Integrating multiple ASR systems into NLP backend with attention fusion

*Takatomo Kano, Atsunori Ogawa, Marc Delcroix (NTT Corporation), and Shinji Watanabe (Carnegie Mellon University)

SPE-8.5
icassp 2022 Singapore

Background

- SLP systems, such as speech translation or summarization, convert a speech signal into a text document, e.g., a translation or summary
- SLP systems can be realized by combining an ASR frontend and NLP backend such as text summarization (TS) and machine translation (MT)

Difficult to achieve perfect ASR → ASR errors propagate to NLP backend

We propose an attention-based ASR hypothesis fusion:

- We exploit results from various ASR systems showing different error tendencies and expect that the correct meaning can be extracted from the multiple ASR results
- The fusion process is implemented within the NLP backend allowing to considers the context and meaning of the sentences
- The fusion mechanism is optimized for the SLP tasks

E.g. speech summarization : Make a long story short

ASR frontend: Input speech → Output transcription

ASR result: We **reuse** the ASR error propagation on NLP …

NLP backend: Input transcription → Output summarization

Summary: Utilize ASR error for…

ASR 1	We reuse the ASR error propagation on NLP backend using multiple ASR hypotheses …
ASR 2	We **reduce** an ASR error propagation on LNG _____ using multiple ASR hypotheses …
ASR 3	We reuse an ASR error proportion on NLP backend using multiple AMR hypotheses …

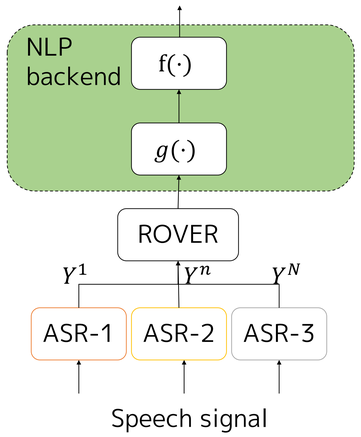
Summary: Reduce ASR error propagation on NLP…

Proposal: Attention-based ASR hypotheses fusion considering the word meaning and context

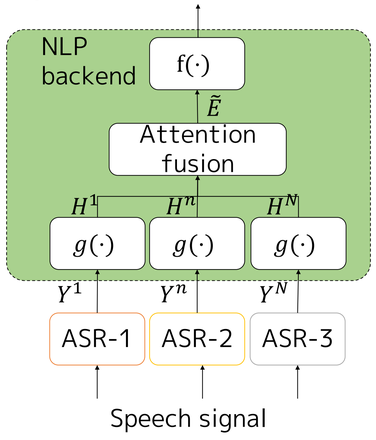
We find out the correct output by comparing the meaning and content of each word in the ASR hypotheses, even if each recognition result individually is wrong

- We utilize pre-trained NLP backends trained on large text only corpora, which can model word meanings and context information
- We combine each hypothesis inside the NLP backend encoder using two attention mechanisms performing alignment and combination
- The attention fusion layer considers the meaning and context of input hypothesis sequence based on NLP intermediate representation

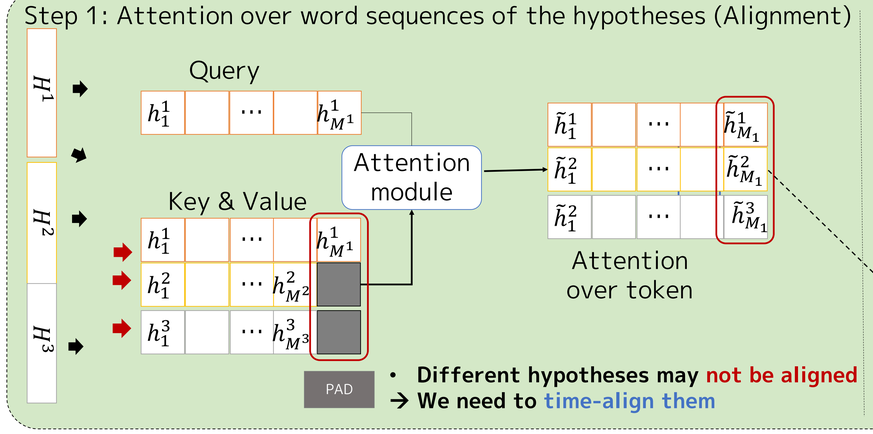
Conventional ASR hypotheses fusion



Proposal

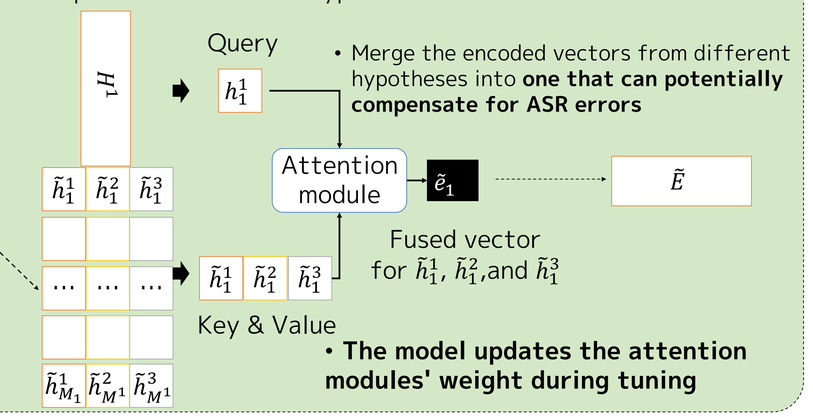


Step 1: Attention over word sequences of the hypotheses (Alignment)



- Different hypotheses may not be aligned
  → We need to time-align them

Step 2: Attention over hypothesis words



- Merge the encoded vectors from different hypotheses into **one that can potentially compensate for ASR errors**
- **The model updates the attention modules' weight during tuning**

Experiments

ASR frontend
Transformer encoder + Transformer decoder

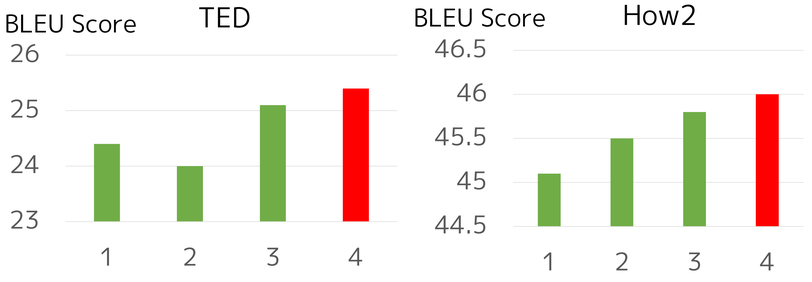
WER [%] for ASR systems with different BPE sizes

	500	5k	10k	20k	30k	30.5k	ROVER
How2	n/a	13.0	13.6	14.1	14.3	14.6	**12.2**
TED	8.5	n/a	8.7	9.5	10.0	10.4	**8.3**

Datasets	How2	TED
Documents	12k	1.5k
Sentences	72k	4k
Languages	En->Pt	En->De
Summarization ratio	16%	5%

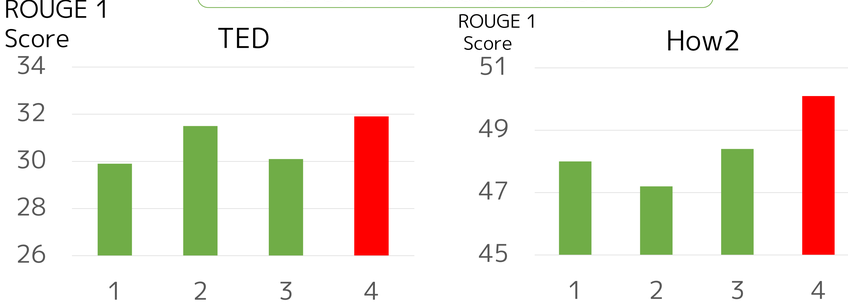
Model settings	MT	TS
Num of hypotheses	5	5
Encoder depth	6	12
Fusion layer	5-th	8-th
Pre-train model	None	BERT-base

NLP backend: Machine Translation
Transformer encoder + Transformer decoder



1.Tuning NLP backend with the best ASR system's transcription
3.Tuning NLP backend with ASR statics, e.g., posterior probability

NLP backend: Text Summarization
BERT encoder + Transformer decoder



2. Input the ROVER system output to NLP backend
4.[Proposed] ASR hypotheses fusion within the NLP backend