Sequence-to-sequence 모델을 이용한 로마자-한글 상호(商號) 표기 변환 시스템

김태현○, 정현근, 김재화, 김정길

사람인HR, 사람인LAB
{taehyun.kim, antkdi, jungkil, jaehwa.kim}@saramin.co.kr

Roman-to-Korean Conversion System for Korean Company Names Based on Sequence-to-sequence learning

Tae-Hyun Kim○, Hyun-Guen Jung, Jae-Hwa Kim, Jeong-Gil Kim
SaraminHR, SaraminLAB

요 약

상호(商號)란 상인이나 회사가 영업 활동을 위해 자기를 표시하는데 쓰는 명칭을 말한다. 일반적으로 국내 기업의 상호 표기법은 한글과 로마자를 혼용함으로 상호 검색 시스템에서 단어 불일치 문제를 발생시킨다. 본 연구에서는 이러한 단어 불일치 문제를 해결하기 위해 Sequence-to-sequence 모델을 이용하여 로마자 상호를 이에 대응하는 한글 상호로 변환하고 그 후보들을 생성하는 시스템을 제안한다. 실험 결과 본 연구에서 구축한 시스템은 57.82%의 단어 정확도, 90.73%의 자소 정확도를 보였다.

주제어: 로마자-한글 상호 표기 변환, Sequence-to-sequence, Machine Learning

1. 서론

상호(商號)란 상인이나 회사가 영업 활동을 위해 자기를 표시하는데 쓰는 명칭을 말한다. 일반적으로 상호는 한글 상호 표기와 로마자 상호 표기가 혼용되며, 이는 상호 검색 시스템에서 단어 불일치 문제를 야기한다.

상호의 표기 변환을 위해 국립국어원에서 제정한 국어의 로마자 표기법이나 외래어 표기법을 이용할 수 있지만 인명, 회사명, 단체명 등은 그동안 써 온 관습적 표기를 허용하고 있으며, 실제로 표기법을 따르지 않는 상호가 대부분이다. 또한, “LG”, “SK C&C”, “NHN Entertainment”와 같이 각 로마자를 그대로 한글 표기하거나 숫자 또는 특수 문자와 조합된 상호가 존재하기 때문에 표기법이나 규칙만으로 상호의 로마자-한글 표기를 변환하는 것에는 많은 어려움이 따른다.

따라서, 본 연구에서는 상호 도메인에 적합한 학습 데이터를 구축하는 방법과 Sequence-to-sequence(이하 Seq2seq) 모델[1]을 이용한 로마자-한글 상호 표기 변환 방법을 제안한다. Seq2seq 모델은 기계 번역 분야에서 주로 사용되는 모델로, 입력 시퀀스를 인코딩 및 디코딩하여 길이가 다른 출력 시퀀스를 생성한다. 제안하는 시스템은 규칙 및 자질 튜닝에 의존하는 기존의 연구와 달리 End-to-end 방식으로 접근하였다.

로마자 상호를 한글 상호로 표기 변환하는 시스템은 질의 확장이나 동의어 사전의 구축 등에 활용이 가능하며, 표기 혼용에 따른 단어 불일치 문제의 해결에 기여할 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 살펴보고, 3장에서는 전체적인 로마자-한글 상호 표기 변환 시스템의 구성을 소개한다. 4장에서는 제안 시스템의 실험 결과를 분석하고, 마지막 5장에서는 결론에 대해 기술한다.

2. 관련 연구

로마자-한글 상호 표기 변환에 대한 직접적인 연구는 없었으나 음차 표기의 연구가 있었다. 음차 표기란 외국어의 발음을 자국어로 표기하는 것으로 본 연구와 유사하다. 하지만 상호 표기의 경우 외국어의 원래 발음을 그대로 따르지 않는 경우가 대부분이라는 차이점이 있다. 음차 표기의 기존 연구로는 확률 모델을 이용한 연구[2]와 최대 엔트로피 모델을 이용한 연구[3], 메모리 기반 학습과 결정 트리를 이용한 연구[4]가 있었다.

상호 표기 변환에 관한 연구로는 한글-로마자 상호 표기 변환을 위한 부분 문자열 분석에 대한 연구가 있었다.[5] 한글-로마자 상호 표기 변환의 경우 많은 모호성(ambiguity)이 존재하며, [5]의 연구와 같이 부분 문자열 분석이 필수적이다. 예를 들어 한글 상호 “하이텍”은 “Hi-tech”, “High-tech” 등 여러 로마자 표기가 가능하다. 하지만 로마자-한글 상호 표기 변환의 경우 상대적으로 표기 변환 시 모호성이 적다. 따라서 본 연구에서는 로마자 상호에서 한글 상호로 표기를 변환하였다.

본 연구에서 이용한 Seq2seq 모델은 기계 번역[6] 분야 이외에도 형태소 분석 및 품사 태깅[7-8], 구구조 구문 분석[9] 등 다양한 자연어 처리 분야에 적용되고 있으며 우수한 성능을 보이고 있다.

3. 시스템 구축

이 장에서는 먼저 로마자-한글 상호 표기 변환의 시스템에 대해서 자세히 살펴보고, 그다음으로 학습 데이터를 구축하는 방법을 소개하겠다.

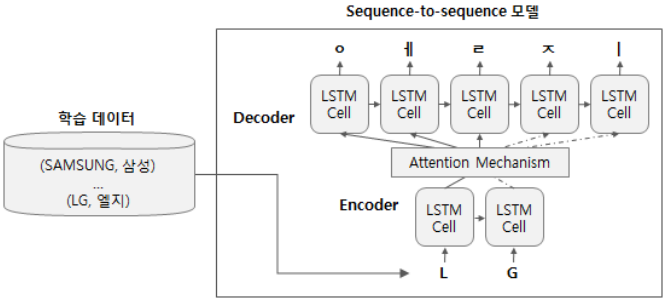
3.1 Sequence-to-sequence 모델을 이용한 로마자-한글 상호 표기 변환

본 연구에서는 로마자-한글 상호 표기 변환을 위해 Seq2seq 모델을 이용하였다. Seq2seq 모델은 입력 시퀀스를 인코딩 및 디코딩하여 길이가 다른 출력 시퀀스를 생성한다. 여기서 모델의 입력 시퀀스는 로마자 상호이며 모델의 출력 시퀀스는 한글 상호이다.

로마자-한글 상호 표기 변환의 문제는 텍스트 요약이나 기계 번역의 문제와 같이 입력 시퀀스와 출력 시퀀스 사이에 일정한 정렬(alignment)이 존재한다. 예를 들어, 로마자 'm', 'b'는 각각 한글 자음 'ㅁ', 'ㅂ'에 명확하게 대응된다. 따라서 Attention Mechanism을[10] 적용함으로써 성능의 향상을 기대하였다.

Attention Mechanism은 디코딩 과정 중 입력을 전역적으로 참고하여 중요한 정보가 있다고 판단되는 특정한 hidden state에 높은 가중치를 주기위한 방법이다.[11]

전체적인 시스템의 구조는 <그림 1>와 같다.



<그림 1> 시스템 구성

3.1.1 한글 표기의 후보 생성

상호 표기는 관습적 표기를 허용하므로 로마자 표기에 대응하는 한글 표기는 다양한 이형태가 존재할 수 있다. 예를 들어 로마자 표기 "RND"는 "알앤디", "알엔디" 등의 이형태가 존재한다. 이러한 이형태는 혼동 자소 때문에 발생하게 되는데 혼동 자소들을 단순하게 치환하는 방법은 불필요한 이형태를 과도하게 생성하여 비효율적이다.

본 연구에서는 모델의 출력 시퀀스에서 가장 모호성이 높은 자소들을 차 순위의 확률 값을 가지는 자소로 치환하여 한글 표기 후보들을 생성하였다.

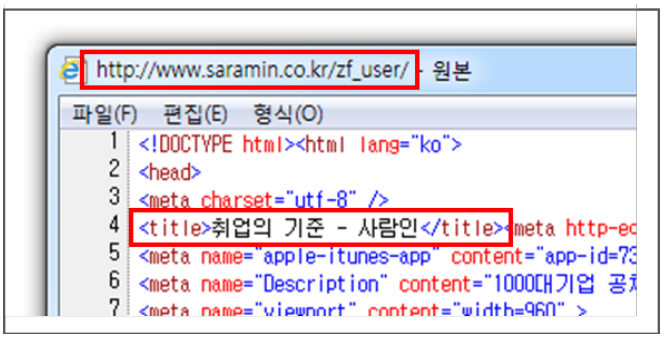
3.2 학습 데이터 구축

이 절에서는 로마자-한글 상호 표기 변환을 위해 Seq2seq 모델의 학습 데이터를 구축하는 방법을 소개한다. 학습 데이터는 <표 1>과 같이 로마자-한글의 쌍으로 구성된다.

<표 1> 학습 데이터의 구성

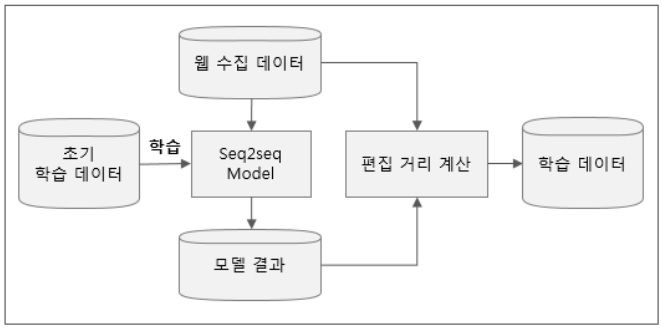
로마자	한글
LG	엘지
SK C&C	에스케이 씨앤씨
NHN Entertainment	엔에이치엔 엔터테인먼트
…	…

본 연구에서는 Seq2seq 모델의 학습 데이터를 구축하기 위해 임의의 웹 사이트의 URL과 TITLE 태그 정보를 이용하였다. URL과 TITLE 태그는 각각 로마자 상호와 한글 상호에 대응하는 경우가 많기 때문이다. 수집한 URL과 TITLE 태그의 예는 <그림 2>와 같다.



<그림 2> URL과 TITLE 태그의 예

학습 데이터를 구축하기 위한 시스템 구성은 <그림 3>과 같다. <그림 3>의 웹 수집 데이터는 URL과 TITLE 태그에서 불필요한 정보를 제거한 로마자-한글 쌍으로 구성된다. 학습 데이터 구축을 위해 우선 초기 학습 데이터를 이용하여 Seq2seq 모델을 학습시킨다. 그다음 웹에서 수집한 로마자-한글 쌍의 데이터 중 로마자 데이터를 학습된 모델의 입력으로 제공하여 한글로 표기 변환된 결과를 얻는다. 마지막으로 모델의 입력에 해당하는 한글 데이터와 모델 결과 사이의 편집 거리(Edit distance) 알고리즘을 계산하여 양질의 학습 데이터만 선별하였다. 편집 거리란 두 개의 문자열이 같아지기 위해 이루어져야 하는 삽입, 삭제, 치환의 최소 연산 개수를 말한다.



<그림 3> 학습 데이터 구축 방법

초기 학습 데이터로는 국립국어원의 외래어 용례와 위키 낱말 사전 등을 이용하였다. 편집 거리는 문자열의 길이로 나누어, 그 값이 0.3 이상인 로마자-한글 쌍들만 학습 데이터로 사용하였다.

3.2.1 한글의 자소 분리

<표 2>는 자소로 분리한 학습 데이터의 예이다.

<표 2> 자소 분리 학습 데이터의 구성

로마자	한글 자소
LG	ㅇㅔㄹㅈㅣ
SK C&C	ㅇㅔㅅㅡㅋㅔㅇㅣ ㅆㅣㅇㅐㄴㅆㅣ
…	…

한글은 19개의 초성과 21개의 중성, 27개의 종성으로 조합이 가능하다. 즉 11,172개의 음절 조합이 가능하며 이는 Vocabulary Size를 증가시켜 학습 데이터에 낮은 빈도로 출현하는 음절에 대한 정확도를 감소시킨다. 따라서, 본 연구에서는 한글을 초성, 중성, 종성의 자소로 분리하여 모델의 학습 데이터를 구축하였다.

4. 실험

4.1 실험 환경

실험 환경은 텐서플로우의 seq2seq 라이브러리를 이용하여 구현하였으며 사용된 데이터 셋은 <표 3>와 같다. 본 연구에서는 금융감독원의 DART 기업명 데이터의 5%인 939쌍을 각각 검증 및 평가 데이터로 사용하였다. 데이터 중 한글 상호와 영문 상호가 다르거나 번역에 해당하는 상호는 데이터 셋에서 제외하였다.

<표 3> 실험 데이터 구성

데이터 셋	갯수
외래어 용례	31898
위키 낱말 사전	783
DART 기업명 데이터	18771
구축한 학습 데이터	74024

Seq2seq 모델은 3-layer, 256 size의 LSTM으로 인코더, 디코더를 구성하였다. drop out은 0.5의 확률로 모든 레이어에 동일하게 적용하였으며 batch size는 64로 학습하였다.

모델의 성능 평가에는 단어 정확도와 자소 정확도를 사용하며, 식은 다음과 같다.

$$단어\ 정확도 = \frac{정답\ 단어\ 수}{전체\ 단어\ 수}$$

$$자소\ 정확도 = \frac{L-(i+d+s)}{L}$$

여기서 L은 원 자소 문자열의 길이를 나타내며, i,d,s는 각각 원 자소 문자열에서 목표 자소 문자열로 변환하기 위해 필요한 삽입, 삭제, 치환의 개수를 나타낸다. 만약 L < (i+d+s)이면 자소 정확도는 0으로 판단한다.[4]

4.2 학습 데이터에 따른 성능 실험

<표 4>은 학습 데이터에 따른 성능 평가 결과이다.

<표 4> 학습 데이터에 따른 실험 결과

데이터 셋	단어 정확도	자소 정확도
외래어 용례 + 위키 낱말 사전	23.85	77.72
DART 기업명 데이터	49.62	87.10
구축한 학습 데이터	53.67	88.46
DART 기업명 데이터 + 구축한 학습 데이터	**57.82**	**90.73**

실험 결과 국립국어원의 외래어 용례나 위키 낱말 사전 데이터의 경우 인명이나 화합물과 같이 일반적으로 상호에 사용하지 않는 데이터들이 많아 성능이 낮은 경향을 보였다. 하지만 본 연구에서 제안한 방법으로 구축한 학습 데이터는 모델의 성능을 향상시키는 것을 볼 수 있다.

4.3 자소 분리에 따른 성능 실험

본 연구에서는 한글을 초성, 중성, 종성의 자소로 분리하여 모델을 학습하였다. <표 5>는 자소 분리에 따른 모델의 성능 평가 결과이다. 모델의 Vocabulary Size는 로마자 50, 한글 자소 60으로 설정하였다.

<표 5> 학습 데이터의 자소 분리 실험 결과

	단어 정확도	자소 정확도
음절	48.56	82.88
자소	**57.82**	**90.73**

실험 결과 한글을 자소 단위로 분리하여 학습한 모델이 음절 단위로 학습한 모델보다 높은 정확도를 보였다. 한글을 자소로 분리함으로써 모델이 예측해야 하는 시퀀스의 길이는 증가하였지만, 빈번하게 출현하지 않는 음절에 대해서도 학습이 가능하기 때문에 모델의 성능을 향상 시킨 것으로 보인다.

<표 6> 자소 분리에 따른 표기 변환 결과

정답		음절 기반	자소 기반
Chulgab	철갑	철비비비	철겁
Bukak …	부각 …	부막 …	부각 …
… Mogul …	… 모굴 …	… 모걸스 …	… 모굴 …
Hawkeyes …	호크아이즈…	홍아이즈…	호크아이즈…
nskorea	엔에스코리아	코리아아아	엔에스코리아

<표 6>의 결과에서 보듯이 음절 단위로 학습한 모델의 경우, ‘…gab’, ‘…kak’, ‘…gul’, ‘hawk…’, ‘ns…’ 등 출현 빈도가 낮은 시퀀스는 제대로 예측하지 못하는 경향을 보였다. 하지만 자소 단위로 학습한 모델의 경우, 대부분 정답과 유사한 음절을 예측하였다. 따라서 로마자-한글 상호 표기 변환의 경우 학습 데이터를 자소로 분리하는 것이 효과적임을 알 수 있다.

4.4 Attention Mechanism 적용 실험

<표 7>은 Attention Mechanism 적용 유무에 따른 성능 평가 결과이다.

<표 7> Attention Mechanism 적용 실험 결과

	단어 정확도	자소 정확도
LSTM Cell	43.45	84.29
LSTM Cell + Attention Mechanism	**57.82**	**90.73**

실험 결과 Attention Mechanism은 모델의 성능을 향상시켰고, 이와 같은 결과는 로마자 표기와 한글 자소 표기 사이에는 명확하게 대응되는 일정한 정렬(alignment)이 존재하기 때문인 것으로 보인다.

4.5 한글 표기의 후보 생성 실험

본 실험에서는 3.1.1 절의 방법으로 생성한 한글 상호 표기의 후보들에 대한 성능을 평가한다. Top-N 단어 정확도는 N개의 후보들 중에 정답이 있을 확률이며, Top-N 자소 정확도는 N개의 후보들의 평균 자소 정확도이다. 실험 결과는 각각 <표 8>과 같다.

<표 8> 한글 표기 후보의 개수에 따른 실험 결과

후보 개수	Top-N 단어 정확도	Top-N 자소 정확도
1	57.82	90.73
2	65.06	87.36
3	67.73	85.22
5	**71.56**	**84.26**
10	72.31	82.54

실험 결과 한글 표기의 후보 개수를 늘릴수록 Top-N 단어 정확도는 증가하지만 Top-N 자소 정확도는 감소하는 결과를 보였다. 후보 개수를 5개 이상 늘릴 때부터 Top-N 단어 정확도의 증가치가 크게 감소하였다. 따라서 로마자 상호에 대한 5개의 한글 상호 표기 후보를 생성하는 것이 효율적이라 볼 수 있다.

5. 결론

본 연구에서는 Sequence-to-sequence 모델을 이용하여 로마자 상호 표기에 대한 한글 상호 표기와 그 후보들을 생성하는 시스템을 소개하였다. 또한 로마자-한글 상호 표기 변환을 위한 양질의 학습 데이터를 구축하는 방법을 제시하였다. 그리고 자소 분리 방법과 Attention Mechanism의 적용을 통해 모델의 성능을 향상시켰다.

하지만 본 연구의 실험 대상에서 번역에 해당하는 상호는 제외하였기 때문에, 이와 같은 경우는 상호의 표기 변환을 지원하지 않는다는 문제점이 남아있다.

향후 본 연구의 시스템을 질의 확장이나 동의어 사전 구축에 이용한다면 단어 불일치 문제의 해결에 기여할 수 있을 것이라 기대한다.

참고문헌

[1] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems, 2014.

[2] 이재성, 최기선, "정보검색을 위한 외래어 자동표기 모델", 한국정보과학회 제4회 학술대회 논문집, pp. 17-24, 1997.

[3] 김태일, "최대 엔트로피 모델을 이용한 다국어 정보검색에서의 영-한 음차 표기 모델", 서강대학교 석사학위 논문, 1999.

[4] 오종훈,배선미,최기선, "글자 및 발음 기반 영-한 음차표기 모델", 한국정보과학회 봄 학술발표논문집, 제31권, 제1호, pp. 925-927, 2004.

[5] 황명진,조선호,권혁철, "한글 상호(商號)를 로마자로 변환하기 위한 고속 부분문자열 분석 알고리즘", 한국정보처리학회 2008년 추계 학술대회 논문집, 제15권, 제2호, pp. 0168-0170, 2008.

[6] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," arXiv preprint arXiv:1406.1078, 2014.

[7] 이건일,이의현,이종혁, "Sequence-to-sequence 기반 한국어 형태소 분석 및 품사 태깅", 정보과학회논문지, 제44권, 제1호, pp. 57-62, 2017.

[8] 정의석,박전규, "seq2seq 주의집중 모델을 이용한 형태소 분석 및 품사 태깅", 제28회 한글 및 한국어 정보처리 학술대회 논문집, pp. 217-219, 2016.

[9] 황현선,이창기, "Sequence-to-sequence 모델을 이용한 한국어 구구조 구문 분석", 제28회 한글 및 한국어 정보처리 학술대회 논문집, pp. 20-24, 2016.

[10] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translation." arXiv preprint arXiv:1409.0473, 2014.

[11] 이현구, 김학수, "주의집중 및 복사 작용을 가진 Sequence-to-Sequence 순환신경망을 이용한 제목 생성 모델", 한국정보과학회논문지, 제44권, 제7호, pp. 674-679, 2017.