Hamiltonian System을 이용한 텍스트 번역 서비스

제안서

팀명: 카페인

팀장: 201910964 박찬진

제안 배경

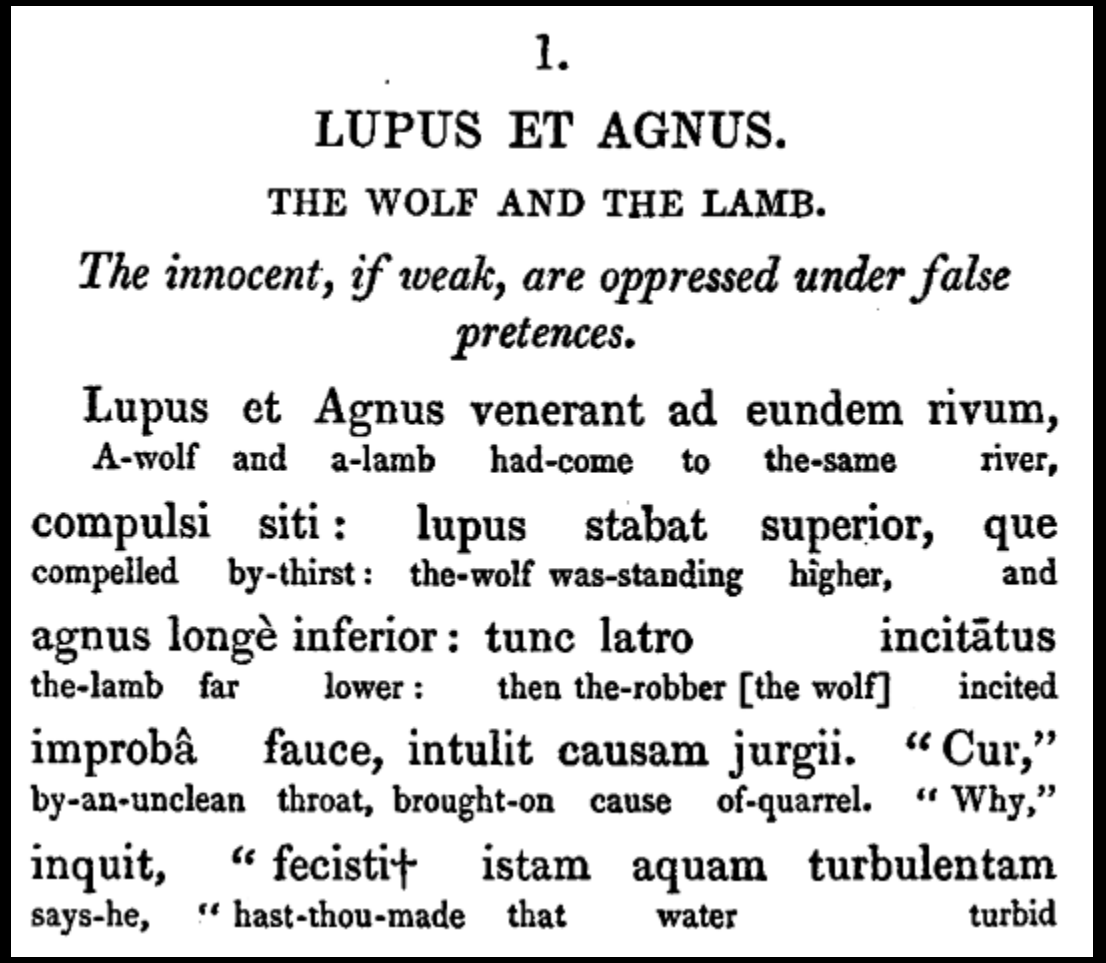
- 외국어를 배우는 사람이 외국어 텍스트를 읽을 때 가장 어려워 하는 것은 **단어**이다.
- 모르는 단어가 나올 때마다 사전을 들춰보거나 검색하는 것은 번거로울 뿐만 아니라 몰입을 방해할 수 있다.
- 그러므로 **단어** 단위로 번역을 해주는 서비스가 있으면 도움이 될 수 있다.

- 최근 발달한 자연어 처리 기술을 사용할 수 있다.

Interlinear Gloss

- 단어에 주석을 다는 것.
- 아일랜드의 James Hamilton (1769-1829)이 언어 교수 용도로 사용하여 Hamiltonian System이라고도 함.
- 텍스트를 읽으며 단어와 문법을 배울 수 있음.

- 예) 상용 앱 Legentibus에서 일부 라틴어-영어 말뭉치 제공



1.

LUPUS ET AGNUS.

THE WOLF AND THE LAMB.

The innocent, if weak, are oppressed under false pretences.

Lupus et Agnus venerant ad eundem rivum,
A-wolf and a-lamb had-come to the-same river,

compulsi siti : lupus stabat superior, que
compelled by-thirst : the-wolf was-standing higher, and

agnus longè inferior : tunc latro incitātus
the-lamb far lower : then the-robber [the wolf] incited

improbâ fauce, intulit causam jurgii. "Cur,"
by-an-unclean throat, brought-on cause of-quarrel. "Why,"

inquit, "fecisti† istam aquam turbulentam
says-he, "hast-thou-made that water turbid

라틴어 "이솝의 우화" 를 영어화자를 위해 Hamilton System을 이용해 주석을 단 책. (1832)
Æsop's Fables, as Romanized By Phædrus: with a Literal Interlinear Translation

기계번역과 다른 점

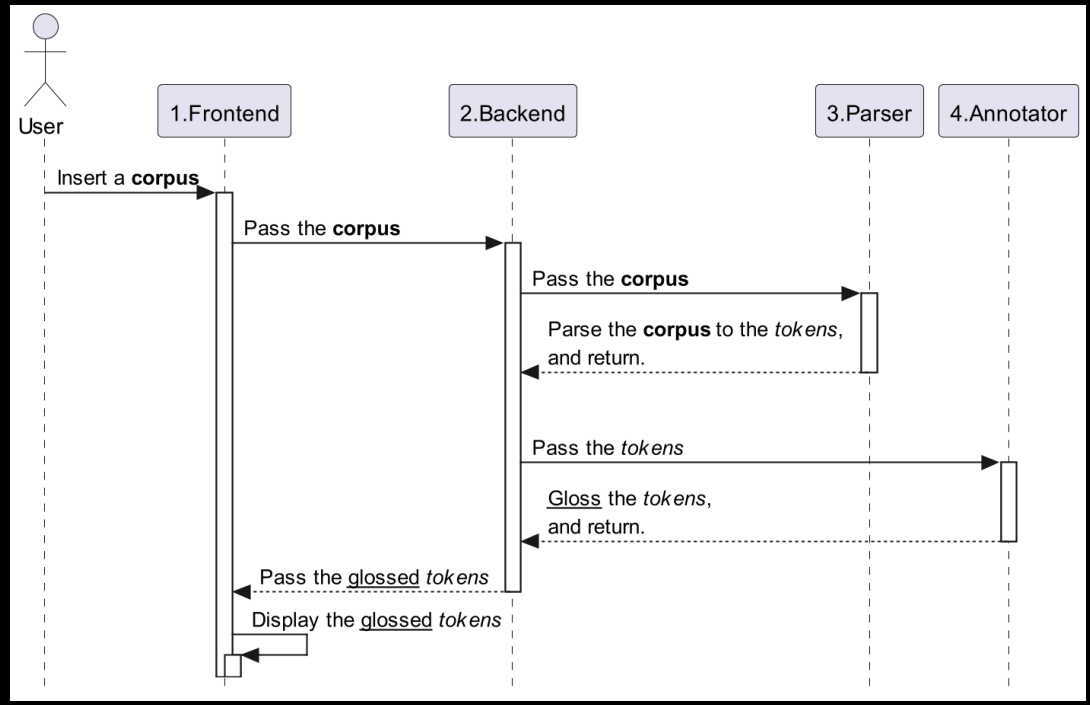
- 기계번역 특성상 출력물이 어색하다.
- 학습자가 번역문과 원문과의 상관관계를 파악하기 힘들어 언어학습에 부적합하다.
- 원문을 제대로 이해하는데 도움이 크게 되지 않는다.

- 해밀턴 시스템은 이미 알고 있는 언어가 아니라 대상 언어 (Target Language)를 직접 읽게 함으로써 효율적으로 언어학습을 할 수 있게 한다.

제안 시스템 구조

- 사용자로부터 말뭉치를 입력 받아 이를 단어 단위로 분해하여 자연어 처리 기술 등을 이용하여 의미를 단 후, 이를 알아보기 쉽게 표시한다.

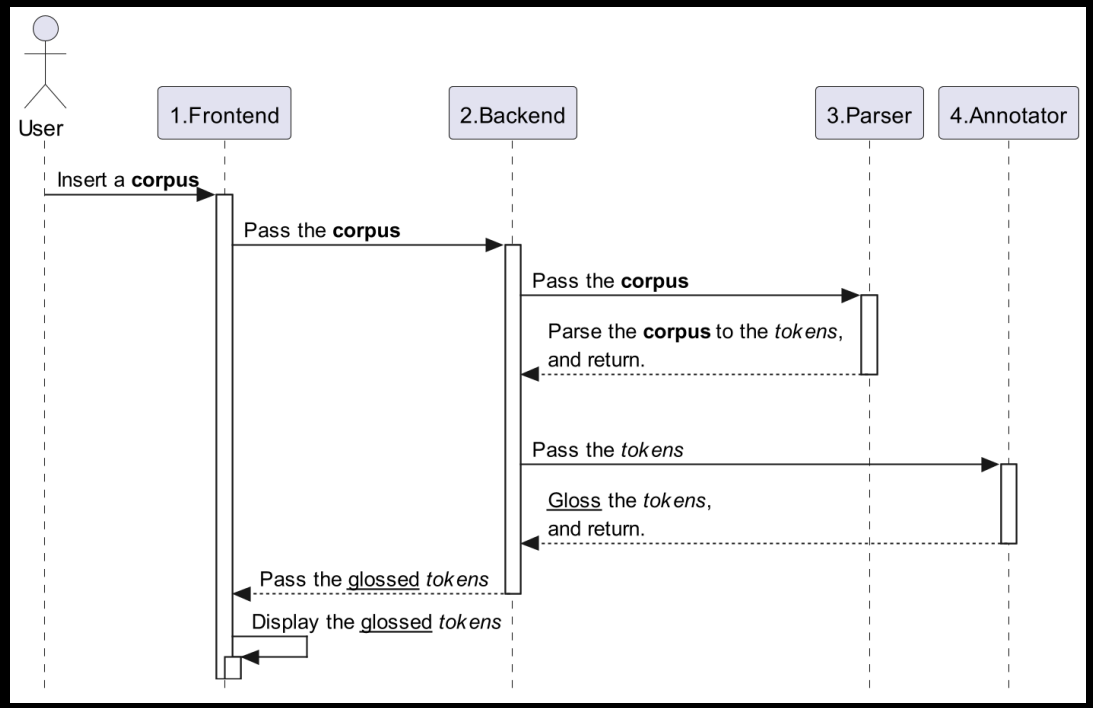
- 구성요소 (계속)
  1. 프론트엔드
  2. 백엔드
  3. 말뭉치 분해기 Parser
  4. 어노테이터 Annotator



1. 프론트엔드

“Lupus arguebat vulpem crimine furti;”

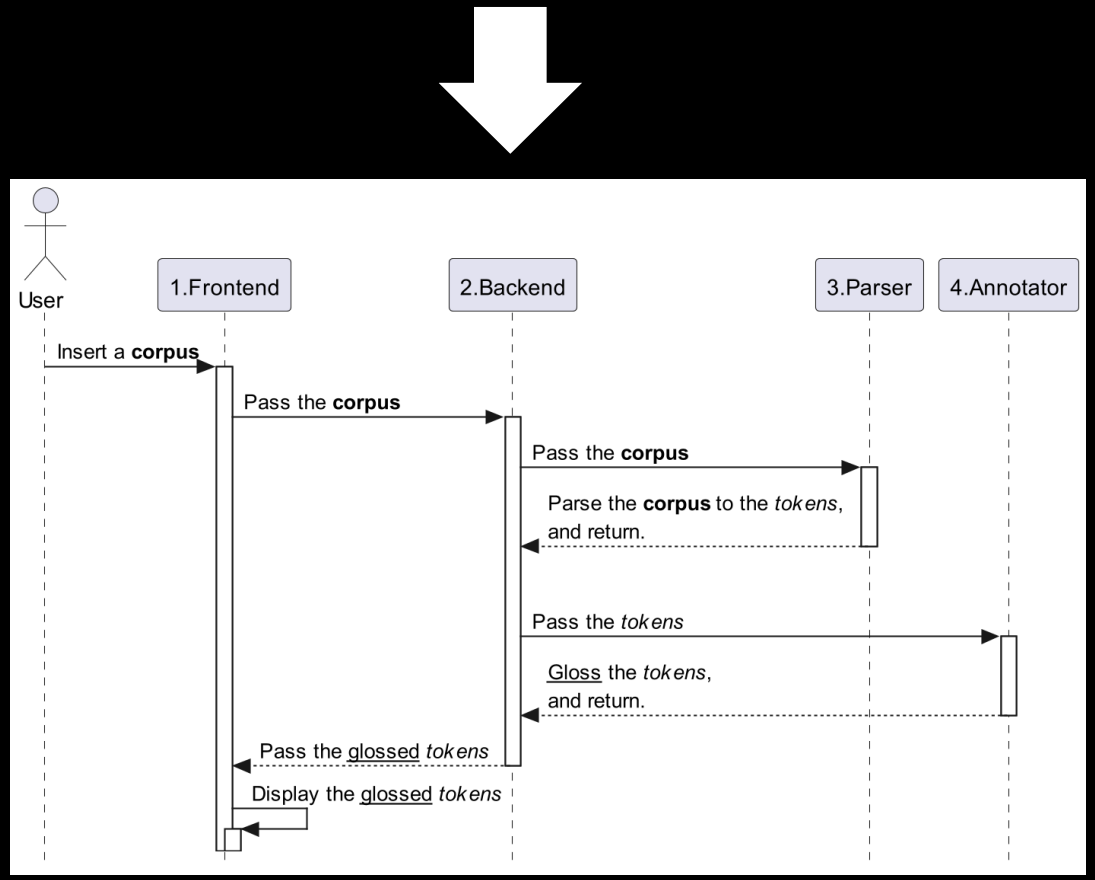
- 사용자로부터 말뭉치를 입력받는다.
- 입력받은 말뭉치를 백엔드로 전달한 뒤, 주석이 달린 결과물을 사용자가 읽기 쉽게 표시한다.



2. 백엔드

- 프론트엔드로부터 전달받은 말뭉치를 Parser에 전달해 토큰Token 단위로 분해한다.
- 분해된 토큰을 Annotator에 전달해 Token 당 주석Gloss을 전달받는다.
- Gloss가 포함된 Token들을 다시 프론트엔드로 전달한다.

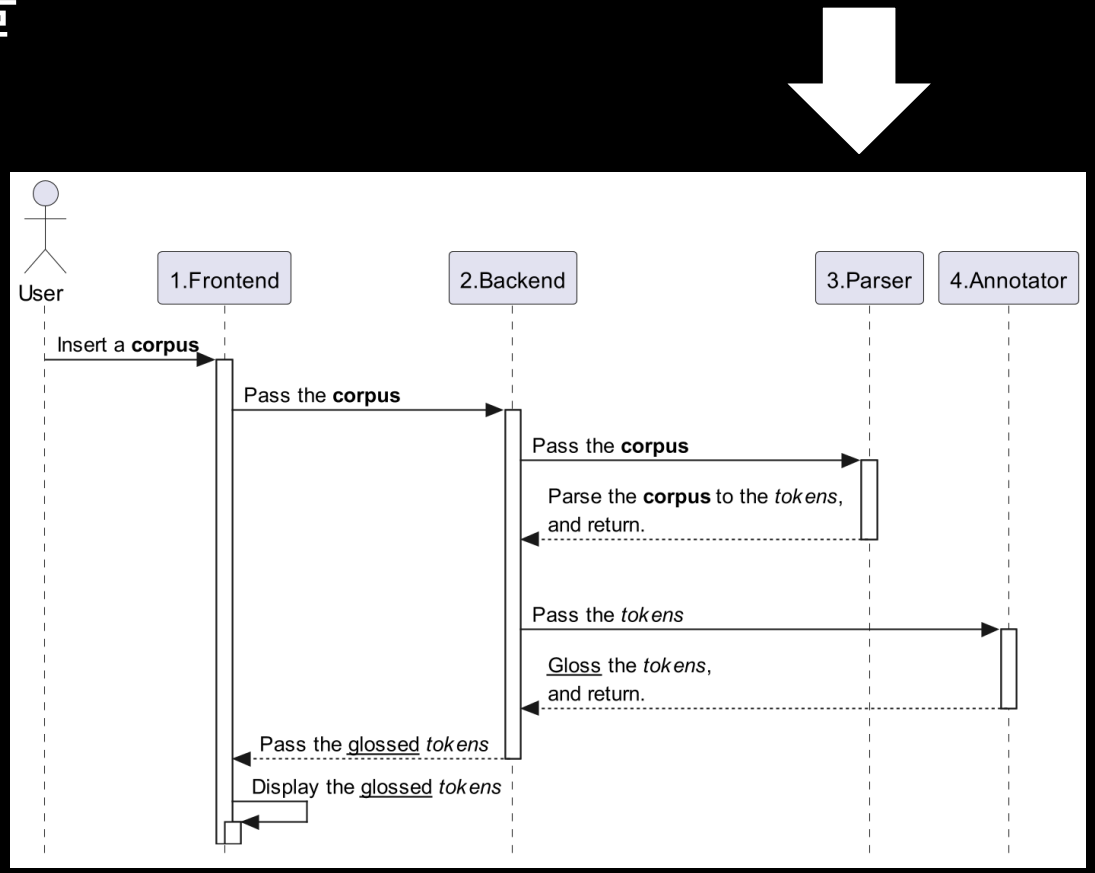
“Lupus | arguebat | vulpem | crimine | furti;”
A-wolf | charged | a-fox | with-the-crime | of-theft



3. Parser

“Lupus | arguebat | vulpem | crimine | furti;”

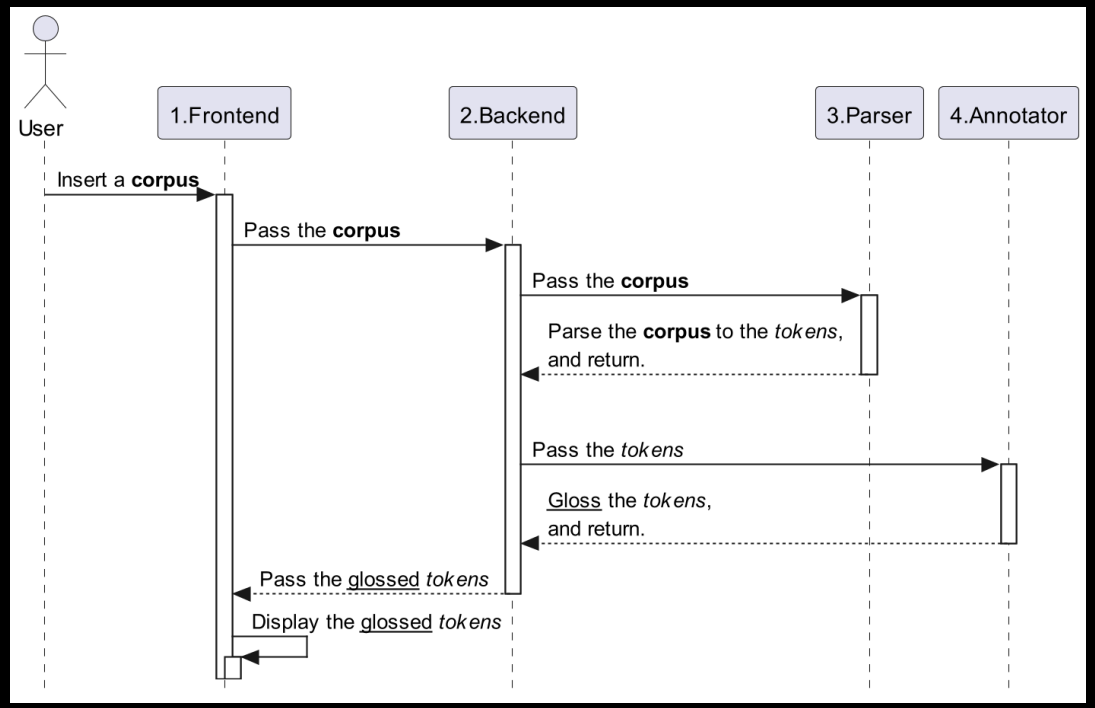
- 백엔드로부터 전달받은 말뭉치를 “단어” 단위로 분해한 Token들을 다시 백엔드로 전달한다.
- 영어 등의 언어는 whitespace를 기준으로 나누면 되지만, 일본어/중국어와 같이 띄어쓰기를 하지 않는 언어는 다른 기준을 사용해야 할 수도 있다.



4. Annotator

“Lupus | arguebat | vulpem | crimine | furti;”
A-wolf | charged | a-fox | with-the-crime | of-theft

- 백엔드로부터 전달받은 Token들에 올바른 Gloss를 삽입한다.
- 동음이의어/다의어 등을 문맥에 따라 번역하여야 한다.
- ChatGPT와 같은 LLM을 사용하거나 전통적인 ML 자연어 처리, 혹은 단순 사전 인색과 같은 방법을 사용할 수 있다.
  - 사전 인색의 경우 문맥이 무시됨.



개발 방법론

- 반복적이고 점진적인 개발방법인 UP (United Process)를 원칙으로 개발할 계획이다.

개발팀 구성

이름	학과	학번	역할
박찬진	컴퓨터과학전공	201910964	팀장 및 백엔드 등
고준식	컴퓨터과학전공	201910921	백엔드
강병규	역사콘텐츠전공	201810001	프론트엔드 및 백엔드
이재웅	역사콘텐츠전공	201810027	프론트엔드
최재영	역사콘텐츠전공	201810035	프론트엔드

기대효과

- 외국어를 배우고자 하는 사용자들이 읽고싶어하는 텍스트를 쉽게 받아들일 수 있게 하여 언어학습에 도움을 준다.