Multilingual Second-Pass Rescoring for Automatic Speech Recognition Systems

Neeraj Gaur, Tongzhou Chen, Ehsan Variani, Parisa Haghani, Bhuvana Ramabhadran, Pedro J. Moreno

Google Inc., USA

Introduction

- Second-pass rescoring is a well known technique to improve the performance of Automatic Speech Recognition (ASR) systems.
- Multilingual first-pass speech recognition models often outperform their monolingual counterparts.
- The rescoring model can be made ***multilingual***.
- First-pass multilingual model does not require a language-id.
- An estimate of the language-id would be available for second-pass rescoring.

Neural Oracle Search (NOS)

- NOS treats the oracle search problem as a sequence classification problem.

$$\text{maximize } P\left(\text{Oracle\_index} = i \,|\, X, H_1, \cdots, H_N\right)$$

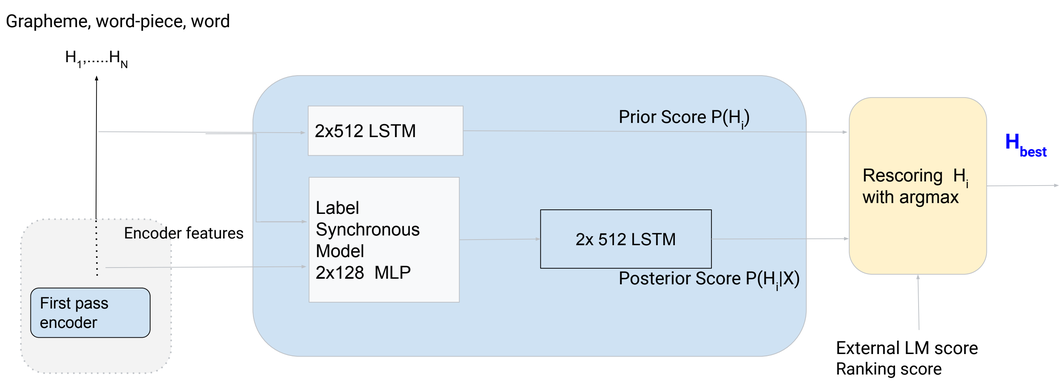
$$P\left(\text{Oracle\_index} = i \,|\, X, H_1, \cdots, H_N\right) = \frac{\exp\left(S(X, H_i)\right)}{\sum_j \exp\left(S(X, H_j)\right)}$$

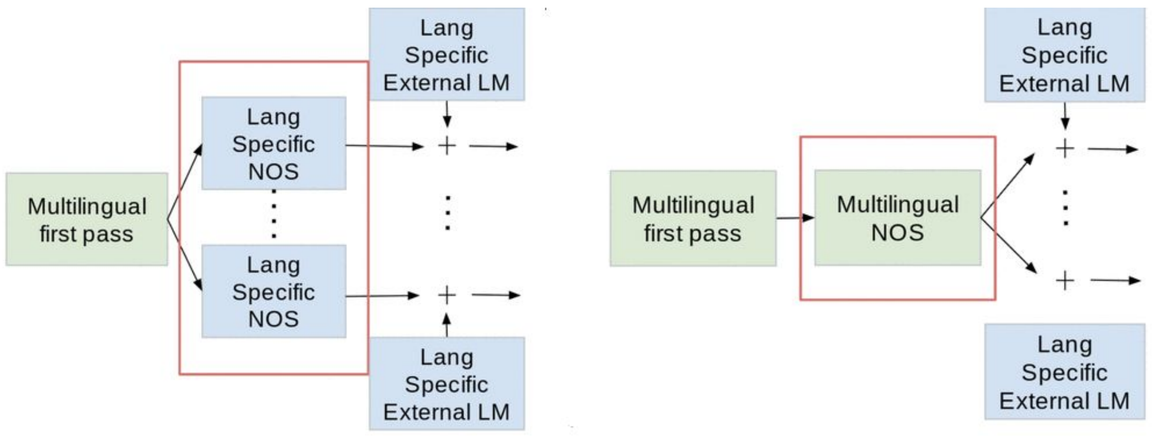
$S(X, H_i)$: Joint score of audio X and hypothesis H_i

X : Acoustic feature representation

H_i : Hypothesis ranked i from 1st-pass

- $S(X, H_i)$ consists of 3 parts:
  - Likelihood score $S_1(X|H_i) = \log P(H_i|X) - \log P(H_i)$.
  - External LM score $S_2(H_i)$.
  - Ranking score $S_3(i)$.





Experimental Setup

- Languages:
  - Nordic: Danish, Finish, Norwegian, Swedish and Dutch.
  - Nordic++: Nordic, US English (higher capacity model).
- First-pass: 17 layer Conformer encoder, 2 layer LSTM decoder trained using RNN-T loss. Encoder dim 512 for Nordic, dim 768 for Nordic++.
- Second-pass: 2 layer LSTM for posterior and prior scores, each of dim 512. 2 layer label-sync attention of dim 128.

Results

Language	First-Pass	+Mono NOS	+Multi NOS
da-dk	8.9	8.5	**8.3**
fi-fi	15.2	14.7	**14.0**
nb-no	11.4	**10.0**	10.1
nl-nl	10.1	9.4	**9.1**
sv-se	11.6	10.9	**10.4**

WER, Nordic First-pass

Language	First-Pass	+Mono NOS	+Multi NOS
da-dk	8.5	7.9	**7.6**
fi-fi	15.0	14.5	**13.8**
nb-no	10.8	**9.8**	**9.8**
nl-nl	9.7	9.1	**8.8**
sv-se	10.9	10.2	**9.8**
en-us	5.6	**5.3**	**5.3**

WER, Nordic++ First pass

- Nordic
  - Avg 6.8% gain by Mono NOS, avg 9.4% gain by Multi NOS.
  - Multi NOS is better almost in all languages.
- Nordic++
  - Avg 6.0% gain by Mono NOS, avg 8.4% gain by Multi NOS.
  - Multi NOS without capacity increase, gives comparable result as Mono NOS in high-resource en-us.
- Why Multi NOS outperforms Mono NOS?
  - Multi NOS makes less errors than Mono NOS when the first-pass picks the oracle hypothesis.

Relative WER Contribution	First-Pass picks oracle	First-Pass picks non-oracle
Mono NOS picks oracle	0%	-19.6%
Mono NOS picks non-oracle	+13.4%	-0.7%
Multi NOS picks oracle	0%	-18.7%
Multi NOS picks non-oracle	+9.6%	-1.3%