형제어 대체를 이용한 개체명 말뭉치 확장

김재균[O1], 김창현[2], 천민아[1], 박혁로[3], 김재훈[1]
한국해양대학교[1], 한국전자통신연구원[2], 전남대학교[3]
jgk20000@naver.com, chkim@etri.re.kr, minah0218@kmou.ac.kr, hyukro@jnu.ac.kr, jhoon@kmou.ac.kr

Named Entity Tagged Corpus Augmentation Using Co-hyponym Replacement

Jae-Kyun Kim[O1], Chang-Hyun Kim[2], Min-Ah Cheon[1], Hyuk-Ro Park[3], Jae-Hoon Kim[1]
Korea Maritime and Ocean University[1], Electronics and Telecommunications Research Institute[2], Chonnam National University[3]

요 약

말뭉치는 기계학습 및 심층학습을 위한 필수 자원이다. 한국어 개체명의 경우 학습에 사용할 잘 정제된 개체명 부착 말뭉치가 충분하지 않다. 말뭉치 정제 작업은 시간적, 경제적으로 많은 비용이 소모된다. 따라서 본 논문에서는 적은 양의 말뭉치를 이용하여 말뭉치를 자동적으로 확장하는 방법을 제안한다. 특별히 소규모 말뭉치에 속하는 문장의 단어에 대한 형제어들을 선정하여 형제어의 확률추출을 기반으로 대체함으로써 새로운 문장을 생성함으로써 말뭉치 확장하는 방법이다. 본 논문에서는 확장된 말뭉치를 이용해서 대부분의 시스템에서 성능이 향상됨을 확인할 수 있었다. 앞으로 단어의 삭제 및 삽입 등 다양한 방법으로 좀 더 다양한 문장을 생성할 수 있을 것으로 생각합니다.

주제어: 형제어, 개체명, 말뭉치 확장

1. 서론

좁은 의미에서 말뭉치(corpus)는 텍스트(text)를 모아 놓은 것이다. 여기에서 텍스트란 폭넓게 해석되는 것으로 일상 대화를 전사해 둔 자료에서부터 신문기사, 소설 등 문자로 작성된 모든 것을 포괄하는 개념이다[1]. 그 중 부착 말뭉치(tagged corpus, 이하 말뭉치)는 최근 활발히 연구되는 기계학습(machine learning) 및 심층학습(deep learning)을 이용한 자연언어 처리 연구에 있어 필수적인 자원이다. 하지만 한국어의 경우 학습에 활용할 잘 정제된 말뭉치가 충분하지 않다는 문제점이 있다. 따라서 말뭉치 자료를 확보하기 위해서는 기존에 존재하는 언어 자료를 수집하고 정제하는 방법으로 새로운 말뭉치를 생성한다. 이러한 말뭉치 정제 작업은 시간적, 경제적으로 많은 비용이 소모된다. 따라서 이러한 비용을 최소화하기 위해 다양한 방식의 기계학습을 이용한 자동화된 말뭉치 구축 연구가 진행되었다[2-4]. 본 논문에서는 학습자료 확장 혹은 데이터 증강(data augmentation) 기법을 이용한 개체명 말뭉치를 확장하는 제안한다.

데이터 증강 기법이란 학습데이터가 부족한 상황에 데이터를 변형시켜 학습데이터를 늘리는 기법이다. 데이터 증강기법은 컴퓨터 비전분야에서 흔히 사용되며 적은양의 학습데이터를 이용하여 더욱 강건한 모델을 학습할 수 있다[5]. 하지만 자연언어 처리 분야에서는 한 단어만 바뀌어도 문장의 의미가 달라지는 문제점 때문에 데이터 증강 기법을 적용하기 어렵다. 따라서 자연언어 처리 연구 분야에서 데이터 증강 기법은 다른 분야에 비해 더딘 발전을 보였다. 이전에 제안되었던 자연어 처리의 데이터 증강기법은 역번역(back translation)[6-7], 데이터 노이징[8], 유의어 교체를 이용한 예측언어 모델[9]등이 있다. 이러한 기법의 경우, 유효한 성능 향상을 보이나 구현 비용이 높아 잘 사용되지 않는다는 문제점이 있다. 이러한 문제점을 완화하여 비교적 간단한 방법으로 성능을 향상시키는 데이터 증강기법인 EDA(Easy Data Augmentation)[10]가 있다. 그러나 EDA 또한 단순한 유의어 대체를 이용하여 문맥과 관련 없이 단어를 대체한다는 문제점이 있다. 본 논문에서는 대체어의 전방 확률과 후방 확률을 이용하여 대체어와 문맥의 연관성을 향상시켰다. 또한 단순 동의어 관계의 단어가 아닌 같은 상위어(hyponym)를 가지는 형제어(co-hyponym) 관계의 단어 사전을 구축하여 문장 생성의 다양성을 높였다.

본 논문은 다음과 같이 구성된다. 2장에서는 본 논문에 사용된 형제어에 대해 설명하고, 3장에서는 본 논문의 모델에 대해 기술한다. 4장에서는 실험 및 평가에 대해 논의한 후, 5장에서 결론과 향후 연구에 대해서 간단히 기술한다.

2. 형제어와 문장 생성

형제어(co-hyponym)란 앞에서 언급한 바와 같이 동일한 상위어를 가지는 단어의 집합이다. 형제어는 동의어와 같이 문장 내에서 유사한 의미를 가지나 동의어보다 더 다양한 단어를 추출할 수 있다. 예를 들어 한국어 어휘지도[11]를 이용하여 추출한 사과에 관한 단어 관계는 그림 1과 같다.

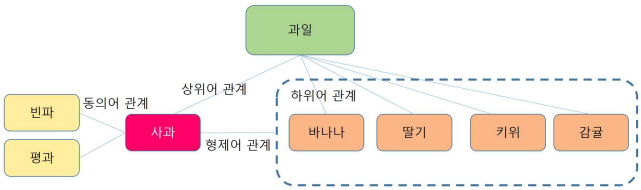

그림 1 사과에 관한 단어관계 지도

그림 1과에서 '사과'는 2개의 동의어 '빈파'와 '평과'를 가지지만 형제어는 71개를 가진다. 이처럼 형제어를 이용한 문장의 의미를 변경되지만 같은 형식의 다양한 어휘를 포함하는 문장을 생성할 수 있을 것이다. 세종 현대 문어 말뭉치[12]를 이용하여 추출한 단어의 평균 동의어와 형제어 수는 표 1과 같다.

표 1. 말뭉치 내의 평균 동의어, 형제어 수

전체 단어 수(|V|)	541,410
동의어 존재 단어 수	16,609
형제어 존재 단어 수	89,090
평균 동의어 수	1.47
평균 형제어 수	554.45

표 1에서 평균 형제어 수는 약 554로 매우 높다. 이는 '사람'[1]과 일반적인 단어는 그 하위어의 개수가 매우 높다. 또한 동의어에 비해 형제어는 약 5배 이상이 존재하므로 훨씬 더 다양한 문장을 생성할 수 있다. 더구나 평균 형제어 수도 약 370배 이상 많이 존재함으로 선정된 단어가 매우 다양한 단어로 대체할 수 있을 것이다.

3.형제어 대체를 이용한 말뭉치 확장

그림 2는 본 논문의 말뭉치 확장 과정이다.

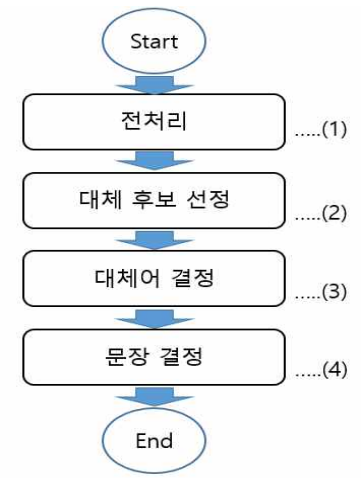

그림 2. 말뭉치 확장 과정

1) '사람'의 하위어는 5699개이다.

그림 2에서 전처리 단계(1)는 형제어 사전과 단어 확률 등을 제작하는 단계이다. 대체 후보 선정 단계(2)는 학습 말뭉치에 속하는 문장에서 대체할 수 있는 단어를 선정하는 단계이다. 대체어 결정 단계(3)는 단계 (1)에서 제작된 사전을 기반으로 단계 (2)에서 선정된 단어를 대상으로 대체어를 결정하는 단계이다. 문장 결정 단계(4)는 문장의 혼잡도(perplexity)를 이용해서 확장 말뭉치에 추가 여부를 결정한다. 이하의 절에서 각 단계를 자세히 기술할 것이다.

3.1 전처리

형제어 사전(co-hyponym)은 2.1절에 언급했듯이 어휘지도[11]를 이용해서 구축된다. 단어의 전방 확률과 후방 확률은 그림 3과 같은 방법으로 구축된다. 전방 확률은 왼쪽에서 오른쪽으로 읽으면서 다음 단어를 예측할 확률이고($P(w_{i+1} \mid w_{i-1}w_i)$), 후방 확률은 오른쪽에서 왼쪽으로 읽으면서 다음 단어를 예측할 확률이다 ($P(w_{i-1} \mid w_i w_{i+1})$).

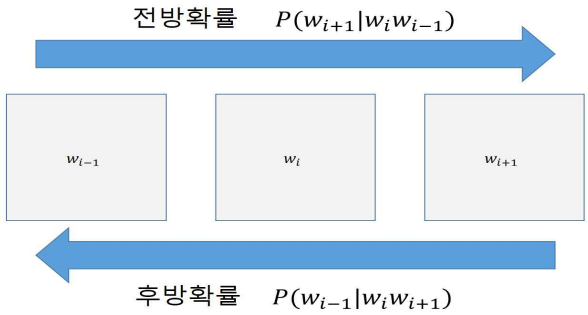

그림 3. 단어의 전방 확률과 후방 확률

3.2 대체 후보 선정

대체 후보 선정 단계에서는 문장에 속한 단어가 형제어 사전에 포함될 경우, 대체 후보로 선정되나, 모든 대체 후보가 형제어로 대체되면 생성된 문장이 매우 부자연스러울 수 있다. 예를 들어 조사와 같은 기능어는 대체후보 선정시 문장 구조에 부자연스러움을 유발한다. 따라서 본 논문에서는 문장 구조상 내용어에 해당하는 명사 단어만을 대체 후보선정 단계에 사용한다. 또한 본 논문에서는 임의추출(random sampling)을 통하여 대체어 후보를 선정하고, 얼마나 많은 대체어를 선택할 것인지는 매개변수 ϵ을 통해서 조절할 것이다. 대체 후보 선정 단계는 알고리즘 1로 요약된다.

알고리즘 1. 대체 후보 선정

Input: $S=\{w_1,w_2,...,w_n\}$, *Co-Hyponyms*

$S' = \varnothing$, $\epsilon = 0.5$
for w **in** S:
  if w *in Co-Hyponyms*
    $r = uniform(0.0, 1.0)$.
    if r $> \epsilon$
      $S' = S' \cup \{w\}$
return S'

알고리즘 1에서 S는 문장이고, *Co-Hyponyms*는 형제어 사전이다. r은 0과 1 사이의 무작위 실수 값이다.

3.3 대체어 결정

대체어 결정 단계는 전방 확률과 후방 확률을 사용하여 3.2에서 선택된 대체 후보 단어의 대체어를 결정하는 단계이다. 알고리즘 2는 그 과정을 보이고 있다.

알고리즘 2 대체어 결정

Input: $co\_hyponym = [(c_1,P(c_1)),...,(c_n,P(c_n))]$, $thresholdp$

$cumulative\_prob = 0$
for i **in range**(n):
  $cumulative\_prob \mathrel{+}= co\_hyponym\_list[i][1]$
  if $cumulative\_prob \geq thresholdp$: **break**
return $random.choice(co\_hyponym[:i][0],$
$p = co\_hyponym[:i][1])$

알고리즘 2에서 3.2절에서 선택된 대체 후보 단어 w_i에 대한 대체어를 선정하는 과정이며, *co_hyponym*은 w_i의 형제어 리스트 $[c_1,...,c_n]$이고, $p(c_i)$는 전방 확률과 후방확률을 이용해서 식 (1)과 같이 계산된다.

$$P(c_i) = 0.5*[P(w_{i+1} \mid w_{i-1}c_i) + P(w_{i-1} \mid c_i w_{i+1})] \quad (1)$$

또한 $c_1,...,c_n$는 $P(c_i)$의 역순으로 정렬되어 있다. 알고리즘 2와 같이 본 논문에서 상위확률추출(nucleus sampling)[13]을 이용하여 대체어를 결정한다. 상위확률추출은 확률 분포 $P(c_i)$에 대해 누적확률이 매개변수인 $thresholdp$를 초과하지 않는 형제어들 중에서 확률적으로 추출하는 것이다. 따라서 상위확률추출 과정은 대체어 후보 리스트에서 단어 등장 확률이 높은 단어들이 높은 확률로 추출된다.

3.4 문장 결정

문장 결정 단계에서는 (2)-(3) 단계를 거쳐 생성된 문장을 최종적으로 말뭉치에 포함할지를 결정하는 단계이다. 본 논문에서는 문장의 혼잡도(perplexity)를 사용한다. 혼잡도란 문장을 생성할 때 다음 단어로 예측할 수 있는 단어의 개수이다[14]. 혼잡도가 낮을수록 다음 단어 예측의 모호성이 낮다는 의미로 성능이 우수하다. 문장의 혼잡도를 구하는 수식은 다음과 같다.

$$PPL(S) = \left(\prod_{1}^{n} P(w_i|w_{i-1})\right)^{-\frac{1}{n}} \quad (2)$$

$$S = w_1, w_2, ..., w_n$$

생성된 문장의 혼잡도가 말뭉치 전체 문장의 혼잡도 평균보다 낮으면 생성 말뭉치에 추가할 것이다.

4. 실험 및 평가

본 논문에서는 엑소브레인 말뭉치[2]와 자체 제작한 한국어 개체명 말뭉치[3]를 합쳐 총 24,086개의 문장을 학습 말뭉치로 사용하고, 전방 확률과 후방 확률은 자체 제작한 원시 말뭉치 14,730,000문장을 사용하였다. 대체어 후보 선정에 사용하는 ϵ은 0.5, 상위확률추출의 임계값 $thresholdp$는 0.8로 설정하였다. 이를 이용하여 생성한 말뭉치의 예는 그림4.와 같다.

생성문장	원문	형태소	개체명	생성문장	원문	형태소	개체명	생성문장	원문	형태소	개체명	생성문장	원문	형태소	개체명
이제	이제	MAG	O	매머드	대형	NNG	O	'	'	SS	O	솔로	솔로	NNG	O
_	_	_	O	_	_	_	O	위	위	NNG	O	국	국	XSN	O
바둑	바둑	NNG	O	건비	건설	NNG	O	&	&	SW	O	으로	으로	JKB	O
애호가	팬	NNG	O	사	사	XSN	O	간	간	NNB	O	_	_	_	O
들	들	XSN	O	의	의	JKG	O	케어	케어	NNG	O	산울림	산울림	NNP	B-ORG
의	의	JKG	O	_	_	_	O	'	'	SS	O	의	의	JKG	O
_	_	_	O	CBSI	CBSI	SL	B-POH	의	의	JKG	O	_	_	_	O
관심	관심	NNG	O	는	는	JX	O	_	_	_	O	'	'	SS	O
은	은	JX	O	_	_	_	O	헬리	헬리	NNP	B-POH	회상	회상	NNG	B-POH
_	_	_	O	금년	올해	NNG	O	코	코	NNG	I-POH	'	'	SS	O
한마디	한마디	NNG	O	_	_	_	O	리스	리스	NNG	I-POH	을	을	JKO	O
로	로	JKB	O	내내	내내	MAG	O	(	(	SS	O	_	_	_	O
_	_	_	O	_	_	_	O	감조	감조	NNG	O	부르	부르	VV	O
'	'	SS	O	100	100	SN	B-PNT	표준물질	추출물	NNG	O	면서	면서	EC	O
조훈현	조훈현	NNP	B-PER	_	_	_	O	)	)	SS	O	_	_	_	O
과	과	JC	O	이상	이상	NNG	O	는	는	JX	O	가면	가면	NNG	O
_	_	_	O	을	을	JKO	O	_	_	_	O	을	을	JKO	O
이창호	이창호	NNP	B-PER	_	_	_	O	헬리코박터	헬리코박터	NNP	B-POH	_	_	_	O
중	중	NNP	O	유지하	유지하	VV	O	토양미생물군		NNG	I-POH	벗	벗	VV	O
_	_	_	O	였	였	EP	O	의	의	JKG	O	은	은	ETM	O
과연	과연	MAG	O	지만	지만	EC	O	_	_	_	O	_	_	_	O
_	_	_	O	_	_	_	O	중세	중식	NNG	O	드라이아이얼음		NNG	B-PER
누	누구	NP	O	11	11	SN	B-DAT	을	을	JKO	O	광주	광주	NNG	I-PER
가	가	JKS	O	월	월	NNB	I-DAT	_	_	_	O	_	_	_	O
_	_	_	O	_	_	_	O	억제	억제	NNG	O	조혜련	조혜련	NNP	B-PER
더	더	MAG	O	92.3	92.3	SN	B-PNT	시키	시키	XSV	O	은	은	JX	O
_	_	_	O	으로	으로	JKB	O	고	고	EC	O	_	_	_	O
세	세	VA	O	_	_	_	O	,	,	SP	O	눈물	눈물	NNG	O
나	나	EF	O	하락하	하락하	VV	O	_	_	_	O	을	을	JKO	O
.	.	SS	O	였	였	EP	O	위	위	NNG	O	_	_	_	O
에	에	JKB	O	다	다	EF	O	_	_	_	O	흘리	흘리	VV	O
_	_	_	O		.	SF	O	홍역막	점막	NNG	O	며	며	EC	O
모	모으	VV	O					을	을	JKO	O				O

그림 4. 생성 말뭉치의 예시

생성 말뭉치를 분석한 결과 그림 4와 같이 자연스러운 문장이 생성되기도 했지만, 그림 5와 같이 조금 부적절한 문장도 생성하였다.

2) http://aiopen.etri.re.kr/service_dataset.php
3) https://github.com/kmounlp/NER

(1)

생성문장	원문	형태소	개체명
남상미	남상미	NNP	B-PER
_	_	_	O
족남	족녀	NNG	O
,	,	SP	O
_	_	_	O
산모	산모	NNG	O
_	_	SO	O
아이	아이	NNG	O
_	_	_	O
모두	모두	MAG	O
_	_	_	O
건강	건강	NNG	O
…	…	SE	O
'	'	SS	O
새로	새롭	VA	O
운	ㄴ	ETM	O
_	_	_	O
가족	가족	NNG	O
_	_	_	O
생겨	생겨	VV+EC	O
_	_	_	O
기쁘	기쁘	VA	O
고	고	EC	O
_	_	_	O
감사	감사	NNG	O
'	'	SS	O
스턴트우먼	배우	NNG	O
_	_	_	O
남상미	남상미	NNP	B-PER
가	가	JKS	O
_	_	_	O
딸	딸	NNG	O
을	을	JKO	O
_	_	_	O
날	날	VV	O
았	았	EP	O
다	다	EF	O

(2)

생성문장	원문	형태소	개체명
인명진	인명진	NNP	B-PER
과	과	JC	O
_	_	_	O
서정원	서정원	NNP	B-PER
_	_	_	O
추악한	추악하+ㄴ	VA+ETM	O
_	_	_	O
밀약	밀약	NNG	O
등짓날	설	XSN	O
CJ	CJ	SL	B-ORG
그룹	그룹	NNG	I-ORG
이	이	JKS	O
_	_	_	O
해외	해외	NNG	O
_	_	_	O
업체	업체	NNG	O
을	을	JKO	O
_	_	_	O
대상	대상	NNG	O
으로	으로	JKB	O

(3)

생성문장	원문	형태소	개체명
경상도	경상도	NNP	B-LOC
_	_	_	O
섬	섬	NNG	O
_	_	_	O
계남	사나이	NNG	O
는	는	JX	O
_	_	_	O
아내	아내	NNG	O
를	를	JKO	O
_	_	_	O
'	'	SS	O
명순이	명순이	NNP	B-PER
'	'	SS	O
라고	라고	JKQ	O
_	_	_	O
고무	불	VV	O
했	했	EP	O
다	다	EF	O
.	.	SF	O

그림 5. 부적절한 문장 생성 예시

부적절한 문장 생성의 경우 그림 5의 (1)과 같이 대체한 단어가 문맥상 의미가 틀린 경우, (2)와 같이 “견해, 주의, 통설을 이르는 말”인 ‘설’을 명절 중 하나인 ‘설’로 분석하여 대체할 단어의 의미 분석이 틀린 경우, 그리고 (3)의 ‘고무’와 같이 대체어의 형제어가 의미 또는 용법이 틀린 경우 등이 있다.

또한 매개변수 ϵ을 조절하여 전체 대체어 중 개체명 부착 단어의 비율을 살펴보았으며 그 결과는 그림 6과 같다.

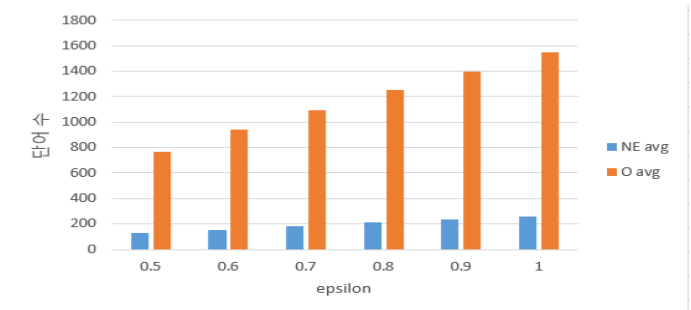

그림 6. 매개변수 ϵ 조절에 따른 개체명 단어 갯수

그림 6은 다른 매개변수를 고정한 후 ϵ 값을 조정하여 말뭉치를 10번 생성한 결과 대체된 개체명부착 단어 개수와 개체명 미부착 단어 개수의 평균값이다. ϵ 값이 증가할수록 개체명 부착 단어의 증가는 적은 반면 개체명 미부착 단어의 대체 횟수는 크게 증가하는 것을 알 수 있다. 이는 형제어 사전에 개체명 부착의 주요 대상인 고유명사의 부족으로 인한 것으로 추측된다.

생성된 말뭉치의 평가를 위해 세 종류의 개체명 인식기를 이용하여 성능을 평가해 보았다. 기존 개체명 부착 말뭉치 6,490 문장, 237,133개의 형태소를 가진 말뭉치에 제안하는 시스템을 사용하여 1,000 문장 30,454개의 형태소를 추가하여 실험하였다. 실험 결과는 표 2와 같다.

표 2. 말뭉치 확장을 이용한 개체명 인식기 성능 평가(단위 %)

시스템	정밀도		재현율		F1	
	기본	확장	기본	확장	기본	확장
[15]	80.08	**80.32**	73.35	**74.75**	76.56	**77.44**
[16]	89.70	**89.82**	87.55	**88.60**	88.12	**89.21**
[17]	**89.87**	89.82	87.71	**88.06**	88.77	**88.93**
평균	86.55	**86.65**	82.87	**83.80**	84.48	**85.19**

% 기본: 확장 전 학습 말뭉치,
확장: 기본 학습말뭉치에 확장된 말뭉치가 추가됨.

표 2에서 시스템은 각각 LSTM-CRF[15], CHARS-LSTM-LSTM-CRF[16], CHARS-CONV-LSTM-CRF[17]이다. 또한 ‘기본’은 확장하지 전의 학습 말뭉치이고 ‘확장’은 확장 전 학습 말뭉치에 본 논문에서 확장된 문장이 추가된 학습 말뭉치를 사용했을 경우이다. 현재는 초벌 실험으로 성능의 개선이 아주 미약하지만 거의 모든 경우에 성능이 개선되어 형제어 외에 삽입이나 삭제 등을 통할 경우, 좀 더 큰 성능의 향상을 기대할 수 있을 것이다.

5. 결론

본 논문에서는 형제어 대체를 이용한 개체명 말뭉치 확장 기법을 제안하였다. 제안하는 시스템을 이용하여 확장한 1,000문장의 개체명 말뭉치를 추가하여 세 종류의 개체명 인식기의 성능을 평가한 결과, 거의 모든 경우에 성능이 향상됨을 볼 수 있었다. 향후 연구로는 대체 이외에 삽입, 삭제 등을 이용하여 더 다양한 방식의 말뭉치 확장 방법을 연구할 계획이다.

감사의 글

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(R7119-16-1001, 지식증강형 실시간 동시통역 원천기술 개발)과 2017년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068187, 한국어 정보처리 원천 기술 연구 개발).

참고문헌

[1] 최재웅, “말뭉치와 언어연구 - 외국의 사례와 경향”, 한국어학, 제63권, pp. 71-102, 2014.
[2] 장경애, 박상현, 김우제, “인터넷 감정기호를 이용한 긍정/부정 말뭉치 구축 및 감정분류 자동화”, 정보과학회논문지, 제42권, pp. 512-521, 2015.
[3] 이준우, 한솔, 류법모, “복수 언어 분석기와 언어 규칙 기반 한국어 형태소 부착 말뭉치 반자동 구

축”,한국정보과학회 학술발표논문집, pp. 431-433, 2019.
[4] 임준호, 곽용재, 박소영, 임해창, “신경망을 이용한 반자동 구문분석 말뭉치 구축도구”, 한국 정보과학회 학술발표논문집, 제30권, pp. 483-485, 2003.
[5] Shorten, Connor, and T. M. Khoshgoftaar, “A survey on image data augmentation for deep learning.”, Journal of Big Data, vol. 6,no. 1, pp. 60 2019.
[6] Adams Wei Yu, David Dohan, Minh-Tang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc C. Le, Qanet: Combining local convolution with global self-attention for reading comprehension, In Proceedings of International Conference on Learning Representations, 2018.
[7] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 86-96, 2016.
[8] Ziang Xie, Sida I. Wang, Jiwei Li, Daniel Levy, Aiming Nie, Dan Jurafsky, and Andrew Y. Ng. , Data noising as smoothing in neural network language models. In Proceedings of International Conference on Learning Representations, 2017.
[9] Sosuke Kobayashi, Contextual augmentation: Data augmentation by words with paradigmatic relation”, In proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies , 2018.
[10] Jason Wei, Kai Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing. pp. 6383-6389, 2019.
[11] 배영준, 옥철영, “한국어 어휘지도(UWordMap)와 API소개”, 제26회 한글 및 한국어 정보처리 학술대회, pp. 27-31, 2014.
[12] 김홍규, “21세기 세종계획 국어 기초자료 구축”, 연구보고서, 2007
[13] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, Yejin Choi, The Curious case of Neural Text Degeneration, In Proceedings of International Conference on Learning Representations, 2020.
[14] Stanley. F. Chen, Douglas Beeferman, and Roni Rosenfeld, Evaluation metrics for language models, Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, pp. 275-280, 1998.
[15] Zhiheng Huang, Wei Xu, and Kai Yu., Bidirectional LSTM-CRF models for sequence tagging., arXivpreprint arXiv:1508.01991, 2015
[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer, Neural architectures for named entity recognition. In proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.260-270, 2016.
[17] Xuezhe Ma and Eduard H. Hovy, End-to-end sequence labeling via bi-directional LSTM-CNNsCRF. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics , Vol.1, pp.1064-1074, 2016.