N-Best Re-ranking에 기반한 한국어 음성 인식 성능 개선

N-Best Reranking for Improving Automatic Speech Recognition of Korean

이 정°[1], 서민택[2], 나승훈[3], 나민수[4], 최맹식[5], 이충희[6]

전북대학교[1,2,3], ㈜엔씨소프트[4,5,6]

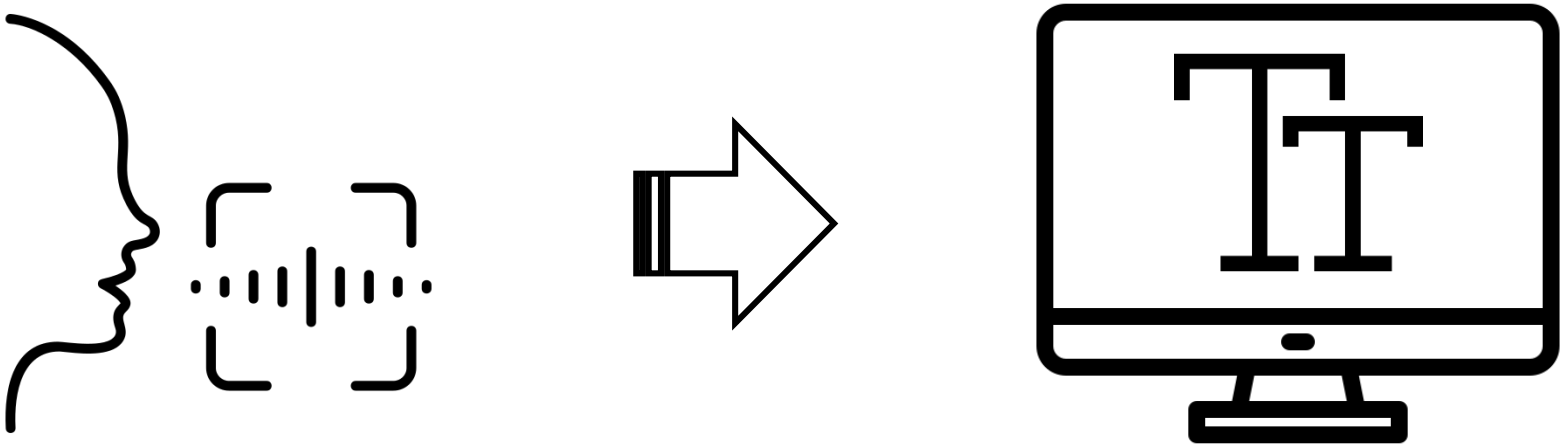
전북대학교 JEONBUK NATIONAL UNIVERSITY

HCLT 2022

NCSOFT

▪ 목차

▪ 서론

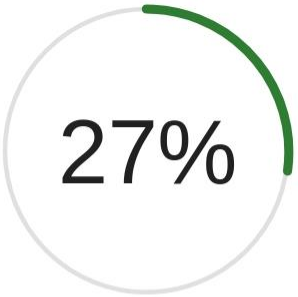


- 자동 음성 인식 (Automatic Speech Recognition, ASR)
  = Speech-to-Text(STT)

사람이 말하는 음성 언어를 텍스트 데이터로 전환하는 일련의 처리나 기술

ASR 기술의 발전으로 음성 기반의 인터페이스를 통한 상호 작용이 확대되는 추세

▪ 서론



of the global online population is
using voice search on mobile.

Think with Google

Global Web Index, Voice Search Insight Report, Global Data n=400,0001, 2018.

- **음성 기반 인터페이스를 통한 상호 작용**

구글이 발표한 자료에 따르면 모바일 검색에서 음성 검색을 활용하는 비율이 27%에 도달

다양한 산업 전반에 걸쳐 적용되고 있는 음성 기반 인터페이스
➔ 최근 CNN 기반의 음성 특징 추출 기법을 포함한 딥러닝 기술과의 결합으로 성능 증대

https://www.thinkwithgoogle.com/marketing-strategies/search/voice-search-mobile-use-statistics/

▪ 관련 연구

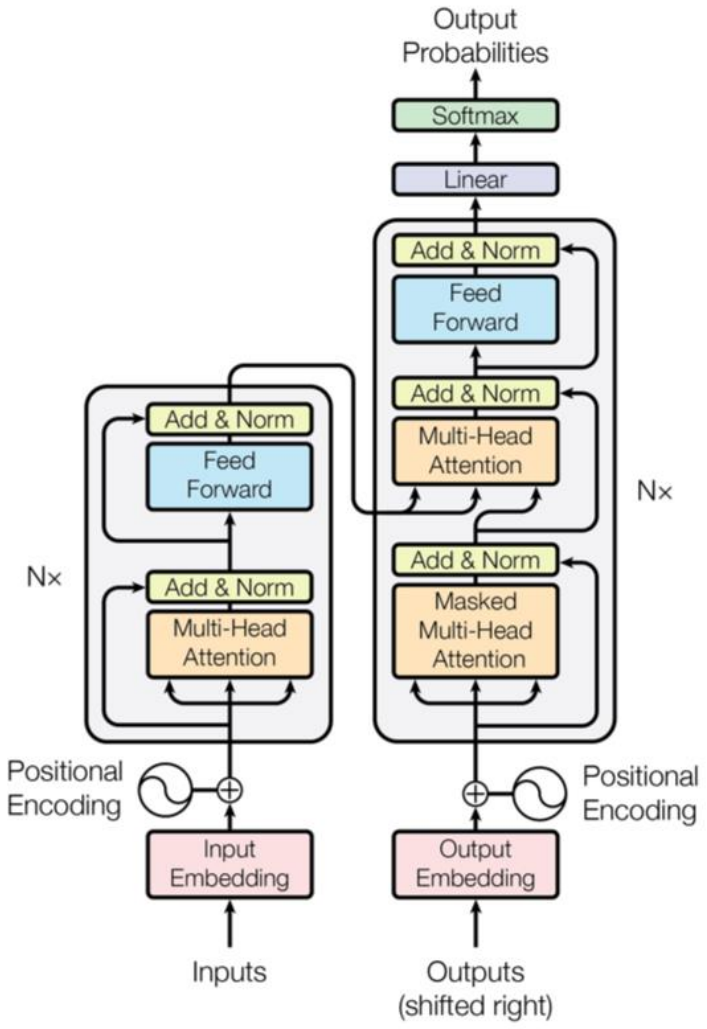


Figure 1: The Transformer - model architecture.

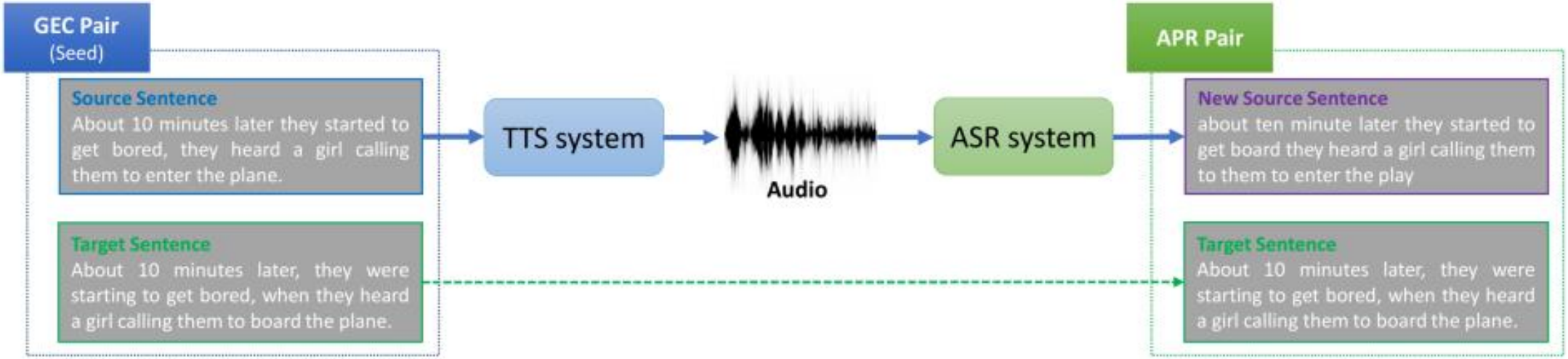
- **Transformer**

  Attention is all you need 논문에서 보이는 Transformer 모델의 등장으로 자연언어처리 기술이 획기적인 성능 향상

  최근 공개되는 대부분의 모델이 Transformer 모델을 기반으로 설계

Attension is all you need [Vaswani et al. `17]

▪ 관련 연구



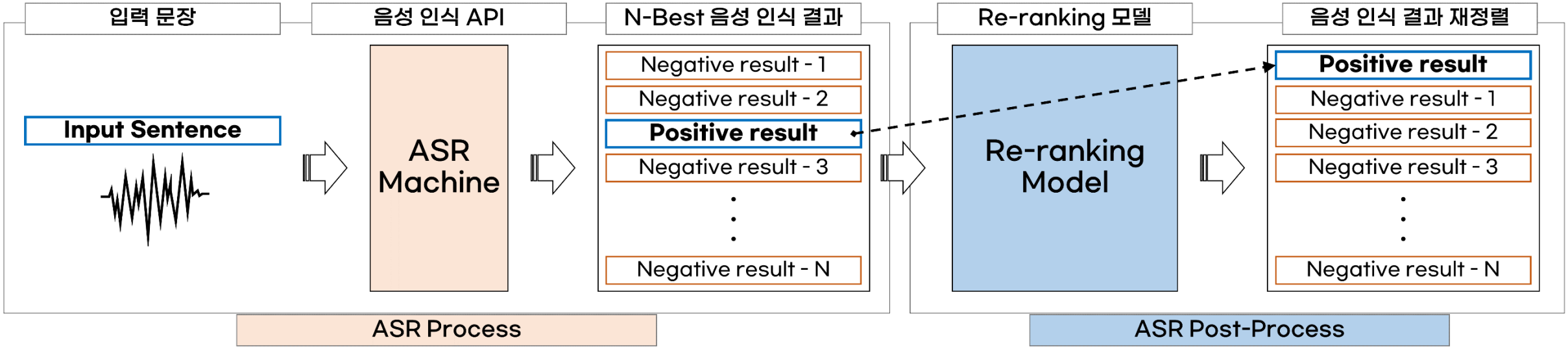
- **텍스트 기반 음성 인식 오류 교정**

  Improving Readability for Automatic Speech Recognition Transcription, Fastcorrect2 등 선행 연구에서 **음성 인식 후처리 과정을 통하여 텍스트 기반의 음성인식 오류 교정**을 보임

  특히 Fastcorrect2 에서는 N-Best 음성 인식 결과 내에서 정렬 알고리즘을 이용해 최적의 문장을 택하여 오류 교정

Improving Readability for Automatic Speech Recognition Transcription [Liao et al. `20]

▪ 모델 소개

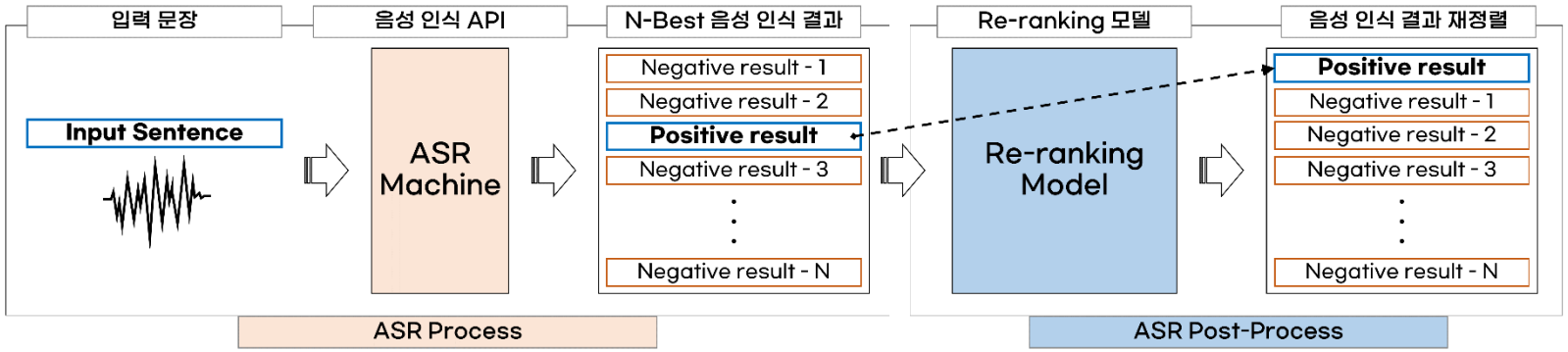


- **모델 소개**

음성 인식 API를 통해 입력한 문장의 수 만큼 N-Best 음성 인식 결과가 주어졌을 때
N-Best 음성 인식 결과로부터 입력 문장과 유사성을 판단하여 재정렬하는 모델

입력 문장에 해당하는 Positive result가 1순위로 재정렬된 N-Best 음성 인식 결과를 도출
기존에 사전 학습된 KLUE-Roberta 모델을 Fine-tuning하여 모델 구축

▪ 모델 소개



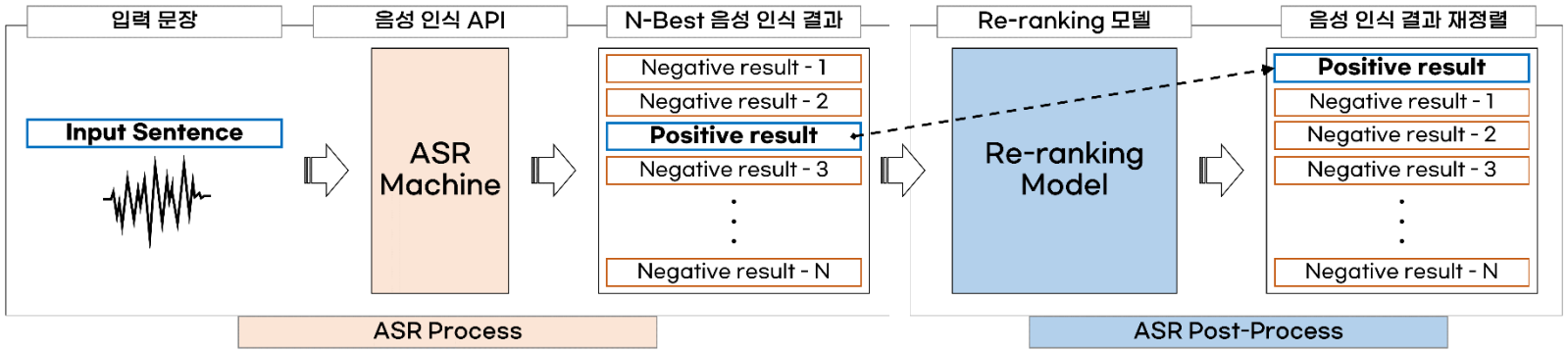
$$y_i = \begin{cases} 1, & \text{if } x_i \in \text{ Input sentences} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

- **모델 설계**

모든 N-Best 음성 인식 결과로부터 총 n개의 음성 인식된 문장 $X = [x_1, x_2, x_3, \cdots, x_n]$와 인식 결과 문장에 따른 레이블 $\mathrm{Y} = [y_1, y_2, y_3, \cdots, y_n]$를 모델의 입력으로 활용

인식 결과 문장 x_i가 입력 문장과 동일한 경우 레이블 y_i를 긍정 문장 1, 이외의 경우는 부정 문장 0으로 정의

▪ 모델 소개



$$H = \mathrm{LM}(x), \quad H \in \mathbb{R}^{|x| \times d} \tag{2}$$

$$\hat{y} = \sigma(\mathrm{W}_1 H_{[\mathrm{cls}]}), \quad H_{[\mathrm{cls}]} \in \mathbb{R}^{d} \tag{3}$$

- **모델 설계**

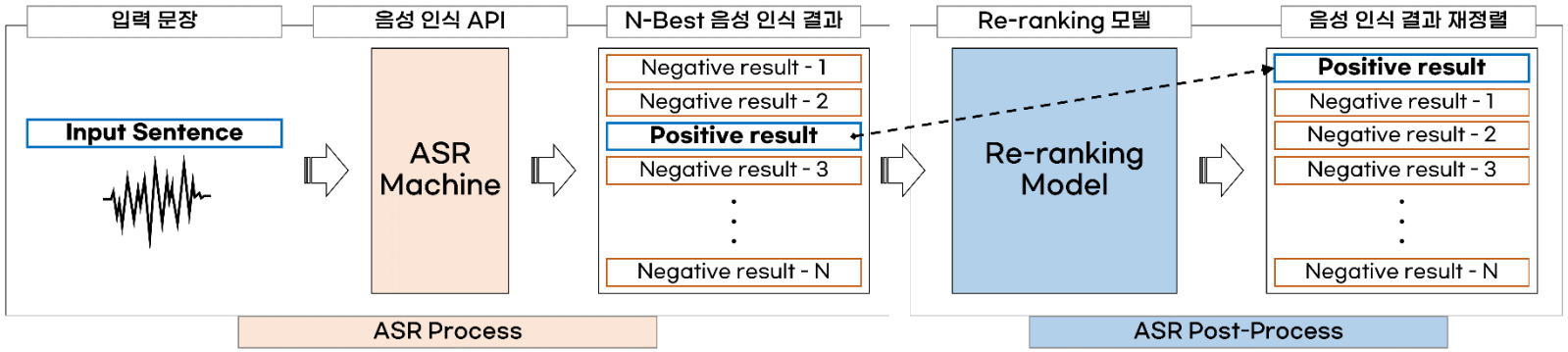
$|x|$: 인식 결과 문장 x의 토큰 길이

d : KLUE-RoBERTa 모델(LM)의 차원

σ : Sigmoid 함수

W_p : d차원의 $H_{[cls]}$를 p차원의 벡터로 projection하는 행렬

▪ 모델 소개



$$H = \mathrm{LM}(x), \quad H \in \mathbb{R}^{|x| \times d} \tag{2}$$

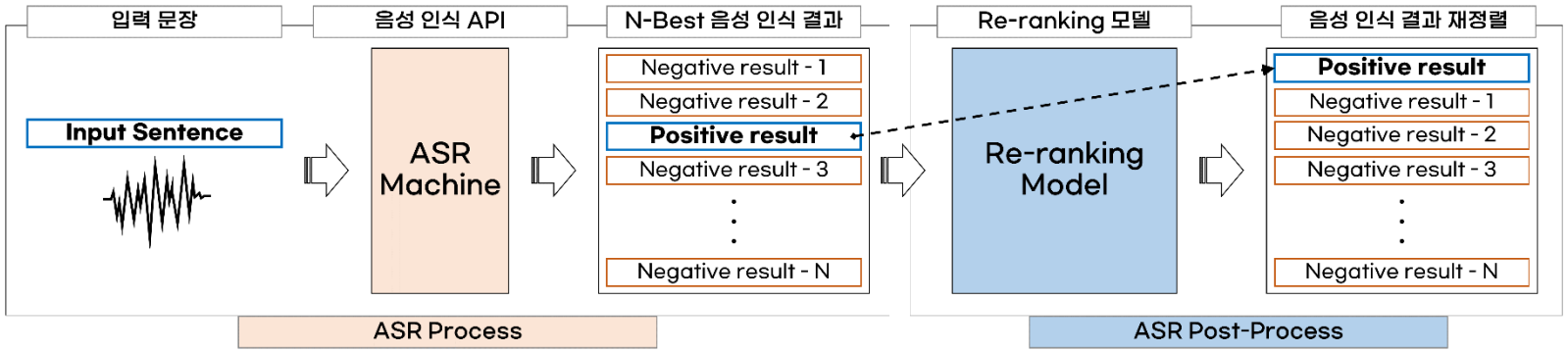
$$\hat{y} = \sigma(\mathrm{W}_1 H_{[\mathrm{cls}]}), \quad H_{[\mathrm{cls}]} \in \mathbb{R}^{d} \tag{3}$$

- **모델 설계**

x를 입력으로 하여 얻어낸 LM의 출력 H 및 Re-ranking 모델의 출력 $\hat{y}_i$

H의 $[CLS]$ 토큰을 1차원에 projection한 뒤 Sigmoid 함수를 취하여 최종 출력을 계산하는 이진 분류 모델로 구성

▪ 모델 소개



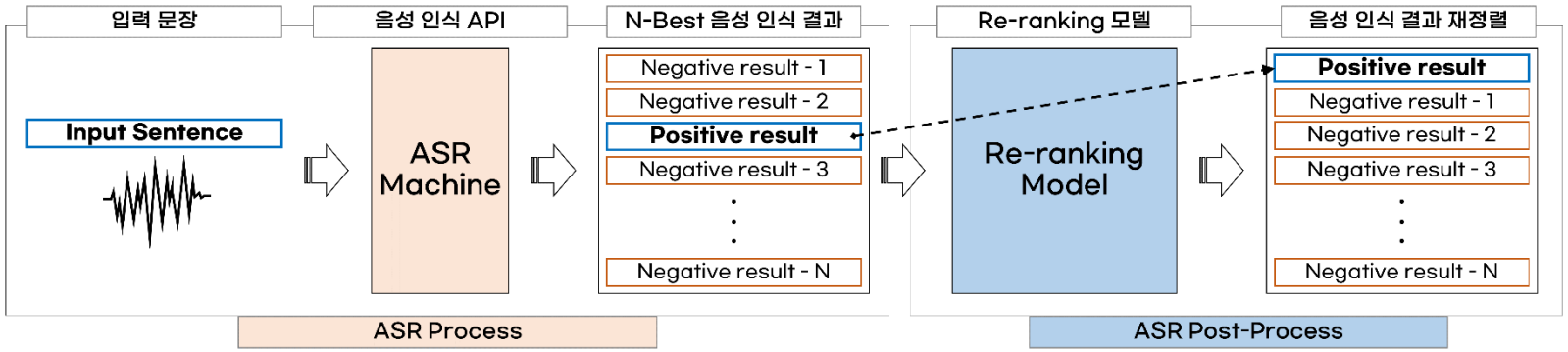
$$L = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)\right] \qquad (4)$$

- **모델 설계**

Re-ranking 모델의 손실 함수

위 손실 함수를 통해 모델은 입력 x_i에 대한 레이블 $\hat{y}_i$을 적절히 분류하도록 학습

▪ 모델 소개



$$X_{\text{best1}} = \arg\max_{i \in N}(P(\hat{y}_i \mid X)) \qquad (5)$$

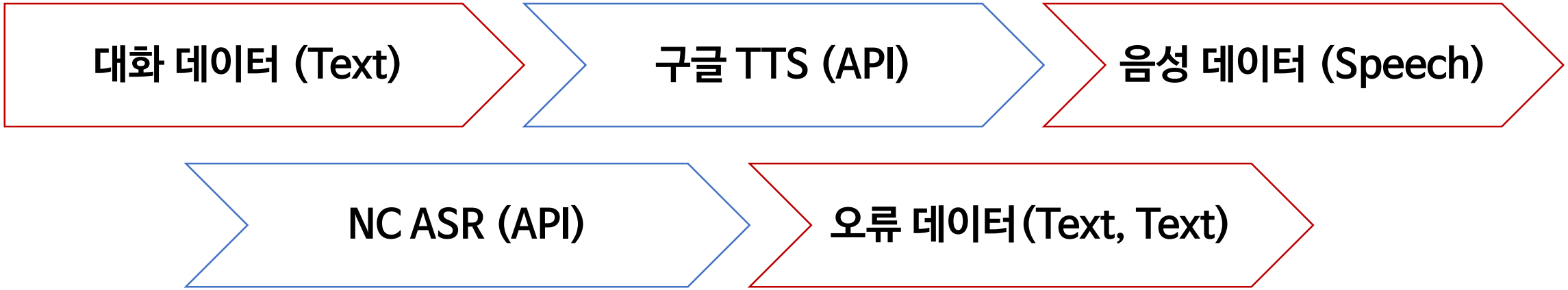
- **결과 추론**

음성 인식 결과 재정렬 단계

Re-ranking 모델을 통해 $X = [x_1, x_2, x_3, \cdots, x_n]$로부터 얻어낸 출력 $\hat{y}_i$에 기반하여 가장 유력하게 긍정 문장으로 분류되는 x_i를 1순위의 음성 인식 결과로 추론

▪ 실험

<데이터 셋 수집 과정>



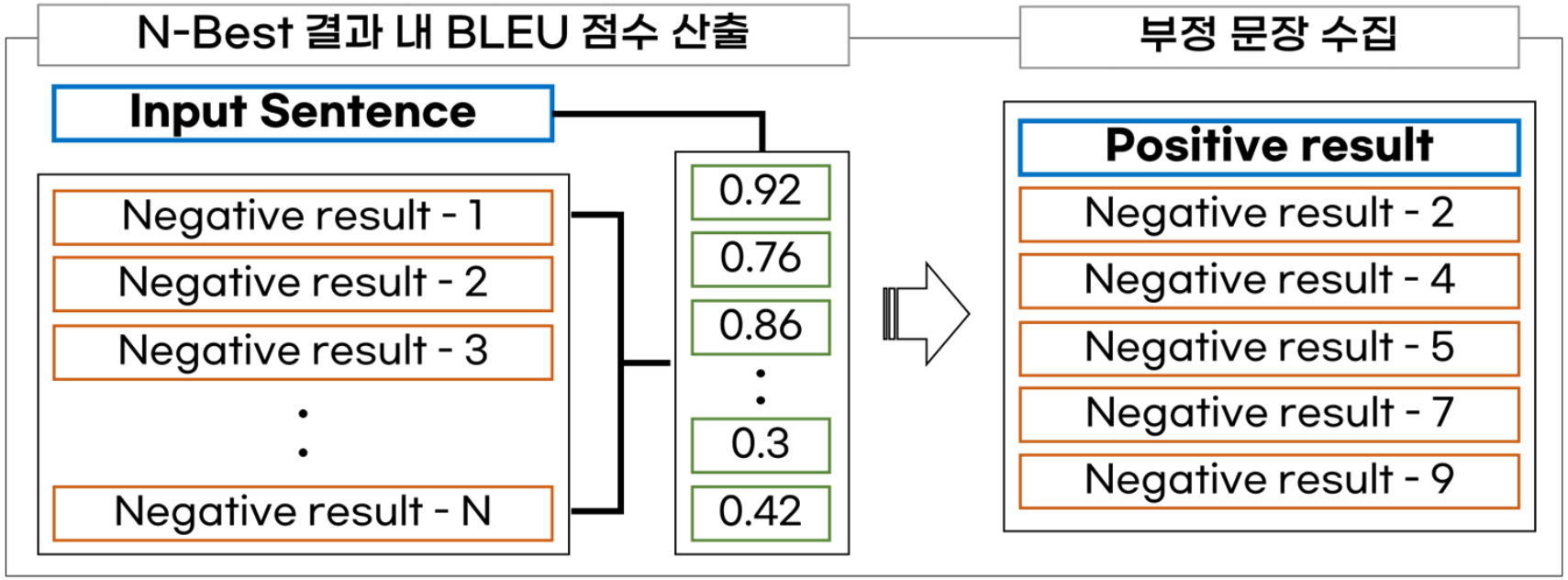
- **데이터 셋**

AI Hub에서 제공하는 〈한국어 대화〉 데이터 활용

소매점, 민원, 관광·여가·오락 분야 등 14개 도메인에서 수집한 총 101,263 개의 대화 데이터 중 2음절 이하의 단순 데이터 및 숫자가 포함된 문장을 제외하여 88,577 개의 데이터 활용

https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=116

▪ 실험

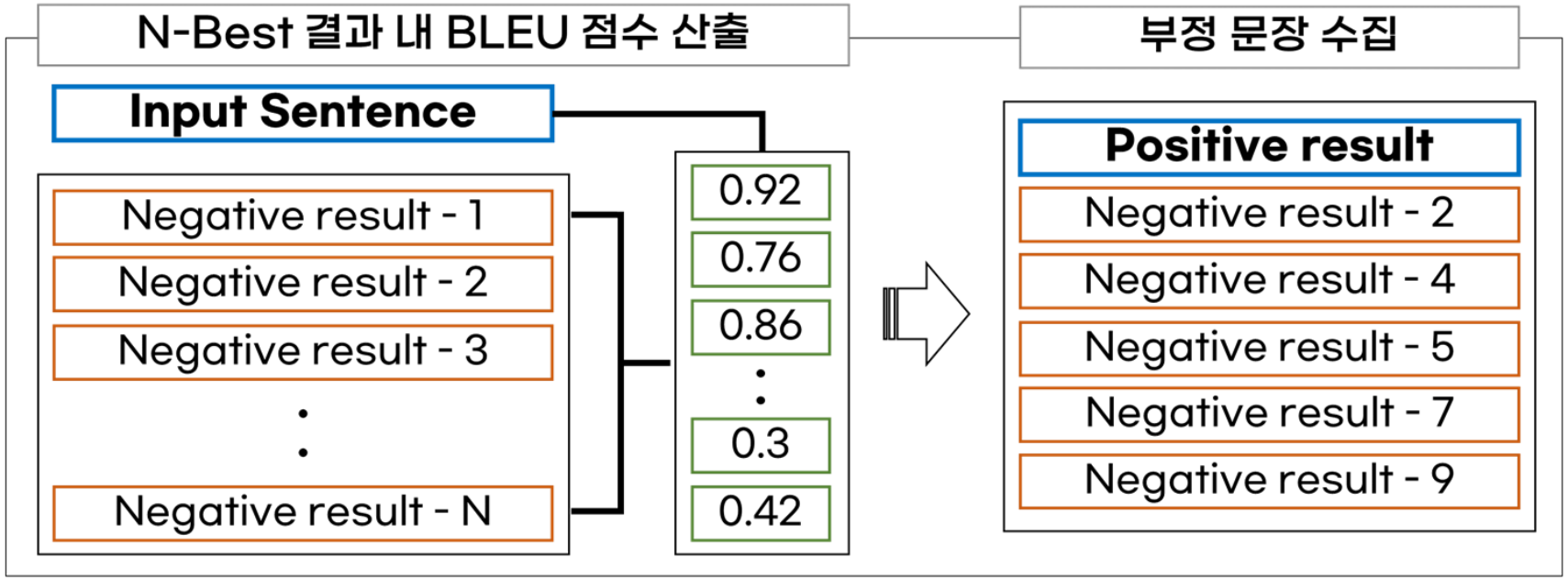


- **실험 설계**

데이터 셋의 특성상 N-Best 음성 인식 결과는 입력 문장을 제외하고 모두 부정 문장으로 구성

부정 문장을 전부 학습 데이터로 활용하지 않고 **부정 문장의 수집 방법 간 차이를 두어 모델 학습 및 성능 비교**

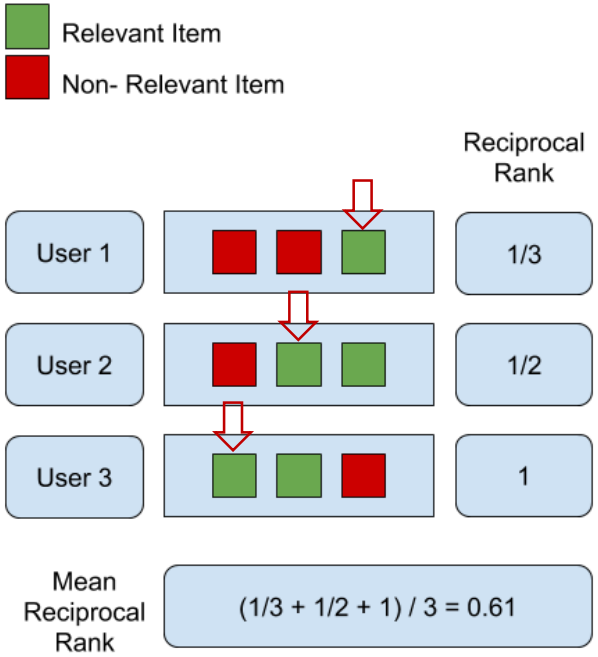
▪ 실험



- Random Negative Sampling
  무작위로 선정된 부정 문장을 수집
- Character-level BLEU score based Negative Sampling
  음절 단위로 입력 문장과의 BLEU 점수를 산출하여 낮은 점수의 문장 순서대로 수집
- Morpheme-level BLEU score based Negative Sampling
  형태소 단위로 입력 문장과의 BLEU 점수를 산출하여 낮은 점수의 문장 순서대로 수집

▪ 실험

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}$$



- **평가 지표**

정보 검색 평가에 주로 활용되는 Mean Reciprocal Rank (MRR) 도입

Relevant Items 중 최상위에 위치한 Item의 Rank를 고려한 계산 방식

재정렬된 N-Best 음성 인식 결과에 대해 MRR 점수 및 Rank 평균 산출하여 모델 평가

▪ 실험

Model	Avg Rank of Positive Results	MRR@10
Baseline	1.892	0.270
Random Negative Sampling	**1.472**	**0.302**
Character-level BLEU score based Negative Sampling	1.674	0.273
Morpheme-level BLEU score based Negative Sampling	1.544	0.292

- **실험 결과**

세 가지 방식의 Re-ranking 모델에 기반하여 재정렬된 N-Best 음성 인식 결과에서 모두 성능 향상

무작위로 부정 문장을 수집하여 학습한 모델의 경우 MRR 점수 기준 약 11.85% 성능 향상

▪ 결론

Model	Avg Rank of Positive Results	MRR@10
Baseline	1.892	0.270
Random Negative Sampling	**1.472**	**0.302**
Character-level BLEU score based Negative Sampling	1.674	0.273
Morpheme-level BLEU score based Negative Sampling	1.544	0.292

- **결론 및 향후 연구**

  제안하는 세 가지 방식의 Re-ranking 모델 전부 성능 향상을 보임

  N-Best 음성 인식 결과 특성상 인접한 순위의 문장 간에는 큰 차이를 보이지 않아 무작위로 수집한 모델이 유리한 것으로 추정

  음성 인식 결과에 따른 Re-ranking의 Upper bound가 존재하여 음성 정보 혹은 문법 교정을 위한 모델을 추가로 활용하는 등 향후 연구를 통해 개선

감사합니다