Low-Resource 환경에서 Multi-Task 학습을 이용한 카자흐어 형태소 분석

Nazira Kaibalina, 박성배
경희대학교 컴퓨터공학과
nazira.kaibalina@khu.ac.kr, sbpark71@khu.ac.kr

Low-Resource Morphological Analysis for Kazakh using Multi-Task Learning

Nazira Kaibalina, Seong-Bae Park
Department of Computer Science & Engineering, Kyung Hee University

요 약

지난 10년 동안 기계학습을 통해 자연어 처리 분야에서 많은 발전이 있었다. Machine translation, question answering과 같은 문제는 사용 가능한 데이터가 많은 언어에서 높은 정확도 성능 결과를 보여준다. 그러나 low-resource 언어에선 동일한 수준의 성능에 도달할 수 없다. 카자흐어는 형태학적 분석을 위해 구축된 대용량 데이터셋이 없으므로 low-resource 환경이다. 카자흐어는 단일 어근으로 수백 개의 단어 형태를 생성할 수 있는 교착어이다. 그래서 카자흐어 문장의 형태학적 분석은 카자흐어 문장의 의미를 이해하는 기본적인 단계이다. 기존에 존재하는 카자흐어 데이터셋은 구체적인 형태학적 분석의 부재로 모델이 충분한 학습이 이루어지지 못하기 때문에 본 논문에서 새로운 데이터셋을 제안한다. 본 논문은 low-resource 환경에서 높은 정확도를 달성할 수 있는 신경망 모델 기반의 카자흐어 형태학 분석기를 제안한다.

1. 서론

카자흐어는 카자흐스탄, 러시아, 중국 중심으로 1천만 명 이상이 사용하는 투르크어이며 교착어이다. 자연어처리에서 카자흐어 연구는 말뭉치 생[1]-[2], 형태소 분석[3]-[4], 번역[5]이 있었다. 하지만 카자흐어의 part-of-speech(POS) 태깅 문제는 데이터셋의 부족으로 많은 주목을 받지 못했다.

형태학은 단어의 구조와 lemma, 어근, 접미사와 같은 단어의 구성을 연구하는 학문이다[6]. Lemma는 단어의 사전 형태이고 lemmatization은 주어진 단어의 lemma를 결정하는 과정이다. 형태학적 분석은 POS를 살펴보고 문맥에 따라 POS 태그를 단어에 할당한다. 그래서 문장의 형태학적 분석은 문장의 의미를 이해하는 기본적인 단계이다. 카자흐어에서는 접미사를 통해 단일 어근으로 수백 개의 단어를 생성할 수 있다. 그래서 동일한 단어 일지라도 문맥에 따라 다른 POS로 분석된다.

신경망 모델은 데이터가 충분한 환경에선 가능한 모든 단어 형태를 학습데이터로 학습한다. 지난 10년간 자연어 처리 분야에서 기계학습의 발전을 통해 많은 성과를 이루었지만[7], 신경망 모델은 학습데이터가 부족하면 성능이 저하된다는 문제점이 있다[8]. 그래서 자연어처리에서 low-resource 환경은 애플리케이션을 위한 충분한 대규모 말뭉치 또는 구축된 언어 자원의 부족을 의미한다[9].

카자흐어는 형태학적 분석을 위해 구축된 대용량 데이터셋이 없으므로 low-resource 환경이다. 이 문제를 해결하기 위해 본 논문에서는 기존 데이터 셋[2]을 재구축하여 보다 구체적인 형태학적 분석이 가능하도록 한다. 제안된 데이터셋에서는 각 단어마다 lemma와 형태소 접미사 태그를 사용하여 표시했다.

본 논문에서는 기존 POS뿐만 아니라 형태소 접미사를 모델이 학습하기 위해 multi-task 학습을 사용한다. Baseline 모델[10]에서, 두 가지 문제, 즉 lemmization과 POS 태깅은 표현을 공유하도록 학습되어 일반화 성능이 향상된다. Baseline 모델의 성능을 향상시키기 위해 본 논문에서는 FastText[11] pre-trained 임베딩을 인코더 계층에 추가한다. 이를

통해 lemmatization 작업의 경우 98.7%, POS 태그 작업의 경우 92%의 정확도를 얻었다. 본 논문은 제한된 양의 학습 데이터로 카자흐어를 위한 형태소 분석기를 제안한다.

2. 관련 연구

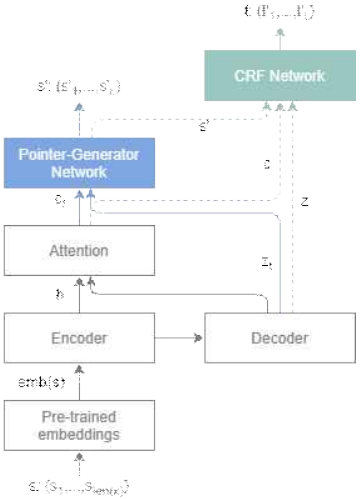
기존에 카자흐어를 위한 형태소 분석 방법 및 모델이 제안되었다. 첫 번째 연구에서는 카자흐어를 위한 data-driven 형태학적 분석과 HMM 기반 모델 제안한다[3]. 다른 연구에서는 형태학적 정보가 카자흐어의 POS 태그 성능에 미치는 영향을 분석한다[12].

형태소 분석을 위해 사용되는 모델 중 하나는 Pointer Generator (PG)[13] 네트워크이다. Pointer를 사용하여 원문에서 단어를 복사하고 generator를 통해 새 단어를 생성한다. 이러한 hybrid pointer-generator 시스템은 새로운 단어를 생성하는 능력을 유지하면서 정보의 정확한 복제를 가능하게 한다. 디코더는 사전 분포에서 단어를 생성할지 아니면 attention 분포에서 단어를 복사할지를 선택한다.

Conditional Random Field (CRF)[14]는 POS 태그 예측 작업에 가장 널리 사용되는 모델 중 하나이다. CRF는 또한 개별 태그가 아닌 태그의 문장 수준 위치에 초점을 맞춰 현재 태그를 예측하는 데 인접 태그 정보를 사용한다. 또한 과거 태그와 미래 태그를 효율적으로 사용할 수 있는 매개 변수로 상태 전환 매트릭스를 가지고 있다.

3. 모델

Baseline 모델은 한국어 형태소 분석을 위해 제안된 모델[10]이다. 제안된 모델은 attention을 사용하는 sequence-to-sequence 모델을 기반으로 하며, recurrent 인코더, recurrent 디코더, attention 및 task-dependent networks의 네 부분으로 구성된다. Input sequence는 pre-trained FastText[11] 임베딩 결과로 제공된다. 인코더는 embedded sequence를 일련의 hidden state로 인코딩한다. Attention은 이러한 hidden state를 일련의 context vector c로 변환한다. 이러한 vector는 형태소 처리 및 POS 태깅에 관련된 source-side 정보를 캡처한다. 디코더는 두 가지 형태소 작업 간에 공유되며, 이로 one-to-one mapping을 가질 수 밖에 없다. 형태소 처리는 PG[13] 네트워크에서 수행되고 POS 태깅은 CRF[13]에서 수행된다.



(그림 1) 제안 모델

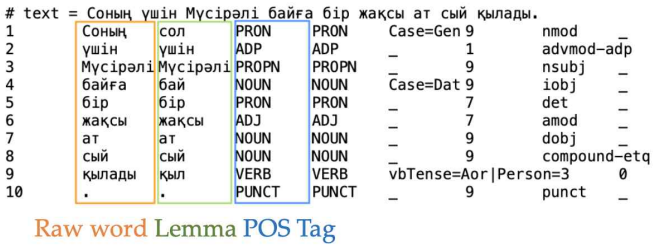
4. 실험

4.1. 데이터셋

이 연구에 사용된 데이터셋은 약 61K개의 문장을 포함하는 Kazakh Dependency Treebank[2]이다. Treebank 데이터셋은 Universal Dependency 2 지침에 따른 주석이 있으며, UD-native CoNLL-U 형식이다. 본 논문에서 제안한 데이터셋은 카자흐어 형태소 분석을 위해 재구축된 데이터셋이며 카자흐어를 위한 소수의 데이터셋 중 하나이다.

기존 Treebank 데이터셋은 두 가지 문제점이 있다. 먼저, 데이터의 양이 매우 적어서 충분한 학습이 어려운 low-resource 환경이다. 둘째, 기존 데이터에 대한 완전한 분석이 없다. 예를 들어, Treebank 데이터셋에는 다양한 접미사와 접미사에 대한 POS 태그에 대한 분석이 전혀 없다. 또한 데이터셋은 주로 온라인 뉴스와 위키백과 기사로 구성되어 있어 일부 단어에서 철자 오류가 존재한다. 그림 2는 기존 Treebank 데이터셋의 예시를 보여준다. 주황색 표시가 된 열은 단어의 원래의 형식이며, 녹색 및 파란색 표시가 된 열은 각각 lemma와 POS 태그로 구성된다.

기존 Treebank 데이터셋을 사용하여 baseline 모델은 5000개 이상의 문장이 누락되어 lemmatization 작업에서만 94.2%의 정확도를 달성했다. Resource-rich 환경에서, lemmatizer은 현재 결과보다 훨씬 높은 99%의 정확도를 달성한다[10]. 이것의 가능한 이유는 기존 데이터셋에 있는 오류와 접미사에 대한 분석 부족이다. 결과를 개선하기 위해 접미사 및 POS 태그의 분석을 통해 형태학적 분석이 처



(그림 2) Kazakh Dependency Treebank 데이터셋 예시

음부터 수동으로 수행되었다. 이를 위해 20개의 태그 셋을 표 1과 같이 구성하였다. 주요 태그는 표준 UD 태그 집합에서 가져온 반면, 접미사 태그는 문법 규칙에 따라 카자흐어의 다양한 형태소 접미사를 그룹화하여 처음부터 생성하였다. 카자흐어는 교착어이기 때문에, 가능한 형태소 접미사의 방대한 집합이 존재한다. 데이터 부족으로 인해 태그의 수를 제한하고 다양한 접미사 families (예: 사례, 분위기 등)를 함께 그룹화하기로 결정했다. 형태학적 분석은 약 100만 개의 토큰에 대해 수행하였다.

<표 1> POS tags 셋

주요 POS 태그	접미사 POS 태그
NOUN	ZHR
PROPN	PLR
VERB	POSS
AUX	SEPT
ADJ	RAI
PRON	TENSE
ADV	ZHAK
INTJ	
CONJ	
PART	
NUM	
SYM	
PUNCT	
FORGN	

그림 3은 변경된 예시가 노란색으로 강조 표시된 데이터 세트의 새로운 형식을 보여준다. 기존 데이터셋의 분석(왼쪽)은 "Raw word Lemma/POS 태그" 형식이었고, 새 분석(오른쪽)은 "Raw word Lemma/POS 태그 + Ending/POS 태그" 형식이다.

4.2 실험 설정

본 논문에서 사용된 모델은 attention을 사용하는 sequence-to-sequence 모델이다. 인코더는 3개의 레이어, hidden 크기 100, 임베딩 dimension 300, 드랍아웃 0.2의 양방향 LSTM이다. 디코더는 3개의 레이

(그림 3) Kazakh Dependency Treebank 데이터셋과 제안된 데이터셋 간의 비교 예시

어, 임베딩 dimension 300, 드롭아웃 0.2, teaching force ration 0.1의 LSTM이다. Optimizer는 adam을 사용했으며 학습률은 0.001, 배치 크기가 64으로 설정했다. 모델은 100 epochs 만큼 학습되었다.

4.3 실험 결과

Lemmatization 및 POS 태그 분류 작업의 정확도는 표 2에서 보여준다. 데이터 셋과 임베딩 설정이 다른 모델의 세 종류의 정확도 결과가 있다. 첫 번째 열의 결과는 baseline 모델을 기존 Treebank 데이터셋을 사용한 정확도이다. 추가 형태학적 분석을 통해 생성된 데이터셋으로 학습된 Baseline 모델은 lemmatization와 POS 태깅 작업에 대해 98%와 90%의 정확도를 보여준다. 마지막으로, 모델의 성능을 향상시키기 위해 pre-trained FastText 임베딩이 [9] 모델의 인코더 부분과 함께 사용되었고 98..7%와 92.%의 정확도를 보여준다. 새로운 데이터 셋으로 학습된 두 모델 모두 기존 statistical POS 태그 모델[3] 성능보다 월등히 높으며, 기존 데이터셋으로 학습된 Baseline 모델보다 우수한 성능을 보여준다.

<표 2> 실험 결과

	이전 데이터셋	새 데이터셋	
		FastText w/o	FastText w
Lemmatization	94.2%	98%	**98.7%**
POS tagging	80%	90%	**92%**

기존 데이터셋은 세분화된 형태학적 분석이 부족했기 때문에 모델이 제대로 학습할 수 없어 정확도가 낮은 결과를 보여준다. 더 구체적인 형태학적 분석을 추가하면 모델이 카자흐어의 문법을 학습할 수 있음을 결과를 통해 알 수 있다. Pre-trained 임베딩은 대규모 데이터셋에 대해 훈련되기 때문에 단어의 의미와 구문적 의미를 매우 잘 포착할 수 있다. 결

과적으로, Pre-trained 임베딩의 사용은 모델의 형태학적 분석에 도움을 준다.

5. 결론

본 논문은 신경망 모델을 사용하여 카자흐어에 대한 형태학적 분석을 제안한다. 추가적인 POS 태깅 작업을 위해 100만 개의 토큰을 분석하여 작은 데이터 세트가 구축했다. 제안된 모델은 pre-trained FastText 임베딩을 추가하여 기본 모델의 성능을 개선했다. 기존의 statistical POS 태그 모델보다 우수한 POS 태깅 성능을 보여주며, lemmatization 작업의 경우 대용량 데이터셋을 통해 학습된 모델과 유사한 성능을 보여준다. 본 논문에서 제안한 카자흐어를 위한 형태소 분석 모델은 가장 좋은 성능을 보여주는 모델이다.

사사

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. 2020R1A4A1018607)과 정보통신기획평가원의 지원(2017-0-01772, 비디오 튜링 테스트를 통과할 수준의 비디오 스토리 이해 기반의 질의응답 기술 개발)을 받아 수행된 연구임.

참고문헌

[1] Olzhas Makhambetov, Aibek Makazhanov, et al., “Assembling the Kazakh Language Corpus,” EMNLP, Seattle, USA, 2013, 1022-1031.

[2] Aibek Makazhanov, Olzhas Makhambetov, et al., “Syntactic annotation of Kazakh: following the universal dependencies guidelines. A report,” 3rd International Conference on Computer Processing in Turkic Languages, Kazan, Tatarstan, 2015, 338 - 350.

[3] Olzhas Makhambetov, Aibek Makazhanov, et al., “Data-Driven Morphological Analysis and Disambiguation for Kazakh,” CICLing, Cairo, Egypt, 2015, 151-163.

[4] Gulshat Kessikbayeva and Ilyas Cicekli, “Rule based morphological analyzer of Kazakh language,” Linguistics and Literature Studies, 4(1), 96-104, 2014.

[5] Assem Shormakova and Aida Sundetova, “Features of machine translation of different systemic languages using an apertium platform (with an example of English and Kazakh languages),” ICCAT, Sousse, Tunisia, 2013, 255-259.

[6] Altan Zeynep, “The Role of Morphological Analysis in Natural Language Processing,” 2002.

[7] Yonatan Belinkov and James Glass, “Analysis Methods in Neural Language Processing: A Survey,” Transactions of the Association for Computational Linguistics, 7, 49-72, 2019.

[8] Anastasopoulos, Antonios and Graham Neubig, “Pushing the Limits of Low-Resource Morphological Inflection,” EMNLP/IJCNLP, Hong Kong, China, 2019, 984-996.

[9] Michael Hedderich, Lukas Lange, et al., “A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios,” ArXiv, 2020.

[10] Hyun-Je Song and Seong-Bae Park, “Korean Morphological Analysis with Tied Sequence-to-Sequence Multi-task Model,” EMNLP/IJCNLP, Hong Kong, China, 2019, 1436-1441.

[11] Edouard Grave, Piotr Bojanowski, et al., “Learning Word Vectors for 157 Languages,” Proceedings of the International Conference on Language Resources and Evaluation, Miyazaki, Japan, 2018, 1-5.

[12] Aibek Makazhanov, Zhandos Yessenbayev, et al., “On certain aspects of Kazakh part-of-speech tagging,” AICT, Astana, Kazakhstan, 2014, 1 - 4.

[13] Abigail See, Peter Liu, and Christopher Manning, “Get To The Point: Summarization with Pointer-Generator Networks,” ACL, Vancouver, Canada, 2017, 1073-1083.

[14] Zhiheng Huang, Wei Xu and Kai Yu, “Bidirectional LSTM-CRF Models for Sequence Tagging,” ArXiv, 2015.