Large Language Models in Speech Recognition

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Author: Martin Tomašovič Supervisor: Ing. Karel Beneš

Excel@FIT 2024

- I chose the *n-best* rescoring method, to use LLM in ASR
- the decoder outputs *n* hypotheses
- the hypotheses are reordered using LLM
- the hypothesis with the best score should be better then the original best hypothesis

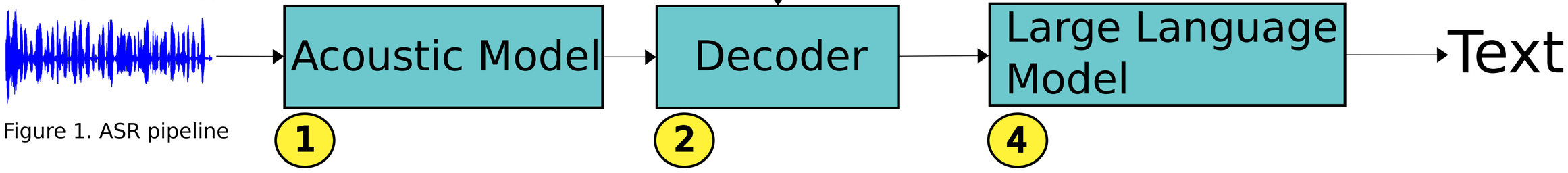


Figure 1. ASR pipeline

1 Wav2Vec 2.0 Base 960h, Whisper Medium or STT En Jasper10x5dr

2 decoders use Beam Search to output multiple hypotheses
Wav2Vec 2.0 and Jasper uses CTC decoder

3 Wav2Vec 2.0 and Jasper use KenLM during decoding

Large Language Models

4
- Masked: BERT, RoBERTa
- Autoregressive: GPT-2, TinyLlama, Falcon, Mistral, MPT, Llama2

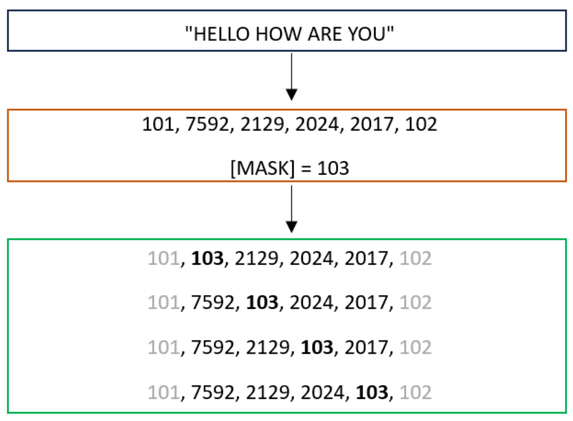


Figure 2. Data processing for masked language model.

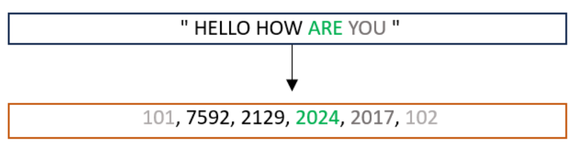


Figure 3. Data processing for autoregressive language model.

Fine-tuning

- Llama2 7B fine-tuned using LoRA on the GigaSpeech XL text
- BERT base and GPT-2 fine-tuned on the LibriSpeech train clean 100h text

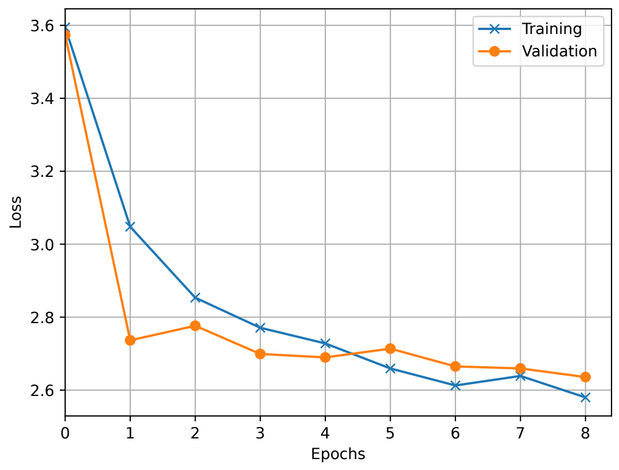


Figure 4. BERT training.

Rescoring

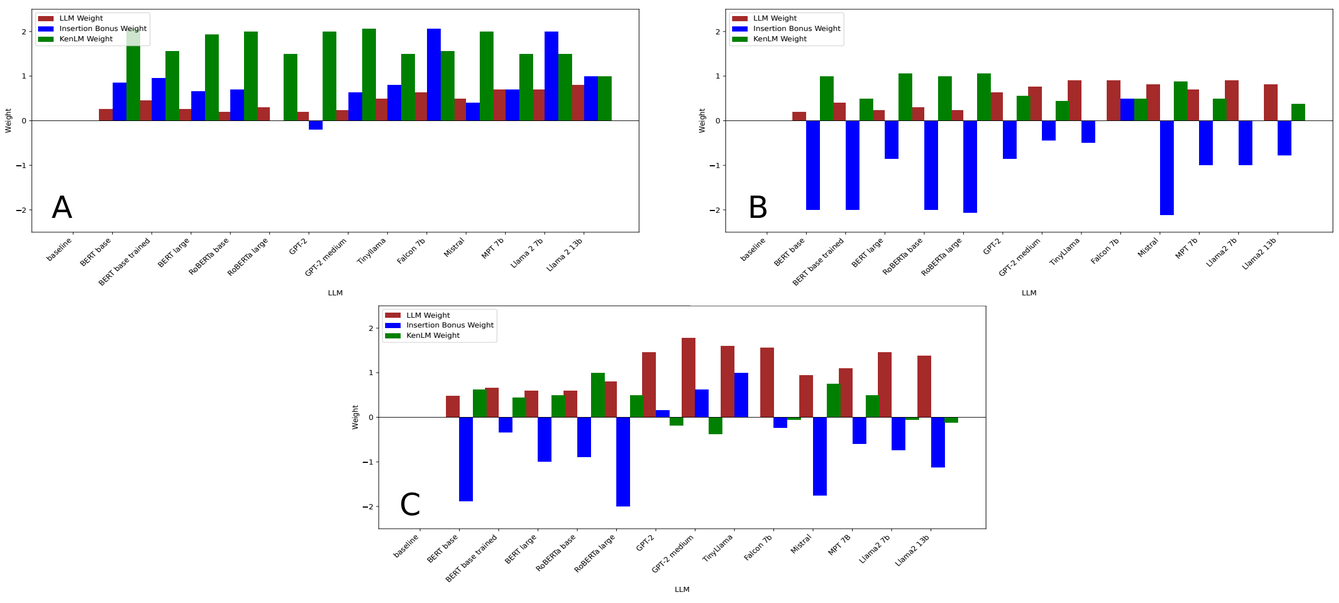


Figure 5. Wav2Vec score weights of LLMs, KenLM and insertion bonus.
On A: LibriSpeech, B: GigaSpeech and C: TED-LIUM dataset.

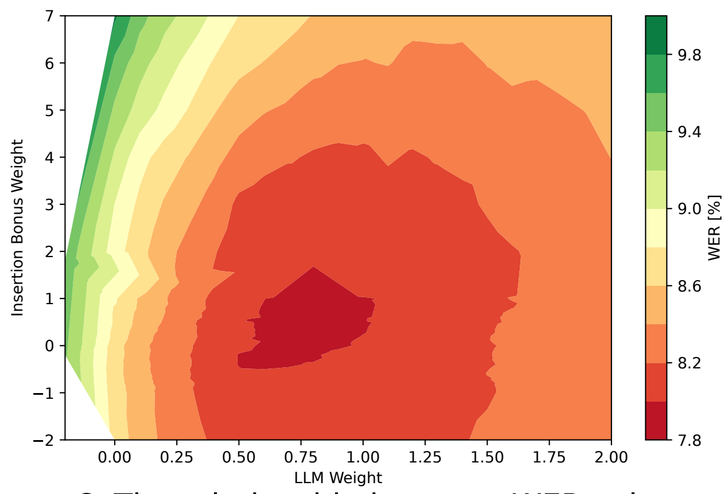


Figure 6. The relationship between WER values and the weights of rescoring LibriSpeech dev other using Llama2 7B with LoRA r = 128, α = 256, hypotheses are obtained from Jasper.

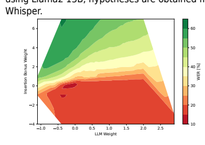
Results

- rescoring does not improve the WER of GigaSpeech hypotheses from Whisper but rescoring Wav2Vec 2.0's and Jasper's hypotheses workes on all datasets

Table 3. Whisper best WER on GigaSpeech dev. β means the weight of LLM, γ is the insertion bonus weight. Results show LLMs do not improve WER.

LLM	Parameters (in B)	β	γ	best WER[%] ↓	best CER[%] ↓
BERT base	0.11	-0.0625	0.1375	12.93	8.1
BERT base trained	0.11	-0.13125	0.03125	13.01	8.01
BERT large	0.34	-0.046875	0.184375	13.05	8.23
RoBERTa base	0.125	-0.015625	0.184375	13.2	8.4
RoBERTa large	0.355	-0.015625	0.184375	13.19	8.42
GPT-2	0.137	-0.1	0.0	12.83	7.96
GPT-2 medium	0.380	-0.084375	0.015625	12.91	7.96
TinyLlama	1.1	-0.084375	0.015625	13.0	8.05
Falcon	7	-0.06875	0.03125	13.02	8.09
Mistral	7	-0.084375	0.046875	13.09	8.14
MPT	7	-0.1	0.05	12.99	8.0
Llama2	7	-0.06375	0.046875	13.05	8.08
Llama2	13	-0.053125	0.046875	13.06	8.2

Figure 7. The relationship between WER values and the weights of rescoring GigaSpeech dev using Llama2 13B, hypotheses are obtained from Whisper.



- the best WER improvement is 4%
- fine-tuned models and the biggest one Llama2 13B performed the best
- all 7B models are very effective in specific experiments
- even though small, TinyLlama 1.1B improved WER competitively with bigger models
- masked models were mostly better in LibriSpeech rescoring, autoregressive of the same size in GigaSpeech and TED-LIUM

Table 1. WER↓ after rescoring Wav2Vec 2.0. KenLM means the KenLM n-gram model is being used during decoding, Lex means with and No Lex without lexicon decoding.

LLM	Parameters (in B)	LibriSpeech dev other			GigaSpeech dev			TED-LIUM dev
		KenLM Lex	Lex	No Lex	KenLM Lex	Lex	NoLex	KenLM Lex
baseline	–	8.59	9.33	9.77	28.49	30.37	31.34	17.81
BERT base	0.11	6.31	7.79	8.13	25.61	27.63	28.93	15.07
BERT base trained	0.11	6.11	7.45	7.87	25.32	27.28	28.59	14.93
BERT large	0.34	6.25	7.94	8.35	25.56	27.69	29.12	14.98
RoBERTa base	0.125	6.38	7.82	8.48	25.32	27.25	28.83	14.85
RoBERTa large	0.355	6.31	7.83	8.53	25.26	27.49	29.1	14.84
GPT-2	0.137	6.54	7.91	8.33	25.19	27.02	28.43	14.63
GPT-2 medium	0.380	6.44	7.64	8.16	25.05	26.93	28.34	14.57
TinyLlama	1.1	6.28	7.44	7.88	24.52	**26.47**	28.19	14.4
Falcon	7	6.12	7.22	7.77	**24.44**	26.58	**28.09**	14.13
Mistral	7	6.11	7.39	8.18	24.75	26.98	28.59	14.76
MPT	7	6.09	7.26	7.78	24.64	26.82	**28.09**	**14.08**
Llama2	7	6.07	**7.16**	**7.58**	24.59	26.62	**28.06**	**14.02**
Llama2	13	**5.92**	**7.07**	**7.51**	**24.44**	**26.5**	**28.07**	**13.99**

Table 2. WER↓ after rescoring Jasper.

LLM	Parameters (in B)	LibriSpeech	GigaSpeech	TED-LIUM	
	dev other	dev	dev	test	
baseline	–	8.69	28.26	14.01	14.67
BERT base	0.11	8.44	27.88	13.53	14.67
BERT base trained	0.11	**8.22**	27.8	13.47	14.21
BERT large	0.34	8.45	27.85	13.38	14.5
RoBERTa base	0.125	8.51	27.7	13.23	14.17
RoBERTa large	0.355	8.44	27.71	13.27	14.27
GPT-2	0.137	8.59	27.77	13.24	14.15
GPT-2 trained 1	0.137	8.56	27.98	13.36	14.41
GPT-2 trained 2	0.137	8.49	27.92	13.33	14.24
GPT-2 medium	0.380	8.46	27.64	13.1	14.19
TinyLlama	1.1	8.45	27.49	13.03	**13.86**
Falcon	7	**8.26**	27.35	**12.91**	**13.96**
Mistral	7	**8.23**	27.52	**12.83**	14.15
MPT	7	**8.26**	27.6	13.02	14.44
Llama2	7	**8.24**	27.46	13.01	**13.81**
Llama2 LoRA r8/α16	7	**7.98**	**27.25**	**12.76**	**13.73**
Llama2 LoRA r128/α256	7	**7.92**	**27.26**	**12.75**	**13.62**
Llama2 LoRA r256/α128	7	**7.94**	**27.27**	**12.76**	**13.68**
Llama2	13	**8.23**	27.46	**12.87**	**13.72**