한국어 SNS 문서에 적합한 문장 경계 인식

염하람○, 김재훈
한국해양대학교, 컴퓨터공학과 및 해양인공지능융합전공
rami7878@naver.com, jhoon@kmou.ac.kr

Robust Sentence Boundary Detection for Korean SNS Documents

Haram-Yeom○, Jae-Hoon Kim

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

요 약

다양한 SNS 플랫폼이 등장하고, 이용자 수가 급증함에 따라 온라인에서 얻을 수 있는 정보의 활용 가치가 높아지고 있다. 문장은 자연어 처리 시스템의 기본적인 단위이므로 주어진 문서로부터 문장의 경계를 인식하는 작업이 필수적이다. 공개된 문장 경계 인식기는 SNS 문서에서 좋은 성능을 보이지 않는다. 본 논문에서는 문어체로 구성된 일반 문서뿐 아니라 SNS 문서에서 사용할 수 있는 문장 경계 인식기를 제안한다. 본 논문에서는 SNS 문서에 적용하기 위해 다음과 같은 두 가지를 개선한다. 1) 학습 말뭉치를 일반 문서와 SNS 문서 두 영역으로 확장하고, 2) 이모티콘을 사용하는 SNS 문서의 특징을 반영하는 어절의 유형을 자질로 추가하여 성능을 개선한다. 실험을 통해서 추가된 자질의 기여도를 분석하고, 또한 기존의 한국어 문장 경계 인식기와 제안한 모델의 성능을 비교·분석하였다. 개선된 모델은 일반 문서에서 99.1%의 재현율을 보이며, SNS 문서에서 88.4%의 재현율을 보였다. 두 영역 모두에서 문장 경계 인식이 잘 이루어지는 것을 확인할 수 있었다.

주제어: 문장 경계 인식, 기계학습, Feature Selection

1. 서론

문장이란 생각이나 감정을 말로 표현할 때 완결된 내용을 나타내는 최소의 단위로, 자연어 처리의 주요 작업들에서 문장이 기본적인 처리 단위가 된다. 문장 단위의 작업을 수행하기 위해서는 우선 문장의 경계를 인식하는 것이 필요하다[1].

최근 다양한 SNS 플랫폼이 등장하고, 이용자의 수가 급증함에 따라 온라인에서 얻을 수 있는 다양한 분야의 정보 활용 가치가 높아지고 있다. 일반적으로 문어체는 비교적 정확한 구두점을 사용하며, 제한된 종결어미를 사용하고 있으므로 비교적 간단하게 문장 경계를 인식할 수 있다. 기존에 공개된 여러 문장 경계 인식기[1]는 일반 문서에 대해 높은 성능을 보인다. 반면 댓글이나 SNS 문서에는 좋은 성능을 보이지 않는다. 그 이유는 SNS 문서의 다음과 같은 특징을 고려하지 않았기 때문이다. 1) 구두점이 자주 생략된다. 2) 특수기호나 자소를 이용한 다양한 이모티콘이 사용된다. 3) 유행어, 줄임말 등 새로운 형태의 종결어미가 자주 사용된다.

본 논문은 문어체뿐 아니라 SNS에서 사용되는 다양한 문체에서도 잘 적용되는 문장 경계 인식기를 제안한다. 제안된 문장 경계 인식기는 SNS 문서에 잘 적용되도록 다음의 두 가지를 개선한다. 첫째로 학습 말뭉치를 확장한다. 학습 말뭉치를 일반 문서로만 제한하지 않고, SNS 문서도 함께 학습할 수 있도록 학습 말뭉치에 SNS 문서를 추가한다. 둘째로 자질 집합을 확장한다. 기존의 문장 경계 인식기의 자질을 이용하여 말뭉치의 범위만 늘리는 것으로는 성능을 높이기에 한계가 있어 각종 자소, 특수기호, 이모티콘을 추가하여 다양한 형식으로 문장이 종결되는 SNS 문서에 적합하도록 개선한다. 이와 같은 개선법을 적용하여 CRF로 학습한 제안된 모델은 일반 문서뿐 아니라 SNS 문서에서도 좋은 성능을 보였다.

본 논문의 구성을 다음과 같다. 2장에서 문장 분리에 관련된 연구를 조사하고, 3장에서 SNS 문서에서의 다양한 고려사항을 자세히 기술한다. 4장에서는 각각 제안된 모델을 평가하고 5장에서 결론을 맺고 향후 연구에 대해 기술한다.

2. 관련 연구

Riley의 연구[2]에서는 구두점 주변 단어의 출현 확률과 구두점이 발견된 어절의 클래스를 자질로 추출하여 AP News의 2,500만 단어를 이용하여 학습한 Decision Tree(C4.5)를 이용하였고, Brown 말뭉치에서 99.8%의 정확률을 보였다. 임희석과 한군희의 연구[3]에서는 후보 구두점 자체, 후보 구두점의 앞/뒤에 나타난 음절, 문장 시작과 후보 구두점이 나타난 위치까지의 거리 등을 자질로 사용하였다. 문장 정확률은 98.82%, 문장 재현율은 99.09%의 정확률을 보였다. 이후 비전문가가 작성한 형식이 자유로운 웹 문서들에서 문장 경계를 인식하는 연구가 수행되었다. 서정연과 김주희의 연구[1]에서는 음절 n-gram을 사용하여 자질을 추출하여 통계 기반 기계

1) https://github.com/likejazz/korean-sentence-splitter

학습 기법으로 CRFs를 이용하였다. 세종말뭉치에서 99.99%의 정확률과 100.0%의 재현율을 보였고, 구두점을 제거한 경우에도 96.20%의 정확률과 87.51%의 재현율을 보였다. 이충희 외[4]에서는 문장 경계로 사용될 수 있는 모든 종결어미를 대상으로 후보를 만들어 자질로 이용하여 문장 경계 인식을 수행하였고, 96.3%의 성능을 보였다. 이주은 외[5]에서는 문장 경계 인식기를 2단계로 나누었다. 1단계로 Random Forest 패키지를 이용하여 문장 경계를 추정하였다, 2단계로 후처리 규칙을 적용하여 성능을 높였다. 문장 경계 인식기는 기계학습 후 93.3%의 정확률과 96.4%의 재현율을 보였으며, 후처리 규칙 적용 후 94.2%의 정확률과 100.0%의 재현율을 보였다.

3. SNS 문서에서의 고려사항

일반적으로 많은 문장 경계 인식기는 문어체를 위주로 제안되었으며 SNS 문서에는 그다지 좋은 성능이 보이지 않는다. 본 논문에서는 문어체뿐 아니라 SNS 문서에 잘 적용되도록 다음의 두 가지를 고려한다. 첫째는 학습 말뭉치의 확장이다. 이처럼 적용 영역이 변경되었을 경우에는 변경된 영역의 학습 자료를 추가함으로써 성능을 개선할 수 있다. 따라서 본 논문에서는 학습 말뭉치에 포함된 문서를 일반 문서로 제한하지 않고, SNS 문서도 포함되도록 학습 말뭉치를 확장한다. 둘째는 자질 집합의 확장이다. 기존의 문장 경계 인식기는 주로 문어체에 적합한 자질 집합을 주로 사용하였다. 본 논문에서는 말뭉치 확장만으로는 성능의 개선에 한계가 있어 각종 자소, 특수기호, 이모티콘을 자질 집합으로 추가하여 다양한 형식으로 종결되는 SNS 문서에 적합하도록 개선한다.

3.1 말뭉치 확장

문장 경계 인식기가 일반 문서와 SNS 문서 두 영역 모두에서 좋은 성능을 나타낼 수 있도록 학습 말뭉치를 두 영역으로 확장한다. 일반 문서의 학습 말뭉치로는 KCC 말뭉치[2](KCC)를 사용하고 SNS 문서의 학습 말뭉치로는 NSMC 말뭉치[3](NSMC)를 사용한다. KCC에는 다양한 형식의 인용문이 포함되어 있다. 하나의 문장 안에 문장의 인용문이 포함되어 있을 경우에는 각각의 인용문도 하나의 문장으로 인식될 수 있도록 부분적으로 수정하여 사용한다. NSMC는 ID 별로 작은 문서로 구성되어 있으며 각 문서는 일반적으로 여러 문장으로 구성되어 있다. 이와 같은 SNS 문서를 문장으로 분리하여 학습 말뭉치로 사용한다. 또한 NSMC에는 영화 리뷰 데이터이므로 '영화', '드라마' 등 명사로 끝나는 문장이 다수 있어서 모델의 편향을 방지하기 위해 이와 같은 문장을 제거한다.

2) http://nlp.kookmin.ac.kr/kcc/
Korean Contemporary Corpus of Written Sentences
3) https://github.com/e9t/nsmc
Naver sentiment movie corpus v1.0

표 1. SNS 문장 분리 및 제거 예시

	원래문장	전처리 후 문장	비고
문장1	쓰레기 영화	-	제거
문장2	뭐 말할것도없다장국영 자살이 안타깝다	뭐 말할것도없다	분리
		장국영 자살이 안타깝다	
문장3	와 진짜 재미없다. 최악	와 진짜 재미없다.	제거
문장4	걍노잼ㅡㅡ 지루해죽는줄알았다.	걍노잼ㅡㅡ	분리
		지루해죽는줄알았다.	

표 1은 SNS 문서를 문장으로 분리할 때의 예를 보여준다. 문장 1은 '영화'라는 명사로 종결되어 제거하였다. 문장 2와 같이 2개의 문장이지만 띄어쓰기 없이 붙어있는 경우 2개의 문장으로 분리하였다. 문장 3은 두 개의 문장으로 분리 후 명사로만 이루어져 있는 문장을 제거하였다. 문장 4와 같이 이모티콘으로 종결된 문장 뒤에 새로운 문장이 나온 경우 분리하였다. 이와 같은 SNS 문장 분리는 연구원들과 함께 진행하였으며 깃허브에 공개할 예정이다.

3.2 자질 추가

일반 문서와 SNS 문서 모두에서 좋은 성능을 보일 수 있게 하도록 8개의 자질 후보를 사용하였다.

(1) **F1 혹은 PUNCT** - 문장종결에 사용된 기호 (!?.):
일반적으로 문장의 종결에 많이 나타나는 구두점 정보인 ". ? !"가 어절의 마지막 음절로 나타났을 경우, 자질로 사용한다.

(2) **F2 혹은 SHAPE** - 문장종결에 사용된 어절의 유형:
문장의 종결에 사용된 어절의 유형을 표 1과 같이 6가지로 나누어 정의한 어절 정보이다. 자소, 특수문자를 문장종결에 자주 사용하는 SNS 문서를 위한 유형으로 자소, mixcase, 기타로 구분한다.

표 2. 어절의 유형과 그 예

어절의 유형(SHAPE)	Example
Number(숫자)	0-9
Hangul(한글)	가-힣
English(영어)	A-Z, a-z
Jaso(자소)	ㅠㅠㅠ, ㅋㅋ
Mixedcase(혼합)	헐…., ♥_♥
Other(기타 특수문자)	

(3) **F3 ~ 5** - 문장 경계 앞 1, 2, 3번째 음절:
문장 경계 앞의 1번째 음절 정보와 1, 2번째 음절

정보, 1, 2, 3번째의 음절 정보이다.

(4) **F6 ~ 8** - 문장 경계 뒤 1, 2, 3번째 음절: 문장 경계 뒤의 1번째 음절 정보와 1, 2번째 음절 정보, 1, 2, 3번째의 음절 정보이다.

표 2는 문장의 일부인 "어용! 재미는"에 대해 자질 F3 ~ 8의 예를 보이고 있다. 자질은 어절 경계에서 추출되는데 색인 0이 기준이며 어절의 마지막 음절이다. 이를 기준으로 왼쪽으로는 음수 색인을 사용하고 오른쪽 즉 다음 어절에 대해서는 양수 색인을 사용한다.

표 3. 자질 F3~8의 예

색인	…	-2	-1	0	1	2	3	…
음절	…	어	용	!	재	미	는	…
문장경계	…	N	N	Y	N	N	N	…
F3 : [!]				F6 : [재]				
F4 : [용!]				F7 : [재미]				
F5 : [어용!]				F8 : [재미는]				

4. 실험 및 분석

4.1 실험 환경

훈련 및 평가를 위해 3장에서 기술한 말뭉치 구축 방법에 따라 일반 문서와 SNS 문서에서 문장을 추출하여 전체 말뭉치를 구성하였다. 전체 말뭉치 중에서 훈련 및 평가 말뭉치를 표 3과 같이 분리하였다.

표 4. 말뭉치 구성

구분	학습 말뭉치	평가 말뭉치
KCC	100,000	10,000
NSMC	20,000	8,000

문장 경계 인식기의 성능 평가 척도는 각 영역에서의 문장 재현율(Re)과 문장 정확률(Pr)을 사용하며 각각 식 (1)과 식 (2)과 같이 정의된다.

$$\text{문장재현율} = \frac{\text{시스템이 추출한 정답 문장 수}}{\text{전체 문장 수}} \quad (1)$$

$$\text{문장정확률} = \frac{\text{시스템이 추출한 정답 문장 수}}{\text{시스템이 추출한 문장 수}} \quad (2)$$

각 영역의 재현율을 하나의 척도로 표현하기 위해서 Macro-AverageRecall(MAR)과 Micro-AverageRecall(mAR)을 사용하며 각각 식 (3)과 (4)과 같이 정의된다.

$$MAR = \frac{KCC\text{문장재현율} + NSMC\text{문장재현율}}{2} \quad (3)$$

$$mAR = \frac{KCC\text{정답 문장 수} + NSMC\text{정답 문장 수}}{\text{전체 문장수}} \quad (4)$$

4.2 학습 말뭉치 재구성

학습 말뭉치는 일반 문서와 SNS 문서의 어떤 순으로 배열되느냐에 따라 모델의 성능에 영향을 준다. 본 논문에서는 일반 문서의 100,000 문장 뒤에 SNS 문서의 20,000 문장이 순차적으로 나오도록 구성한 것(C1)과 임의적으로 문장을 추출하여 같은 비율을 이루도록 구성한 것(C2)의 성능을 비교해 보았으며 그 결과는 표 4와 같다.

표 5 학습 말뭉치의 재구성에 따른 성능 평가

말뭉치	KCC Re	KCC Pr	NSMC Re	NSMC Pr	MAR	mAR
C1	0.988	0.983	0.880	0.932	0.934	0.940
C2	**0.991**	**0.989**	**0.884**	**0.939**	**0.938**	**0.943**

실험 결과, 일반 문서와 SNS 문서 모두에서 C2가 C1보다 좋은 성능을 보였다. 이후 시행될 실험에서 C2를 사용하여 실험을 진행하였다.

4.3 자질의 기여도

기계학습 방법에서는 어떤 자질을 사용하느냐에 따라 성능이 달라지므로 각 자질이 성능에 끼치는 영향을 분석해 보았다[4]. 자질의 기여도는 모든 자질에서 자질을 하나씩 제외하는 자질 삭제 방법(feature ablation study)을 이용한다. F3은 필수 자질로 삭제 자질에 포함되지 않았으며 그 결과는 표 5와 같다.

표 6 자질의 기여도 분석

삭제 자질	KCC Re	KCC Pr	NSMC Re	NSMC Pr	MAR	mAR	Impr. (%)
∅ 기준	**0.991**	**0.989**	**0.884**	**0.939**	**0.938**	**0.943**	-
PUNCT	0.967	0.983	0.825	0.931	0.896	0.904	-3.9
SHAPE	0.967	0.985	0.810	0.932	0.889	0.897	-4.6
F5	0.989	0.982	0.845	0.923	0.917	0.925	-1.8
F8	0.983	0.987	0.812	0.934	0.898	0.898	-4.5
F4~5	0.981	0.966	0.786	0.933	0.884	0.894	-4.9
F7~8	0.983	0.989	0.814	0.935	0.899	0.908	-3.5
F6~8	0.996	0.989	0.824	0.923	0.910	0.919	-2.4

표 5의 삭제 자질 ∅는 모든 자질을 그대로 사용할 경우이며 이를 기준으로 성능이 어떻게 변하는지를 관찰하였다. 표 5의 마지막 열(Impr.)은 성능의 변화를 의미한다. 표 5에서 보는 바와 같이 F4~5를 제외했을 경우 성능 저하가 가장 크게 나고, F5를 제외한 경우, 성능 변

화가 가장 작다. 이는 자질 F4의 기여도가 가장 높은 것을 의미하며 F5의 기여도가 비교적 낮다는 것을 보여준다. 실험 결과 모든 자질을 이용했을 때 가장 좋은 성능을 보였다. 이는 어떤 자질을 삭제해도 성능의 저하가 발생하므로 모든 자질이 유용함을 알 수 있다.

4.4 모델별 성능 비교

제안된 모델과 비교하기 위해 NLTK 모듈[4]과 KSS 모듈[5]을 사용하였다. NLTK에서 제공하는 tokenize를 사용하여 제안된 모델과 같은 학습 말뭉치를 이용해 학습하여 비교하였다. KSS 모듈은 종결형에 사용되는 [다/요/.?!] 음절의 이전/이후 음절을 매칭하여 문장을 구분하며, 이전/이후 패턴을 미리 정의해두고 패턴을 2음절까지 확장하여 문장의 경계를 정의한다.

표 7 문장 경계 인식기의 성능 평가

모델	KCC Re	KCC Pr	NSMC Re	NSMC Pr	MAR	mAR
KSS	0.817	0.990	0.338	0.949	0.578	0.604
NLTK	1.000	0.984	0.333	0.937	0.667	0.704
제안된 모델	**0.991**	**0.989**	**0.884**	**0.939**	**0.938**	**0.904**

일반 문서의 경우, KSS 모듈은 81.7%의 재현율을, NLTK의 경우 100%의 재현율을, 제안 모델은 99.1%의 재현율을 보여준다. NLTK는 일반 문서에 대한 문장 경계 인식에서 가장 높은 성능을 보였다. 그에 반해 SNS 문서 말뭉치에 대해서는 KSS가 33.8%, NLTK가 33.3%의 성능을 보였다. 이는 다른 문장 경계 인식기가 비형식적인 문장이 많은 SNS 문서에서 문장 경계 인식을 수행하지 못하고 있음을 보여준다. 제안된 모델은 SNS 문서의 문장 경계 인식 성능은 88.4%로 가장 높은 성능을 보였다.

5. 결론 및 향후 연구

본 논문에서는 기계학습 방법을 이용하여 문어체로 이루어진 일반 문서와 비형식적인 문장으로 구성된 SNS 문서 문장 경계 인식 모델을 제안하였다. 학습 말뭉치의 영역을 일반 문서와 SNS 문서로 확장하고, SNS 문서에서 자주 보이는 형태의 단어 유형을 새로 정의해 이를 자질로 사용하여 효과적으로 문장 경계 인식을 수행하였다. 제안한 모델은 일반 문서의 문장 경계 인식 성능으로 99.1%의 재현율을 보였으며, SNS 문서에서도 88.3%의 높은 재현율을 보였다.

향후 연구로는 후처리 규칙을 이용하여 재현율을 더 높여 다양한 내용의 말뭉치를 사용하여 더 넓은 분야에서 이용 가능한 문장 경계 인식기를 제안하고자 한다.

4) https://www.nltk.org/api/nltk.tokenize.html

5) https://github.com/likejazz/korean-sentence-splitter

참고문헌

[1] 김주희, 서정연, “비형식적인 문서에 강건한 문장 경계 인식”,한국컴퓨터종합학술대회 논문집, vol. 37, no. 1, pp. 266-270, 2010.

[2] M. D. Riley, “Some Applications of Tree-based Modeling to Speech and Language”, Proceedings of the DARPA Speech and Natural Language Workshop, pp. 339-352, 1989

[3] 임희석, 한군희, “메모리 기반의 기계 학습을 이용한 한국어 문장 경계 인식”, 한국콘텐츠학회논문지 vol. 4, no. 4, pp. 133-139, 2004

[4] 이충희, 장명길, 서영훈, "웹 문서를 위한 개선된 문장경계인식 방법", 정보과학회논문지, 소프트웨어 및 응용, 제37권, 제6호, pp. 455-465, 2010.

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