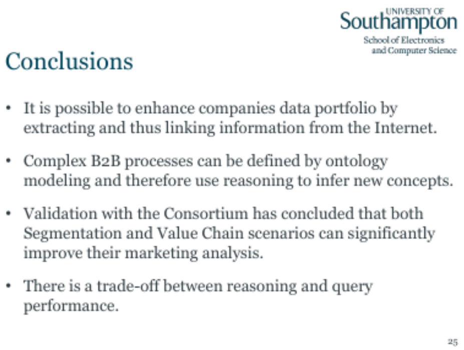
7.2 GPT-4 based evaluation

The task of generating transcripts from slides shares similarities with open-ended text generation, making evaluation challenging. Moreover, this task necessitates attention to the relevance of generated transcripts to the source slide and the overall presentation, further adding to the complexities of evaluation.

With the recent advancements in LLMs, recent studies propose directly using LLMs for evaluating natural language generation models without reference texts (Liu et al., 2023). Notably, GPT-4 has demonstrated near-human performance on various benchmarks (OpenAI, 2023). Therefore, in this work, in addition to conventional metrics, we

Table 3: Comparison to base Alpaca, using conventional metrics.

Compared with	*Metrics*	*Base Alpaca*	*Finetuned LLaMA*	*Finetuned Alpaca*
Source slide	ROUGE-1-F2	**0.360**	0.310	0.325
	ROUGE-2-F2	**0.224**	0.136	0.148
	ROUGE-L-F2	**0.300**	0.250	0.264
Golden transcript	ROUGE-1-F1	0.243	0.303	**0.307**
	ROUGE-2-F1	0.045	**0.072**	**0.072**
	ROUGE-L-F1	0.142	0.173	**0.178**
	Perplexity	167.74	**21.889**	22.091



Conclusions

- It is possible to enhance companies data portfolio by extracting and thus linking information from the Internet.
- Complex B2B processes can be defined by ontology modeling and therefore use reasoning to infer new concepts.
- Validation with the Consortium has concluded that both Segmentation and Value Chain scenarios can significantly improve their marketing analysis.
- There is a trade-off between reasoning and query performance.

25

```
ALPACA (F2-ROUGE-L = 0.76)

This slide concludes the presentation. We have seen how semantic
web technologies can be used to enhance companies' data
portfolio by extracting and linking information from the
internet. We have also seen how complex B2B processes can be
defined by ontology modeling and thus use reasoning to infer new
concepts. Finally, we have seen how the Consortium has concluded
that both Segmentation and Value Chain scenarios can
significantly improve their marketing analysis. There is a
trade-off between reasoning and query performance however.
=====================================
FINETUNED LLAMA (F2-ROUGE-L = 0.28)

The conclusion, and the main result we can extract from it, is
that it is possible to improve companies' data portfolios by
extracting data from the Internet and therefore the link to the
information, for example, in the segmentation scenario. This has
allowed us to define complex business processes that can be
modeled using ontology and therefore use reasoning to derive new
concepts. The validation of the consortium has concluded that
both the scenarios are significantly improved in their market
analysis. There is a trade-off between reasoning and
performance, query.
=====================================
FINETUNED ALPACA (F2-ROUGE-L = 0.62)

I think that we have achieved our goal here. We have extracted
the data from the website and have linked it to the company. We
have developed and tested a model that allows us and the company
to create complex B2B processes and to use reasoning to define
the concepts that have not previously been defined. The
validation has been very positive for this and the two
scenarios. We were able to significantly improve the analysis of
the marketing segmentation and the value chain. The performance
of the model is still a bit of a trade-off but I think it's
worth it.
```

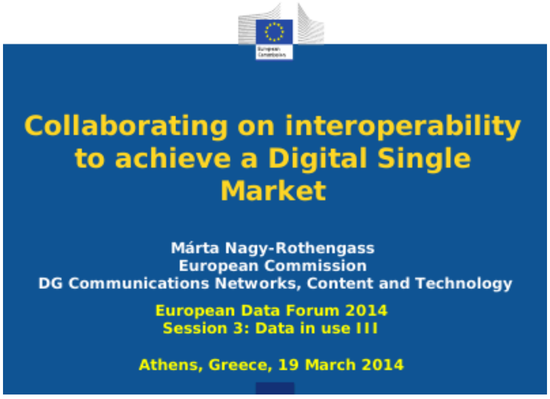
Figure 4: Comparison of generated transcripts between the original Alpaca and the finetuned LLaMA and Alpaca models, with the slide on the left serving as the input. Despite achieving lower ROUGE scores, the finetuned models exhibit the capability to paraphrase the slide content.

used GPT-4 to evaluate results of slide-to-transcript generation. First, from the test set, we randomly sampled out 5 slide decks, corresponding with 94 slides. For each slide, providing GPT-4 with the same context and input as utilized during the finetuning of LLaMA and Alpaca models, we asked GPT-4 to rank transcripts produced by the three models based on several criteria including naturalness, relevance to the source materials, coherence with the surrounding context, and consideration of the global context.

The results as shown in Figure 5 indicated that GPT-4 consistently produces well-reasoned rankings. When comparing the responses of the finetuned models with the base Alpaca, out of 94 samples in the test set, GPT-4 preferred approximately 36 transcripts from finetuned LLaMA over those of the base Alpaca. Conversely, 54 transcripts from finetuned Alpaca were favored over those of the base Alpaca, as illustrated in Figure 6. This improvement in favor of finetuned Alpaca might be attributed to our prompt design, which is based on the Alpaca prompt format. As Alpaca was already finetuned using this prompt format, it appears to exhibit more effective learning capabilities compared to LLaMA.

8 Discussion

Firstly, the majority of our dataset is sourced from *videolectures.net*, a website primarily hosting academic lectures. Consequently, the transcripts exhibit a highly academic tone. Despite our efforts to cull low-quality transcripts, a few samples still suffer from stuttering or excessive repetition, which adversely impacts the quality of the generated transcripts. In order to address this problem, in future work, we plan to explore additional sources of high-quality data.



Collaborating on interoperability to achieve a Digital Single Market

Márta Nagy-Rothengass
European Commission
DG Communications Networks, Content and Technology

European Data Forum 2014
Session 3: Data in use III

Athens, Greece, 19 March 2014

Candidate 1: Base Alpaca *(Rank 3)*	Candidate 2: Finetuned LLaMA *(Rank 1)*	Candidate 3: Finetuned Alpaca *(Rank 2)*
The aim of this session is to explore ways to increase data sharing between public administrations, businesses and citizens to benefit from the Digital Single Market, the European Commission's strategy to create a single market for digital goods and services. To this end, the European Commission's DSM Action Plan is promoting the development of interoperability solutions to facilitate data sharing. \n\nThe European Data Forum is a great place to start collaborating on this effort. The European Data Forum brings together experts from various fields to discuss how to make data more accessible, better quality and more secure while protecting privacy and creating economic value. During this session, Marta Nagy-Rothengass from the European Commission's DG Communications Networks, Content and Technology will explain the importance of Data Interoperability in the context of the European Commission's DSM Action Plan.	I'm Marta Nagy-Rothengass with the European Commission. I am a project manager at DG Connectivity, Networks, Content and Technology. I'm here today to introduce the European Data Forum and I have a very short introduction on the European data policy in general terms. The European data policy is part of the Commission's strategy for a digital single market. This strategy, launched in the fall of this year, is designed to unlock the potential of digital technologies to promote economic growth and social and cultural development of our economy and to create new jobs. The digital single market is the European Commission's main priority for the next five years. It will focus on areas in which we believe that the EU can deliver the most value for our economy and our society by using digital technologies in a smarter, more inclusive and more effective way, and this includes data.	We are going to talk about collaborating on interoperability to achieve a digital single market, and I will start with a little bit of context. We have to consider the digital single market from the perspective of the European Commission's 2015-2018 policy framework and in the context of the 2014-2020 policy framework. In the context of the Digital Single Market, the Commission has set a goal and it has a clear goal.

GPT-4 Reasoning
Candidate 2 is ranked first as it provides a natural sounding introduction to the speaker and the topic, it captures the content of the slide well, and it suits the context of being the first slide in the presentation. Candidate 3 is ranked second as it also provides a decent introduction to the topic and sounds relatively natural. However, it doesn't mention important specifics like the speaker's name and the European Data Forum. Candidate 1 is ranked last because it doesn't mention the speaker's name or properly acknowledge that this is the start of the presentation. It launches into a discussion about the topic that is better suited to a later slide.

Figure 5: GPT-4 is able to evaluate and compare the produced transcripts while incorporating reasoning abilities.

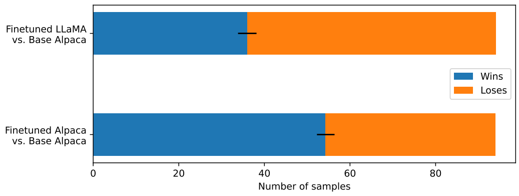


Figure 6: Comparison of produced transcripts, assessed by GPT-4.

Secondly, in this paper, our experiments have been limited to finetuning large language models, thereby inheriting their weaknesses, including the issue of hallucination. Furthermore, as large language models solely process textual input, they do not encompass other modalities commonly found in slides, such as images, tables, and diagrams. To address these non-textual modalities, the integration of visual-language models is essential (Zhang et al., 2023). Thus, we aim to conduct future experiments involving such models in future research.

Lastly, we lack a reliable and rigorous method to evaluate the generated transcripts. An evaluation system for a context-based open-ended text generation remains an open challenge demanding further research.

9 Conclusion

In this study, we introduced the slide-to-transcript generation task and curated a new dataset consisting of slide-decks paired with their corresponding transcripts. We concentrated our investigation on the performance of large language models, specifically LLaMA and Alpaca. In order to effectively feed one slide at a time to these models, we devised a prompt format that encompasses not only the content of the current slide but also the surrounding context. Through our GPT-4 based evaluation, we found that finetuning Alpaca using our dataset results in transcripts that are more preferable in comparison to the base Alpaca model. Those preferences are measured in terms of naturalness, relevance to the source materials, and cohesiveness with the global context.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yu Chen and Wei Jyh Heng. 2003. Automatic synchronization of speech transcript and slides in presentation. In *2003 IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 2, pages II–II. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Thierry Dutoit. 1997. *An introduction to text-to-speech synthesis*, volume 3. Springer Science & Business Media.

Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. 2022. Doc2ppt: Automatic presentation slides generation from scientific documents.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yue Hu and Xiaojun Wan. 2014. Ppsgen: Learning-based presentation slides generation for academic papers. *IEEE transactions on knowledge and data engineering*, 27(4):1085–1097.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

OpenAI. 2023. Gpt-4 technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. In Michael W. Berry and Jacob Kogan, editors, *Text Mining. Applications and Theory*, pages 1–20. John Wiley and Sons, Ltd.

Athar Sefid and Jian Wu. 2019. Automatic slide generation for scientific papers. In *Third International Workshop on Capturing Scientific Knowledge co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), SciKnow@ K-CAP 2019*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Masao Utiyama and Koiti Hasida. 1999. Automatic slide presentation from semantically annotated documents. In *Coreference and Its Applications*.

Sida Wang, Xiaojun Wan, and Shikang Du. 2017. Phrase-based presentation slides generation for academic papers. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.