

한국어 학습자 말뭉치의 모어 판별

허희정[○], 정승연, 김한샘[●]

연세대학교 국어국문학과[○], 언어정보학협동과정, 언어정보연구원
{Heuijung.hur, withjess278}@gmail.com, khss@yonsei.ac.kr

Native Language Identification for Korean Learner Corpus

Heuijung Hur[○], Seung Yeon Chung, Han-Saem Kim[●]

Department of Korean Language and Literature, Yonsei University[○]
Interdisciplinary Graduate Program of Linguistics and Informatics, Yonsei University
Institute of Language and Information Studies, Yonsei University[●]

요약

모어 판별이란 제 2 언어를 습득하는 학습자들이 생산한 목표 언어에 기반하여 학습자들의 제 1 언어를 자동적으로 확인하는 작업을 말한다. 모어 판별 과제를 성공적으로 수행하기 위한 방법을 다룬 다양한 연구들이 진행되어 왔으나, 한국어를 대상으로 진행된 모어 판별 연구는 그 수가 극히 적다. 본 연구에서는 한국어 학습자 텍스트를 대상으로 머신 러닝, 딥 러닝의 다양한 문서 분류 모델을 실험하고, 이를 통해 한국어 학습자 텍스트 모어 판별을 위해 적합한 모델을 구축하기 위해 필요한 조건을 찾아보고자 하였다.

주제어: 모어 판별, 머신 러닝, 딥 러닝, Native Language Identification

1. 서론

이 연구는 한국어 학습자 말뭉치의 모어 판별(Native Language Identification, NLI)의 여러 방법론을 비교하고, 문서 분류 과제를 해결하는 여러 방법을 모어 판별에 적용하고 그 결과를 비교하여 한국어 모어 판별에 적합한 방법을 탐구하는 것을 목표로 한다. 모어 판별이란 제 2 언어를 습득하는 학습자들이 생산한 목표 언어에 기반하여 학습자들의 제 1 언어를 자동적으로 확인하는 작업을 말한다[1]. 이는 문서에 나타난 언어 자질을 이용하여 주어진 문서가 속하는 분류를 맞추는 작업으로, 저자 판별과 유사한 과제라고 볼 수 있다. 그러나 개별 저자의 독특한 스타일이 아닌 어떤 저자 집단이 공유하는 특징을 밝혀내고자 한다는 점에서 저자 판별보다 난이도가 높은 과제라고 할 수 있다 [2]에 따르면, 모어 판별은 학습자 오류에 좀 더 초점이 맞추어진 피드백을 제공하기 위해 교육적으로 이용될 수 있을 뿐만 아니라, 글쓰기 지도 시스템(writing tutor system)을 구현하는 데에도 이용될 수 있다, 또한, 법언어학에서 필요로 하는 authorship profiling에도 이용될 수 있어, 그 적용의 폭이 매우 넓다고 할 수 있다[4].

모어 판별은 영어 학습자 작문을 대상으로 한 [5]를 시작으로, 그 방법론과 연구 대상을 점차 확장시켜왔다. 그러나 2021년 현재, 한국어 학습자 말뭉치를 대상으로 한 연구로는 [6]이 유일하다. 해당 연구는 모어 판별을 저자 판별 과제의 연장선 상에 있는 것으로 파악하고, 통계적인 방법을 이용하여 한국어 학습자 말뭉치의 모어 판별에 유의미한 자질을 알아보려고 하였다.

본 연구에서는 국립국어원의 한국어 학습자 말뭉치 자료를 이용하여 모어 판별을 위한 실험 말뭉치를 구축하

고, 이를 이용하여 일반적인 문서 분류에 사용되는 다양한 테크닉을 비교하는 한편 한국어 모어 판별에 적합한 방법론을 탐색하는 것을 목표로 한다.

2. 관련 연구

모어 판별을 다루는 대부분의 논문에서 모어 판별에 대한 최초의 연구로 언급하고 있는 것은 [5]이다. 이 논문에서는 데이터로는 International Corpus of Learner English(ICLE)의 러시아어, 체코어, 불가리아어, 프랑스어, 스페인어 모어 학습자의 영어 작문 자료를, 분류 방법으로는 서포트 벡터 머신(Support Vector Machine, SVM)을 사용하였다. 분류 과정에서 사용된 피쳐로는 일반적으로 사용되는 기능어, 문자 n-gram, 학습자 작문에서 발견되는 오류, 저빈도 POS bigram 등이 있었으며, 이 모든 피쳐를 동시에 사용하였을 때 80.2%의 정확도를 얻을 수 있었다. [3], [7]는 [5]과 같은 데이터와 같은 알고리즘(SVM)을 사용하되, 판별에 사용하는 언어 자질의 폭을 다양화 하였다. [3]에서는 문자 bigram을 확장하여 unigram부터 trigram을 실험하는 한편 모어 판별에서 기능어와 내용어의 모어 판별에서의 영향을 가늠해보려고 하였다. 반면 [7]에서는 빈도가 높은 학습자 오류와 다양한 크기의 문자 n-gram, POS n-gram 등을 사용하였다. 이 시기의 모어 판별 연구에서는 분류 알고리즘으로 SVM이 가장 활발하게 사용되었으나, [8],[9]와 같이 분류 과정에 LDA 등의 모델을 사용하거나, 어휘 자질 대신 통사 자질을 사용하여 모어 판별 과제를 수행하고자 하는 시도가 있었다.

한편 이 시기의 모어 판별 연구는 실험에 사용되는 데이터, 방법론이 지나치게 다양하여, 연구 결과를 서로 비교하고 평가하기 어렵다는 한계가 있었다. 이러한 배경 속에서, 서로 다른 11개 모어 배경을 가진 영어 학습자의 TOEFL 시험 작문으로 이루어진 TOEFL-11 말뭉치를 이용하여 모어 판별 연구를 촉진시

키고자 하는 목적으로 2013 NLI Shared Task가 진행되었다. 2013 NLI Shared Task의 진행 과정과 결과를 다룬 [2]에 따르면, 대부분이 SVM을 사용하였으며, 단어 n-gram, POS n-gram, 문자 n-gram 등이 분류에 주로 이용되는 피쳐였다.

2015년을 전후하여, 영어 외의 다른 언어를 대상으로 모어 판별 연구가 이루어졌다. 중국어, 아랍어, 핀란드어, 러시아어, 노르웨이어 등을 학습하는 학습자의 자료를 대상으로 연구가 이루어졌으며, 이 일련의 연구들은 분류 알고리즘으로 SVM을 사용하되, 분류 대상 언어의 특성에 맞춰 조금씩 조정된 피쳐들을 사용하였다. ([10], [11], [12], [13], [14], [15])

2017 Shared Task에서는 모어 판별의 연구 범위가 문어뿐만 아니라 구어까지 확장되었다. 또한, 그동안 좋은 성능을 보인 SVM 외에 다른 방법을 접목해보고자 하는 시도들이 생겨나기도 하였다. 2017 NLI Shared Task의 결과를 다룬 [16]에 따르면, 전반적으로 복수의 분류기 (multiple classifier)를 이용하는 게 가장 효과적이며, 동시에 어휘 n-gram이 가장 효과적인 피쳐임을 알 수 있었다고 한다. 또한, 모어 판별에 딥 러닝을 사용하고자 하는 시도들이 있었으나, 전통적인 머신 러닝에 비해 정확도가 떨어지는 양상을 보였다.

이후 연구에서도 모어 판별의 대상이 되는 언어권, 연구 범위는 확장되어도 방법론 자체는 머신 러닝 모델을 많이 사용하는 경향이 지속되어 왔으나, 동시에 딥 러닝 모델을 모어 판별에 접목시켜보려는 연구가 계속해서 진행되었다. 예를 들어, [17]은 모어 판별에 단어 임베딩을 사용하여 분류 성능을 높이고자 하였다. 최근의 연구 중에는 [18]가 각 언어권 별로 훈련한 복수의 GPT-2 모델 이용, 86.6~ 94.2%의 정확도를 보인 바 있다.

앞선 연구를 살펴보면, 딥 러닝을 사용한 모어 판별 연구는 사용하는 자료, 방법 등에서 아직 탐구할 영역이 더 남아 있는 것으로 볼 수 있을 것이다.

3. 실험 방법

3.1. 실험 말뭉치의 구성

본 연구에서 사용한 실험용 말뭉치는 국립국어원의 한국어 학습자 말뭉치의 형태주석 말뭉치 중 중국어, 일본어, 베트남어, 영어권 학습자 작문으로 구성되어 있다. 한국어 학습자 말뭉치란, ‘제 2언어 혹은 외국어로 한국어를 배우는 학습자들이 산출한 텍스트 및 담화 자료를 수집하여 컴퓨터가 읽을 수 있는 형태로 전산화한 언어 자료’를 의미한다([19]). 지난 2015년부터 구축되어 온 이 말뭉치는 원시 말뭉치 기준 약 300만 어절 규모로 구성되어 있으며, 총 142개국의 93개 언어권 배경 학습자로부터 수집한 기관 학습자 자료, 이주민 자료, 국외 학습자 자료 등으로 구성되어 있다. 이 말뭉치의 중요한 구조상 특징 중 하나는 원시 말뭉치 중 일부가 형태 주석이 되어 있고, 또 형태 주석된 말뭉치 중 일부가 오류 주석이 되어 있는 피라미드식 구조로 되어 있다는 점이다. 본 연구에서는 전체 말뭉치 자료 중 형태주석 말뭉치 자료를 사용하였으며, 93개 언어권 자료 중 가장 큰 비중을 차지하고 있는 중국어권, 일본어권, 영어권, 베트남어권 한국어 학습자의 국내 기관 자료를 사용하였다.

앞선 NLI 관련 연구를 살펴보면, 실험 대상 말뭉치를 구성할 때 중요한 요인으로 말뭉치를 구성하는 모어 집단 사이의 균형성뿐만 아니라 말뭉치를 구성하는 주제의 균형성, 학습자 집단의 숙달도 분포 등이 있었다([20]). 이에, 본 연구에 사용할 실험 말뭉치를 구성하기 위해서 각 언어권 집단 사이의 균형을 가장 먼저 고려하되, 한국어 학습자 말뭉치의 표본 정보를 참조하여 주제 별로 균형을 이룰 수 있도록 각 언어권 별로 말뭉치 샘플을 임의로 추출하였다. 언어권의 균형성, 주제의 균형성, 숙달도의 균형성을 모두 성취하기가 쉽지 않은 관계로, 본 연구에서는 언어권과 주제에 좀 더 중점을 두었다.

그 결과, 다음 [표1]과 같이 말뭉치를 구성하였다.

	전체 어절 수	평균 어절 수	표본 개수
중국어	60,991	104.62	594
일본어	60,863	109.27	565
영어	60,051	106.28	557
베트남어	58,877	99.12	583
	240,782		

[표 1] 말뭉치 구성

3.2. 실험 방법

머신 러닝 모델로는 SVM, softmax, XGBoost를 사용하였다. SVM은 마진을 최대화하는 최적의 결정 경계 (Decision Boundary)를 정의하고 이를 이용해 분류를 수행하는 모델로, 앞서 살펴본 바와 같이 모어 판별 관련 연구에서 가장 빈번하게 사용되는 분류 알고리즘인 동시에 꾸준히 높은 정확도를 보인 분류 알고리즘이기도 하다. softmax는 softmax 함수를 이용해 X를 3개 이상의 선택지 중 하나로 분류하는 다중 분류 (Multi Classification)를 수행하는 알고리즘이다. Softmax와 XGBoost는 문서 분류 과제를 위해 자주 사용되는 분류 알고리즘이기도 하다. 이때, 모어 판별 모델을 학습시키기 위한 피쳐로는 형태소 unigram, bigram, POS unigram, bigram을 이용하였으며, 학습을 위한 자질값으로는 빈도와 TF-IDF 값을 이용하였다.

딥 러닝 모델로는 CNN, LSTM, MLP, Universal Sentence Encoder, Sentence BERT 등을 이용하였으며, 이때 복수의 단어 임베딩과 문장 임베딩을 이용하여 모어 판별에 가장 적합한 실험 방법을 찾아보고자 하였다. 합성곱 신경망이라고도 알려져 있는 CNN의 경우, 손글씨 분류와 같은 과제에서 좋은 성능을 보이는 모델로, 본 연구에서는 Conv2d, Conv1d 합성곱 신경망 모델을 모두 이용하여 모델링하였으며 Word2Vec 단어 임베딩과 glove 단어 임베딩을 모두 사용하여 실험하였다. 게이트가 추가된 순환신경망 모델의 일종인 LSTM의 경우, CNN과 마찬가지로 glove 단어 임베딩을 사용하여 실험을 진행하였다. MLP는 피드포워드

(FeedForward) 뉴럴 네트워크로도 불리며 퍼셉트론으로 이루어진 층을 여러 개 순차적으로 쌓아놓은 모델로, 오차역전파(Backpropagation)을 통해 학습을 진행한다. 상기한 딥 러닝 모델은 모두 문서 분류 과제를 위해 일반적으로 사용되는 모델이다.

이와 더불어, 본 연구에서는 Universal Sentence Encoder, Sentence BERT를 이용하여 모어 분류 과제에 있어 문장 임베딩과 단어 임베딩이 어떤 영향을 미치는지 알아보고자 하였다. Universal Sentence Encoder는 구글에서 내놓은 문장 인코더로 트랜스포머 모델을 SNLI 데이터셋을 이용하여 문장 임베딩을 프리 트레이닝한 모델이다. Sentence BERT는 BERT가 의미적으로 유사한 문장들을 잘 판별하도록 Siamese, Triplet 뉴럴 네트워크 형태로 프리 트레이닝한 문장 인코더이다. 이들은 문장 단위로 의미를 인코딩하여 분류 태스크를 수행하였을 때 NLI 태스크의 성능이 어떻게 바뀌는지 살펴보기 위해 사용하였다.

4. 실험 및 결과

	softmax	SVM	MLP ¹
Morpheme Unigram	0.65	0.63	0.62
Morpheme Unigram + TF-IDF	0.64	0.66	0.61
Morpheme Bigram	0.65	0.64	0.65
Morpheme Bigram + TF-IDF	0.64	0.68	0.66
Morpheme Unigram + Bigram + TF-IDF	0.68	0.69	0.68
POS unigram	0.46		
POS bigram	0.50		
Random Baseline	0.25		

[표 2] 머신 러닝 결과물

본 연구에서 사용한 여러 모델 중에서 가장 좋은 결과를 보인 것은 형태소 bi-gram을 자질로, 빈도의 TF-IDF 보정치를 자질 값으로 취한 SVM 모델을 사용한 경우로, 정확도가 69.6%로 나타났다. 또한, 머신 러닝 모델의 결과물을 서로 비교했을 때에도, 형태소 bi-gram을 자질로, 빈도의 TF-IDF 보정치를 자질 값으로 취한 SVM 모델이 가장 좋은 정확도를 보였다.

	glove	Word2Vec
LSTM	0.32	
CNN(Conv ID)	0.30	

¹ MLP는 분류 상으로는 딥러닝 모델에 속하는 모델이지만, 본 연구에서 사용한 다른 딥러닝 모델과는 달리 glove, word2vec을 사용하는 대신 머신 러닝 모델에서 사용한 것과 같은 자질을 사용하였다.

CNN(Conv 2D)	0.59	0.53
Hierarchical LSTM	0.51	
Universal Sentence Encoder + CNN Conv1d	0.55	
Universal Sentence Encoder + LSTM	0.34	
Sentence Transformer + FF Layer	0.54	
Sentence Transformer + Conv1d	0.57	

[표 3] 딥 러닝 결과물

반면, 딥 러닝 알고리즘 중에서 가장 좋은 결과를 보인 것은 MLP 모델로, 정확도가 68%로 가장 높게 나타났다. 이는 영어 학습자 말뭉치를 사용한 이전 연구에 비하면 다소 미흡한 수준이지만, 최종적으로 도달한 성능은 비슷하다는 점에서 흥미롭다고 할 수 있다.

앞서 살펴본 바와 같이 본 연구에서 사용한 데이터셋의 규모가 크지 않으므로, 보다 엄밀한 모델 비교를 위해 가장 정확도가 높게 나온 MLP, SVM, Softmax의 세 모델에 대해 5-fold cross validation을 실시하였다. 그 결과, 0.8/0.2로 Train-Test Split했을 때 세 모델들의 CV Score의 평균 값은 각각 0.7149, 0.7192, 0.7170으로 이들 사이에 큰 차이는 없는 것으로 나타났다.

머신 러닝, 딥 러닝 각각의 모델들 중 가장 높은 정확도를 보인 SVM 모델과 MLP 모델의 classification report를 살펴보면 다음과 같다. 두 모델 모두, 일본어 - 베트남어 - 중국어 - 영어 순으로 정확도가 높게 나타나는데, 이는 언어간 유사성과 모어 판별 과제의 난이도에 대해 더 탐구할 수 있는 부분이 있음을 시사한다.

	precision	recall	F1-score	support
중국어	0.65	0.68	0.67	111
영어	0.57	0.62	0.59	104
일본어	0.82	0.77	0.79	118
베트남어	0.74	0.69	0.72	127
Accuracy			0.69	460
Macro avg	0.69	0.69	0.69	460
Weighted avg	0.70	0.70	0.70	460

[표 4] SVM classification report

	precision	recall	F1-score	support
중국어	0.62	0.66	0.64	111
영어	0.54	0.59	0.56	104

일본어	0.80	0.79	0.79	113
베트남어	0.74	0.67	0.70	132
Accuracy			0.68	460
Macro avg	0.68	0.67	0.67	460
Weighted avg	0.68	0.68	0.68	460

[표 5] MLP classification report

전반적으로 살펴보면 머신 러닝을 사용하였을 때에는, 사용한 모델에 관계 없이 정확도가 비교적 안정적으로 나타나는 것을 알 수 있었다. 반면에 딥 러닝 모델을 사용했을 때에는 알고리즘, 임베딩의 종류에 따라 차이가 다소 크게 나타나는 현상을 보였다. 이는 앞선 연구에서 나타난 것과 비슷한 경향을 보인다고 할 수 있다. 다시 말해, 딥 러닝 알고리즘을 모어 판별이라는 과제에 적용하기 위해서는 과제 특성에 맞는 모델 설계가 매우 중요하다는 의미로 받아들일 수 있을 것이다.

머신 러닝 모델링 시 POS 태그 unigram, bigram 피처보다는 형태소 관련 피처가 성능이 훨씬 잘 나오는 것을 알 수 있었다. 이는 모어 판별 과정에서 어휘 자질이 미치는 영향이 더 크다고 판단할 수 있는 근거가 될 수 있으며, 앞서 살펴본 여러 선행 연구에서 어휘 관련 자질들이 좋은 성능을 보인 것과 일맥상통하는 결과이기도 하다.

머신 러닝 모델링 결과를 살펴보았을 때, 가장 성능이 잘 나온 피처는 형태소 unigram, bigram 피처의 TF-IDF 값을 사용한 경우였다. TF-IDF는 문서 집단 내에서 특정 문서에 더 자주 등장하는 단어에 가중치를 주는 방식으로, 이러한 값을 취했을 때 모어 판별의 성공률이 올라간다는 점은 모어 판별 과제에 있어 내용어, 기능어가 어떤 영향을 주는지 살펴봐야 할 필요성이 있음을 시사한다. 이 부분에 대해서는 추후 연구에서 다루고자 한다.

딥 러닝의 경우 LSTM 같은 순환 신경망 모델보다는 CNN이, 더 잘 작동한 것을 확인할 수 있었다. 이는, 텍스트 전체 어절(token)들을 재귀적으로 처리하는 방식보다는, 형태소들의 국소적인 연쇄만으로 자질을 추출하여 분류 태스크에 적용하는 것이 모어 판별 태스크에 보다 효율적이기 때문일 것이다.

또한, 일반적인 문서 분류에서 단어 단위 임베딩, 피처보다 좋은 성능을 보이는 문장 임베딩 모델들이 오히려 정확도가 떨어지는 것을 알 수 있었다. 이 역시 모어 판별 태스크에는 문장 단위로 의미를 인코딩하여 자질을 얻는 방식보다는, 형태소 단위의 자질을 이용해 분류 태스크를 수행하는 것이 더 효율적이기 때문으로 보인다. 다시 말해, 모어 판별의 경우 일반적인 문서 분류와는 다른 접근 방법을 필요로 한다는 것을 알 수 있다. 특히, 본 연구에서 사용한 말뭉치의 경우 언어 교육 기관에서 작성된 시험 작문, 과제 작문 등으로 이루어져 있다. 이는 기관의 교육 과정에 따라 학습자가 사용하도록 유도된 표현, 어휘 등이 학습자의 문장 구성에 영향을 미칠 수 있음을 의미한다. 이러한 경우에는 문장보다 더 작은 단위의 임베딩이 더 효과적일 것이다.

5. 결론 및 향후 연구

본 연구에서는 한국어를 대상으로는 상대적으로 연구가 덜 이루어진 분야인 모어 판별의 여러 방법을 실험하고, 이를 통해 성공적인 한국어 학습자 텍스트의 모어 판별 모델이 갖추어야 할 조건을 탐색하고자 하였다. 실험은 머신 러닝 알고리즘인 SVM, Softmax, XGBosst와 딥 러닝 알고리즘인 CNN, LSTM, MLP, Universal Sentence Encoder, Sentence BERT 을 이용하여 이루어졌으며, 단어 임베딩과 문장 임베딩을 모두 사용하여 어떤 방식이 모어 판별이라는 특수한 과제에 더 적합한 방식인지 확인해보고자 하였다. 실험 결과, 머신 러닝 모델 중 가장 높은 정확도를 나타낸 것은 형태소 bigram의 빈도의 TF-IDF 보정치를 자질값을 취한 SVM 알고리즘으로, 68.3%의 정확도를 나타냈다. 반면에, 딥 러닝 모델 중 가장 높은 정확도를 나타낸 것은 68%의 정확도를 보인 MLP 모델이었다. 특히, 딥 러닝 모델의 경우 일반적인 문서 분류 과제와 달리 문장 임베딩에 비해 단어 임베딩이 더 나은 성적을 보이는 것을 알 수 있었다. 이를 통해 n-gram 단위의 국소적인 어절들의 나열에서 피처를 추출하는 방식이 전체 텍스트를 순차적으로 처리하는 것보다 모어 판별 과제에 더 효율적이라는 사실을 알 수 있었다.

본 연구는 국내 한국어 학습자의 인구 통계학적 비율을 반영하고, 또 다양한 언어권의 언어를 포함시키고자 하는 과정에서 학습자의 숙달도 간 균형을 충분히 반영하지 못했다는 한계가 있다.

추후 연구에서는 모어 판별의 성공률을 높일 수 있는 모델을 제안하는 한편, 모어 판별 결과물 자체에 대한 분석을 통해 한국어 학습자의 언어에 나타나는 특징을 분석하고 더 나아가 한국어 모어 화자와 가까운 언어 능력을 구성하는 요소가 무엇인지 분석해보고자 한다.

참고문헌

- [1] Jarvis, S. & Paquot, M., "Learner corpora and native language identification" in Granger, S., Gilquin, G., & Meunier, F. (eds.) The Cambridge Handbook of Learner Corpus Research. Cambridge: Cambridge University Press. pp. 606-627. 2015
- [2] Tetreault, J., Blanchard, D., & Cahill, A. A report on the first native language identification shared task. In Proceedings of the eighth workshop on innovative use of NLP for building educational applications. pp. 48-57. 2013.
- [3] Tsur, O., & Rappoport, A. Using classifier features for studying the effect of native language on the choice of written second language words. In Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition. pp. 9-16. 2007.
- [4] Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. Author profiling for English emails. In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics. Vol. 263. p. 272. 2007.
- [5] Koppel, M., Schler, J., & Zigdon, K. Automatically determining an anonymous author's native language. In International Conference on Intelligence and Security Informatics. pp. 209-217. 2005.
- [6] 유소영, 허희정, & 한송화. 한국어 학습자의 자동 L1 판별을 위한 언어 자질 연구. 언어사실과 관점, 52, 385-417. 2021

- [7] Wong, S. M. J., & Dras, M. Contrastive analysis and native language identification. In Proceedings of the Australasian Language Technology Association Workshop 2009. pp. 53-61. 2009.
- [8] Wong, S. M. J., Dras, M., & Johnson, M. (2011). Topic modeling for native language identification.
- [9] Brooke, Julian, and Graeme Hirst. "Robust, lexicalized native language identification." *Proceedings of COLING 2012* (2012): 391-408.
- [10] Malmasi, S., & Dras, M. Chinese native language identification. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers. pp. 95-99. 2014.
- [11] Malmasi, S., & Dras, M. Arabic native language identification. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP). pp. 180-186. 2014.
- [12] Malmasi, S., & Dras, M. Finnish native language identification. In Proceedings of the Australasian Language Technology Association Workshop 2014. pp. 139-144. 2014.
- [13] Remnev, Nikita. "Native Language Identification for Russian." *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2019.
- [14] Malmasi, S., Dras, M., & Temnikova, I. Norwegian native language identification. In Proceedings of the International Conference Recent Advances in Natural Language Processing. pp. 404-412. 2015.
- [15] Malmasi, Shervin, Mark Dras, and Irina Temnikova. "Norwegian native language identification." *Proceedings of the International Conference Recent Advances in Natural Language Processing*. 2015.
- [16] Malmasi, S., Evanini, K., Cahill, A., Tetreault, J., Pugh, R., Hamill, C., ... & Qian, Y. A report on the 2017 native language identification shared task. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 62-75. 2017.
- [17] Franco-Salvador, M., Kondrak, G., & Rosso, P. Bridging the native language and language variety identification tasks. *Procedia computer science*, 112, pp. 1554-1561. 2017.
- [18] Lotfi, E., Markov, I., & Daelemans, W. A Deep Generative Approach to Native Language Identification. In Proceedings of the 28th International Conference on Computational Linguistics. pp. 1778-1783. 2020.
- [19] 국립국어원 한국어 학습자 말뭉치 나눔터, <https://kcorpus.korean.go.kr>
- [20] Malmasi, Shervin, and Mark Dras. "Multilingual native language identification." *Natural Language Engineering* 23.2 (2017): 163-215.