자동 띄어쓰기에서 글쓴이 의도를 반영한 자질의 활용

이정욱○, 김재훈
한국해양대학교, 컴퓨터공학과 및 해양인공지능융합전공
wjddnr96177@naver.com, jhoon@kmou.ac.kr

Exploiting Features of Writer's Intent in Automatic Spacing

Jeong-wook Lee○, Jae-Hoon Kim

Dept. of Computer Engineering and Interdisciplinary Major of Maritime AI Convergence,
Korea Maritime & Ocean University

요 약

띄어쓰기에 대한 오류는 한국어 처리 전반에 영향을 주므로 자동 띄어쓰기는 필수적인 요소이다. 글쓴이의 대부분은 띄어쓰기 오류를 범하지 않으므로 글쓴이의 의도가 띄어쓰기 시스템에 반영되어야 한다. 그러나 대부분의 자동 띄어쓰기 시스템은 모든 띄어쓰기 정보를 제거하고 새로이 공백문자를 추가하는 방법으로 띄어쓰기를 수행한다. 이런 문제를 완화하기 위해서 본 논문에서는 기계학습에서 글쓴이의 의도가 반영된 자질을 추가하는 방법을 제안한다. 실험을 위해서 CRFs(Conditional Random Fields)를 사용하여 기존 시스템과 사용자의 의도를 반영한 띄어쓰기 시스템과의 성능을 비교하고 분석한다.

주제어: 자동 띄어쓰기, 글쓴이 의도, 자질 추출, CRF

1. 서론

띄어쓰기는 문장의 내용을 쉽게 이해하고 뜻을 정확하게 전달하기 위해 어절 간의 경계를 구분하는 문제이다. 동일한 음절을 사용한 문장이라도 띄어쓰기에 따라 의미가 달라진다. 예를 들어 "엄마는서울시어머니합창단"의 경우, 띄어쓰기에 따라 "엄마는 서울시 어머니합창단"이라는 문장과 "엄마는 서울 시어머니 합창단"이라는 문장을 생성할 수 있다. 이와 같이 띄어쓰기에 따라 문장의 뜻이 완전히 달라지기 때문에 올바른 띄어쓰기는 문장을 이해하는데 필수적이다. 따라서 대부분의 자연어 처리 연구는 올바른 띄어쓰기를 전제하고 있고, 띄어쓰기에 오류가 있을 경우, 띄어쓰기 시스템의 성능에 영향을 줄 수 있다.

대부분의 한국어 자동 띄어쓰기 연구들은 원시 코퍼스(raw corpus)의 띄어쓰기가 완벽하다는 가정 하에 이를 학습데이터로 사용한다. 학습데이터에 띄어쓰기 정답만 존재하기 때문에 글쓴이가 의도한 띄어쓰기 정보를 자질로 사용할 수 없다. 이에 기존의 자동 띄어쓰기 시스템은 입력문장의 띄어쓰기를 제거하고 문장에 공백문자를 삽입하는 방식으로 작동된다[1-3]. 이런 시스템들은 올바른 입력 문장에도 공백문자를 잘못 삽입하는 문제가 발생한다는 단점이 있다. 이러한 문제를 해결하기 위해 [4]는 글쓴이의 띄어쓰기를 최소한으로 수정할 수 있도록 개선하였고, [5]는 글쓴이 의도의 N-gram 확률을 통해 띄어쓰기를 개선하였다.

본 논문에서는 글쓴이 의도의 N-gram 정보를 기계학습의 자질로 추가하는 방법을 제안하고, 띄어쓰기 속도와 정확도를 고려하여 CRF 모델에 적용한다. 글쓴이의 의도가 자동 띄어쓰기 시스템에 미치는 영향을 분석하기 위해서 다양한 실험을 수행한다. 띄어쓰기 정보를 자질로 추가함으로써 99.15의 정확도를 보였으며, 이 결과는 기존의 CRF 모델보다 1.1%p가 향상되었다.

본 논문의 구성은 2장에서 기존의 띄어쓰기 시스템을 살펴보고, 3장에서는 글쓴이 의도의 자질 집합에 대해서 설명한다. 4장에서 실험을 통해서 글쓴이 의도의 영향을 분석하고, 5장에서 결론을 맺는다.

2. 관련 연구

기존의 한국어 띄어쓰기 연구는 크게 규칙 기반 방식과 확률 및 통계 기반 방식으로 나눌 수 있고, 최근 기계학습과 딥러닝 기술에 대한 관심이 높아짐에 따라 이를 사용한 자동 띄어쓰기 연구도 활발히 진행되고 있다.

규칙 기반 방식은 전문적인 언어학적 지식을 갖춘 사람이 직접 규칙을 작성하고 유지한다는 점에서 한계가 있다[6]. 통계 기반 방식은 음절 N-gram을 사용하여 데이터에서 추출한 확률정보를 통해 띄어쓰기를 교정한다. 기존의 통계 기반 방식의 연구는 글쓴이가 의도한 띄어쓰기 정보를 무시하고 입력문장에 공백을 제거한 후 새로운 띄어쓰기 생성하기 때문에 띄어쓰기에 따라 의미가 달라지는 문장에 대해 하나의 결과만 생성한다는 단점이 있다. 최근 띄어쓰기 문제를 순차적 레이블링 문제로 보고 CRFs, Structual SVM 등 기계학습 모델을 적용한 연구가 진행되고 있다[4]. 더 나아가 높은 성능을 보이는 딥러닝 모델을 사용하여 순차적 레이블링 문제에 적합한 RNN(Recurrent Neural Network), LSTM(Long Short Term Memory) 등에 적용하는 연구가 진행 중이다[7-8].

3. 글쓴이 의도를 위한 자질 집합

표 1. 자동 띄어쓰기에서 음절의 N-gram 자질 집합

x_t	x_t의 형태	x_{t-1}	x_t	x_{t+1}	$x_{t-2,t-1}$	$x_{t-1,t}$	$x_{t,t+1}$	$x_{t+1,t+2}$	$x_{t-2,t}$	$x_{t-1,t+1}$	$x_{t,t+2}$
줄	한글	$	줄	이	$$	$줄	줄이	이길	$$줄	$줄이	줄이길
이	한글	줄	이	길	$줄	줄이	이길	길어	$줄이	줄이길	이길어
길	한글	이	길	어	줄이	이길	길어	어서	줄이길	이길어	길어서

※ x_t = 현재음절 · $ = <EOS>

일반적인 자동 띄어쓰기 시스템이 자질 집합은 음절의 n-그램을 사용한다. 표 1은 문장 “줄이 길어서 늦었다.”에 대한 음절의 n-그램 자질 집합의 일부를 보이고 있다. 여기서 x_t은 현재 음절이고, 문장의 앞이거나 뒤에서 기호 $(EOS, End of Sentence)를 추가한다. $x_{i,j}$ $(i>j)$는 음절 x_i에서 x_j까지의 문자열이다. 자질 집합은 크게 음절의 형태(x_t의 형태), 1-그램(x_{t-1}, x_t, x_{t+1}), 2-그램($x_{t-2,t-1}$, $x_{t-1,t}$, $x_{t,t+1}$, $x_{t+1,t+2}$), 3-그램($x_{t-2,t}$, $x_{t-1,t+1}$, $x_{t,t+2}$)으로 분류한다. 음절의 형태는 한글, 영어, 숫자와 같은 종류를 나타낸다.

한편, 글쓴이 의도를 위한 자질 집합을 생성하기 [5]에서 제안한 공백 정보 표기 방법을 이용한다. 여기서 s_t는 음절 x_t의 공백 정보, 즉 띄어쓰기 정보이다. s_t가 1이면 음절 x_t앞에 공백이 있음을 의미하고 0이면 공백이 없음의 나타낸다.

표 2. 입력 문장에 대한 공백 띄어쓰기 정보의 표기

x_t	$	$	줄	이	길	어	서	늦	었	다	.	$	$
s_t	1	1	1	0	1	0	0	1	0	0	1	1	1

표 3은 표 2를 기반으로 띄어쓰기 자질을 추출한 예이다. 공백정보의 자질 집합은 2-그램($s_{t,t+1}$), 3-그램($s_{t-1,t+1}$, $s_{t,t+2}$), 4-그램($s_{t-2,t+1}$, $s_{t-1,t+2}$, $s_{t,t+3}$)으로 분류된다. 자질 $s_{t,t+1}$은 현재 음절 x_t의 앞과 뒤의 띄어쓰기 정보를 의미하고, 자질 $s_{t-1,t+1}$은 음절 x_{t-1}의 앞에서 현재 음절 x_t의 뒤까지의 띄어쓰기 정보를 의미한다. 나머지 자질도 이와 같은 의미로 해석될 수 있다.

제안된 시스템에서 사용될 자질 개수는 총 17개이며, 문장에 대한 n-그램과 띄어쓰기 정보를 자질로 사용하여 문장을 올바르게 띄어쓰기 할 수 있도록 모델링한다.

4. 실험 및 평가

4.1 실험 환경

훈련 및 평가를 위해 세종 말뭉치로부터 700,000 문장을 임의로 추출하였다. 이 중에 500,000문장을 훈련을 위해 사용하였고, 나머지 200,000문장을 평가를 위해 사용하였다. 평가 척도는 띄어쓰기에서 널리 사용되는 정확도(accurary)를 사용한다.

4.2 글쓴이 의도 자질의 성능 평가

표 4는 글쓴이의 의도 자질에 대한 성능을 보이고 있다. 표 4에서 n-그램 자질과 의도 자질은 각각 3장에서 언급한 자질 집합이다.

표 4. 의도 자질을 반영한 성능 평가

의도 자질	자질 집합	학습문장의 오류율	정확률
사용 않음	n-그램 자질		98.05
사용함	n-그램 자질 + 의도 자질	0%	100.00
		10%	99.15
		20%	99.04
		30%	98.92
		40%	97.32

의도 자질의 경우, 원시 코퍼스의(raw corpus)의 띄어쓰기가 완벽하다고 가정하므로 학습 과정에서 n-그램 자

표 3. 띄어쓰기를 위한 글쓴이의 의도 자질

	$s_{t,t+1}$	$s_{t-1,t+1}$	$s_{t,t+2}$	$s_{t-2,t+1}$	$s_{t-1,t+2}$	$s_{t,t+3}$
줄	10	110	101	1110	1101	1010
이	01	101	010	1101	1010	0100
길	10	010	100	1010	0100	1001

질 집합의 정보를 무시하고 오로지 의도 자질 집합으로만 학습된다. 평가 말뭉치가 학습 말뭉치와 완전히 다른 성격의 문장들이라 할지라도 평가 문장에 대한 의도 자질 집합으로 시스템의 결과를 얻을 수 있다(표 4의 0%의 오류율 참조). 이와 같은 시스템은 입력 문장에 오류가 포함되어 있다면 오류 그 자체를 결과로 출력하므로 자동 띄어쓰기 시스템이라고 말할 수 없다. 이 문제를 해결하는 가장 좋은 방법으로는 SNS로부터 문장을 수집하여 띄어쓰기 말뭉치를 구축하고 그 결과를 이용해서 학습하는 것이다. 그러나 현실적으로 띄어쓰기 말뭉치가 없을 뿐 아니라 이를 구축하는 데는 많은 비용과 시간이 소요된다. 따라서 본 논문에서는 학습 말뭉치에 속한 문장에 임의의 오류(10%, 20%, 30%, 40%)를 생성하여 성능을 평가하였다(표 4 참조). 표 4에서 볼 수 있듯이 학습 말뭉치의 오류가 증가하면 상대적으로 정확률은 떨어진다. 어느 정도의 오류를 생성하여 학습해야 실제 환경에 가장 적합할까 하는 새로운 문제가 발생하였다. 객관적으로 한국어 문서에서 띄어쓰기 오류율이 있을 경우에 그 결과를 사용할 수 있다. 외국인 학습자에 대한 띄어쓰기 오류에 대한 분석 결과는 있으나[11] 보편적인 한국인에 대한 띄어쓰기 오류는 찾을 수 없었다. 따라서 본 논문에서는 [5]에서와 같이 10%로 설정하였다. 오류율 10%에 대해서는 본 논문의 의도대로 글쓴이의 의도를 포함한 띄어쓰기 결과가 제대로 작동한다는 것을 알 수 있었고 n-그램 자질만 사용했을 때 보다 1.1%p 높은 99.15의 정확도를 얻을 수 있었다.

4.3 띄어쓰기 시스템의 비교 분석

표 5는 제안된 시스템과 기존의 띄어쓰기 시스템과 비교하고 분석하고자 한다. 연구마다 실험 환경이 다소 차이가 있어서 성능에 대한 객관적인 비교는 다소 어려울 것이다.

표 5. 띄어쓰기 시스템 성능 평가 비교

띄어쓰기 시스템	정확률	글쓴이 의도 반영
제안된 시스템	99.15	반영함.
n-그램 확률[5]	99.05	반영함.
structural SVM[4]	99.64	반영함.
CRF[9]	98.84	반영 안 함.
BiLSTM-CRF[10]	97.17	반영 안 함.
BERT[7]	98.14	반영 안 함.

표 5에서 보는 바와 같이 전체적으로 의도 자질을 반영한 경우가 더 좋은 결과를 보임을 알 수 있었다. 제안된 시스템은 기본적인 자질은 [5]에서 제시한 자질을 CRF에 적합하도록 개선하였다. [5]는 n-그램 확률의 선형 결합(linear combination)을 이용하고 있으며 각 항목의 가중치를 경험적으로 조정하고 있다. [4]는 가장 좋은 성능을 보이고 있으며 의도를 분별학습(discriminative learning)를 통해서 모델을 학습한다. [9]는 CRF 모델을 사용하지만 의도 자질은 사용하지 않는다. [10, 7]은 심층학습 모델을 사용하고 있다. [10]은 종단 간 심층신경망을 사용하고 있으며 특별한 학습 말뭉치를 구축할 필요가 없다는 특징이 있다. [7]은 심층학습의 전이학습을 이용하고 있다.

4.4 띄어쓰기 시스템의 성능 비교를 위한 T-test

[8]에서 제시한 음절의 n-그램 자질 집합을 사용한 CRF모델과 본 논문에서 제안한 CRF모델의 비교를 위해 K-fold cross-validated independent t-test를 사용하였다. K-fold cross-validated independent t-test은 두 모델의 성능을 비교하기 위한 일반적인 방법이며 귀무가설은 '기존 모델의 성능이 더 높다'로 설정하였다. 두 모델의 성능 평균은 각각 98.9, 97.2로 제안한 모델의 성능 평균(Macro-average)이 더 높았고 95% 신뢰수준에서 검정하였을 때 p-value가 유의 수준인 0.05보다 작은 0.00125로 귀무가설이 기각되어 대립가설이 채택되었다. 결과적으로 기존의 CRF모델[8]과 비교하여 본 논문에서 제안된 CRF모델이 유의미한 성능 차이를 보인다는 것을 알 수 있다.

5. 결론

기존의 띄어쓰기 모델은 입력문장에 대해 공백을 제거하여 글쓴이의 의도를 제외하고 공백을 다시 삽입하는 방법으로 시스템을 구현했다. 본 논문에서 제안된 모델은 글쓴이의 의도를 자질로 사용하여 띄어쓰기가 민감한 부분을 잘 처리함을 알 수 있었다. 제안된 모델은 심층학습 모델에 비해서도 큰 성능의 차이가 없을 뿐 아니라 모델의 크기나 처리 속도 등에서 큰 장점을 보인다. [8]에서 제시한

참고문헌

[1] 강승식, "한글 문장의 자동 띄어쓰기를 위한 어절 블록 양방향 알고리즘", 정보과학회논문지: 소프트웨어 및 응용, 제27권, 제4호, pp. 441-447, 2000

[2] 이호준, 박종철, "음절단위 결합범주문법을 이용한 한국어 문장의 자동 띄어쓰기", 제14회 한글 및 한국어 정보처리 학술대회 논문집, pp. 47-54, 2002

[3] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회 논문지(B), 제23권, 제9호, pp. 991-1000,1996

[4] 이창기, "사용자가 입력한 띄어쓰기 정보를 이용한 Structural SVM 기반 한국어 띄어쓰기", 정보과학회 논문지: 컴퓨팅의 실제 및 레터, 제20권, 제5호, pp. 301-305, 2014

[5] 박서연, 옥철영, "사용자의 입력 의도를 반영한 음

절 N-gram 기반 한국어 띄어쓰기 및 붙여쓰기 오류 교정 시스템", 정보과학회 컴퓨팅의 실제 논문지, 제27권, 제3호, pp.145-150, 2021
[6] 심광섭, "말뭉치와 형태소 분석기를 확용한 한국어 자동 띄어쓰기", 정보과학회논문지, 제42권, 제1호, pp. 68-75, 2015
[7] 황태욱, 정상근, "BERT를 이용한 한국어 자동 띄어쓰기", 정보과학회 학술발표논문집, pp. 374-376, 2019
[8] 윤호, 김재훈, "CRFs와 Bi-LSTM/CRFs의 비교 분석: 자동 띄어쓰기 관점에서", 한글 및 한국어 정보처리 학술대회 논문집, pp. 189-192, 2018
[9] 심광섭, "CRF를 이용한 한국어 자동 띄어쓰기", 한국인지과학회, 제22권, 제2호, pp.217-233, 2011
[10] 이현영, 강승식, "종단 간 심층 신경망을 이용한 한국어 문장 자동 띄어쓰기", 정보처리학회논문지. 소프트웨어 및 데이터 공학, 제8권, 제11호, pp.441-448, 2019
[11] 황란아, 심현주, “한국어 학습자의 띄어쓰기 오류 분석 연구: 띄어쓰기 교육 항목의 구체적인 제시를 위하여”, 언어과학연구, 제82권, pp. 429-451, 2017.