Machine Translation Research

Natural Language Processing & AI Lab
Korea University

박찬준

자세한 연구 내용은
보안상 공개하지 않음

Korean Spelling Correction

Multimedia Tools and Applications
An International Journal

Korean Spelling Correction

- **The task of Spelling correction is to detect Spelling errors in a given sentence, and automatically correct them accordingly in grammatical fashion.**

Korean Spelling Correction

- 사람인, 잡코리아 등 자기소개서 맞춤법 검사
- 음성인식 후처리 시스템
- **NLU** 입력 시 전처리 시스템
- 워드 문서 작업 맞춤법 검사

Korean Spelling Correction

- **Commercial System**

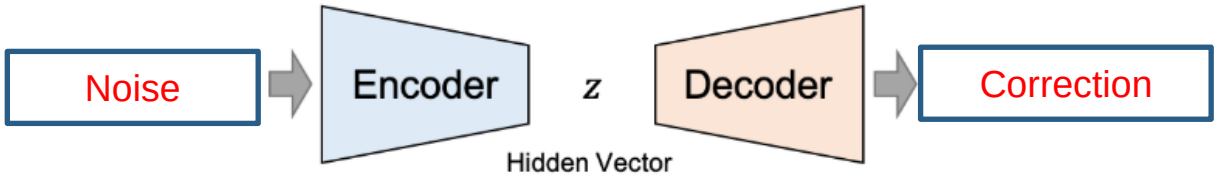
kakao

Translating Incorrect sentences to Correct sentences

기계번역: 원시언어(Source Sentence)를 타겟언어 (Target Sentence)로 번역하는 시스템

원시언어: 오류문장
타겟언어:교정문장

단일 말뭉치로 한국어 맞춤법 병렬 말뭉치를 구성하는 방법을 제안.



Automatic Noise Generation

- Grapheme to Phoneme noise generation
- Edit distance based Alphabetical spelling noise generation
- Real-time translation system error based noise generation

Grapheme to Phoneme Noise Generation

- **Grapheme to Phoneme(G2P) is a technology which converts the sentence according to its pronunciation.**
- **A large amount of misspelled words are caused by people writing text according to the how they are pronounced. Inspired from this we created a noise generation method using the G2P technology.**
- **As the noise generation method used here is based on the phonetic features in linguistics, it means that the given output is generated according to the phonetic features of phonological fluctuation rules and parsing results.**
- **Thus does not require to construct a separate linguistic set of rules. If a Sequence to Sequence model is made with data constructed as above, rules can be learnt implicitly without programming a separate set of rules manually.**

Grapheme to Phoneme
Noise Generation

- 신을 신고 얼른 동사무소에 가서 혼인 신고 해라 ➔ 시늘 신꼬 얼른 동사무소에 가서 호닌 신고 해라

- 나의 친구는 계산이 아주 빠르다 ➔ 나의 친구는 계사니 아주 빠르다

Edit distance based Alphabetical spelling noise generation

- **Edit distance is an algorithm that calculates the similarity between two sequences**
- **Through the means of insertions, deletions, or substitutions.**
- **A noise generation method based on the characteristics of the edit distances indicates that when a source sentence "Hello" is given as input, the output will be "hallu". In our case 2 noises are generated per sentence.**
- **When a human user commits an error, one does not usually commit the error over the whole sentence in general, but rather scatters it across the sentence in few localized areas.**

Edit distance based Alphabetical spelling noise generation

- **Deletions**

예시: 안녕하세요 ➔ 아녀하세요 **(‘ㄴ’ 삭제, ‘ㅇ’ 삭제)**

- **Insertions**

예시: 안녕하세요 ➔ 안녕한세용 **(‘ㄴ’ 추가 ‘ㅇ’ 추가)**

- **Substitions**

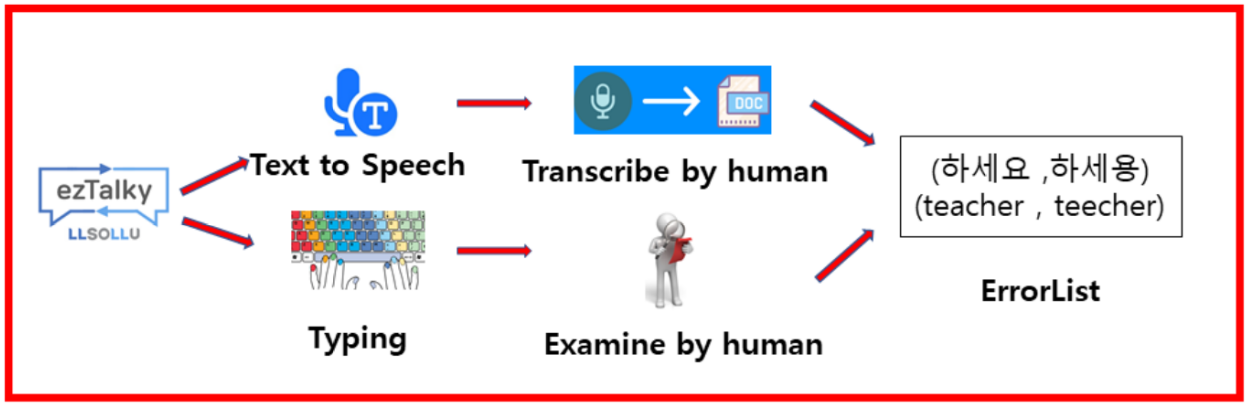
예시: 안녕하세요 ➔ 안령하에요 **(‘ㄴ’ to ‘ㄹ’ 교체, ‘ㅅ to ‘ㅇ’ 교체)**

Real-time translation system error based noise generation

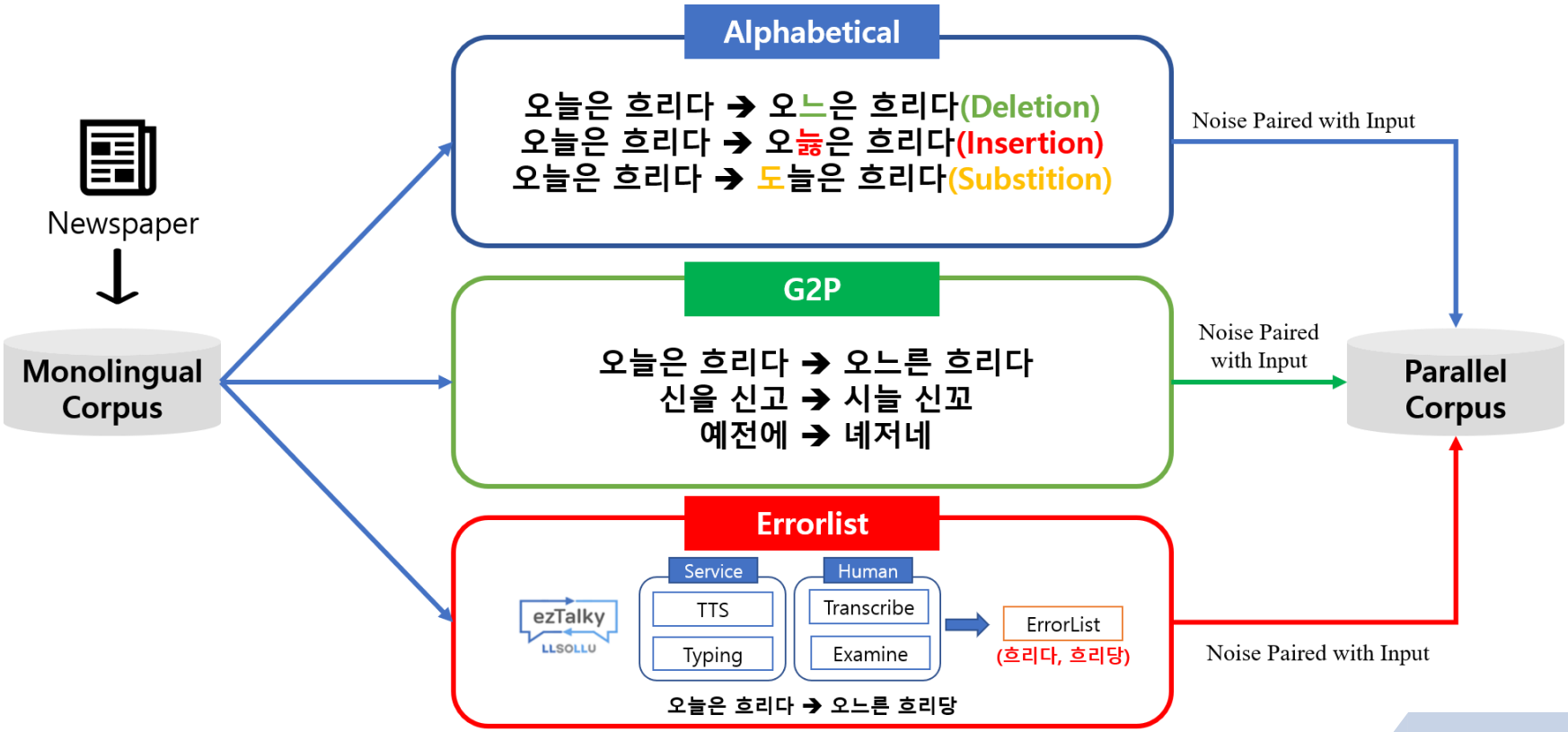
- By using the data from ezTalky a commercialized translation assistant serviced by SYSTRAN , we created a parallel set of word error units such as (thire, their).
- we define this parallel set as the ‘error list’. The error list is a highly reliable set of data based off of actual incidents that happened during ezTalky’s time of service.
- Because of this the keyboard edit distance error is included in the data.
- We constructed a total of 45,711 list of error pairs for the error list. Thus, whenever a matching word from the error list is given as input, its counterpart noise is automatically generated.

Real-time translation system error based noise generation

Input		Output
집에 갔어요	갔어 가써 가요 가욤 가요 가욤 가자 가잣 가자 가짱 가자 가쟈 가자 가장 갈아 가타 갈다 가탕 갈애 가태 갈아 가탱 갈아 가터 갈아 가테	집에 가써요
미칠거같아		미칠거가테
집에가자		집에가쟝



Automatic Noise Generation



Methodology Review

- As an independent resource construction methodology, one can utilize it to generate an exponential amount of data from a monolingual corpus. Which means this methodology allows the generation of an unlimited amount of Korean parallel corpora.

- Additionally, due to the Grapheme to Phoneme approach, the proposed method reflects the characteristics of linguistics.

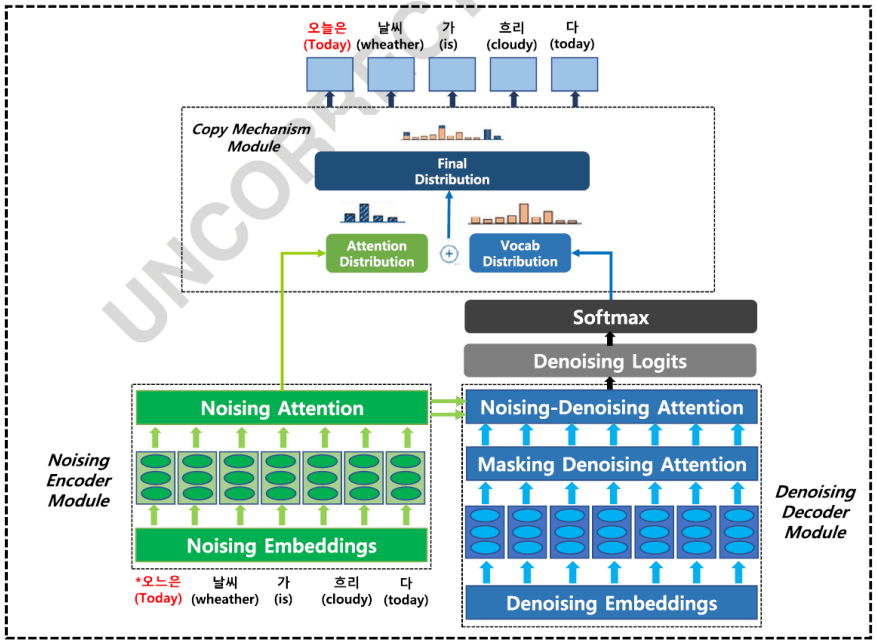
- And as the noise is generated on a random alphabetical unit, it allows it to cover a comprehensive amount of spelling errors.

- Finally, we presented the ‘Error list’, a noise generation method based from actual error incidents in real-life. And as the 'Error list' is generated from an ongoing service, it includes the concept of crowd sourcing.

Methodology Review

Noise	Source	Target
G2P	궁내 처음 심장마비 환자 소생으 뤼한 저체온 치료를 시행해 주모글 바닫따 (It was de first time in Korea to perform the theraputic hypothermia protocall to resurrect patience who sufferd a stroke)	국내 처음 심장마비 환자 소생을 위한 저체온 치료를 시행해 주목을 받았다. (It was the first time in Korea to perform the therapeutic hypothermia protocol to resurrect patients who sufferd a stroke.)
Alphabetical	시범 경ㄱ부터 그런 조지이 나타났다 (Such sgns have appeared since the conest)	시범 경기부터 그런 조짐이 나타났다. (Such signs have appeared since the contest.)
ErrorList	교사들은 학겨 수업이 ebs 문제풀이식으로 이뤄지면서 학생 실력이 하향 평준화됐다고 주장한다. (Teachers insist that the student's academic abilities are degrading ever since skool classes are conducted based on the EBS problem solving method)	교사들은 학교 수업이 ebs 문제풀이식으로 이뤄지면서 학생 실력이 하향 평준화됐다고 주장한다. (Teachers insist that the student's academic abilities are degrading ever since school classes are conducted based on the EBS problem solving method.)

Model



- **Denoising Auto-Encoder (DAE)**
- **We generate noises to the monolingual corpus with our proposed noise generation methods introduced previously. We label these noise generation methods as function noise(x).**
- **Thus, function noise(x) convolutes the given sentences from the monolingual corpus.**
- **The DAE model attempts to correct this. Thus our proposed model, trains itself by correcting the incorrect sentence created by function noise(x).**

Experiments

Dataset	Size
Training (Total)	3.0M
G2P Noise	1.0M
Edit Distance Noise	1,0M
Error Lists Noise	1.0M
Validation	5,000

Information	**Source**	**Target**
Average Length	55.42	55.33
Average Token	13.13	13.12
Max Length	247	247
Min Length	17	17
Max Token	57	57
Min Token	4	4

Experiments

Model	GLEU	BLEU	Precision	Recall	F1
LSTM	74.57	76.42	70.60	71.78	71.19
Conv2Conv	68.17	69.57	70.73	73.35	72.01
Transformer	93.22	94.27	95.42	93.74	94.58
+ShareVocabulary	82.42	83.41	84.12	89.94	86.93
+Copy Attention - Dot	94.76	95.40	96.18	95.49	95.83
+Copy Attention - General	95.27	95.88	96.53	95.93	96.23
+Copy Attention - MLP	95.59	96.13	96.81	96.32	96.56

GLEU(Generalized Language Evaluation Understanding) is similar to BLEU but it differs in that it also considers the source information and is a performance metric specialized for Grammar Error correction systems

Additional Effect

자동 문장 분리 효과

입력	죄송합니다 모든 좌석이 매진됐습니다
출력	죄송합니다. 모든 좌석이 매진됐습니다.

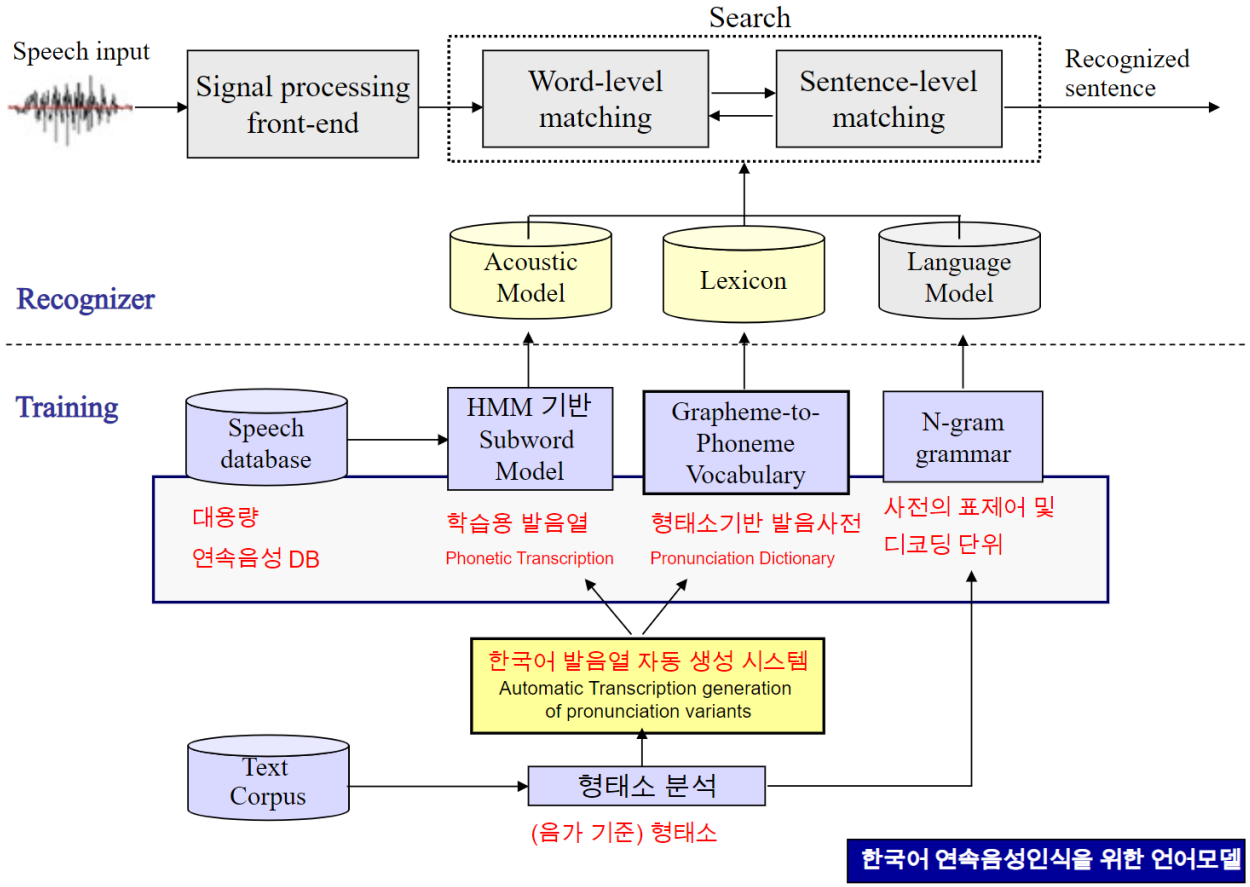
자동 기호 부착 효과

입력	여기 가까운 식당이 어디있습니까
출력	여기 가까운 식당이 어디 있습니까?

Demo

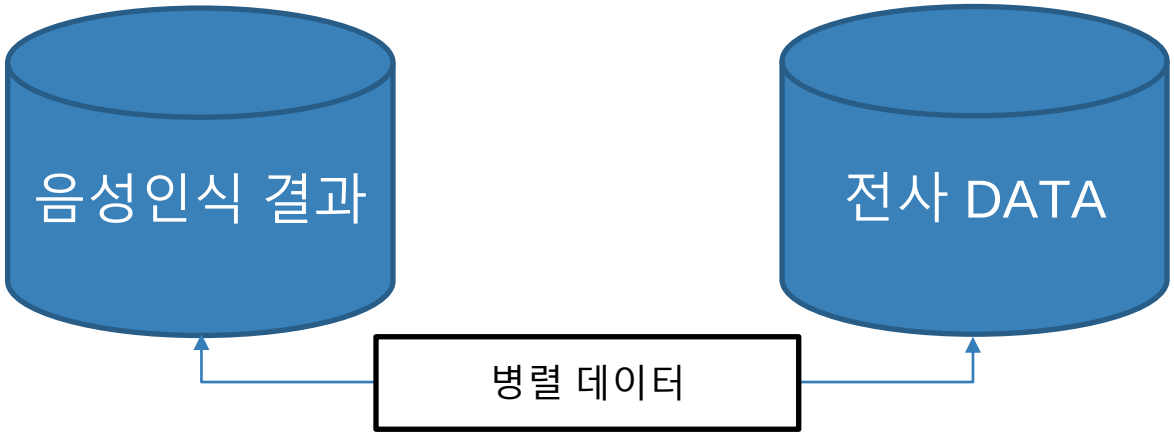
http://nlplab.iptime.org:32288/

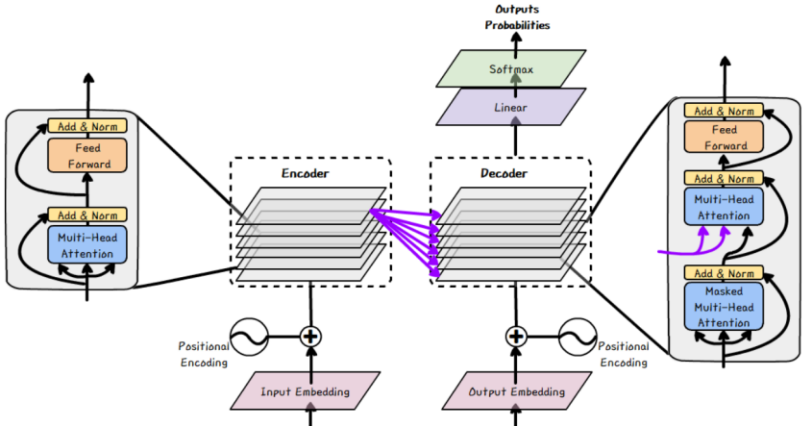
STT와 결합


Search
Speech input
Signal processing front-end
Word-level matching
Sentence-level matching
Recognized sentence
Acoustic Model
Lexicon
Language Model
Recognizer
Training
Speech database
HMM 기반 Subword Model
Grapheme-to-Phoneme Vocabulary
N-gram grammar
대용량 연속음성 DB
학습용 발음열
Phonetic Transcription
형태소기반 발음사전
Pronunciation Dictionary
사전의 표제어 및 디코딩 단위
한국어 발음열 자동 생성 시스템
Automatic Transcription generation of pronunciation variants
Text Corpus
형태소 분석
(음가 기준) 형태소
한국어 연속음성인식을 위한 언어모델

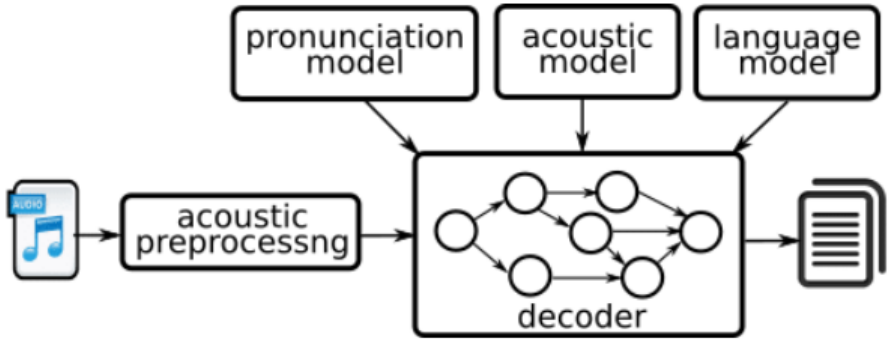
복잡

STT와 결합

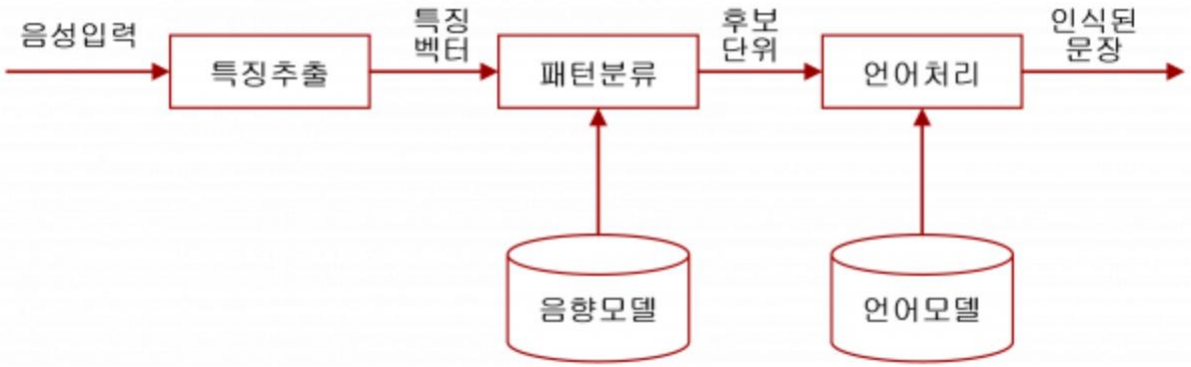

음성인식 결과
전사 DATA
병렬 데이터


Outputs Probabilities
Softmax
Linear
Encoder
Decoder
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Positional Encoding
Input Embedding
Positional Encoding
Output Embedding

STT와 결합



KALDI



Post – Processing Module

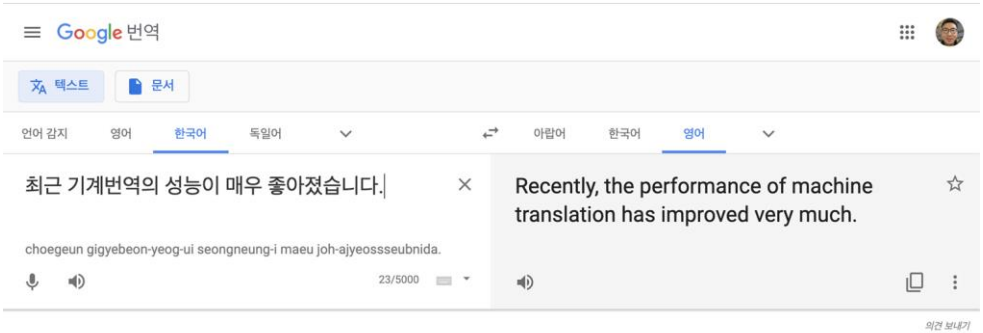
Ancient Korean Neural Machine Translation

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Ancient Korean Neural Machine Translation

고전번역: 조선왕조실록, 승전원일기와 같은 고어를 번역하는 것을 의미함

기계번역: 소스문장(Source Sentence)을 타겟문장(Target Sentence)으로 컴퓨터가 번역하는 시스템을 의미하며 이를 고전번역에 적용할 경우 소스문장에 고어 타겟 문장에 한국어가 적용될 수 있음



Ancient Korean Neural Machine Translation

- 우리 고전에 대한 사회적 수요가 증가하고 있음
- 기존 방식의 고전번역의 한계
- 인공지능 기술의 발전

Ancient Korean
Neural Machine Translation

NMT기반 고전번역의 장점은?

- 기존 고전번역사들의 업무 효율성 강화
- 빠른 시간에 번역 가능
- 플랫폼을 통한 번역결과물의 DB화 및 지식증강형 Infinite Training모델 구축
- 품질 편차를 최소화하고 일관된 번역 품질을 만들어 낼 수 있음.
- 미번역된 문서에 대한 번역도 가능하다. (규장각 도서 등)

Ancient Korean Neural Machine Translation

- 본 논문은 Neural Machine Translation(NMT)를 이용하여 문제를 해결하고자 함.
- 현재 한국 고전번역 관련하여 연구 발표된 논문이 많지 않은 실정임.
- 그러나 일본의 KuroNET, 그리스의 Greece epigraphy에 대한 연구 등 고어에 대한 연구에 움직임이 시작되고 있음.
- 따라서 본 논문은NMT 모델인 Transformer를 이용하여 한국 고전번역 모델을 만듬.

KuroNet: Pre-Modern Japanese Kuzushiji Character Recognition with Deep Learning

Tarin Clanuwat*
Center for Open Data in the Humanities
National Institute of Informatics
Tokyo, Japan
tarin@nii.ac.jp

Alex Lamb*
MILA
Université de Montréal
Montreal, Canada
lambalex@iro.umontreal.ca

Asanobu Kitamoto
Center for Open Data in the Humanities
National Institute of Informatics
Tokyo, Japan
kitamoto@nii.ac.jp

Ancient Korean Neural Machine Translation

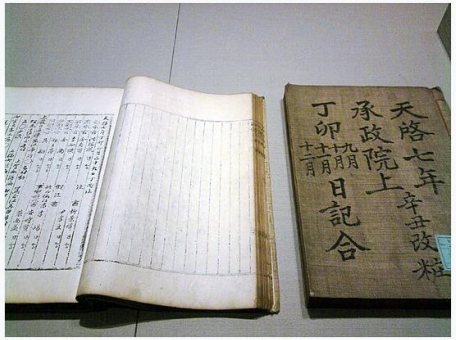
- **To the best of our knowledge, this is the first study on a system for ancient Korean machine translation.**
- **We propose SVER BPE specifically for ancient translation.**
- **We present the results of experiments on various decoding strategies such as beam search, n-gram blocking, and ensemble models.**

Ancient Korean Neural Machine Translation

한국고전번역원
Institute for the Translation of Korean Classics

승정원일기
(承政院日記)

대한민국의 국보



종목 국보 제303호
(1999년 4월 9일 지정)

수량 3,243책

시대 조선, 대한제국

주소 서울특별시 관악구 서울대학교 규장각 한국학연구원

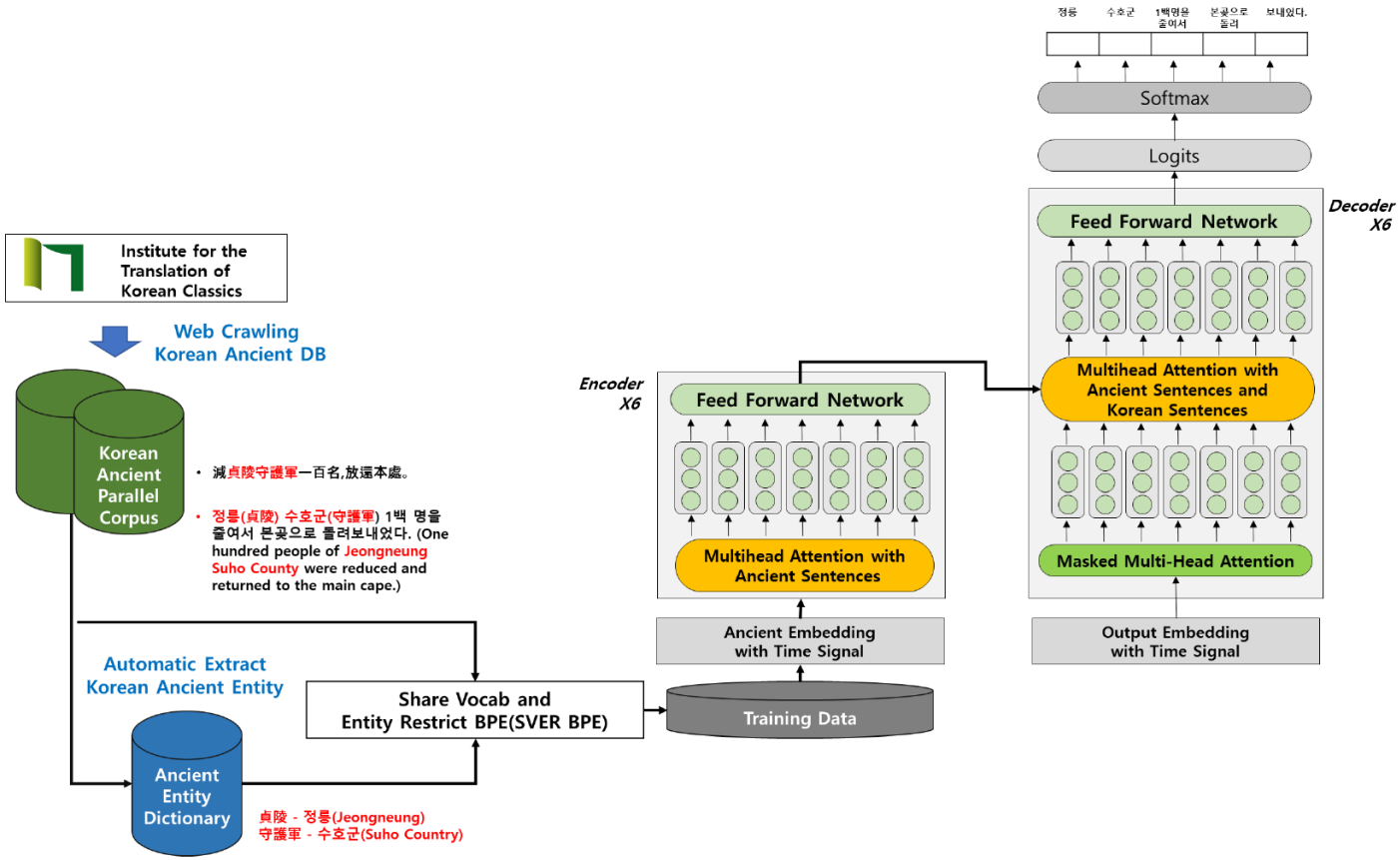
정보 문화재청 국가문화유산포털 정보

Veritable Records of the Joseon Dynasty

Veritable Records of Jeongjo

Korean name	
Hangul	조선왕조실록
Hanja	朝鮮王朝實錄
Revised Romanization	*Joseon Wangjo Sillok*
McCune–Reischauer	*Chosŏn Wangjo Sillok*[1]
North Korea name	
Hangul	조선봉건왕조실록[2]
Hanja	朝鮮封建王朝實錄
Revised Romanization	*Joseon Bonggeon Wangjo Sillok*
McCune–Reischauer	*Chosŏn Bonggŏn Wangjo Sillok*

Ancient Korean Neural Machine Translation



Ancient Korean Neural Machine Translation

(A) Prepare an ancient Korean bilingual corpus.
(B) Combine two corpora to share vocabulary.
(C) Set the subword vocabulary size. In the model we constructed, we set it to 32,000.
(D) Split the words into sequence characters. However, do not separate the entities.
(E) Merge the most frequent adjacent pair of characters.
(F) Repeat step E for a fixed number of times or the defined size of the vocabulary is reached.

Experiments

Ancient Sentence	減貞陵守護軍一百名,放還本處 "
Korean sentence	정릉(貞陵) 수호군(守護軍) 1백 명을 줄여서 본곳으로 돌려보내었다.
English(Translate)	One hundred people of Jeongneung Suho County were reduced and returned to the main cape.
Ancient Sentence	丁未/上坐經筵,令侍講官裵仲倫,講《貞觀政要》 "
Korean sentence	임금이 경연(經筵)에 앉아서 시강관(侍講官) 배중륜(裵仲倫)으로 하여금 《정관정요(貞觀政要)》
English(Translate)	The king sat down at the contest and asked the Shigangguan Bae Joong-ryun to be called "The Junggwanjeongyo".
Ancient Sentence	甲戌/召見回還三使臣于熙政堂 "
Korean sentence	돌아온 세 사신(使臣)을 희정당에서 소견(召見)하였다.
English(Translate)	The three reapers who returned were found at the Heejeongdang.
Ancient Sentence	上移御于昌慶宮之儲承殿 "
Korean sentence	임금이 창경궁(昌慶宮)의 저승전(儲承殿)에 이어(移御)하였다.
English(Translate)	The king succeeded Changgyeonggung Palace's demise.

	Ancient-Train	Ancient-Val	Ancient-Test	Korean-Train	Korean-Val	Korean-Test
Size	52,778	5,000	3,000	52,778	5,000	3,000
Average	39.12	38.39	38.83	92.78	90.72	91.79
Max	167	137	141	350	311	301
Min	3	4	4	5	7	7

Ancient Korean
Neural Machine Translation

- C기반 디코더로 인한 빠른 속도, 실제 상용화를 고려한 CPU 기반 서비스 가능(Scale Up 용이)
- 속도를 고려한 GPU 서비스도 가능하게 설계함
- 글자수 제한 없음, 친숙한 UI

http://nlplab.iptime.org:32242/

Experiments

Model	BLEU	Token Per Second
Sentencepiece-LSTM-Attention	24.39	2758
Sentencepiece-Transformer	22.69	982
BPE-LSTM-Attention	25.18	2029
BPE-Transformer	24.43	1122
Char- LSTM-Attention	23.66	8785
Char- Transformer	16.24	1466
Entity Restrict- LSTM-Attention	14.74	3013
Entity Restrict- Transformer	15.12	1174
(Our) Share Vocab and Entity Restrict BPE - LSTM Attention	29.40	5004
(Our) Share Vocab and Entity Restrict BPE -Transformer	**29.68**	1379

Conclusion

- 본 논문은 한국 고전번역기에 대한 연구를 진행하였음
- 이를 통해 인간 번역자의 수고를 덜어줄 수 있으며 그만큼 번역에 드는 시간을 단축함과 동시에 번역의 질을 올릴 수 있는 토대를 마련함
- 추후 고전번역 데이터의 Alignment 작업을 추가적으로 진행한 후 번역의 단위에 따라서 어떠한 성능 변화가 있는지에 대한 연구를 진행할 예정임

감사합니다.

박찬준
bcj1210@naver.com
Parkchanjun.github.io