# 문법성 품질 예측에 기반한 음성 인식 오류 교정

## Grammatical Quality Estimation for Error Correction in Automatic Speech Recognition

서민택 나승훈
전북대학교

나민수 최맹식 이충희
㈜엔씨소프트

HCLT22

# Introduction

- 최근 음성 인식의 기술 발전으로 좋은 품질의 결과를 제공해주지만 이상적인 결과를 제공하지는 않음.
- 실제 사용 시 API로 제공하는 경우가 많아서, 사용자 입장에서 수정하기 쉽지 않음.
- 따라서 Decoder에 독립적인 모듈을 통해 ASR(Automatic Speech Recognition) 모델에 종속되지 않고, 오류 교정을 하는 모델을 연구하고자 함.

# 데이터 구축

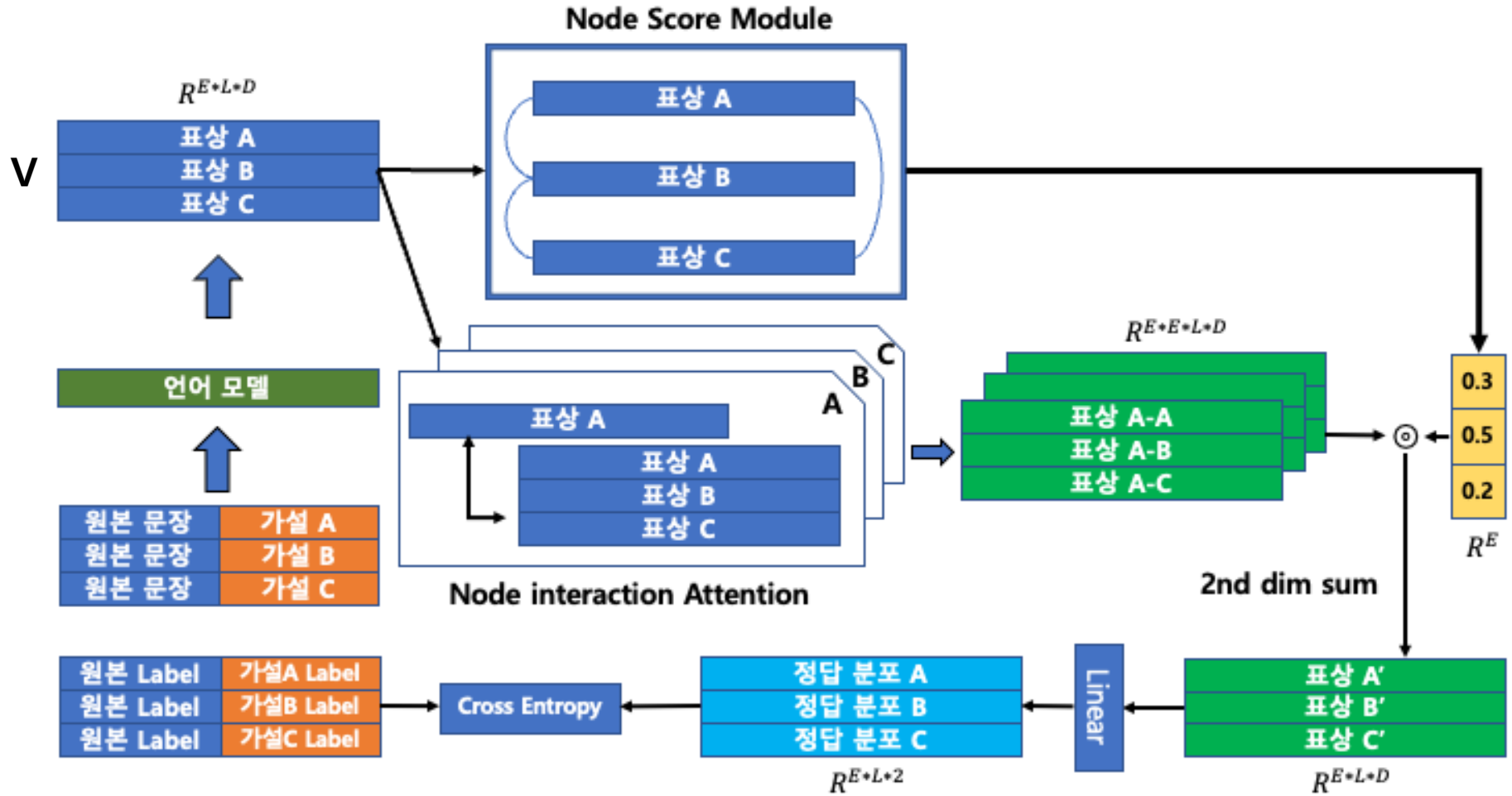
- AIHUB 데이터를 통해서 대화 대본 말뭉치 확보
- TTS(Text-To-Speech)를 통해서 (대본, 음성) 쌍 획득
- STT(Speech-To-Text)를 통해서 (대본 , 음성 , Top N 결과) 쌍 획득

# VERNet (Verification Network)

**(et al., Liu 2021 NACCL)**

- **ASR의 Decoder 결과와 나머지 N개의 결과를 이용하여 토큰을 검증 이때 Top1의 결과를 원본, 나머지 n-1개의 결과를 가설로 설정**
- **모델입력 :** $x_i = [CLS] + \text{원본} + [SEP] + \text{가설}_i + [SEP]$ $(i \leq n-1)$
- **다음과 같은 입력 구성을 통해 원본문장과 가설이 서로 Attention 하도록 한다.**
- **모델의 입력은 Decoder의 표상을 이용하지 않기 때문에 ASR모델과 독립적이다.**

# VERnet Architecture



# 모델 세부 연산

- **Node Score Module**

$$\mathbf{M} = V * W * V^T$$
$(M \in \mathbb{R}^{(N-1 \times L \times L)}, W \in \mathbb{R}^{(H \times H)})$

$$\mathbf{A_s} = \mathbf{Sum_{dim1}}(\mathbf{M}) * \frac{\mathbf{1}}{L}$$
$$\mathbf{A}_H = \mathbf{Sum_{dim2}}(\mathbf{M}) * \frac{\mathbf{1}}{L}$$
$(A \in \mathbb{R}^{(N-1 \times L)})$

$$K_S = \sum_{l}^{L} \mathbf{Diag}(\mathbf{A}_S) \odot (\mathbf{W} * \mathbf{V})$$
$$K_H = \sum_{l}^{L} \mathbf{Diag}(\mathbf{A}_H) \odot (\mathbf{W} * \mathbf{V})$$
$(M_s \in \mathbb{R}^{(N-1)}, Diag : \mathbb{R}^N \rightarrow \mathbb{R}^{(N \times N)})$

$$\mathbf{M_s} = [K_s; K_H; K_s \odot K_H] * W^S \quad (M_s \in \mathbb{R}^{(N-1)})$$

- **Node interaction Attention**

$$\mathbf{N_i} = V * v_i^T = [n_{l,l'}]$$
$(N \in \mathbb{R}^{(N-1 \times L \times L)}, [v_i] = V)$

$$\mathbf{N_{i_s}} = softmax_{i'}(n_{l,l'})$$
$(N_{i_s} \in \mathbb{R}^{(N-1 \times L)}, [v_i] = V)$

$$\mathbf{N_{i_h}} = \mathbf{N_{i_s}} * \mathbf{V^T}$$
$(N_{i_h} \in \mathbb{R}^{(N-1 \times L \times H)})$

$$\mathbf{o}_i = \sum_n^{N-1} \mathbf{N_{i_h}} \odot softmax(M_s)$$
$(o_i \in \mathbb{R}^{(L \times H)})$

**L : 문장 길이**
**N-1 : 가설,원본의 결합 수**
**H : 표상의 차원수**

# 실험 구성

- 문장의 토큰을 정답, 문법 오류 두 가지로 구별하는 문제로 구성 (어절단위 Label)
- 한국어로 학습한 Roberta를 Backbone 모델로 설정
- Baseline은 가설과 원본을 구별하지 않고 단일 문장 토큰 이진 분류 학습
- AIHUB '한국어 대화' 데이터에서 분류 구별 없이 랜덤으로 8:0.5:1.5의 비율로 학습, 개발, 시험 데이터로 분리
- 음성 인식에서 숫자를 글자(1개 ->한 개) 치환하는 경우는 오류로 보기 부적합 하여 이런 경우는 제외함.

# 성능 비교

| Dev | Baseline | VERNet |
|---|---|---|
| Precision | 70.87 | 83.39 |
| Recall | 59.05 | 69.75 |
| $F_{0.5}$ | 68.14 | 80.25 |

| Test | Baseline | VERNet |
|---|---|---|
| Precision | 69.43 | 83.13 |
| Recall | 56.58 | 69.23 |
| $F_{0.5}$ | 66.42 | 79.92 |

$$F_{0.5} = \frac{((0.5)^2 + 1) * precision * recall}{(0.5)^2 * precision + recall}$$

**가설을 추가하여 표상 간 상호작용을 하는**
**VERNet은 한국어에서도 성능이 많이 향상됨**

**이후 추가적 Decoder 결합이나, Beam search 같은 알고리즘을 통해**
**추가적 학습을 한다면 오류 교정에서 성능 향상 기대**