Artficial Intelligence Laboratory

The SIGMORPHON 2020 Shared Task on Multilingual Grapheme-to-Phoneme Conversion

논문리뷰

정보융합공학과 AI전공

정주경

부산대학교
PUSAN NATIONAL UNIVERSITY

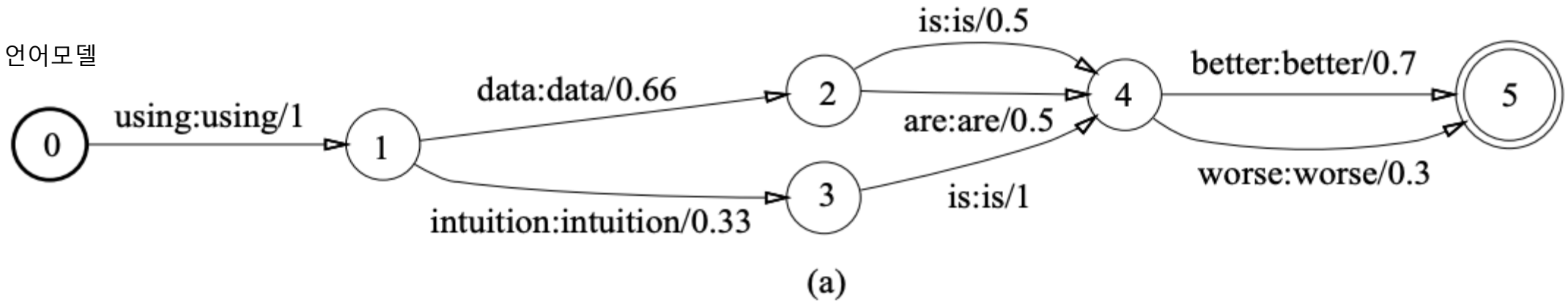
목 차

Introduction

- 음성 기술(자동 음성 인식, Text-To-Speech)은 쓰여진 단어와 발음 사이의 매핑 필요
- grapheme(문자소: 의미 나타내는 최소 문자 단위)과 phoneme(음소: 의미 구별 가능한 음성 상의 최소 단위)간의 매핑 표현
- 매핑은 글을 읽고 쓸 줄 아는 사람이 필요한 규칙을 간단히 열거할 수 있을 정도로 일관성 있음
- 이 규칙의 순서는 finite-state transducer(유한 상태 변환기)로 컴파일 할 수 있음

Introduction - WFST

Weighted Finite-State Transducers



(a)

- 노드가 상태(state) -> 시점(time), 엣지 위에 적힌 정보는 “입력 레이블:출력 레이블/스코어“
- 입력 경로(path): “using data is better”, “using data are better”, …, “using intuition is worse”
- WFST는 단어 시퀀스를 입력 받아 대응하는 출력 레이블 시퀀스에 관련된 경로들의 확률 합 리턴

⇒ WFST는 입력이 주어졌을 때 가능한 모든 출력 경로들에 관련된 확률들을 빠르고 효과적으로 계산하는 것이 목적

Introduction

- 현대 음성 엔진은 Weighted-Finite-State Transducer로 표현된 생성 모델
- conditional random fields(조건부 랜덤 필드)
- recurrent neural network(순환 신경망)
- transformer를 기반으로 하는 discriminative model을 사용한 grapheme-to-phoneme 변환

- grapheme-to-phoneme(G2P)변환 음성 기술 연구의 대다수는 영어 또는 글로벌 언어에 초점
- 본 논문에서 dataset, evaluation metrics, strong baseline이 있는 다국어 G2P 변환 작업 제시
- 무료 사용가능한 발음 사전 모음인 WikiPron 데이터 이용

Data

- 15개의 언어/스크립트 쌍 선택

- 10개는 페르키아어(이집트 상형문자)에서 유래한 알파벳 체계
- 이들 중 7개는 라틴 문자 변형

- 히라가나 대부분의 문자는 음절 전체를 의미

- 히라가나와 마찬가지로 한글도 음절문자

Language	ISO 639-2	Example training data pair	
Armenian	arm	մեծապանակ	m ɛ t͡s ɑ kʰ ɑ n ɑ k
Bulgarian	bul	североизток	s ɛ v ɛ r o i s t o k
French	fre	hébergement	e b ɛ ʁ ʒ ə m ɑ̃
Georgian	geo	ფორმიანი	pʰ ɔ r m ɪ ɑ n ɪ
Modern Greek	gre	καθισμένες	k a θ i z m e n e s
Hindi	hin	कैलकुलेटर	k ɛ: l k ʊ l e: ʈ ə ɾ
Hungarian	hun	csendőrök	t͡ʃ ɛ n d ø: r ø k
Icelandic	hin	þýskaland	θ i s k a l a n t
Korean	kor	말레이시아	m ɐ ɭ ɭ ɐ̞ i ɕʰ i ɐ
Lithuanian	lit	galinčiais	g a: lʲ ɪ nʲ tʲ ʃʲ ɛ j s
Adyghe	ady	бзыукъолэн	b z ə w qʷ a l a n
Dutch	dut	aanduiding	a: n d œ y̯ d ɪ ŋ
Japanese hiragana	jpn	どちらさま	d o̞ t͡ɕ i ɾ a̠ s a̠ m a̠
Romanian	rum	bineînțeles	b i n e ɨ n t s e l e s
Vietnamese	vie	duyên phận	z w i ə n ɨ̌ f ə n ˧̰ ʔ

Table 1: Languages, language codes, and example training data pairs for the shared task.

Methods

- 기본 데이터는 wiktionary(위키낱말사전)의 대규모 다국어 grapheme-phoneme 쌍 리소스인 WikiPron에서 파생
- WikiPron은 case-folding(대소문자 구분 없이 모두 대문자나 소문자로 변경)을 적용
- segments library를 사용해 강세, 음절 경계 marker, 국제 음성 문자로 인코딩된 발음 문자열 제거
- 다중 발음(동음이의어, 자유 발음 변형) 제외 => local context에 따라 다른 발음

Methods

- 각 언어의 데이터 4,500개 다운샘플링
- 이 데이터를 'medium-resource'설정으로 간주
- 학습(80%, 3600), 검증(10%, 450), 테스트(10%, 450)
- 일부 언어 lemma 발음(표제어, 인용 형태) 및 Inflection variant 모두 포함

Inflection morpheme(굴절 형태소): 단어 형태를 굴절

ex) He likes to play soccer : like"–s" 굴절 형태소

Language	ISO 639-2	Example training data pair	
Armenian	arm	մեծապանակ	m ɛ t͡s ɑ kʰ ɑ n ɑ k
Bulgarian	bul	североизток	s ɛ v ɛ r o i s t o k
French	fre	hébergement	e b ɛ ʁ ʒ ə m ɑ̃
Georgian	geo	ფორმიანი	pʰ ɔ r m ɪ ɑ n ɪ
Modern Greek	gre	καθισμένες	k a θ i z m e n e s
Hindi	hin	कैलकुलेटर	k ɛ: l k ʊ l e: ʈ ə ɾ
Hungarian	hun	csendőrök	t͡ʃ ɛ n d ø: r ø k
Icelandic	hin	þýskaland	θ i s k a l a n t
Korean	kor	말레이시아	m ɐ̠ l ʲ ɭ ʲ e i ɕʰ i ɐ̠
Lithuanian	lit	galinčiais	g a: lʲ ɪ nʲ tʲ ʃʲ ɛ j s
Adyghe	ady	бзыукъолэн	b z ə w qʷ a l a n
Dutch	dut	aanduiding	a: n d œ y̯ d ɪ ŋ
Japanese hiragana	jpn	どちらさま	d o̞ t͡ɕ i ɾ a̠ s a̠ m a̠
Romanian	rum	bineînțeles	b i n e ɨ n t s e l e s
Vietnamese	vie	duyên phận	z w i ə n ˧˧ f ə n ˨ʔ

Table 1: Languages, language codes, and example training data pairs for the shared task.

Evaluation

WER(word error rate)

- 예측 표기 시퀀스와 정답 표기 시퀀스가 동일하지 않은 단어의 백분율
- WER가 낮을수록 더 좋은 성능
- shared task에 대한 기본 메트릭 선택

PER(phoneme error rate)

$$\mathrm{PER} := 100 \times \frac{\sum_i^n \mathrm{edits}(p, r)}{\sum_i^n |r|}$$

- 예측 표기와 정답 음소 사이의 정규화 거리를 측정하는 척도
- 예측과 정답 표기 사이의 minimum edit distance 합, 정답 표기 길이의 합으로 나눔
- p는 예측 발음 시퀀스, r은 정답 발음 시퀀스, edit(p, r)은 둘 사이의 Levenshtein 거리(삽입, 삭제 및 대체 횟수)
- 값은 100이 넘을 수 있어 실제 백분율은 100을 초과할 수 있음
- PER이 낮을 수록 더 좋은 성능

Baselines

Pair n-gram model

- graphemes를 나타내는 상태와 phone(음소)을 출력으로 갖는 hidden Markov model의 finite-state 근사치
- 유일한 hyperparameter인 Markov model을 각 언어에 대해 별도로 조절

Encoder-decoder LSTM

- single-layer bidirectional LSTM encoder + single unidirectional LSTM decoder로 구성
- 'small': 128차원 embedding layer, 512 unit hidden layer
- 'large': 256차원 embedding layer, 1024 unit hidden layer
- 두 경우 모두, 디코더는 입력과 출력 모두 single embedding layer 공유

Encoder-decoder Transformer

- hidden layer반복을 multi-head self-attention layer로 대체하는 neural sequence-to-sequence model
- 4개의 인코더 레이어와 4개의 디코더 레이어로 구성, character-level task에 맞게 조정된 pre-layer normalization
- hyperparameter는 LSTM모델과 비슷

System descriptions

CLUZH

- 취리히 대학에서 target 문자열 자체가 아닌 일련의 edit action을 출력하는 모방 학습 기반 transducer 모델
- 각 edit action cost는 Weighted Finite State Transducer(WFST)에서 도출
- 품사, 어원 및 형태론적 세분화 같은 외부 어휘 정보가 시스템을 개선할 것이라 제안
- 전처리 과정에서 한글 문자를 각각 하나의 phoneme에 해당하는 jamo(자모: 하나의 음절을 자음과 모음으로 분석하여 적을 수 있는 낱 글자)로 분해

모방학습(역강화학습[IRL: Inverse Reinforcement Learning])

- IRL: agent의 policy나 action 이력을 통해 그 action을 설명하는 Reward function을 구하는 알고리즘

⇒ 복잡한 상황에서 다양한 보상 요소를 반영하여 최적의 정책 찾음

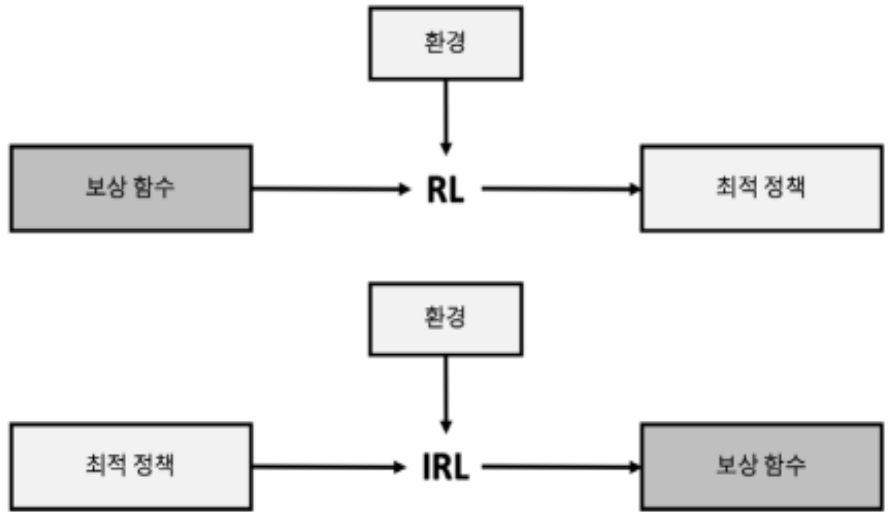


그림 1 RL과 IRL 개념 비교

System descriptions

CU

- 콜로라도 대학교에서 서로 다른 random seed로 만든 여러 transformer model ensemble, Majority voting 사용
- 멀티태스킹 학습의 형태 실험

bidirectional 모델 학습시켜 grapheme-to-phoneme, phoneme-to-grapheme 예측

CUZ

- 콜로라도 대학교에서 'slice-and-shuffle' data augmentation
- grapheme과 phoneme 사이에 문자 수준 일대일 정렬
- 빈번한 subsequence 쌍을 서로 연결해 임시 학습 예제 생성
- 증강된 데이터에 대해 양방향 인코더 LSTM 모델 제출

System descriptions

UA

- 엘버타 대학교에서 non-neural discriminative string transduction model(DTLM), transformer 사용
- grapheme-to-phoneme, phoneme-to-grapheme 모두 활용 data augmentation
- 100개의 학습 예제를 사용한 저자원 시나리오에서 좋은 성능
- DTLM이 transformer보다 학습 빠름
- DTLM, transformer, data augmentation한 transformer 사용

UBCNLP

- 브리티시컬럼비아 대학교에서 첫번째로 다국어 모델로 입력 시퀀스에 언어 식별 토큰 추가
- 여러 개의 checkpoint ensemble
- 위키피디아 텍스트에 대한 self-training 추가
- data augmentation이 점수 향상X

System descriptions

UZH

- 취리히 대학 모든 언어에 공유되는 단일 인코더-디코더 parameter set 사용
- UZH-1: large embedding, hidden layer, batch, dropout 확률이 높은 large transformer 모델
- UZH-2: 다른 6개 언어에 대한 WikiPron 데이터로 모델 보강
- UZH-3: 이전 두 모델 앙상블, posterior probability 더 높은 모델의 예측을 사용해 두 구성 요소 모델 예측 선택

- 앙상블은 대부분의 언어에서 baseline 모델 능가
- 전처리 과정에서 한글을 자모로 분해
- 결과가 46% 상대적인 단어 오류 감소

System descriptions - DeepSPIN

Sparse Attention

- transformer의 Scaled Dot Product는 softmax
- Softmax function은 0이 나올 수 없어 확률이 적은 단어에 0의 attention score가 주어지지 못함
- Softmax의 sparse한 버전으로, sparse한 데이터셋에 적용했을 때 좋은 성능을 보인 normalization 기법

α-entmax

- spread-out이 요구될 때는 softmax가 사용될 수 있도록 α-entmax 채택
- α값이 커질 수록 sparsity, 경사가 급해짐
- 기존 $Att(Q,K,V) = \pi\left(\frac{QK^T}{\sqrt{d}}\right)V$ 에서 π를 α-entmax $(z_i)_j$ 로 치환

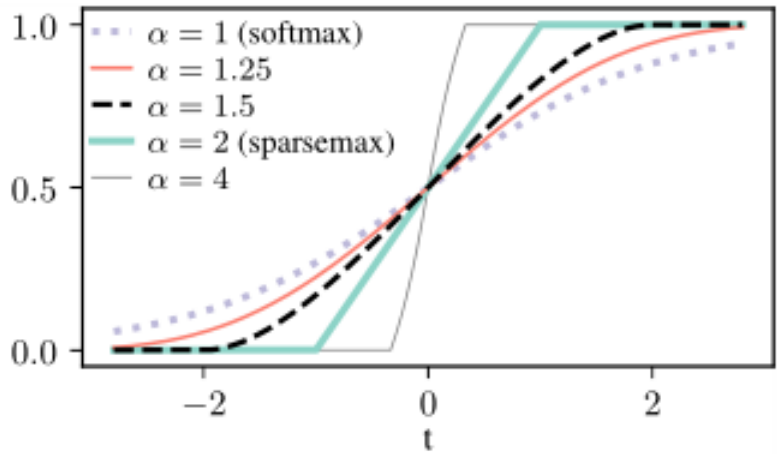


Figure 3: Illustration of entmax in the two-dimensional case α-entmax$([t, 0])_1$. All mappings except softmax saturate at $t = \pm 1/\alpha - 1$. While sparsemax is piecewise linear, mappings with $1 < \alpha < 2$ have smooth corners.

System descriptions

DeepSPIN

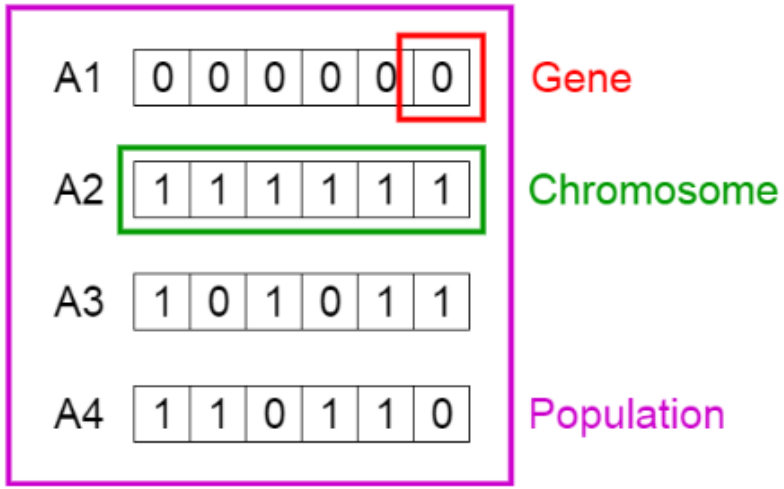
- Superor Tecnico와 Unbabel에서 sparse attention model 기반
- 모든 언어의 데이터에 공동으로 학습하는 다국어 모델, 언어별 hyperparmeter tuning X
- 입력 시퀀스에 언어 식별 토큰 추가 대신, 별도의 학습된 'language embedding'이 모든 인코더, 디코더 상태에 연결되는 single multilingual neural model
- LSTM 또는 Transformer기반 encoder-decoder의 Seq2Seq 모델의 final layer에 sparsity(희소성)을 적용하는 hyperparameter 사용
- 한글 문자를 사전처리하여 각각 단일 음소에 해당하는 자모로 분해

System descriptions - IMS

Genetic Algorithm

- 생물체가 환경에 적응하면서 진화해 나가는 모습 모방, 파라미터의 최적화 문제를 풀기 위한 방법

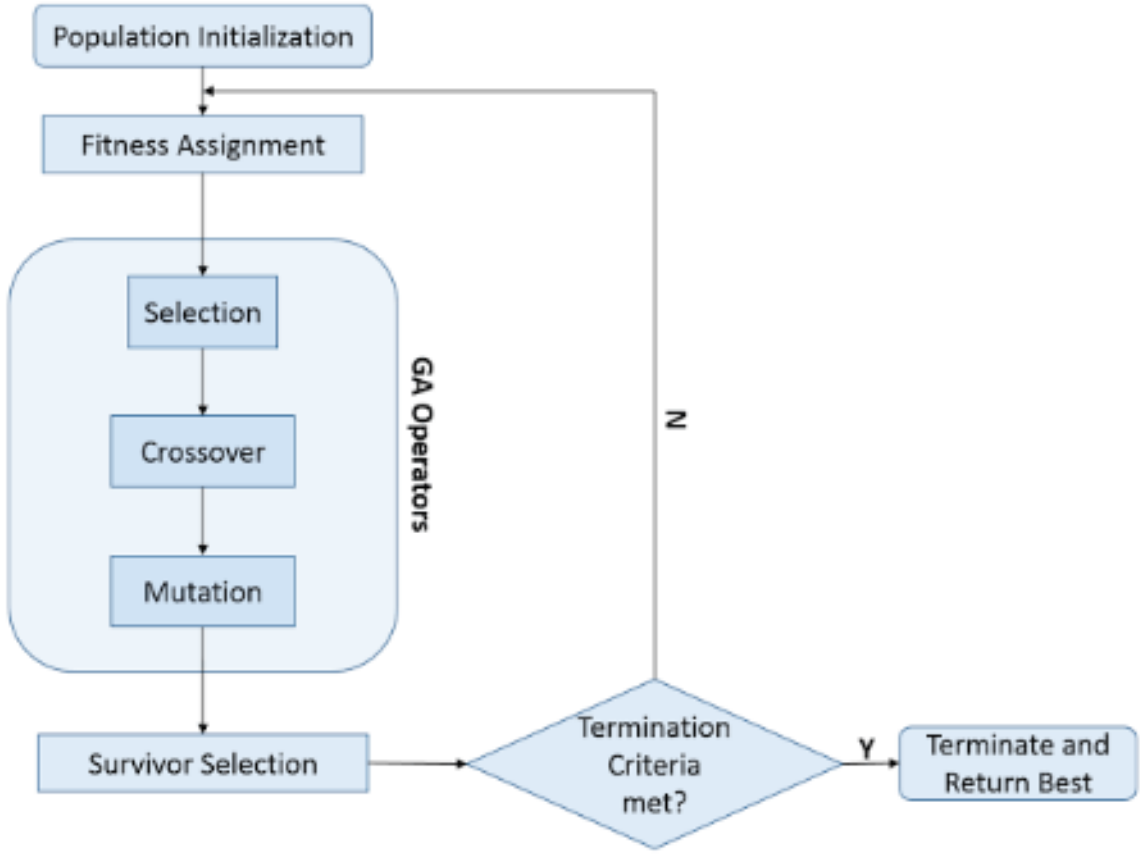
- 염색체(Chromosome): 여러 개의 유전자를 담고 있는 하나의 집합
  하나의 파라미터 표현
- 유전자(Gene): 염색체를 구성하고 있는 요소로 하나의 유전 정보
- 자손(Offspring): 특정 시간(t)에 존재했던 염셱체들
  (ex. A, B 두 염색체 있다고 가정)로부터 생성된 새로운 염색체들
  (C, D라고 가정).
  C, D는 A, B 염색체들의 자손. 부모 염색체의 비슷한 유전 정보 가짐
- 정확도(Fitness): 특정한 염색체가 표현하는 파라미터가 해결하려는 문제에 얼마나 적합한지



유전 알고리즘의 구성요소

System descriptions - IMS

Genetic Algorithm



System descriptions

IMS

1. pair n-gram model기반 Finite-State-Transducer(FST) baseline
2. 디코더가 인코딩된 입력에 attention하고 input vector를 통해 output phonemes를 예측하는 vanilla Seq2Seq model
3. Hard monotonic attention model
디코더는 포인트를 사용해 입력 벡터를 선택하여 phoneme을 생성하거나 포인터를 다음 위치로 이동하여 예측
4. Hard monotonic attention model과 tagging model 하이브리드
각 grapheme에 대해 정렬된 짧은 phonemes 시퀀스를 예측

- 슈투트가르트 대학에서 self-training과 4개 model ensemble
- 앙상블 구성 요소는 유전자 알고리즘 사용
- 200개 학습 예제를 사용해 시뮬레이션된 저자원 설정 제외하곤 데이터 증강이 성능에 크게 영향X
- 사전 처리 단계로 일본어와 한국어 텍스트를 로마자로 표기, 외부 단어 빈도 목록 사용

⇒ 일본어 히라가나와 한글은 음절문자로 하나의 grapheme이 일반적으로 여러 개의 phonemes에 해당

⇒ 로마자로 표기함으로써 (1) 알파벳 크기를 줄이고 (2) 원어와 타깃의 길이를 1:1에 가깝게 하여 정렬 품질 향상

Results

Baseline results

- encoder-decoder LSTM 15개 언어 중 9개 우수
- Transformer는 4개 언어 우수
- 2개(그리스어, 헝가리어) WER기준 성능 일치

- Pair n-gram은 neural baseline보다 성능 떨어짐

- 추가 학습, hyperparameter tuning -> transformer ↑

- 나머지 언어에선 one-best ranking
- 4개 언어(ice, ady, kor, jap) best WER, best PER X

	Pair n-gram		LSTM		Transformer	
	WER	PER	WER	PER	WER	PER
arm	18.00	3.90	14.67	3.49	**14.22**	**3.29**
bul	41.33	9.05	**31.11**	**5.94**	34.00	7.89
fre	13.56	3.12	**6.22**	**1.32**	6.89	1.72
geo	37.78	6.48	**26.44**	**5.14**	28.00	5.43
gre	21.78	4.05	**18.89**	3.30	**18.89**	**3.06**
hin	12.67	2.82	**6.67**	**1.47**	9.56	2.40
hun	6.67	1.51	**5.33**	**1.18**	**5.33**	1.28
ice	17.56	3.62	**10.00**	2.36	10.22	**2.21**
kor	52.22	15.88	46.89	**16.78**	**43.78**	17.50
lit	23.11	4.43	**19.11**	**3.55**	20.67	3.65
ady	32.00	7.56	**28.00**	6.53	28.44	**6.49**
dut	23.78	3.97	16.44	2.94	**15.78**	**2.89**
jap	9.56	2.07	7.56	**1.79**	**7.33**	1.86
rum	11.56	3.55	**10.67**	**2.53**	12.00	2.62
vie	8.44	1.79	**4.67**	**1.52**	7.56	2.27

Table 2: Results for the three baseline systems.

Results

Submission results

- 각 언어 WER 기준 best baseline, best submission

	Best baseline		Best submission	
arm	14.22	transformer	**12.22**	CLUZH
bul	31.11	LSTM	**22.22**	IMS
fre	6.22	LSTM	**5.11**	DeepSPIN-3
geo	26.44	LSTM	**24.89**	IMS
gre	18.89	LSTM, transformer	**14.44**	CU-2, CUZ
hin	6.67	LSTM	**5.11**	CLUZH, IMS
hun	5.33	LSTM, transformer	**4.00**	CLUZH
ice	10.00	LSTM	**9.11**	CLUZH, UBCNLP-2
kor	43.78	transformer	**24.00**	DeepSPIN-1, DeepSPIN-2
lit	19.11	LSTM	**18.67**	CLUZH
ady	28.00	LSTM	**24.67**	DeepSPIN-4
dut	16.44	transformer	**13.56**	IMS
jap	7.33	transformer	**4.89**	DeepSPIN-4
rum	10.67	LSTM	**9.78**	DeepSPIN-3
vie	4.67	LSTM	**0.89**	DeepSPIN-2

Table 3: The best baseline(s) and submission(s) WERs for each language.

DeepSPIN1	RNN-ENTMAX-1.5
DeepSPIN2	RNN-SPARSEMAX
DeepSPIN3	Transformer-ENTMAX-1.5
DeepSPIN4	Transformer-SPARSEMAX

Results

Submission results

- 전체적인 best macro-averaged WER, PER
- Baseline에선 LSTM
- IMS: LSTM baseline대비 WER 절대적 3%(상대적 18%) 감소
  PER 절대적 1%(상대적 31%) 감소
- CLUSZH, DeepSPIN-3는 2위 3위
- CU, UCBNLP, UZH 또한 baseline LSTM 대비 WER 감소

	WER	PER
Pair n-gram	22.00	4.92
LSTM	16.84	3.99
Transformer	17.51	4.30
CLUZH	14.13	2.82
CU-1	14.52	3.24
CUZ	20.87	5.23
DeepSPIN-3	14.15	2.92
IMS	**13.81**	**2.76**
NSU-1	63.56	20.76
UA-2	17.47	4.26
UBCNLP-1	14.99	3.30
UZH-3	16.34	3.27

Table 4: Macro-averaged results for the baselines and the best submission from each team.

Discussion

- 언어를 발음하기 어렵게 만드는 한 가지는 데이터 희소성

- 한국어 가장 높은 baseline 오류율
- 1. 한글은 음절 문자, 알파벳이나 alphasyllabrary(아부기다: 음절 문자와 자모 문자의 특성을 모두 지닌 표시 체계) 보다 더 큰 grapheme
- 2. 한글은 상대적으로 심오하고 추상적인 맞춤법

한글은 형태소 수준, 리투아니아어와 헝가리어는 음소 수준

- 3. 한국어는 음절 경계를 넘나드는 음운론적 과정이 많음

대상 음절 bigram을 관찰해야만 학습 가능

⇒ 한글을 로마자로 만들거나 자모로 분해

: 전처리, data augmentation, ensemble, multi-task learning(phoneme-to-grapheme 변환), self training 사용