Distance LSTM-CNN with Layer Normalization을 이용한 음차 표기 대역 쌍 판별

이창수○, 천주룡, 김주근, 김태일, 강인호

네이버 검색

{ changsu.lee, juryong.cheon, joogeun.kim, eiji.kim, once.ihkang } @ navercorp.com

Verification of Transliteration Pairs Using Distance LSTM-CNN with Layer Normalization

Changsu Lee○, Juryong Cheon, Joogeun Kim, Taeil Kim, Inho Kang

Naver Corporation

요 약

외국어로 구성된 용어를 발음에 기반하여 자국의 언어로 표기하는 것을 음차 표기라 한다. 국가 간의 경계가 허물어짐에 따라, 외국어에 기원을 두는 용어를 설명하기 위해 뉴스 등 다양한 웹 문서에서는 동일한 발음을 가지는 외국어 표기와 한국어 표기를 혼용하여 사용하고 있다. 이에 좋은 검색 결과를 가져오기 위해서는 외국어 표기와 더불어 사람들이 많이 사용하는 다양한 음차 표기를 함께 검색에 활용하는 것이 중요하다. 음차 표기 모델과 음차 표기 대역 쌍 추출을 통해 음차 표현을 생성하는 기존 방법 대신, 본 논문에서는 신뢰할 수 있는 다양한 음차 표현을 찾기 위해 문서에서 음차 표기 후보를 찾고, 이 음차 표기 후보가 정확한 표기인지 판별하는 방식을 제안한다. 다양한 딥러닝 모델을 비교, 검토하여 최종적으로 음차 표기 대역 쌍 판별에 특화된 모델인 Distance LSTM-CNN 모델을 제안하며, 제안하는 모델의 Batch Size 영향을 줄이고 학습 시 수렴 속도 개선을 위해 Layer Normalization을 적용하는 방법을 보인다.

주제어: 음차판별, 음차검증, 정보검색, 딥러닝

1. 서론

외국어로 구성된 용어를 발음에 기반하여 자국의 언어로 표기하는 것을 음차 표기라 정의한다[1][2]. 국가 간의 경계가 허물어짐에 따라 외국어에 기원을 두는 용어(starbucks)를 설명하기 위해 뉴스, 블로그 등의 다양한 웹 문서에서는 외국어 표기와 한국어 표기를 혼용하여 사용하는 경우가 나날이 증가하고 있다. 특히 정보 검색에서는 외국어 표기(starbucks) 하나만을 사용하여 검색을 수행하면 한국어 표기(스타벅스)만으로 문서화되어 있는 양질의 문서들이 검색 결과에서 제외되어 원하는 검색 결과를 얻을 수 없는 문제가 발생한다. 이러한 문제를 해결하기 위해서는 외국어 표기와 더불어 사람들이 많이 사용하는 다양한 음차 표기(한국어 표기 등)를 함께 검색에 활용하는 것이 중요하다.

주어진 외국 용어 및 외국어 표기에 대한 다양한 음차 표기를 얻기 위한 연구로는 음차 표기 모델(Transliteration Model), 음차 표기 대역 쌍 추출(Transliteration Pairs Extraction), 음차 표기 대역 쌍 판별(Transliteration Pairs Verification) 등의 연구가 있었으며, 이 연구들은 주로 검색 품질 향상을 위해 질의 확장 및 언어 자원을 구축하는 목적으로 연구되었다[3].

음차 표기 모델(Transliteration Model)은 외국어 표기를 입력으로 하여 자국어의 표기를 자동으로 생성하는 방법이다. 주로 검색 품질 향상을 위해 번역사전에 존재하지 않는 외국 단어의 음차 표기를 자동으로 생성하기 위한 연구로 진행되었으며, 최근에는 딥러닝을 이용한 음차 표기 모델이 가장 좋은 결과를 보였다[1][4][5][6]. 하지만, 음차 표기 모델을 통해 생성된 음차 표현은 일반적으로 품질이 좋지 않으며, 검색 시스템에서 사용자들이 자주 사용하는 음차 표기가 아닌 경우가 많아, 상용 검색 시스템에 적용하기에는 문제가 있었다.

음차 표기 생성 방법과 달리 문서에서 음차 표기를 추출하기 위한 연구로, 음차 표기 대역 쌍 추출(Transliteration Pairs Extraction) 연구가 있었다[7]. 음차 표기 대역 쌍 추출은 이중 언어 문서에서 외국어와 외국어에 대응되는 음차 표기된 용어들을 자동으로 추출하기 위한 방법이다[3]. 주로 패턴 및 기계학습 등을 이용하여 음차 표기 대역 쌍 후보를 추출하고, 서로 다른 언어로 구성된 음차 표기 대역 쌍 후보를 하나의 언어 형태로 변환한 후, 편집거리 알고리즘을 이용하여 음성적 유사도를 계산하는 방법으로 음차 표기 대역 쌍을 추출했다[3][7]. 하지만 음성적 유사도를 적용하여 음차 대역 쌍을 추출하는 방법은 “starbucks - 스타박스”와 같이 음성적으로 유사하지만, 다른 의미를 가지는 용어(스타박스)가 동일한 음차로 추출되는 문제가 있었으며, 이처럼 잘못 추출된 음차 정보를 검색 시스템에 적용하게 되면 검색 품질에 심각한 문제를 야기하게 된다.

정보 검색에 활용하기 적합하며, 신뢰할만한 음차 표기 대역 쌍을 얻기 위한 연구로는 음차 표기 대역 쌍 판별(Transliteration Pairs Verification) 연구가 있었다. 이는 서로 다른 언어로 구성된 음차 표기 대역 쌍 후보

가 정확한 음차 관계인지 판별하는 것이며, 특히 인명 등과 같이 난도가 높은 음차 표기 대역 쌍(James Rodriguez - 하메스 로드리게스)을 정확하게 판별하기 위해 제안되었다[8][9].

본 논문에서는 검색 품질 향상을 위해, 웹 문서에서 자주 사용되는 다양한 음차 표현을 찾고 이를 언어 자원으로 구축하기 위한 음차 표기 대역 쌍 판별 모델을 제안한다. 기존에 제안되었던 음차 표기 모델과 음차 표기 대역 쌍 추출 모델은 저품질 문제와 더불어 검색 시스템에 적합하지 않는 음차 표기를 생성, 추출하는 문제가 있어 활용하기 어려우며, 두 음차 표기가 정확한 음차 관계인지 판별하는 음차 표기 대역 쌍 판별 연구에 기초하여 문제를 해결한다.

따라서, 본 연구에서는 기존의 음차 표기 모델과 음차 표기 대역 쌍 추출 모델에서 특히 문제가 되는 난도가 높은 음차 표기 후보를 정확히 판별하기 위한 음차 표기 대역 쌍 판별 모델을 구축하는 것을 목적으로, 다양한 딥러닝 모델을 구축하고 딥러닝 모델 간의 비교, 검토를 통하여 최종적으로 음차 표기 대역 쌍 판별에 특화된 모델인 Distance LSTM-CNN with Layer Normalization 모델을 제안한다.

평가를 위해 정보 검색에서 사용자들이 자주 사용하는 음차 표기 후보를 웹 문서에서 수집하되, "AOA - 초아"와 같이 판별 난도가 낮은 음차 표기 대역 쌍 후보는 판별 모델의 차별성을 증명할 수 없으므로 제외하고, 주로 난도가 높은 음차 표기 후보들을 추출하여 데이터 셋을 구축함으로써, 제안하는 방법의 실용성을 검증한다.

본 논문에서 제안하는 음차 표기 대역 쌍 판별에 특화된 모델인 Distance LSTM-CNN with Layer Normalization은 한국어-영어간 음차판별을 위해 변형된 KODEX 방법과의 비교 결과, 약 35%의 품질 향상을 보였으며 [4]에서 높은 품질을 보이는 음차 생성 모델을 음차 판별 모델로 변형한 Sequence-to-Sequence with Attention 모델과의 비교에서도 약 3.5%의 품질 향상을 보였다. 또한, 두 질의의 관련성을 판별하는 연구에서 높은 품질을 보이는 딥러닝 모델인 Distance LSTM 모델보다 약 3% 정도의 품질 향상을 보여 최종적으로 89.70%의 품질을 보였다. 마지막으로 Layer Normalization을 적용한 모델이 적용하지 않은 모델과 비교해 약간의 품질 향상과 더불어 수렴 속도가 약 3배 빨라짐으로써 Layer Normalization 효과를 확인할 수 있었다.

본 논문은 다음과 같이 구성되어 있다. 1장의 서론에 이어 2장에서는 관련 연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 음차 표기 대역 쌍 판별 모델에 대해 설명한다. 4장에서는 실험 결과를 비교, 검토하며 마지막 5장에서는 결론에 대해 살펴 본다.

2. 관련 연구

2.1 음차 표기 모델(Transliteration Model)

음차 표기 모델은 외국어 표기를 입력으로 자국어 표기를 생성하는 연구로써, 주로 확률 및 기계학습을 이용하여 연구되어왔다[1][4][5][6]. [1]에서는 확률 기반 음차 표기 모델을 제안했으며, 발음 단위를 음소에 매핑할 수 있도록 매핑 테이블을 정의하고, 확률 모델을 이용하여 주어진 영어 단어에 대한 가장 높은 확률을 가진 한국어 음차 표기를 생성하였다. [5]는 최대 엔트로피를 이용하여 음차 표기를 생성하는 방법을 제안하였으며, [6]은 음성적 정보와 자소/음소의 문맥정보를 이용, 결정 트리 및 메모리 기반 학습 모델에 활용하여 한국어 음차 표기를 생성하는 방법을 제안했다. 최근 [4]에서는 기계번역에 주로 사용되는 Sequence-to-Sequence 모델을 활용하여 영어를 기반으로 다양한 언어의 음차 표기를 생성하는 모델을 제안하였다.

2.2 음차 표기 대역 쌍 추출(Transliteration Pairs Extraction)

음차 표기 대역 쌍 추출은 이중 언어 문서에서 외국어와 그에 대응되는 음차 표기된 용어를 자동으로 추출하는 방법이다[3][7]. 주로 패턴 및 기계학습을 활용하여 음차 표기 대역 쌍 후보를 추출하고, 추출된 외국어 및 자국어 표기가 정확한 음차 관계인지 여부를 계산하여 음차 표기 대역 쌍을 추출하였다. [3]은 패턴을 이용하여 이중 언어 문서에서 음차 표기 대역 쌍 후보를 추출하고, 음차 표기 모델(Transliteration Model)을 활용하여 외국어를 입력으로 한국어 음차 표기를 생성한 후, 생성된 음차 표기와 대역 쌍 후보에서 추출된 한국어 음차 표기와의 음성적 유사도를 계산하여 음차 표기 대역 쌍을 추출하는 방법을 제안하였다. 그리고, [7]은 영어-일본어 음차 표기 대역 쌍 추출을 위해 잡음 채널 오류 모델과 학습 가능한 편집거리 함수를 이용하여 음차 표기 대역 쌍을 추출하였다.

2.3 음차 표기 대역 쌍 판별(Transliteration Pairs Verification)

음차 표기 대역 쌍 판별은 외국어 표기와 자국어 음차 표기로 구성된 음차 표기 대역 쌍 후보가 정확한 음차 관계인지 판별하는 것이며, 주로 기계번역과 교차언어 정보 검색과 같이 다국어 자연어 처리 작업에 활용되었다[9]. [8]에서는 동일한 언어 간 다양한 외국어 음차 표기 대역 쌍(디지털-디지탈) 비교를 위해 두 입력을 특정 기호로 인코딩하여 비교하는 KODEX 알고리즘을 제안하여, 한국어간 다양하게 표현되는 음차가 동일한지 판별하였다. [9]는 음차 판별이 어려운 인명의 음차 표기 대역 쌍을 판별하기 위해 Discrete Variant Hidden Markov Model (HMM) Alignment 기법을 제안하였으며, 국가간 발음이 달라 음차 판별 난도가 높은 인명 음차 표기 대역 쌍 판별에서 좋은 결과를 얻었다.

3. 음차 표기 대역 쌍 판별 모델

이 장에서는 정보 검색에 적합한 음차 표기 언어 자원을 구축하는 것을 목적으로, 웹 문서에서 추출되는 다양한 음차 표기 대역 쌍 후보가 정확한 음차 관계인지 판별하는 딥러닝 기반 음차 표기 대역 쌍 판별 모델을 구축하는 방법을 보인다.

3.1 Sequence-to-Sequence with Attention 모델

Sequence-to-Sequence 모델은 입력 문장을 $x = x_1, \ldots, x_T$ 로 인코딩 한 후, 디코더를 통해 $P(y|x)$ 를 최대

화하는 출력 문장 $y = y_1, \ldots, y_k$ 을 생성하는 모델이며, 주로 기계 번역 및 챗봇에 적용되어 만족할 만한 결과를 보였다[10]. [4]에서는 영어 표기를 기반으로 다양한 언어(아랍어, 일본어, 중국어 등)의 음차 표기를 생성하기 위해 Sequence-to-Sequence 모델을 적용하여 성공적인 결과를 얻었다. 이와 같은 연구에 기초하여 본 연구에서는 음차 표기 대역 쌍 판별을 위한 Sequence-to-Sequence 모델을 구축하였다. 그리고 LSTM Cell을 추가하여 RNN 내부에 3개의 게이트(Input, Output, Forget)와 1개의 메모리 공간으로 정보를 갱신 혹은 제거를 통해 멀리 있는 정보가 희미해지는 RNN의 그래디언트 소멸 문제(Gradient Vanishing Problem)를 해결하도록 하였다[11]. 또한 단계별 중요도가 고려 될 수 있도록 주의(Attention) 기법을 이용하여 모델을 구성했다.
[그림 1]은 음차 표기 대역 쌍 판별을 위해 구축한 Sequence-to-Sequence with Attention 모델의 구성도 이다.

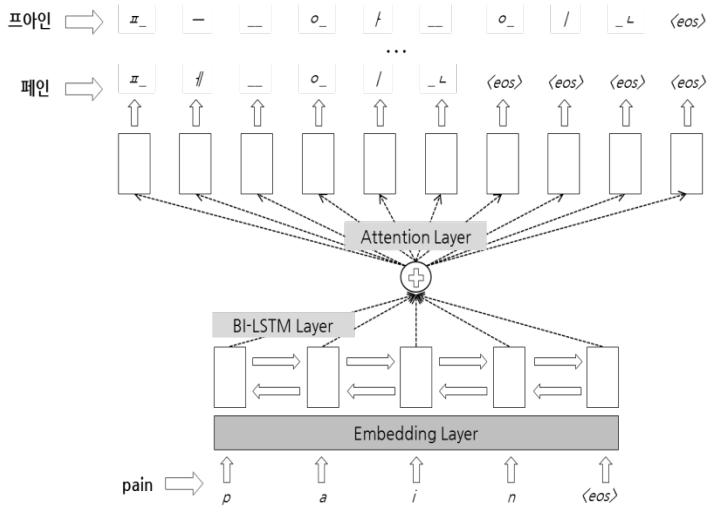

그림 1. Sequence-to-Sequence with Attention 모델

본 연구에서의 목적은 웹 문서에서 추출된 다양한 음차 표기 대역 쌍 후보가 적합한 음차 관계인지 여부를 판별하는 것이므로, 상위 1개의 결과만 추출되는 기존 디코더를 상위 N개의 음차 표기 생성 결과를 추출할 수 있도록 변경했다. 평가를 위해 정답이 부착되어있는 음차 표기 대역 쌍 후보와 모델에서 추출한 음차 표기 추출 결과 상위 N개 중 동일한 음차 표기가 존재하는지 확인하는 방법으로 음차 표기 대역 쌍 판별 모델을 구축했다.
[그림 2]는 음차 표기 대역 쌍 후보를 판별하는 예이다.

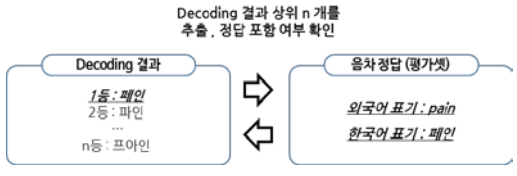

그림 2. 디코더를 변경한 음차 표기 대역 쌍 판별 방법

3.2 Distance LSTM 모델

정보 검색에 적합한 음차 표기 언어 자원을 구축하기 위해서는 음차 표기 대역 쌍 후보가 형태적, 의미적으로 동일한지 여부를 판별해야 한다. 이와 관련된 최신 연구들은, 두 문장의 관련성을 판별하기 위해 구축된 SNLI[12] 및 Quora Question Pairs[13] 데이터를 이용한 연구들이며, 주로 질의-응답 및 동의질의 판별 문제를 해결하기 위해 제안되었다[14][15][16][17]. [15][16][17]은 두 질의가 동일한 의미를 지니는지 판별하기 위해 다양한 정렬(Alignment) 및 주의(Attention) 기법을 적용했다. 하지만, 음차 표기의 경우 음차간 동일한 시퀀스로 매칭되어야 하는 특성이 있으며, 정렬 및 주의 기법을 적용하게 되면 "one way - 웨이 원"과 같이 동일한 단어로 구성되지만 다른 시퀀스를 가진 음차 표기를 동일하게 판별하는 문제가 있었다. 동일한 시퀀스를 유지해야 하는 음차 표기 특성을 반영하여 본 논문에서는 [14]의 Distance LSTM 모델을 이용하여 음차 표기 대역 쌍 판별 모델을 구축했다. 그리고 거리벡터(Distance vector) 연산에서 빼기(Subtract) 및 곱하기(Multiply) 연산이 가장 좋은 품질을 보인다는 [16]의 연구를 반영하여, Distance LSTM 모델의 품질 개선을 위해 거리벡터 연산(Distance vector)에 빼기(Subtract) 및 곱하기(Multiply) 연산이 적용되도록 변경하였다. 결과적으로 다른 거리 벡터 연산과 비교해 가장 좋은 결과를 얻을 수 있었다.
[그림 3]은 음차 표기 대역 쌍 판별을 위해 구축한 개선된 Distance LSTM 모델의 구성도이다.

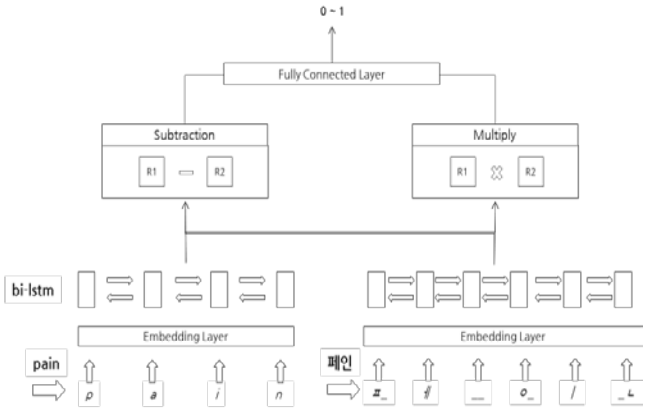

그림 3. Distance LSTM 모델

3.3 Distance LSTM-CNN with Layer Normalization 모델

Convolutional Neural Networks(CNN) 모델은 입력 문장의 주위 문맥 정보를 Convolutional Layer을 통해 추출하고, Pooling Layer를 거쳐 중요 자질만을 추상화하는 특징을 가지는 구조이다. [18]은 Convolutional Layer와 Max Pooling Layer로 구성된 단순한 구조의 CNN 모델을 제안하여 문장 및 문서 분류에서 좋은 결과를 얻었다.

음차 표기의 경우, 두 입력이 동일한 시퀀스로 매칭되어야 하는 특성이 있으므로, 순서를 유지하면서 주위 문맥(음차)의 중요 정보를 추상화하는 CNN 구조를 추가하는 것은 음차 표기 대역 쌍 판별을 위한 중요한 자질이 될 수 있으며, 음차 간 긍정적인 연관 관계를 부여할 수 있다고 가정했다. 이와 같은 가정을 기반으로 음차 판별에 적합한 CNN 구조를 추가, 적용했다.

음차 판별에 적합한 중요 주위 문맥 정보를 추가하기 위한 CNN 구조는 아래처럼 정의하여 적용하였다.

$$C_h = \text{MaxPooling}\left(\text{Conv1d}_{fsize}(\text{x}_1, \dots \text{x}_{n-h+1:n})\right)$$

x는 입력이며, n은 음차 길이, h는 현재 음차를 기준으로 몇 개의 인접한 음차를 조합을 생성할 것인지의 개수(윈도우 크기), Conv1d는 1차원 합성곱, fsize는 필터의 수, 마지막으로 MaxPooling은 벡터를 추상화하기 위해 최대 Pooling을 수행하는 함수이다.

정의된 CNN 구조를 이용, h(윈도우 크기)를 1, 2, 3으로 설정한 후, 거리벡터(Distance vector) 연산인 빼기(Subtract) 및 곱하기(Multiply) 벡터 연산에서 출력된 벡터 각각에 대해 [그림 4]와 같이 CNN 구조를 추가했다. 정의한 CNN 구조를 추가함으로써 인접한 음차 조합 자질을 생성하고, 중요 음차 자질만을 추상화하는 방법으로 음차 판별에 적합한 중요 인접 음차 정보가 모델에 반영되도록 하였다.

[그림 4]는 Distance LSTM 모델에 CNN 구조를 추가한 Distance LSTM-CNN 모델의 구성도이다. 또한, Batch Size에 의존하지 않으며, 학습 수렴 속도를 빠르게 하고, 일부 연구에서는 품질 향상이 있는 계층 정규화(Layer Normalization) 기법을 LSTM에 추가적으로 적용했다[19].

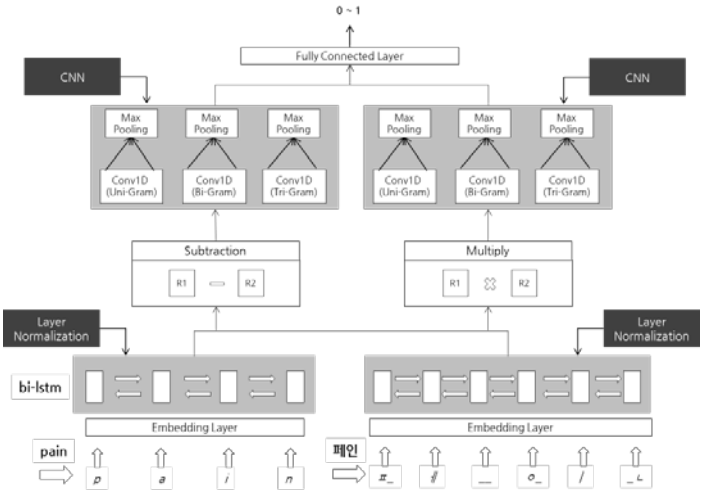

그림 4. Distance LSTM-CNN with Layer Normalization 모델

4. 실험

4.1 실험 환경

모든 모델은 공정한 비교를 위해 동일하게 파라미터를 설정했다. 일반적으로 딥러닝은 많은 파라미터를 가지고 있으며, 파라미터 값의 설정에 따라 조금씩 품질 차이가 나므로, 가장 좋은 품질을 보이는 파라미터를 실험을 통해 찾아 설정하였다.

입력 층은 외국어는 알파벳 단위, 한국어는 자소 단위로 입력을 수행했다. 입력 층은 서로 다른 언어로 이루어져 있으며 음차 표기를 위한 교차언어 데이터는 존재하지 않으므로 임베딩 벡터(Embedding vector)에 대한 전처리(Pre-Training)는 따로 수행하지 않고 학습 데이터에 의해 임베딩 벡터가 학습되도록 구성했다. 임베딩 벡터 차원과 은닉 계층 차원은 128 차원을 사용, 과적합 방지를 위한 Dropout 비율은 0.2, 그리고 활성(Activation) 함수는 마지막 계층에서만 Sigmoid를 사용했으며, 나머지 계층에서는 Relu를 사용했다. 또한, Fully Connected Layer는 1개의 계층만 쌓아 실험을 진행했다.

추가적으로, Distance LSTM-CNN에서 CNN을 위한 필터(Filter)차원은 128차원으로 설정했으며, Dropout 비율은 0.5를 사용했다. Sequence-to-Sequence with Attention 모델은 가장 좋은 품질을 보인 버킷 개수 1개로 고정했으며, 상위 N개는 30개를 추출했다.

마지막으로, 딥러닝 모델들은 오류 역전파 알고리즘에 의해 학습되는데 본 논문에서는 Adam 기법을 이용하여 파라미터를 최적화했다.

4.2 학습 및 평가 데이터

웹 문서에서 자주 사용되는 음차 표기 후보를 추출하기 위해 웹 문서 및 사용자 검색 질의를 이용, “AOA(초아)”, “memento(메멘터)” 등의 패턴을 적용하여 실험을 위한 음차 표기 데이터를 추출했다. 추출된 음차 후보들은 다양한 난도를 가질 수 있으며, “AOA - 초아” 같은 형태는 쉽게 음차 판별이 가능하므로 최대한 반영하지 않고 “memento - 메멘터”와 같이 모호하고 어려운 형태만 추출할 수 있도록 변형된 KODEX 알고리즘을 이용, 다음과 같은 절차를 거쳐 데이터를 추출했다.

1. 웹 문서에서 자주 사용되는 음차 표기 패턴(wikipedia(위키피디아), (concern(관심))을 찾아 음차 표기 대역 쌍 후보를 추출
2. 영어-한국어간 음차 판별을 위해 변형된 KODEX 알고리즘을 이용하여, 음차 표기 대역 쌍 후보간 유사도를 계산하고, 편집 거리가 0, 1인 난도가 높은 음차 데이터를 추출(예시 : concern - 관심, memento - 에멘토 등)
3. 3명의 검수자가 추출된 음차 표기 대역 쌍 후보를 검수, 정답과 오답을 판별하여 검수 데이터 구축

실험에 사용된 검수 데이터는 음차 15,093개, 비음차 15,093개이며 학습, 개발, 평가 셋을 각각 8:1:1 비율로 나누어 평가 셋을 통해 품질을 측정하였다.

실험에 사용된 평가 셋에 포함된 데이터 예시는 [표 1]과 같다.

표 1. 웹 문서로부터 추출된 난도가 높은 음차 후보

외국어	한국어	음차여부
ibm	아이비엠	1
bas	버스	0
glam	그램	0
you light up my life	유라이트업마이라이프	1
lcd tv	엘씨디티비	1
civil war	씨빌워	1

[표 1]을 보면 일반단어, 개체명, 약어성, 복합명사 등 난도가 비교적 높은 다양한 형태의 음차 표기 대역쌍이 추출된 것을 확인 할 수 있다.

4.3 품질 비교

음차 표기 대역 쌍 판별 모델의 품질 평가를 위해 변형된 KODEX 및 Sequence-to-Sequence with Attention 모델, 그리고 Distance LSTM 모델 및 Distance LSTM 모델에 CNN 구조를 추가한 Distance LSTM-CNN(+LN) 모델과의 결과를 비교하였다. 평가 척도는 정확률(Precision), 재현율(Recall)을 사용했으며, Distance LSTM 및 Distance LSTM-CNN(+LN) 결과는 0~1 사이의 확률 값으로 나오게 되며 0.5 이상의 값이 나올 경우 음차로 판별했다. 용도에 따라 임계 값을 변경하는 것에 의해 정확률과 재현율을 조절하여 결과를 유연하게 추출할 수 있다.

[표 2]는 음차 표기 대역 쌍 판별 모델의 평가 결과를 보여준다.

표 2. 모델별 음차 표기 대역 쌍 판별 결과

모델	정확률	재현율	F1 점수
변형된 KODEX	57.02	55.69	53.67
Seq2Seq with Attention	86.18	86.19	86.18
Distance LSTM	87.30	86.87	86.89
Distance LSTM-CNN	89.75	89.52	89.55
Distance LSTM-CNN(+LN)	89.69	89.70	89.70

[표 2]에서 볼 수 있듯이, 변형된 KODEX를 이용하여 편집거리가 작은 데이터를 추출하였기 때문에 난도가 높은 데이터가 대부분이며, 이러한 이유로 인해 변형된 KODEX의 품질이 53.67%로 비교적 낮은 것을 확인할 수 있다. 또한, 다국어 음차 생성에서 높은 품질을 보이는 음차 생성 모델을 음차 판별 모델로 변형한 Sequence-to-Sequence with Attention 모델은 86.18%의 품질을 보였으며, 두 질의의 관련성을 판별하는 연구에서 좋은 품질을 보이는 Distance LSTM 모델을 개선한 모델이 86.89%로 더 높은 품질을 보이는 것을 확인할 수 있었다. 그리고 음차의 시퀀스적인 특성을 반영하는 CNN 구조를 추가한, 음차 표기 대역 쌍 판별에 특화된 Distance LSTM-CNN 모델은 89.55%로 Distance LSTM 모델과 비교해 약 3%의 품질 향상을 보였다. 마지막으로 계층 정규화(Layer Normalization)를 추가한 Distance LSTM-CNN(+LN)모델과 Distance LSTM-CNN 모델과의 비교 결과 품질이 미미하게 향상(89.70%) 되었으며, epoch가 60 -> 21로 약 40번의 epoch 가 절약됨으로써 학습 시 수렴 속도를 개선하는 계층 정규화(Layer Normalization)의 효과를 확인할 수 있었다.

4.4 결과 검토

변형된 KODEX와 같이 음성적 유사도의 편집거리를 이용하는 기존 방법은 대부분 자음만을 이용하여 음차 표기 대역 쌍 여부를 판별한다. 예를 들어 “korea - 고려(X)”의 경우 ‘k’는 ‘ㄱ’과 매칭되며, ‘o’는 모음이므로 무시, ‘r’과 ‘ㄹ’이 매칭되고, ‘e’, ‘a’는 모음이므로 무시하여 최종적으로 음차 표기가 동일하다고 판단하는 문제가 있었으며, 이외에도 ‘ibm - 아이비엠(0)’ 등 약어성 단어를 판별하지 못하는 문제가 있었다.

반면, 딥러닝을 이용한 방법들은 모음 및 시퀀스 정보를 추가로 반영하기에 “youtube - 유트브(X)” 등의 난도가 높은 잘못된 음차 표기까지 정확히 판별할 수 있었다.

Sequence-to-Sequence with Attention 모델은 생성모델의 특성상 긴 단어나 복합명사인 ‘for the first time - 포더퍼스트타임(0)’, ‘you light up my life - 유라이트업마이라이프(0)’, ‘identification - 아이덴티피케이션(0)’와 같은 음차 표기에 대해서 낮은 품질을 보였으며, 편집거리 알고리즘을 이용하여 해당 문제를 일부 해결할 수 있지만 잘못된 음차 표기까지 추출될 수 있는 문제가 있었다.

마지막으로, Distance LSTM-CNN(+LN) 모델은 순서를 유지하면서 주위 문맥의 중요 정보를 추상화하는 자질을 추가, 적용함으로써 Distance LSTM 모델에서 잘못 판단하는 난도가 높은 “bubble love - 버블리러브(X)”, “vans - 반즈(X)”와 같은 음차 표기 대역 쌍 후보를 정확히 판별하는 것을 확인할 수 있었다.

5. 결론

본 논문에서는 검색 품질 향상을 위해 문서에서 자주 사용되는 다양한 음차 표현을 언어 자원으로 구축하기 위한 딥러닝 기반 음차 표기 대역 쌍 판별 모델을 제안했다. 제안하는 모델을 평가하기 위해 정보 검색에서 사람들이 자주 사용하는 음차 표기 데이터를 웹 문서에서 수집하고, 수집된 음차 표기 데이터 중 판별 난도가 높은 데이터 위주로 데이터 셋을 구축하여 모델의 실용성을 평가했다. 평가 결과, 최종적으로 제안하는 음차 표기 대역 쌍 판별에 특화된 모델인 Distance LSTM-CNN(+LN)은 한국어-영어간 음차판별을 위해 변형된 KODEX 방법과의 비교 결과, 약 35%의 품질 향상을 보였으며, 다국어 음차 생성에서 높은 품질을 보이는 음차 생성 모델을 음차 판별 모델로 변형한 Sequence-to-Sequence with Attention 모델과의 비교에서도 약 3.5%의 품질향상을 보였다. 또한 두 질의의 관련성을 판별하

는 연구에서 높은 품질을 보이는 딥러닝 모델인 Distance LSTM 모델을 개선한 모델과 비교해, 음차의 중요 자질을 부각시키는 CNN 구조를 추가함으로써 약 3% 정도의 품질 향상을 보여, 최종적으로 89.70%의 품질을 보였다. 마지막으로 제안하는 모델에 계층 정규화(Layer Normalization) 기법을 적용함으로써 미미한 품질 향상과 더불어 학습 시 약 40번의 epoch가 절약됨으로써 학습 시 수렴 속도를 개선하는 계층 정규화(Layer Normalization)의 효과를 확인할 수 있었다.

참고문헌

[1] 이재성, “다국어 정보검색을 위한 영-한 음차 표기 및 복원 모델”, 박사학위논문, 한국과학기술원 전산학과, 1999.

[2] 이희승, 안병희, 고친판 “한글 맞춤법 강의”, 신구문화사, 1994.

[3] 오종훈, 배선미, 최기선 “자동 음차 표기를 이용한 영-한 음차 표기 대역 쌍의 자동 추출”, 정보과학회논문지(B), 제31권, 제1호, pp. 928-930, 2004.

[4] Andrew Finch, Lemao Liu, Xiaolin Wang, and Eiichiro Sumita. Target-bidirectional neural models for machine transliteration. In Proc. of NEWS, pages 78-82, 2016.

[5] 김태일 ,“최대 엔트로피 모델을 이용한 다국어 정보검색에서의 영-한 음차 표기 모델” , 석사학위논문 , 서강대학교, 1999.

[6] 오종훈, 최기선, “자소 및 음소 정보를 이용한 영어-한국어 음차 표기 모델”, 정보과학회논문지(B), 제32권, 제4호, pp. 312-326, 2005

[7] Brill E., Gary Kacmarcik, Chris Brockett, Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs. NLPRS 393-399, 2001.

[8] 강병주, 이재성, 최기선, “외국어 음차 표기의 음성적 유사도 비교 알고리즘”, 정보과학회논문지(B), 제26권, 제10호, pp. 1237-1246, 1999.

[9] Jan, EA-EE., Ge, Niyu, Lin., Shih-Hsiang., Roukos, salim., Sorensen, Jeffrey. A novel approaxh for proper name transliteration verification, ISCSLP, 89-94, 2010.

[10] Sutskever, I. Vinyals, O., Le. Q. V. Sequence to sequence learning with neural networks. , In Proc. Advances in Neural Information Processing Systems, 27, 3104-3112, 2014.

[11] Hochreiter, S. & Schmidhuber, J. Long short-term memory., Neural Comput. 9, 1735-1780, 1997.

[12] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326, 2015.

[13] https://data.quora.com/First-Quora-Dataset Release-Question-Pairs

[14] Lili Jiang, Shuo Chang, and Nikhil Dandekar. "Semantic Question Matching with Deep Learning.", 2017.

[15] Ankur P Parikh, Oscar Tackstrom, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In Proceedings of EMNLP, 2016.

[16] Shuohang Wang and Jing Jiang. 2016. A Compare-Aggregate Model for Matching Text Sequences. CoRR abs/1611.01747, 2016.

[17] Wang, Z.; Hamza, W.; and Florian, R. Bilateral multiperspective matching for natural language sentences. In IJCAI, 2017

[18] Y. Kim, Convolutional Neural Networks for Sentence Classification, Conference on Empirical Methods in Natural Language Processing, 2014.

[19] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016.