

“ 논문에 최적화된 번역 프로그램 ”

Translation Program Optimized for Thesis

숙명여자대학교 컴퓨터과학전공
2021년도 1학기 졸업작품 발표

2021년도 1학기 졸업작품 발표

- 1 팀명 : TP
(Translation Program)
- 2 작품명 : 논문에 최적화된 번역 프로그램
(Translation Program Optimized for Thesis)
- 3 지도교수 : 박동철교수님
- 4 팀원 : 통계학과 김채현

CONTENTS

1 주제 선정 동기

4 Challenging

2 Project 소개

2.1 개발언어 & 환경

2.2 Modeling

2.2.1 Naïve Model

2.2.2 Advanced Model

2.3 주요기능

5 Future Work

3 시연영상

papago 웹사이트 번역 GYM | 사전

한국어 > 영어 > 한국어

영어 > 한국어

One **epoch** is one pass through the training set, NLL is the average conditional log-probabilities of the sentences in either the training set or the development set.

원 에폭 이즈 원 패스 쓰루 더 트레이닝 셋 에널렐 이즈 디 애버리지 컨디셔널 로그 프라버 빌리티즈 어브 더 센턴서즈 인 아이더 더 트레이닝 셋 오어 더 디벨롭먼트 셋.

164 / 5000

번역하기

한국어 > 영어

하나의 **시대**는 훈련 세트를 통과하는 하나의 단계이며, NLL은 훈련 세트 또는 개발 세트에 있는 문장의 평균 조건부 로그 확률이다.

번역 수정 | 번역 평가

자동완성

Google 번역

텍스트 문서

언어 감지 영어 한국어 독일어 >

한국어 > 영어 > 일본어 >

One **epoch** is one pass through the training set, NLL is the average conditional log-probabilities of the sentences in either the training set or the development set.

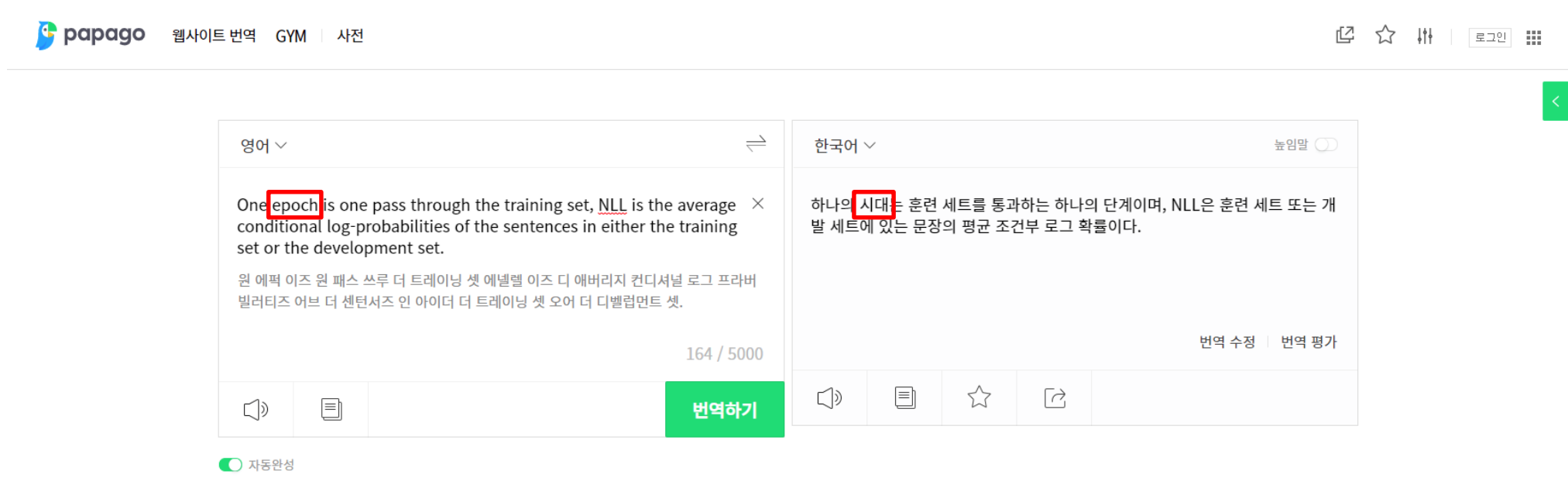
한 **시대**는 훈련 세트를 한 번 통과하는 것이고, NLL은 훈련 세트 또는 개발 세트에 있는 문장의 평균 조건부 로그 확률입니다.

han sidaeun hunlyeon seteuleul han beon tong-gwahaneun geos-igo, NLLeun hunlyeon seteu ttoneun gaebal seteueissneun munjang-ui pyeong-gyun jogeonbu logeu hwaglyul-ibnida.

164 / 5000

의견 보내기

프로젝트의 필요성



- 해외 논문을 주로 읽는 학부생/대학원생 대상
 - 기존 번역기 (Papago, Google Translator)의 경우 모든 영어를 한글로 번역
 - 인문학 / 사회과학 / 자연과학 / 공학 / 의학학 / 농수해양학 / 예술체육학 각 분야에 따라 의미가 달라지는 단어 多
- ↓
- 논문 분야에 맞게 최적화된 번역 프로그램 필요

개발 언어 & 환경

개발 환경 : Windows 10. i7-5820K, RAM 16.0GB

개발 언어 :

Modeling



Python 3.9.1

필요 Library : spaCy
NLTK

Pytorch : torch
torchtext

Server



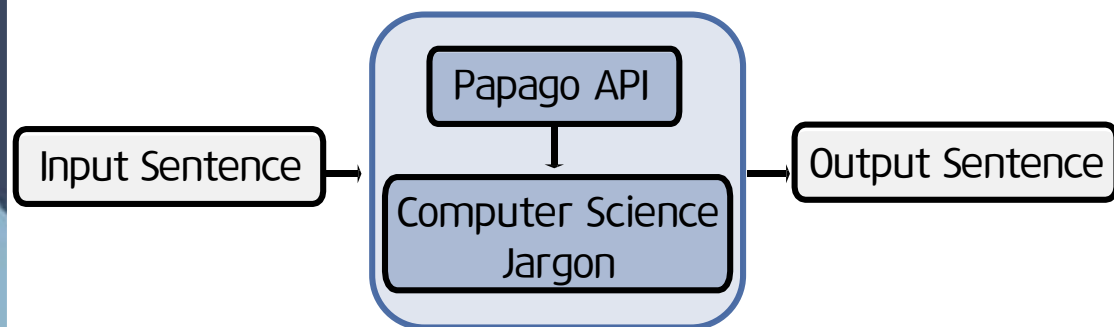
Node.js 14.15.5
NPM 6.14.11

Web

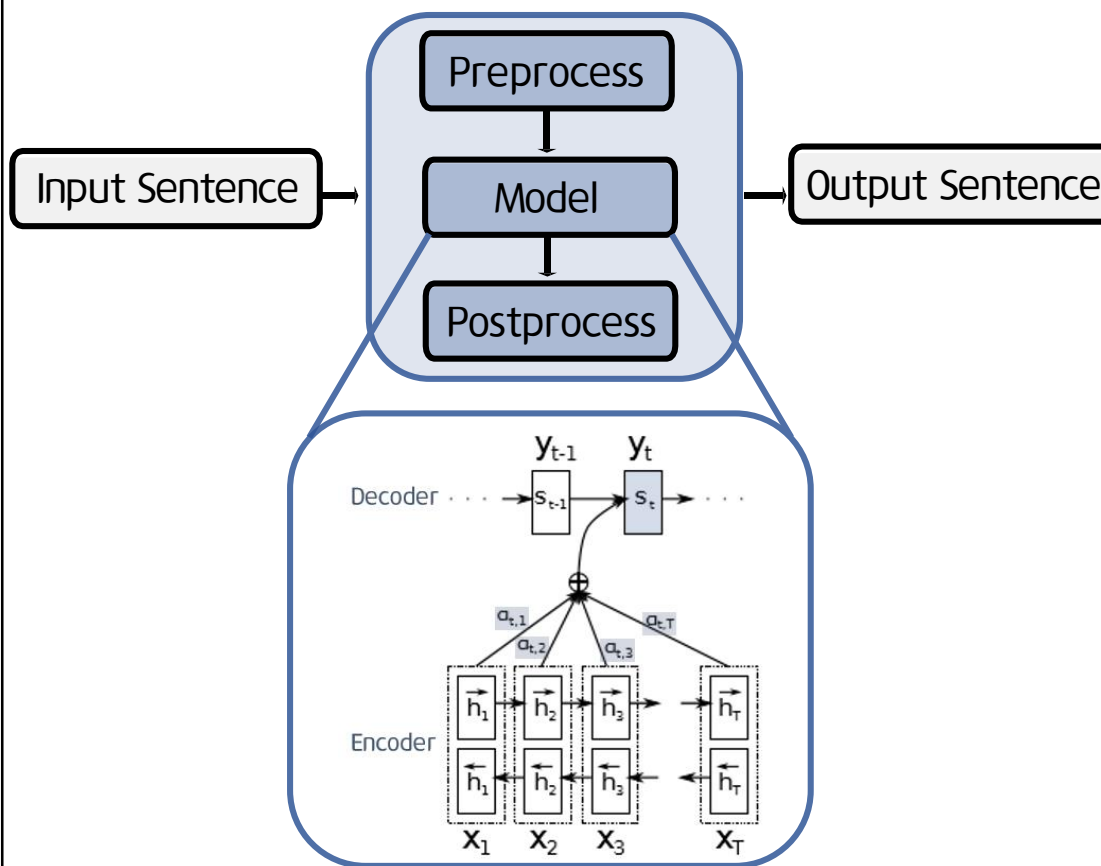


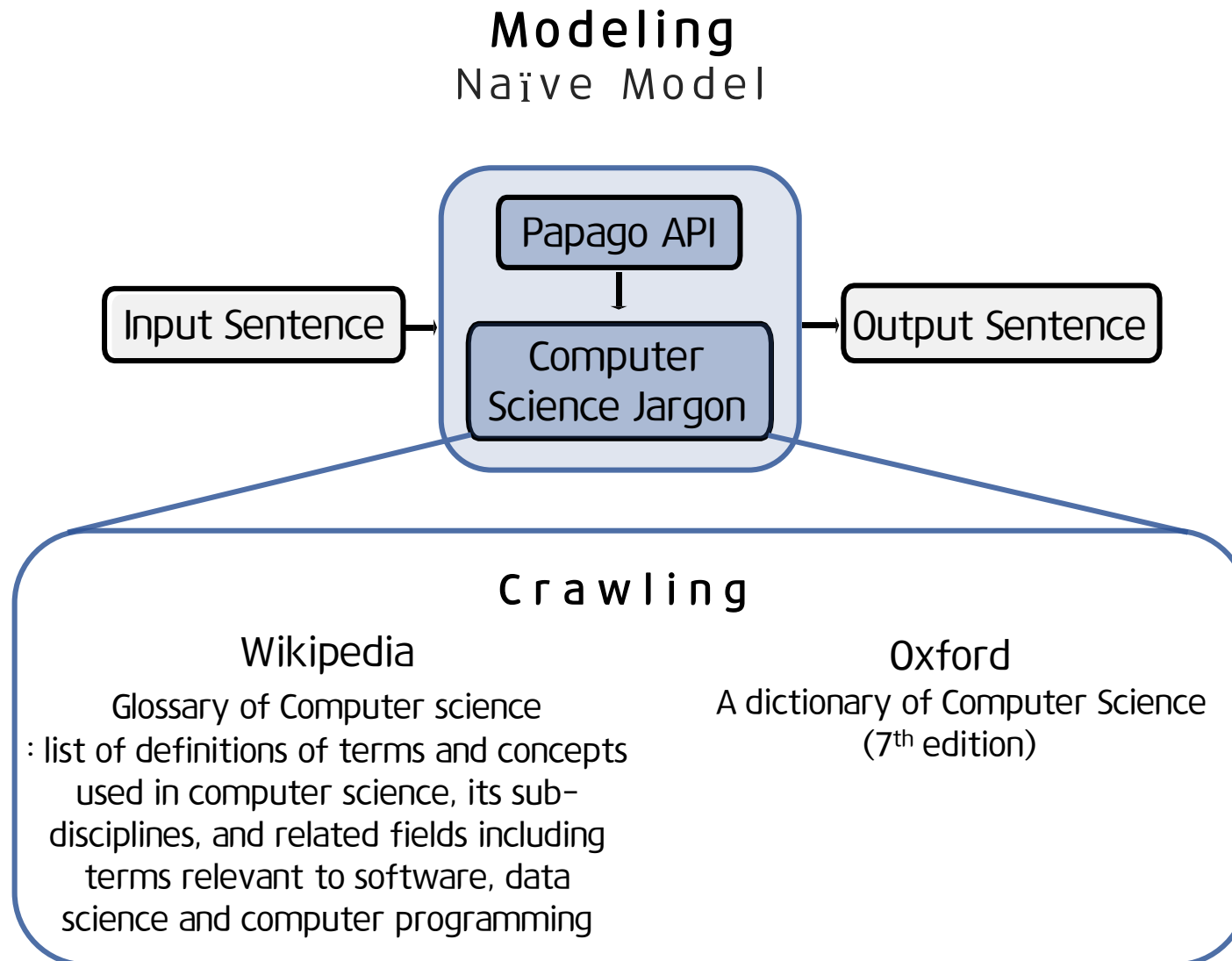
Modeling

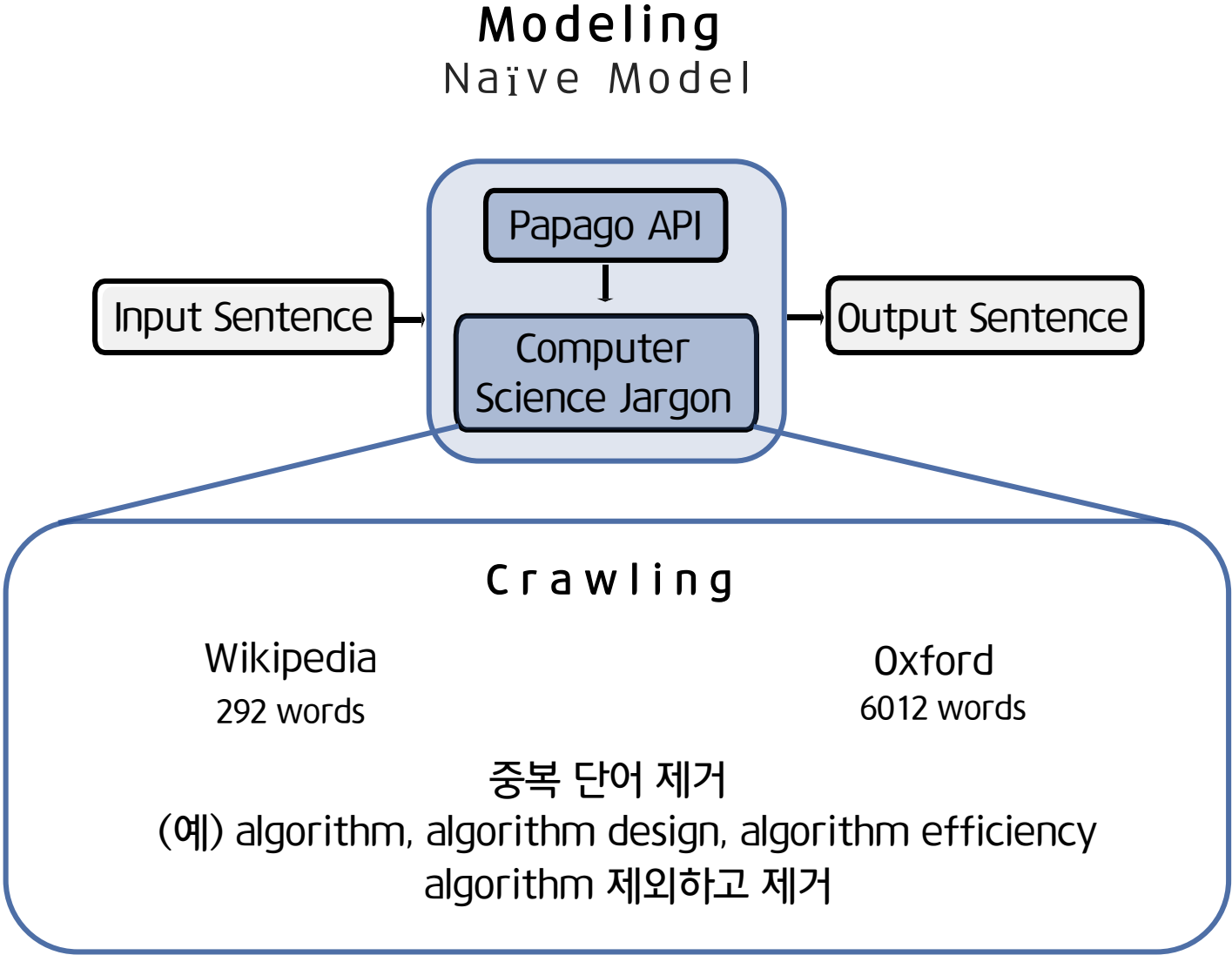
Naïve Model



Advanced Model

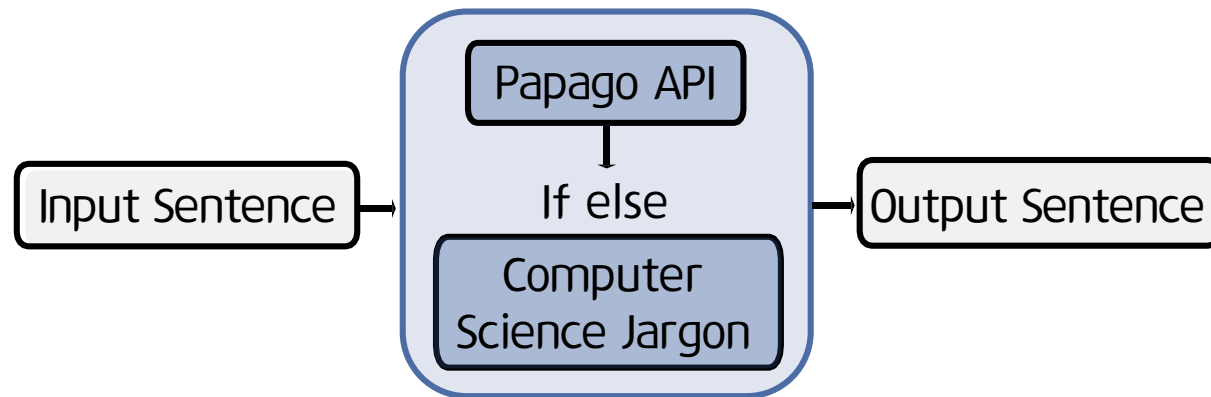






Modeling

Naïve Model



문 제 점

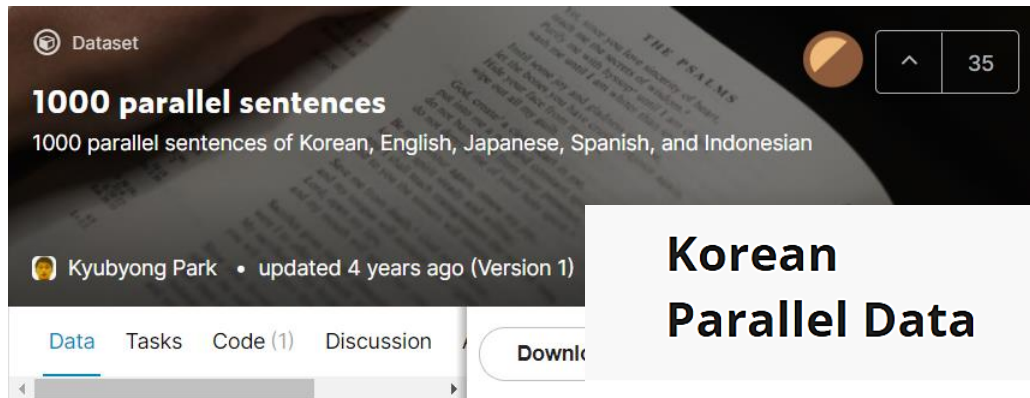
- 문맥에 따라 한국어 번역이 더 자연스러운 단어들의 경우에도 computer science jargon인 경우 영어단어로 남겨둠
- computer science jargon들이 명사가 아닌 경우 번역 부자연 문제

Modeling

Advanced Model

D a t a s e t

- Korean Parallel Text Corpora : 1,000 parallel sentences
- Korean-English parallel corpus : 700 training 700 test sentences
- ParaCrawl English-Korean : 4,002,441 parallel sentences



Korean Parallel Data

KOREAN-ENGLISH
PARALLEL CORPUS
[SITEMAP](#)

Korean-English parallel
corpus



Search this site

Modeling

Advanced Model

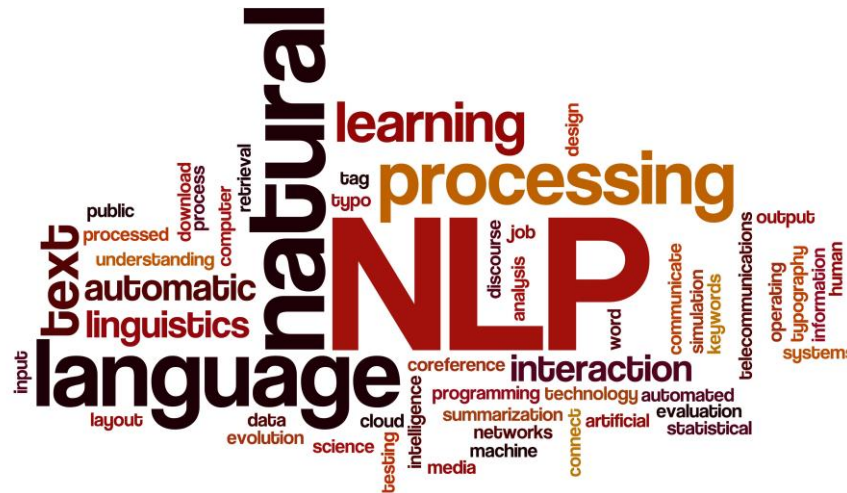
Preprocessing

- 유니코드문자 : ASCII로 변환
- 모든 영어 : 소문자
- 대부분의 구두점 삭제



데이터 준비를 위한 전체 과정 :

- 텍스트 파일을 읽고 줄로 분리, 줄을 쌍으로 분리
- 텍스트 Normalization, 길이와 내용으로 필터링
- 쌍을 이룬 문장들로 단어 리스트 생성
(Lang 인스턴스)



Modeling Advanced Model

Model

 PyTorch
Attention Mechanism



ELMO
: Embeddings from
Language Model



BERT : Pre-training of Deep Bidirectional
Transformers for Language Understanding

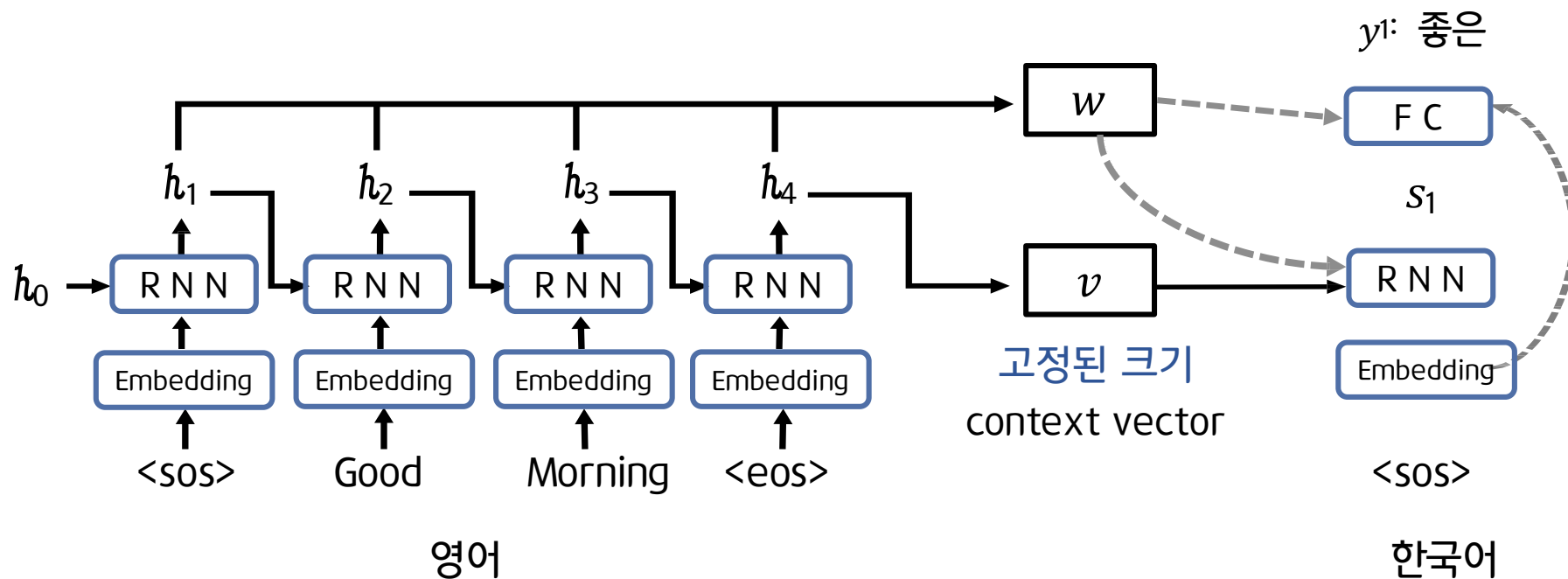
 OpenAI
GPT : Generative Pre-Training

Modeling

Advanced Model

Model

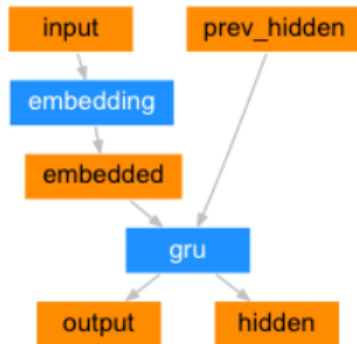
- Seq2Seq 모델에 Attention Mechanism 사용
- Decoder는 Encoder의 모든 출력 참고



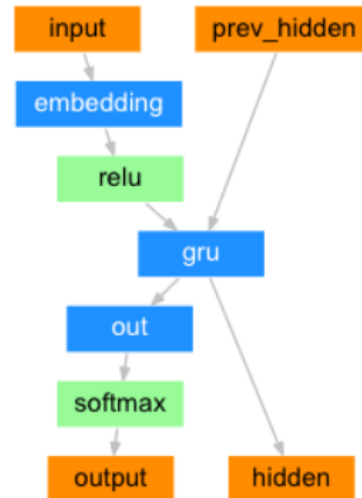
Modeling

Advanced Model

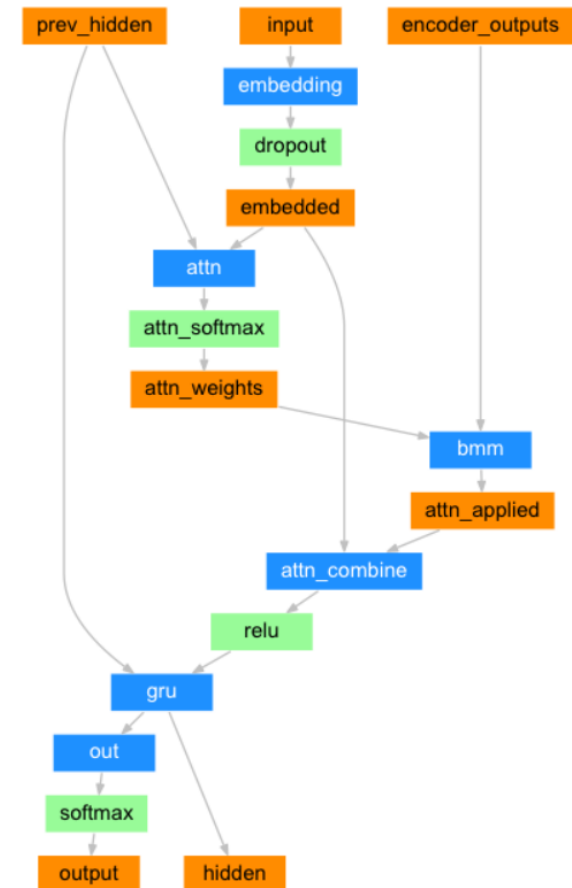
Encoder



Decoder



Attention



전체 학습 과정 :

- 타이머 시작
- Optimizers와 criterion 초기화
- 학습 쌍의 세트 생성
- 도식화를 위한 빈 손실 배열 시작

주요기능

Translation Program for Papers

Translator

☐ 인문학 ☐ 사회과학 ☐ 자연과학 ☐ 공학 ☐ 의학 ☐ 농수해양학 ☐ 예술체육학

English ⇌ Korean

Type the text to translate.

Be a translator

☐ 인문학 ☐ 사회과학 ☐ 자연과학 ☐ 공학 ☐ 의학 ☐ 농수해양학 ☐ 예술체육학 ← 분야 선택

English ⇌ Korean

Type the text to fix.

Submit

Translator : seq2seq with attention 기반
논문에 최적화된 번역 수행

Be a translator : 사용자가 train data 추가
← 분야 선택



- 한국어-영어 데이터 쌍 수집 어려움
- 모델 선정 어려움
- GPU가 없는 로컬에서 모델링 진행

Efforts to Overcome Challenges

- 다양한 한국어-영어 dataset 사용
(Korean Parallel Text Corpora, Korean English Parallel Corpus, Paracrawl English Korean)
- 자연어처리 모델 논문 리뷰

- Knowledge Distillation을 이용해 동일한 성능을 가지는 작은 모델 생성
- 질 좋은 한국어-영어 데이터 쌍 이용
- 사용자로부터 번역이 어색한 문장 → 정확한 문장 多
- 다양한 자연어처리 모델 (BERT, GPT3) 활용
- GPU를 활용한 모델링 진행



- 학습 속도 개선
- 공학 뿐 아니라 인문학, 사회과학, 자연과학, 의약학, 농수해양학, 예술체육학 등 분야 확장
- 번역 모델 성능 ↑

감사합니다