AI and Romanization: Possibilities and Limitations

HYOUNGBAE LEE & DAE CHUL SON
# Traditional Algorithms vs. AI & Machine Learning

<table>
<thead>
<tr>
<th>TRADITIONAL ALGORITHMS</th>
<th>AI &amp; MACHINE LEARNING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based</td>
<td>Data-driven</td>
</tr>
<tr>
<td>Explicit instructions</td>
<td>Learning from patterns</td>
</tr>
<tr>
<td>Deterministic</td>
<td>Probabilistic</td>
</tr>
<tr>
<td>Manual update</td>
<td>Self-adaptive</td>
</tr>
</tbody>
</table>
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

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

900,000 generated by Son’s Romanizer
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 ūi minyo

한국의 동요
Base Model Translated: Han'guk ūi tongyo

한국의 김치
Base Model Translated: Han'guk ūi kimch'i

한국의 역사와 김치
Base Model Translated: Han'guk ūi yōksa wa kimch'i

김치의 역사와 우리 민족
Base Model Translated: Kimch'i ūi yōksa 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 haksan’g 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
AIs are good at analyzing ... 
- Meaning of words and sentences (\(= \text{Semantics}\))
- Grammatical structure of sentences (\(= \text{Syntax}\))
- Internal structure of words (\(= \text{Morphology}\))
- the physical production and perception of speech sounds (\(= \text{Phonetics}\))

But not that good at understanding ... 
- Abstract system of sounds and their underlying patterns (\(= \text{Phonology}\))
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?

<table>
<thead>
<tr>
<th>強力 strong</th>
<th>強: kyō</th>
<th>強力犯 violent criminal</th>
<th>強: gō</th>
<th>強力敏郎 強: gō</th>
</tr>
</thead>
<tbody>
<tr>
<td>強い strong</td>
<td>強: tsuyo</td>
<td>強い stubborn</td>
<td>強: kowa</td>
<td>赤土正強 強: take</td>
</tr>
<tr>
<td>強いる to force</td>
<td>強: shi</td>
<td>強か severely</td>
<td>強: shitata</td>
<td>杉浦強司 強: tsuyo</td>
</tr>
<tr>
<td>強ち (not) necessarily</td>
<td>強: anaga</td>
<td>強情 stubbornness</td>
<td>強情: gōjō</td>
<td>松平貴強 強: masu</td>
</tr>
<tr>
<td>強情 to nag</td>
<td>強情: neda</td>
<td>強情る to nag</td>
<td>強情: yusu</td>
<td>志村弘強 強: yuki</td>
</tr>
<tr>
<td>強情る to extort</td>
<td>強情: yusu</td>
<td>強谷幸雄</td>
<td>強力敏郎 強: gō</td>
<td>伊藤強 強: tsuyoshi</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>岩田強 強: tsutomu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>海東強 強: takeshi</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>赤石強司 強: tsuyoshi</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>杉浦強司 強: tsuyo</td>
</tr>
</tbody>
</table>
You
Translate 強ちそうとは限らない:

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

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."

Gemini
Sure. The Japanese phrase 強ちそうとは限らない (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.
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.