

AI and Romanization: Possibilities and Limitations

HYOUNGBAE LEE & DAE CHUL SON

Traditional Algorithms vs. AI & Machine Learning

TRADITIONAL ALGORITHMS

Rule-based

Explicit instructions

Deterministic

Manual update

AI & MACHINE LEARNING

Data-driven

Learning from patterns

Probabilistic

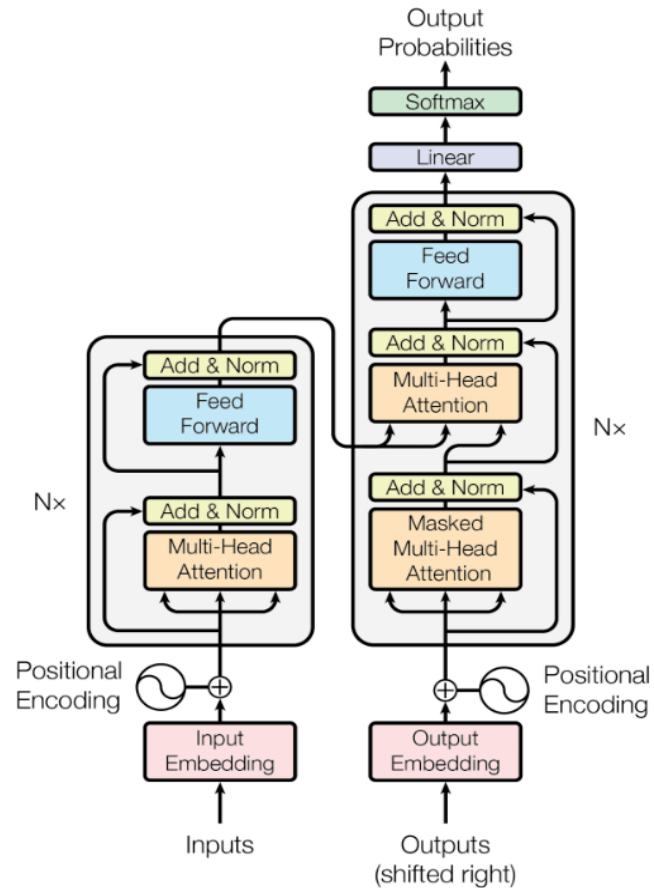
Self-adaptive

Reviewed models

- RNN(Recurrent Neural Network)
- BERT(Bidirectional Encoder Representations from Transformers)
- Transformer

In the end, I chose the Transformer model

The Transformer model architecture



Data collection

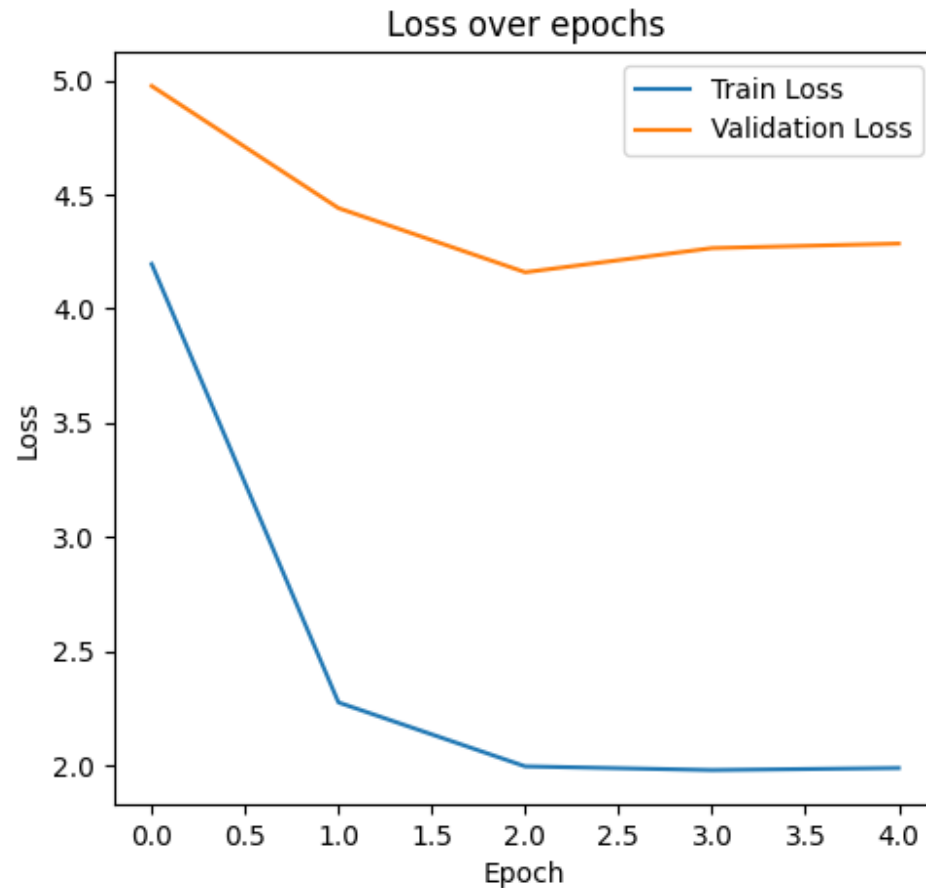
Provided by Princeton Library

- Person name data : 16,000
- Book title data : 26,000

900,000 generated by Son's Romanizer

영국 노동계급의 형성(상),Yōngguk nodong kyegūp ūi hyōngsōng (sang)
영국노동계급의 형성(하),Yōngguk nodong kyegūp ūi hyōngsōng (ha)
컴퓨터 네트워크,K'ōmp'yut'ō net'ūwōk'ū
1000제 해법수학 1-가,1000 che haepōp suhak 1-ka
신중성 종합불어 3(대학원독해),Sinjungsōng chonghap Purō 3 (taehagwōn tokhae)
화용도,Hwayong to
불자의 마음가짐과 수행법,Pulcha ūi maūm kajim kwa suhaeng pōp
나의 어머니 육영수,"Na ūi ōmōni Yuk, Yōng-su"
고전문학과 표현교육론,Kojōn munhak kwa p'yohyōn kyoyungnon
토익 튜터(TRAINING편),T'oik T'yu T'ō (TRAINING p'yōn)
텝스 튜터(TRAINING),T'epsū T'yu T'ō (TRAINING)
텝스 튜터(TRAINING)(T:3),T'epsū T'yu T'ō (TRAINING) (T : 3)
베이징(CITYPACK 3),Peijing (CITYPACK 3)
돌과 피리 1,Tol kwa p'iri 1
돌과 피리 3,Tol kwa p'iri 3
돌과 피리 2,Tol kwa p'iri 2
수랏 샤프드 요가,"Su, nat Sya Pū Tū yoga"
조선비침(상),Chosōn pich'ōp (sang)
조선비침(하),Chosōn pich'ōp (ha)
나모 웹에디터 3.05로 하루만에 내 집 만들기(S/W포함),Namo wep e Tit'ō 3. 05 no ha Numan e nae chip mandūlgi (S / W p'oham)
헌법의 핵(사시2차대비),Hōnpōp ūi haek (sasi 2-ch'a taebi)
기적의 영어발음 표기법,Kijōk ūi Yōngō parūm p'yogipōp
신문을 보면 21세기가 보인다,Sinmun ūl pomyōn 21-segi ka poinda
서글픈 고정관념,Sōgūlp'ūn kojōng kwannyōm
서글픈 고정관념(TAPE포함),Sōgūlp'ūn Kojōng kwannyōm (TAPE p'oham)
형사소송법(경찰채용시험대비),Hyōngsa sosong pōp (kyōngch'al ch'aeyong sihōm taebi)
국어(경찰채용시험대비),Kugō (kyōngch'al ch'aeyong sihōm taebi)
추리캠프(야호 2),Ch'uri Ch'aemp'ū (yaho 2)
국사(경찰채용시험대비),Kuksa (kyōngch'al ch'aeyong sihōm taebi)
미로캠프(야호 1),Miro Ch'aemp'ū (yaho 1)
인터넷 라이프(월간),Int'ōnet laip'ū (wōlgan)
많은 목소리 Y2K,Manūn mul sorī Y 2 K
세무와 철세,Semu wa chōlse
전공 미술(New 중등교원 임용시험 대비),Chōn'gong misul (New chung tūng kyowōn imyong sihōm taebi)
빠른독해 바른독해 빠바 기초세우기,Pparūn tokhae parūn tokhae ppaba kich'o seugi
대운만세력,Taeun man seryōk
아르스 마그나,Arūsū magūna
내겐 이대학 이학과가 최고,"Naegen Yi, Tae-hak i hakkwa ka ch'oego"
내겐 이대학 이학과가 최고,"Naegen Yi, Tae-hak i hakkwa ka ch'oego"
가족법(핵심경리),Kajokpōp (haeksim chōngni)
한글엑셀 2000,Han'gūl Ekse1 2000
박빙 영어어법(수진본)(중문),Pakping Yōngō ōpōp (sujinbon) (chungmun)
SOUTHERN ACCENTS(2004.2),SOUTHERN ACCENTS (2004. 2)
환경건강학,Hwan'gyōng kōn'gang hak
ROCKET,ROCKET
폭풍의연덕(22. 영어로읽는세계명작스프링문고),P'okp'ung ūi ōndōk (22. Yōngō no ingnūn segye myōngjak sūp'ūring mun'go)
21세기를 위한 젊은이의 준비,21-segi rŭl wihan chōlmūni ūi chunbi
백년의 고독 1,Paek yōn ūi kodok 1
백년의 고독 2,Paek yōn ūi kodok 2
중독된 사랑,Chungdoktoen sarang
사명대사(별빛문고 35),Sa myōngdaesa (pyōl pit mun'go 35)
사주문답,Sa chu mundap
죽는것도 나에게는 이득이 됩니다(바울선교일대기),Chungnūn kōt to na ege nūn idūk i toemnida (Paul sōn'gyo iltaegi)
3년 후에도 그리워진다면,3-yōn hu e to kŭriwō chindamyōn
문제해결의 수학적 전략,Munje haegyōl ūi suhakchōk chōllyak
한번 물러서면 천번 물러선다,Han pōn mullōsōmyōn ch'ōn pōn mullōsōnda
조선시대수군 2,Chosōn sidae sugun 2

Problems discovered during learning



울학교?

176,679 search results appeared for '울학교 족보' (Our School Genealogy).

Total number of vocabulary trained: 72,985 (includes a significant number of personal and proper nouns due to the nature of bibliographic records, not just the refined corpus).



Sentences composed of words likely to be abundant in the training data

한국의 민요

Base Model Translated: Han'guk ŭi minyo

한국의 동요

Base Model Translated: Han'guk ŭi tongyo

한국의 김치

Base Model Translated: Han'guk ŭi kimch'i

한국의 역사와 김치

Base Model Translated: Han'guk ŭi yŏksa wa kimch'i

김치의 역사와 우리 민족

Base Model Translated: Kimch'i ŭi yŏksa wa uri minjok

민속문화

Base Model Translated: Minsok munhwa

초등학교 진단평가

Base Model Translated: Ch'odŭng hakkyo chindan p'yŏngka

고등학교 중간고사 범위

Base Model Translated: Kodŭng hakkyo chunggan kosa pŏmwi

Sentences composed of words unlikely to be present in the training data

저는 학생이에요

Base Model Translated: Chŏ nŭn haksaeng ieyo

모델 아키텍처 다이어그램

Base Model Translated: model ak'it'ekch'ŏ Chang, Hon

트랜스포머 모델 연구

Base Model Translated: Ha, ka model yŏn'gu

로마자 변환

Base Model Translated: Yi ki taegye

개발자의 마음 하늘의 마음

Base Model Translated: Kaebal ŭi maŭm hanŭl ŭi maŭm

저자의 마음 하늘의 마음

Base Model Translated: minjok ŭi maŭm hanŭl ŭi maŭm

Everybody has a plan until ...

Original plan:

- Initial Training:
Collect quality romanization data and let AI learn from them
- Reinforcement Learning Loop:
Correct errors and let AI learn more from the corrections
Repeat the process as needed
- Apply the same method to test Japanese data

... it does not work well, and here's why

“Data! Data! Data!”

he cried impatiently.

“I can’t make bricks without clay.”

- Sherlock Holmes

The Adventure of the Copper Beeches



Not Just Any **Data!**

Quantity: More data, better models

Quality: GIGO (Garbage in, garbage out)

Diversity: Avoid bias. Include irregularities.



Ideal scenario: Prepare a **large quantity of high-quality, diverse data**

Ideal: Pure AI Approach

Quantity & Diversity: Synthetic data generation

Quality: Rigorous cleaning & error detection

⇒ Obtain a large quantity of high-quality, diverse data



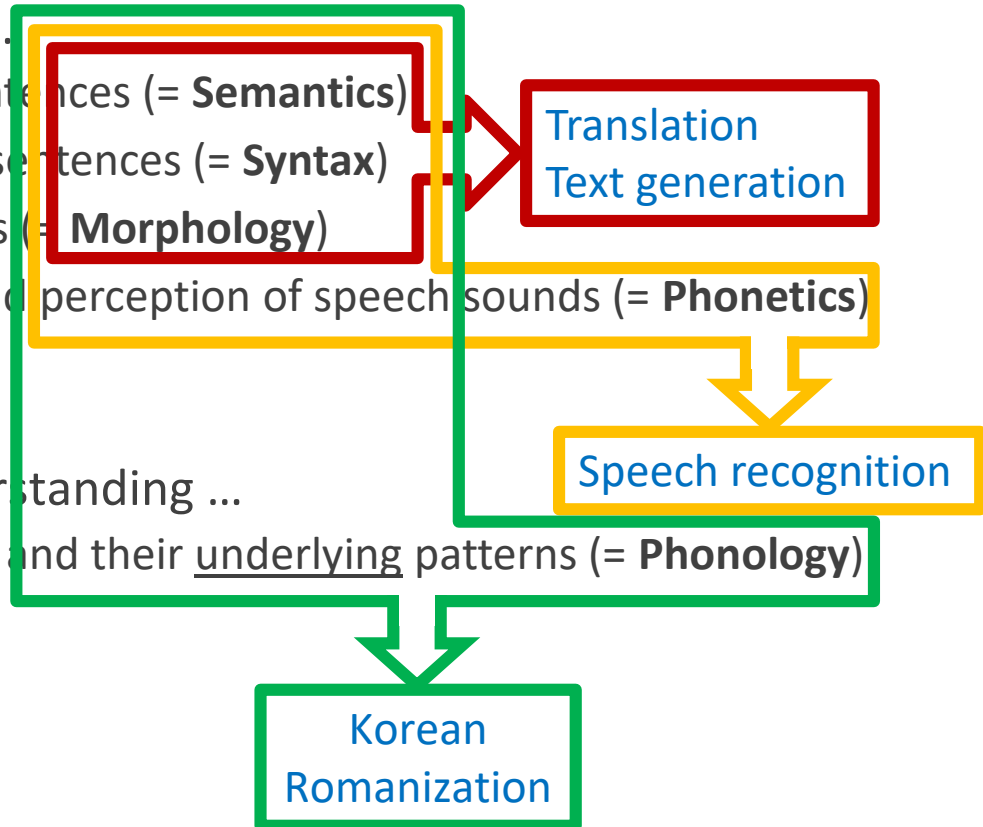
AI: Strengths and Weakness

AI is good at analyzing ...

- Meaning of words and sentences (= **Semantics**)
- Grammatical structure of sentences (= **Syntax**)
- Internal structure of words (= **Morphology**)
- the physical production and perception of speech sounds (= **Phonetics**)

But not that good at understanding ...

- Abstract system of sounds and their underlying patterns (= **Phonology**)



More Practical: Hybrid Approach

AI and machine learning:

- Dependency parsing (e.g. subject-verb, verb-object)
- Constituency parsing (e.g. noun phrases, verb phrases)
- Part-of-Speech (POS) Tagging (e.g. nouns, verbs, particles, etc.)
- Named Entity Recognition (NER) Tagging (people, organizations, places, etc.)

Traditional algorithm programming (including K-Romanizer)

- Apply sound change rules
- Handle exceptions

A case of Japanese: Can AI romanize 強 correctly?

強力 strong

強: kyō

強力犯 violent criminal

強: gō

強い strong

強: tsuyo

強い stubborn

強: kowa

強い to force

強: shi

強く severely

強: shitata

強い (not) necessarily

強: anaga

強情 stubbornness

強情: gōjō

強情 to nag

強情: neda

強情 to extort

強情: yusu

伊藤強

強: tsuyoshi

岩田強

強: tsutomu

海東強

強: takeshi

赤土正強

強: take

赤石強司

強: kyō

杉浦強司

強: tsuyo

松平貴強

強: masu

志村弘強

強: yuki

強谷幸雄

強: sune

強力敏郎

強: gō

FR You
Translate 強ちそうとは限ぎらない:

ChatGPT
強ちそうとは限りません translates to "It is **not necessarily** certain" in English.

FR You
How to pronounce 強ち in that sentence?

ChatGPT
The pronunciation of 強ち in this context is "shikarichi." Each character is pronounced as follows:

- 強 (し or ごう): "shi" or "gou"
- ち: "chi"

So, when combined, it is pronounced as "shikarichi."

強ち: (not) necessarily

Correct reading:

anagachi

ChatGPT



Translate: 強ちそうとは限ぎらない

Show drafts ▾



Sure. The Japanese phrase 強ちそうとは限ぎらない (pronounced "tsuyoshi sou to wa kagiranai") can be translated into English in a few different ways, depending on the context. Here are a few possible translations:

- **Not necessarily**
- It's not always the case
- That's not always true
- It's not guaranteed
- It's possible, but not certain

For example, you might say 強ちそうとは限ぎらない in response to someone who says something like "Hard work always pays off." In this case, you would be saying that it's not always the case that hard work leads to success.

Gemini

Data! Data! Data!

We can't make bricks without clay.

Thank you!

Hyoungbae Lee (hyoungl@princeton.edu)

Korean Studies Librarian, Princeton University

Dae Chul Son (sdc@kongnpark.com)

Chief Technology Officer, Kong & Park, Inc.