악성 댓글에 사용된 문자의 형태를 고려한 한국어 자연어처리를 위한 전처리 기법

김해수[1], 김미희[2]
[1]한경대학교 컴퓨터응용수학부
[2]한경대학교 컴퓨터응용수학부, 컴퓨터시스템연구소
e-mail:{ww232330, mhkim}@hknu.ac.kr

Preprocessing technique for natural language processing considering the form of characters used in malicious comments

Hae-Soo Kim[1], Mi-hui Kim[2]
[1]School of Computer Engineering & Applied Mathematics, Hankyong National University
[2]School of Computer Engineering & Applied Mathematics, Computer System Institute Hankyong National University

요 약

최근 악플에 대한 논란이 끊이지 않고 있어 이것을 해결하기위한 방법으로 자연어 처리를 이용하고 있다. 특히 소셜 미디어, 온라인 커뮤니티에서 많이 발생하고 있고 해당 매체에서는 한글을 그대로 사용하지 않고 그들의 은어를 섞어서 사용하며 그중에서 한글이 아닌 문자를 섞어서 만들어낸 문장도 있다. 이러한 문장은 기존의 모델에 학습된 데이터의 형태와 다르며 한글이 아닌 문장이 많을수록 모델의 예측이 부정확해진다는 단점이 있어 본 논문에서는 인공지능을 이용한 이미지 분류와 띄어쓰기, 오타 교정을 이용한 전처리 기법을 제안한다.

1. 서론

자연어 처리는 인간의 언어를 컴퓨터가 해석할 수 있도록 도와주는 기술로 문장을 통해 구문, 감정 분석, 문서 분류 및 번역과 자동 대화문 생성에 쓰이고 있으며 최근 악플에 대한 논란이 끊이지 않고 특히 소셜 미디어, 커뮤니티에서 많이 발생하고 있어 이것을 해결하기위한 방법으로도 사용하고 있다. 하지만 사용자들이 인터넷상에서 사용하는 언어는 일반적인 한국어와는 다르게 인간의 시각에서만 이해할 수 있는 형태로 사용하고 있으며 사전에 없는 단어를 만들어 사용하거나 기존에 있던 의미와는 다르게 의미를 만들어 사용하여 커뮤니티에서 사용되는 은어를 이용하는 경우가 대부분이다. 많은 커뮤니티 이용자들은 흔히 야민정음이라는 명칭으로 사용되는 형태로 한국어를 사용하고 있다.[1].

[1]에서 설명된 것과 같이 한글이 아닌 문자를 섞는다거나 새로운 형태의 단어를 만들어서 사용하고 있다.

인공지능 모델의 입력 데이터가 일관된 문장의 형태를 가지게 해야 한다. 따라서 위와 같이 다른 형태의 문장들을 일반적인 문장의 형태로 변화시켜주는 과정이 필요하다.

본 논문에서는 CNN(Convolution Neural Network)[2]과 N-Gram[3] 방식을 이용해 한글이 아닌 문자를 이용하여 만든 문장을 한글로 변환하고 오타 및 띄어쓰기 교정을 하는 전처리 기법을 제안한다. 이를 통해 여러 문자가 들어가 일관되지 않은 문장 데이터를 전처리과정을 통해 일관된 데이터로 만들어 인공지능 모델에서 악플 탐지를 효과적으로 수행할 수 있게 하기 위함이다.

2. 관련연구

2.1 한국어 자연어처리

[4]의 연구에서는 악성 댓글의 데이터 특성상 텍스트의 길이가 변칙적이기 때문에 비선형적인 특성을 제대로 학습할 수 있는 CNN과 입력 데이터의 신호를 그대로 분류에 활용할 것인지, 네트워크를 통해 학습하여 최종 결과값에 반영할 것인지 결정하는 하이웨이 네트워크를 결합하여 악성 댓글을 분류하였고 [5]의 연구에서는 입력 시퀀스에 대한 정보를 포함하는 positional embedding과

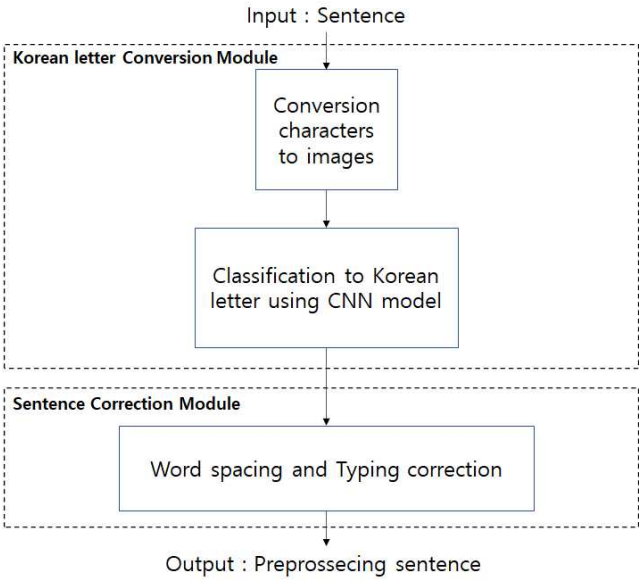
RNN(Recurrent Neural Network)과는 다르게 병렬 연산이 가능한 multi-head attention 기법을 채택한 transformer 모델을 이용하여 악플을 분류했다.

이처럼 악플을 탐지하는 기법들이 제안되었으나 전처리 부분에 중점을 두어 최근 커뮤니티에서 사용되는 한글의 형태를 고려하지 않았다.

본 논문에서는 이미지 분류를 이용해 한글이 아닌 문자를 한글로 변환하고 띄어쓰기 및 오타를 교정하는 과정으로 데이터를 일관된 형태로 전처리하는 기법을 제안한다.

3. 제안하는 전처리 기법

본 장에서는 제안하는 전처리 기법을 설명한다. 전처리과정은 (그림 1)과 같다. 한글로 변환하는 모듈(Korean letter Conversion Module)에서는 입력되는 문장에 있는 문자들을 이미지로 만들고(Conversion characters to images) CNN을 이용한 이미지 분류기법을 이용해 한글로 변환(Classification to Korean letter using CNN model)하는 과정이 이루어진다. 문장 교정 모듈(Sentence Correction Module)에서는 띄어쓰기 및 오타 교정(Word spacing and Typing correction)이 진행된다.



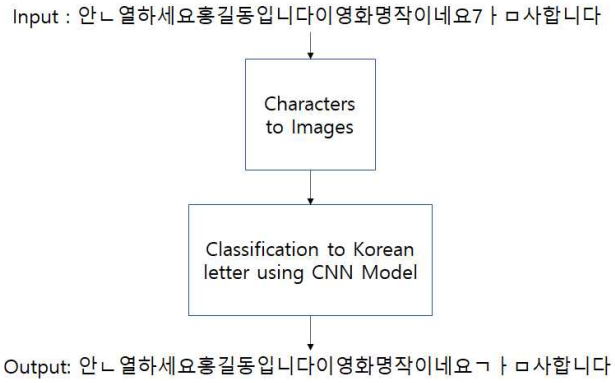
(그림 1) 제안하는 전처리 기법

3.1 한글로 변환 모듈(Korean letter Conversion Module)

해당 모듈은 인간의 시점에서 한글처럼 보이게 사용해 실제로는 한글과 연관이 없는 문자들을 사용하여 만든 문장들을 한글로만 이루어진 문장으로 변환하는 모듈이다.

음절들을 자소 단위로 분리하고 분리된 자음과 모음을 이미지로 변환한다. 그중에 한글이 아닌 문자를 이미지의 형태로 만들고 CNN을 이용한 이미지 분류를 통해 가장 비슷하게 생긴 한글로 변환한다.

(그림 2)는 숫자 7을 사용한 문장을 한글로 변환하는 모듈에서 이루어지는 과정을 통해 한글 ㄱ으로 변환하는 예시를 보여준다.



(그림 2) 한글이 아닌 문자를 한글로 변환하는 과정의 예시

또한 자음, 모음으로 이루어진 데이터만 사용하는 것이 아니라 음절 단위로 만들어진 데이터도 사용하기 때문에 쭈를 한자 卒로 쓴 것처럼 하나의 음절을 다른 문자로 사용하는 경우 해당 과정으로 변환할 수 있다.

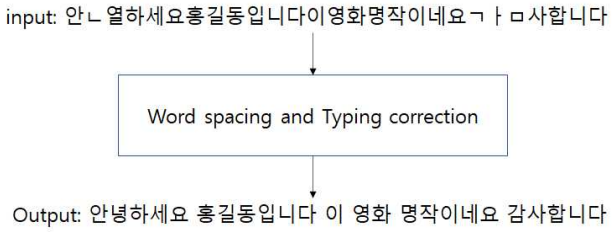
CNN 모델을 학습할 때 사용한 데이터 셋 중 음절 데이터는 AI Hub에서 제공한 한국어 글자체를 사용하고 자음과 모음이 나뉜 데이터는 여러 폰트를 이용해 자체적으로 생성한 데이터 셋을 사용한다.

3.2 문장 교정 모듈(Sentence Correction Module)

해당 모듈은 단어 사전과 N-Gram 방식을 통해 오타 교정 및 띄어쓰기한다.

오타 교정은 해당 단어가 단어 사전에 없으면 문자 추가, 삭제, 교체 및 인접한 문자 뒤바꾸기를 통해 편집 횟수가 최소인 단어가 교정된 단어가 되고 편집 횟수가 같은 단어가 여러 개라면 빈도수가 많은 단어가 교정된 단어가 되는 방식이다.

띄어쓰기는 N-Gram 방식과 RNN을 이용해서 기록된 단어들의 빈도수를 통해 작동하는 방식이며 문맥을 파악하지 못하고 기록된 단어를 통해서만 띄어쓰기를 판별하는 단점을 RNN을 통해 해결한다.



(그림 3) 오타 교정 및 띄어쓰기 과정의 예시

(그림 3)은 문장 교정 모듈에서 이루어지는 과정을 통해 기대되는 출력의 예시를 보여준다.

4. 결론 및 향후 연구

본 논문에서는 온라인 커뮤니티에서 사용되는 한글의 형태를 고려한 전처리 기법을 제안하였다. 해당 기법을 통해 다양한 형태의 문자가 사용되는 문장들을 일관된 형태로 인공지능 모델에 입력하고 한글이 아닌 문자로 인한 모델의 성능저하를 최소화할 수 있을 것이다. 향후 연구에서는 전처리 기법의 개선 및 성능을 평가하고 최종적으로 전처리 된 데이터를 모델에 학습하여 악플 탐지의 성능 향상을 보이고자 한다.

5. Acknowledgement

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2018R1A2B6009620), 교신저자 김미희.

참고문헌

[1] 강옥미 , 야민정음과 급식체의 해체주의 표현연구, 인문학연구, vol., no.56, pp. 325-349 (25 pages), 2018.

[2] K. O'Shea, R. Nash, "An Introduction to Convolutional Neural Networks,", https://arxiv.org/abs/1511.08458, 2015.

[3] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. "Class-Based n-gram Models of Natural Language," Computational Linguistics, 18(4), 467 - 480, 1992.

[4] 이현상, 이희준, 오세환. "하이웨이 네트워크 기반 CNN 모델링 및 사전 외 어휘 처리 기술을 활용한 악성 댓글 분류 연구". 정보시스템연구, vol. 29, pp.103-117, 2020.

[5] 윤현서, 유선용. "Transformer 기반 비윤리적 문장 탐지" 디지털콘텐츠학회논문지 22, no.8, 1289-1293, 2021.