띄어쓰기 및 문장 경계 인식을 위한 다중 손실 선형 결합 기반의 다중 클래스 분류 시스템

김기환○, 서지수, 이경열, 고영중
동아대학교, 컴퓨터공학과
{kimgihwan3364, jisu.ponyo, lky3568, youngjoong.ko}@gmail.com

Multi-class Classification System Based on Multi-loss Linear Combination for Word Spacing and Sentence Boundary Detection

GiHwan Kim○, Jisu Seo, Kyungyeol Lee, Youngjoong Ko
Department of Computer Engineering, Dong-A University

요 약

띄어쓰기와 문장 경계 인식은 그 성능에 따라 자연어 분석 단계에서 오류를 크게 전파하기 때문에 굉장히 중요한 문제로 인식되고 있지만 각각 서로 다른 자질을 사용하는 문제 때문에 각각 다른 모델을 사용해 순차적으로 해결하였다. 그러나 띄어쓰기와 문장 경계 인식은 완전히 다른 문제라고는 볼 수 없으며 두 모델의 순차적 수행은 앞선 모델의 오류가 다음 모델에 전파될 뿐만 아니라 시간 복잡도가 높아진다는 문제점이 있다. 본 논문에서는 띄어쓰기와 문장 경계 인식을 하나의 문제로 보고 한 번에 처리하는 다중 클래스 분류 시스템을 통해 시간 복잡도 문제를 해결하고 다중 손실 선형 결합을 사용하여 띄어쓰기와 문장 경계 인식이 서로 다른 자질을 사용하는 문제를 해결했다. 최종 모델은 띄어쓰기와 문장 경계 인식 기본 모델보다 각각 3.98%p, 0.34%p 증가한 성능을 보였다. 시간 복잡도 면에서도 단일 모델의 순차적 수행 시간보다 38.7% 감소한 수행 시간을 보였다.

주제어: 띄어쓰기, 문장 경계 인식, 다중 클래스 분류 시스템, 다중 손실 선형 결합

1. 서론

자연어처리에서 전처리(Preprocessing)란 형태소 분석(Part-of-Speech Tagging), 개체명 인식(Named Entity Recognition), 의존 구문 분석(Dependency Parsing), 의미역 결정(Semantic Role Labeling)과 같은 자연어 분석 단계 이전에 수행하는 과정이다. 띄어쓰기란 한국어 문장에서 어절과 어절 사이를 구분하는 것을 말하며, 문장 경계 인식이란 문서 안에서의 문장과 문장을 구분하는 것을 의미한다.

이러한 띄어쓰기와 문장 경계 인식은 그 성능에 따라 자연어 분석 단계에서 오류를 크게 전파하기 때문에 굉장히 중요한 문제로 인식되고 있어 띄어쓰기와 문장 경계 인식은 연구가 활발히 진행되고 있다. 그러나 띄어쓰기[1-7]와 문장 경계 인식[8-11]이 서로 다른 자질을 사용하는 문제로 인하여 각각 다른 모델을 사용해 순차적으로 해결하였다.

그러나 띄어쓰기와 문장 경계 인식은 하나가 성립되면 다른 하나가 성립할 수 없는 상호 배타적 관계로 완전히 다른 문제라고는 볼 수 없으며, 두 모델을 순차적으로 수행하면 앞선 모델의 오류가 다음 모델에 전파될 뿐만 아니라 시간 복잡도가 높아진다는 문제점이 있다.

따라서 본 논문에서는 띄어쓰기와 문장 경계 인식을 하나의 문제로 보는 다중 클래스 분류 시스템을 제안한다. 본 논문에서는 띄어쓰기와 문장 경계 인식을 한 번에 처리하여 시간 복잡도 면에서 단일 모델의 순차적 수행 시간보다 38.7% 감소를 보였고, 다중 손실 선형 결합(Multi-loss Linear Combination)을 사용하여 띄어쓰기와 문장 경계 인식이 다른 자질 사용하는 문제를 해결하여 성능 면에서 띄어쓰기 기본 모델과 문장 경계 인식 기본 모델 성능보다 각각 3.98%p, 0.34%p 증가한 성능을 보였다. 이를 통해 제안 모델이 타당함을 입증하였다.

본 논문의 구성은 2장에서 관련 연구를 소개하고, 3장에서는 띄어쓰기와 문장 경계 인식을 한 번에 처리하는 모델을 제안하고, 4장에서 실험을 통해 제안한 방법의 결과를 분석하고, 마지막 5장에서 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

한국어 띄어쓰기와 문장 경계 인식에 대한 연구는 규칙 기반, 통계 기반, 기계학습 기반 방식이 있다. 띄어쓰기 선행연구에서는 상호 정보(Mutual Information)와 같은 통계적인 정보[1-5]를 추출하여 띄어쓰기를 처리하는 방법이 연구되었다. 그리고 띄어쓰기와 품사 부착을 연속적인 문제로 보고 동시에 수행하는 다중 태스크 학습 방법도 연구되고 있다[6,7]. [6]에서는 Structural SVM(Support Vector Machine)을 이용해서 띄어쓰기 태그 및 품사 태그 부착을 동시에 수행 한다. [7]에서는 품사

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2015R1D1A1A01056907)

태그 부착 문제에 적합하다고 알려진 양방향 LSTM-CRF 모델에 음절 임베딩 외의 N-Gram 자질과 N-Gram 명사 자질을 추가하여 띄어쓰기 태그 및 품사 태그 부착을 동시에 수행한다.

문장 경계 인식 선행 연구에서 언어학적 지식을 많이 이용해야 하는 규칙 기반 방법이 많이 연구되었다[8,9]. [8]에서는 정규표현식을 이용하여 구두점의 유무로 문장 경계를 판단하였고 [9]에서는 자동으로 규칙을 추출하는 TBL(Transformation Based Learning)을 이용하여 경계를 판단한다. 그러나 규칙 기반 방법은 작성 및 유지가 힘들고 규칙을 정의 하는데 시간이 오래 걸린다는 문제점이 있다. 최근 연구에서는 구두점과 문장 종결 어미 주변의 단어에 대한 통계 정보와 후처리 규칙을 정의하여 기계학습 기반 방법으로 문장 경계 인식을 해결하고 있다[10, 11]. [10]에서는 Random Forest 알고리즘을 이용한 기계학습과 블로그 데이터만의 특징적인 문장 경계 형태에 후처리 규칙을 적용한다. [11]에서는 문장 경계의 후보로 삼을 수 있는 구두점을 추출하고 기계학습 기법을 사용하여 문장 경계의 자질을 학습한 문장 경계 인식 모델을 제안하였다.

자연어처리에서 띄어쓰기 문제는 음절 사이의 상호 정보를 자질로 사용하여 해결하고 있고 문장 경계 인식 문제는 구두점과 종결 어미 주변의 단어에 대한 통계, 규칙 정보를 자질로 사용하여 해결하고 있다.

본 논문에서는 띄어쓰기와 문장 경계 인식을 한 번에 처리하여 시간 복잡도 문제를 해결하고, 다중 손실 선형 결합을 사용하여 띄어쓰기와 문장 경계 인식이 다른 자질 사용하는 문제를 해결하는 다중 클래스 분류 시스템을 제안한다.

3. 제안 방법

본 논문에서 제안하는 모델은 띄어쓰기와 문장 경계 인식을 다중 클래스 분류 모델로 처리하여 시간 복잡도 문제를 해결하고, 다중 손실 선형 결합 방법을 사용하여 띄어쓰기와 문장 경계 인식이 다른 자질을 사용한다는 문제점을 해결하는 시스템을 제안한다. 그리고 자질 추가 방법 실험을 통해 제안 모델이 자질의 정보를 더욱 잘 반영한다는 것을 증명하였다. 제안하는 모델의 구성은 [그림 1]과 같다.

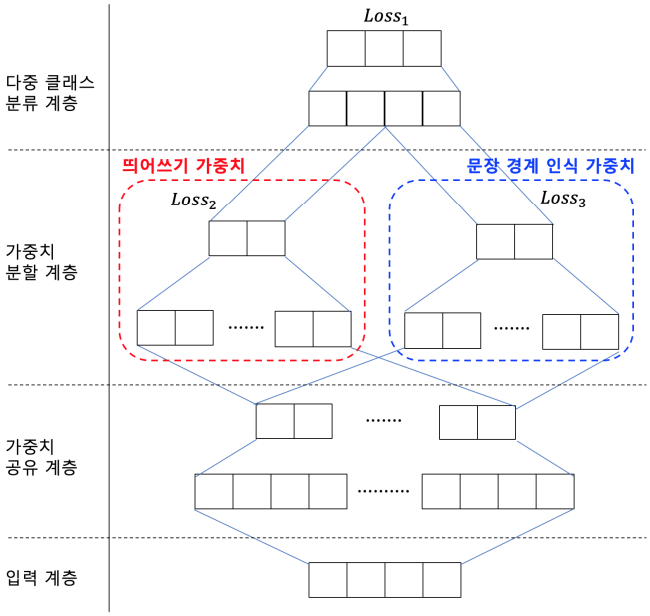

그림 1. 다중 손실 선형 결합을 이용한 공동 학습 모델

3.1 띄어쓰기와 문장 경계 인식 단일 모델

띄어쓰기와 문장 경계 인식 단일 모델은 멀티 레이어 퍼셉트론(Multi-Layer Perceptron)을 이용한 이진 분류 모델을 사용한다. 각각의 모델의 구성은 동일한 구성을 가지며 모델의 입력은 띄어쓰기와 문장 경계 인식에서 입력으로 공통적으로 많이 사용되는 연속된 음절단위 자질을 사용한다. 이때 연속된 음절이 같은 글자여도 그 사이 정보가 다른 경우가 있다. 예를 들어, '외과의사'와 '대부분의 사람'에서 현재 음절이 '의'라고 했을 때, 현재 음절과 다음 음절인 '의사'는 서로 연속된 음절이지만 두 문장에서 '의'와 '사' 사이의 띄어쓰기 정보는 다르다. 이처럼 연속된 음절이 같은 음절 쌍이여도 음절 사이 정보는 다를 수 있다고 판단하여, 연속된 음절 앞과 뒤 음절을 하나씩 추가하여 총 4개의 음절을 입력으로 사용한다.

3.2 다중 클래스 분류 모델

다중 클래스 분류는 여러 개의 클래스를 한 번에 분류하는 것을 의미한다. 앞선 띄어쓰기와 문장 경계 인식의 단일 모델을 합쳐 다중 클래스 분류 모델을 사용하여 시간 복잡도 문제를 해결한다. 다중 클래스 분류 모델의 구성은 단일 모델과 같으나 결과 부분에서 여러 개의 클래스를 한 번에 분류를 한다.

그러나 다중 클래스 분류 모델은 띄어쓰기와 문장 경계 인식이 같은 자질을 사용하기 때문에 띄어쓰기와 문장 경계 인식이 각각 다른 자질을 사용한다는 문제점을 해결 할 수 없고 띄어쓰기 손실과 문장 경계 인식 손실을 정확하게 반영할 수 없다는 단점이 있다.

3.3 다중 손실 선형 결합을 이용한 공동 학습 모델

다중 손실 선형 결합을 이용한 공동 학습(Joint Learning) 모델을 사용하여 다중 클래스 분류 모델의 단점을 해결한 모델을 제안한다. 다중 손실 선형 결합 방법을 이용해 띄어쓰기와 문장 경계 인식이 다른 자질을 사용하는 문제점을 해결했다. 공동 학습 모델에는 가중치 공유계층과 가중치 분할계층, 다중 클래스 분류계층이 있다. 공유계층에서는 띄어쓰기와 문장 경계 인식의 공통된 자질을 계산하고 가중치 분할계층에서는 띄어쓰기 가중치와 문장 경계 인식 가중치를 통해 각 태스크별 특정한 자질을 계산한다. 마지막으로 띄어쓰기 가중치와 문장 경계 인식 가중치에서 계산된 결과를 이용하여 다중 클래스 분류계층에서 최종 클래스 분류 결과가 나온다. 이때 최종 손실 계산은 크로스엔트로피(CrossEntropy)로 한다. $loss1$은 마지막 다중 클래스 분류 계층의 손실을 계산하고 $loss2$와 $loss3$는 띄어쓰기 가중치와 문장 경계 인식 가중치의 손실을 계산한다. $loss1$와 $loss2$, $loss3$을 더해 최종 손실을 구한다. 이때 마지막 다중 클래스 분류의 손실은 띄어쓰기 가중치와 문장 경계 인식 가중치의 손실 보다 결과에 큰 영향을 미치기 때문에 최종 손실에서 반영 비율(α)을 높게 반영했다. 공동 학습 모델의 다중 손실 계산은 다음 식으로 계산했다.

$$loss_{total} = \alpha \cdot loss1(\mathrm{P}redict_{All}) + \frac{1-\alpha}{2} \cdot loss2(\mathrm{P}redict_{wordspace}) + \frac{1-\alpha}{2} \cdot loss3(\mathrm{P}redict_{sentenceboundary})$$

3.4 자질 추가 방법론

본 논문에서 제안하는 모델은 띄어쓰기와 문장 경계 인식을 서로 다른 이진 분류 과정을 통해 띄어쓰기와 문장 경계 인식 각각의 자질을 추가 하는데 장점이 있다. 기존의 우수함이 증명된 띄어쓰기 자질[4]과 문장 경계 인식 자질[11]을 사용하여 다중 클래스 분류 모델과 다중 손실 선형 결합을 이용한 공동 학습 모델에 자질의 다양한 추가 방법을 통해 제안 모델의 장점을 증명한다.

4. 실험 및 성능 평가

4.1 실험환경

본 논문에서는 세종 말뭉치에서 무작위로 추출한 30만 문장 중 24만 문장을 학습에 사용했으며 6만 문장을 평가에 사용하였다. [표 1]은 학습 데이터 및 평가 데이터에 대한 통계 수치이다. 평가를 위해 모델의 띄어쓰기 성능 평가는 정확도(Accuracy)를 사용하였고 모델의 문장 경계 인식 성능 평가는 정확률(Precision), 재현율(Recall), F1-Measure를 사용하였다.

표 1. 학습 데이터 및 평가 데이터 통계

	학습 데이터	평가 데이터
문장 수	240,000	60,000
어절 수	3,957,922	990,279
음절 수	13,223,240	3,313,534

4.2 띄어쓰기 성능 평가

자질 추가 전 띄어쓰기 모델 성능 평가 결과와 자질 추가 후 띄어쓰기 모델 성능 평가 결과는 [표 2]와 [표 3]과 같다. 자질 추가 전 띄어쓰기 성능 평가 결과에서 이진 분류 모델보다 다중 손실을 이용한 공동 학습 모델의 성능이 정확도 1.03%p 증가 하였다. 자질 추가 후 이진 분류 과정에서 자질 추가한 모델이 이진 분류 단일 모델의 성능 보다 정확도 3.98%p 증가 하였다.

표 2. 자질 추가 전 띄어쓰기 모델 성능 평가 결과(%)

모델	정확도
이진 분류 단일 모델	91.32
기본 다중 클래스 분류 모델	91.63
다중 손실을 이용한 공동 학습 모델	92.35

표 3. 자질 추가 후 띄어쓰기 모델 성능 평가 결과(%)

모델	정확도
기본 다중 클래스 분류 모델 + 입력 계층에 자질 추가	92.09
다중 손실을 이용한 공동 학습 모델 + 입력 계층에 자질 추가	93.15
다중 손실을 이용한 공동 학습 모델 + 띄어쓰기 가중치 분할 계층에 자질 추가	95.30

4.3 문장 경계 인식 성능 평가

자질 추가 전 문장 경계 인식 모델 성능 평가 결과와 자질 추가 후 문장 경계 인식 모델 성능 평가 결과는 [표 4]와 [표 5]와 같다. 자질 추가 전 문장 경계 인식 성능 평가 결과에서 이진 분류 모델보다 다중 손실을 이용한 공동 학습 모델의 성능이 F1-Measure 0.04%p 증가 하였다. 자질 추가 후 이진 분류 과정에서 자질 추가한 모델이 이진 분류 단일 모델의 성능 보다 F1-Measure 0.34%p 증가 하였다.

표 4. 자질 추가 전 문장 경계 인식 모델 성능 평가 결과(%)

모델	정확률	재현율	F1
이진 분류 단일 모델	94.61	96.04	95.32
기본 다중 클래스 분류 모델	94.92	95.47	95.20
다중 손실을 이용한 공동 학습 모델	94.25	96.50	95.36

표 5. 자질 추가 후 문장 경계 인식 모델 성능 평가 결과(%)

모델	정확률	재현율	F1
기본 다중 클래스 분류 모델 + 입력 계층에 자질 추가	93.51	94.09	93.79
다중 손실을 이용한 공동 학습 모델 + 입력 계층에 자질 추가	95.15	95.77	95.45
다중 손실을 이용한 공동 학습 모델 + 문장 경계 인식 가중치 분할 계층에 자질 추가	95.05	96.29	95.66

4.4 모델 별 평균 수행 시간 평가 결과

[표 6]은 각 모델의 평균 소요 시간 결과이다. 다중 손실을 이용한 공동 학습 모델에 자질 추가를 해서 성능 향상이 있었음에도 이진 분류 모델 순차적 수행 시간보다 38.7% 줄어든 수행 시간을 보였다.

표 6. 띄어쓰기와 문장 경계 인식 평균 수행 시간

모델	평균 수행 시간
이진 분류 단일 모델 순차적 수행	8분 49초
기본 다중 클래스 분류 모델	4분 24초
다중 손실을 이용한 공동 학습 모델	5분 7초
다중 손실을 이용한 공동 학습 모델 + 자질 추가	5분 32초

5. 결론

선행 연구에서 띄어쓰기와 문장 경계 인식은 서로 다른 자질의 사용 때문에 서로 다른 모델을 사용했다. 그러나 띄어쓰기와 문장 경계 인식은 하나가 성립되면 다른 하나가 성립할 수 없는 상호 배타적 관계로 완전히 다른 문제라고는 볼 수 없다. 본 논문에서는 띄어쓰기와 문장 경계 인식을 합친 다중 클래스 분류 모델을 사용하여 띄어쓰기와 문장 경계 인식 각각의 모델의 순차적 수행 시간보다 38.7% 줄어든 수행 시간을 보였다. 또한 다중 손실 선형 결합을 사용하여 띄어쓰기와 문장 경계 인식이 서로 다른 자질을 사용한다는 문제점을 해결하여 띄어쓰기 이진 분류 모델과 문장 경계 인식 이진 분류 모델 성능보다 각각 3.98%p, 0.34%p 성능 향상을 보여 최종적으로 띄어쓰기 정확도 95.30%, 문장 경계 인식 F1-Measure 95.66%을 보였다.

향후 성능 향상을 위해 학습 데이터를 확장하고 자질에 대한 연구를 추가적으로 진행할 것이다.

참고문헌

[1] 심광섭, “CRF를 이용한 한국어 자동 띄어쓰기”, 인지과학 제22권 제2호, pp.217-233, 2011.

[2] 이창기, 김현기, “Structural SVM을 이용한 한국어 자동 띄어쓰기”, 한국정보과학회 학술발표논문집 제39권 제1호(B), pp.270-272, 2012.

[3] 최성자, 강미영, 권혁철, “통계 정보를 이용한 한국어 자동 띄어쓰기 시스템의 성능 개선”, 한국정보과학회 학술발표논문집 제31권 제1호(B), pp.883-885, 2004.

[4] 심광섭, “음절간 상호 정보를 이용한 한국어 자동 띄어쓰기”, 정보과학회논문지(B) 제23권 제9호, pp.991-1000, 1996.

[5] 김신일, 양선, 고영중, “메모리 제약적 기기를 위한 음절 패턴 기반 띄어쓰기 시스템”, 정보과학회 논문지 : 소프트웨어 및 응용, 제37권 제8호, pp.653-658, 2010.

[6] 이창기, “Structural SVM을 이용한 한국어 띄어쓰기 및 품사 태깅 결합 모델“, 정보과학회논문지 : 소프트웨어 및 응용 제40권 제12호, pp.826-832, 2013.

[7] 김선우, 최성필, “Bidirectional LSTM-CRF 기반의 음절 단위 한국어 품사 태깅 및 띄어쓰기 통합 모델 연구”, 한국정보과학회논문지 제45권 제8호, pp.792-800, 2018.

[8] Gregory Grefenstette, Pasi Tapanainen “What is a word, What is a sentence? Problems of Tokenization”, The 3rd International Conference on Computational Lexicography, pp.79-87, 1994.

[9] E.Stamatatos, N.Fakotakis, G.Kokkinakis, “Automatic Extraction of Rules for Sentence Boundary Disambiguation”, The ECCAI Advanced Course on Artificial Intelligence, 1999.

[10] 이주은, 구민서, 김선홍, 신효섭, “블로그 데이터에 대한 문장 경계 인식”, 한국HCI학회 학술대회, pp.1221-1223, 2014.

[11] 박수혁, 임해창, “기계학습 기법을 이용한 문장경계인식”, 한국정보처리학회 춘계학술발표대회 논문집 제15권 제1호, 2008.