은닉 마르코프 모델을 이용한 한국어 개체명 말뭉치 생성

김재균[O†], 김창현[‡], 천민아[†], 박호민[†], 윤호[†], 남궁영[†], 최민석[†], 김재훈[†]
한국해양대학교[†], 한국전자통신연구원[‡]
jgk20000@naver.com, chkim@etri.re.kr, minah0218@kmou.ac.kr, homin2006@hanmail.net,
4168615@naver.com, young_ng@kmou.ac.kr, ehdqus5136@naver.com, jhoon@kmou.ac.kr

Generating Korean NER Corpus using Hidden Markov Model

Jae-Kyun Kim[O†], Chang-Hyun Kim[‡], Min-Ah Cheon[†], Ho-Min Park[†],
Ho Yoon[†], Young Nam-Goong[†], Min-Seok Choi[†], Jae-Hoon Kim[†]
Korea Maritime and Ocean University[†], Electronics and Telecommunications Research Institute[‡]

요 약

기계학습을 이용하여 개체명 인식을 수행하기 위해서는 많은 양의 개체명 말뭉치가 필요하다. 이를 위해 본 논문에서는 문장 자동 생성을 통해 개체명 표지가 부착된 말뭉치를 구축하는 방법을 제안한다. 기존의 한국어 문장 생성 연구들은 언어모델을 이용하여 문장을 생성하였다. 본 논문에서는 은닉 마르코프 모델을 이용하여 주어진 표지열에 기반 하여 문장을 생성하는 시스템을 제안한다. 제안하는 시스템을 활용하여 자동으로 개체명 표지가 부착된 3,286개의 새로운 문장을 생성할 수 있었다. 학습말뭉치 문장과 약 70%의 차이를 보이는 새로운 문장을 생성하였다.

주제어: 문장 생성, 은닉 마르코프 모델, 언어모델, 말뭉치 생성

1. 서론

최근 개체명 인식 분야에서도 기계학습 및 심층학습을 이용한 연구가 활발히 이루어지고 있다[1]. 하지만 한국어의 경우 학습에 활용할 수 있는 개체명 말뭉치가 충분하지 않다는 문제점이 있다. 말뭉치 자료를 확보하기 위해서는 기존에 존재하는 언어 자료를 수집 및 정제하거나 새롭게 말뭉치를 생성하는 작업이 필요하다. 이러한 작업에는 시간적, 경제적으로 많은 비용이 든다. 따라서, 본 논문에서는 개체명 말뭉치를 확보하기 위한 하나의 방안으로 문장 생성을 통해 개체명 말뭉치를 구축하는 방법을 제안한다.

자연언어처리에서 문장 생성이란 컴퓨터를 이용해 인간이 이해할 수 있는 문장을 만들어 내는 것이다[2]. 문장 생성은 기계 번역[3],문서 요약[4], 이미지 주석 생성[5] 등 다양한 분야에 활용되고 있다.

본 논문에서는 개체명 말뭉치 구축을 위해 기존의 문장 생성과 달리 언어 정보가 포함된 문장을 생성하고자 한다. 이를 위해 기존의 개체명 말뭉치에서 자주 나타나는 표지열을 분석한 뒤 이를 선별하여 문장 생성 모델의 입력으로 한다. 이러한 방법론에 적합한 모델로 은닉 마르코프 모델을 이용하여 한국어 개체명 표지가 부착된 문장을 생성하는 시스템을 제안한다. 제안한 방법을 통해 한국어 개체명 표지가 부착된 문장 3,286개를 자동으로 생성하였다. 제안하는 시스템으로 실험에 사용한 학습말뭉치 문장과 약 70%의 차이를 보이는 새로운 문장을 생성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 언어모델과 은닉 마르코프 모델에 대해서 간단히 설명한다. 3장에서는 은닉 마르코프 모델을 이용해 문장을 생성하는 방법에 대해 상술한다. 4장에서 실험 및 평가에 대해 언급한 후, 5장에서 결론과 향후 연구에 대해서 간단히 기술한다.

2. 관련 연구

2.1 언어모델

언어모델이란 주어진 단어열의 확률분포를 이용하여 다음 단어를 예측하는 방법이다. 언어모델을 수식화하면 식 (1)과 같다. 식 (1)에서 w는 하나의 단어이며, $w_1, w_2, \ldots, w_t$는 주어진 단어열이 된다. w_{t+1}는 주어진 단어열 다음에 올 단어이다[15].

$$P(w_{t+1}|w_1, \ldots, w_t) \quad (1)$$

위의 언어모델은 모든 단어열의 정확한 확률분포를 추정할 수 없다는 문제가 있다. 이를 해결하기 위해 통계를 기반으로 한 n-gram 언어모델이 제안되었다. n-gram이란 n개의 연속적인 단어의 나열을 의미한다. n-gram 언어모델에서는 다음에 나올 단어의 예측은 오직 n-1개의 단어에 의존하여 다음에 나올 단어를 예측한다. 이를

식으로 나타내면 식 (2)와 같다.

$$P(w_{t+1}|w_1,...,w_t) = P(w_{t+1}|w_{t-n+2},...,w_t) \quad (2)$$

식 (2)를 조건부확률의 정의에 따라 풀어보면 식 (3)과 같이 n-gram 확률을 (n-1)-gram의 확률로 나눈 것으로 정리된다.

$$P(w_{t+1}|w_1,...,w_t) = \frac{P(w_{t-n+2},...,w_{t+1})}{P(w_{t-n+2},...,w_t)} \quad (3)$$

이때 $P(w_{t+1},w_t,...,w_{t-n+2})$와 $P(w_t,...,w_{t-n+2})$는 최대우도 추정(maximum likelihood estimation) 방법[16]에 따라 식 (4)와 같이 추정할 수 있다.

$$\frac{P(w_{t-n+2},...,w_{t+1})}{P(w_{t-n+2},...,w_t)} \approx \frac{count(w_{t-n+1},,...,w_{t+1})}{count(w_{t-n+2},...,w_t)} \quad (4)$$

n-gram 언어모델은 충분한 데이터가 있을 때 효과적이다. 실생활에 사용하는 문장이라도 학습말뭉치에 존재하지 않거나 다음으로 나올 수 있는 단어가 학습말뭉치에 없는 경우에는 정확한 모델링을 하지 못하는 희소성 문제가 발생한다[17]. 희소성 문제를 완화하는 방법으로 평활화(smoothing)기법이 있다. 평활화 기법으로는 라프라스 평활화(Laplace smoothing), 캐이츠 평활화(Katz smoothing)등이 있다[18].

2.2 은닉 마르코프 모델

은닉 마르코프 모델은 순차적인 데이터($o_1,...,o_t$)에 대해 현재 상태 s_t를 추정하는 확률 모델이다. 이는 현재 상태 s_t는 직전 상태 s_{t-1}에만 영향을 받는다는 마르코프 가정(Markov assumption)을 바탕으로 한다[15]. 이를 수식으로 표현하면 식 (5)과 같다. 식 (5)에서 볼 수 있듯이 초기 상태 s_1에서 직전 상태 s_{t-1}까지에 대해 현재 상태 s_t로의 전이 확률은 직전 상태 s_{t-1}에서 현재 상태 s_t로의 전이 확률에 근사한다.

$$P(s_t|s_1,s_2,...s_{t-1}) \simeq P(s_t|s_{t-1}) \quad (5)$$

은닉 마르코프 모델은 아직 정해지지 않은 상태인 은닉 상태와 그 상태에 기반한 관측열이 존재하며, 각 상태는 가능한 출력 관측열의 확률분포를 가진다. 이러한 은닉 마르코프 모델을 위해서는 마르코프 체인(Markov chain)의 전이 확률뿐만 아니라 방출 확률이 필요하다. 식 (6)은 임의의 측정 가능한 집단 A에 대하여 은닉상태 s_t과 결과물o_t이 있을 때 방출확률을 구하는 방법을 나타낸다.

$$P(o_t|s_1,...s_t) \simeq P(o_t|s_t) \quad (6)$$

3. 은닉 마르코프 모델을 이용한 한국어 개체명 말뭉치 생성

일반적으로 개체명 표지를 나타내기 위해서 BIO 표기법을 사용한다. 본 논문에서는 개체명 인식 분야에서 사용되는 BIO 표기법 대신에 BIT 표기법[19]을 사용하여 한국어 개체명 말뭉치를 자동으로 생성하는 방법을 제안한다. 일반적으로 널리 사용되는 BIO 표기법은 개체명이 시작되는 단어의 표지에 B(Beginning)-를 붙이고, 개체명에 포함된 그 외의 단어의 표지에는 I(Inside)-를 붙이며, 개체명과 개체명 사이의 모든 단어의 표지를 O(Outside)로 간주하는 방법이다. BIO 표기법으로 표현된 말뭉치는 O 표지가 90% 이상을 차지하므로 O 표지에 대한 혼잡도가 높아지는 문제와 불균형학습 문제가 발생된다. 이러한 문제를 완화하기 위해서 본 논문에서는 BIO 표기법 대신에 BIT 표기법을 사용한다[14]. BIT 표기법이란 BIO 표기법에서 O 표지를 T(Tag) 표지로 변환하는 방법이며 본 논문에서 T 표지는 품사 표지를 나타낸다. 그림 1은 제안한 BIT 표기법을 이용하여 CoNLL 형식으로 나타낸 한국어 개체명 말뭉치의 예이다.

START	START	START	START
한국	한국	NNP	B-ORG
_	_	_	I-ORG
정부	정부	NNG	I-ORG
가	가	JKS	JKS
_	_	_	_
독자	독자	NNG	NNG
_	_	_	_
대북	대북	NNG	NNG
제재	제재	NNG	NNG
를	를	JKO	JKO
_	_	_	_
발표한	발표하+ㄴ	VV+ETM	VV+ETM
_	_	_	_
날	날	NNG	NNG
이	이	VCP	VCP
다	다	EF	EF
.	.	SF	SF
END	END	END	END

그림 1. BIT 표기법으로 표현된 한국어 개체명 말뭉치의 예

개체명 표지열(tag sequence)이 주어졌을 때, 문장 생성 시스템은 식 (7)과 같이 현재 상태에서 관측 가능한 결과물을 표지 t_i로, 생성할 단어 w_i를 은닉상태로 정의한 은닉 마르코프 모델로 정의한다.

$$w_i = \arg max_{w_i} P(w_i|t_i) \bullet P(w_i|w_{i-1}) \quad (7)$$

그림 2는 한국어 말뭉치 생성 과정을 도식화한 것이며, 이 과정을 통해서 그림 1과 같은 한국어 개체명 말

뭉치를 생성한다.

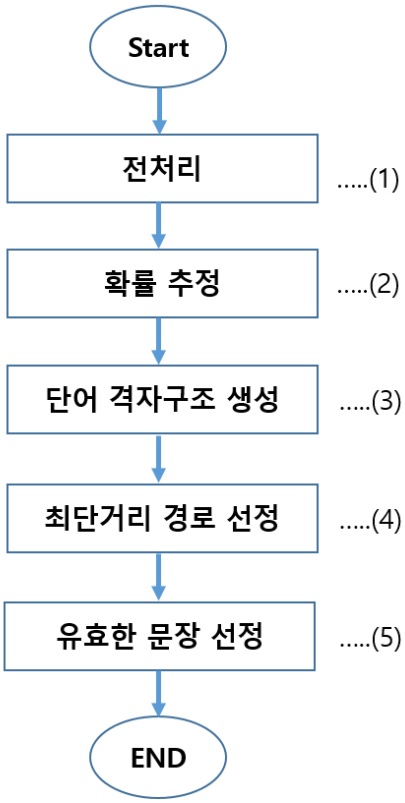

그림 2 한국어 개체명 말뭉치 생성 과정

3.1 전처리(Preprocessing)

그림 2의 (1)에서는 정제되지 않은 말뭉치로부터 그림 1과 같은 CoNLL형식의 말뭉치로 정제한다. 정제된 말뭉치는 한 문장을 "START", "END" 기호로 구분하며, 각 행은 어절, 형태소, 품사 표지, BIT 표지로 구성되어 있다.

또한, 그림 2의 (1)에서는 정제된 말뭉치로부터 확률 추정에 사용할 단어-출현빈도쌍 사전과 단어-태그 사전을 생성한다.

3.2 확률 추정

그림 2의 (2)에서 준비된 말뭉치로부터 단어 간의 전이 확률 행렬과 단어-표지 쌍의 방출 확률을 추정한다. 이때 발생하는 전이 확률의 희소성 문제를 해소하기 위해 캐이츠 평활화[20]를 활용한다. 제안하는 모델에 사용된 캐이츠 평활화는 식 (8)과 같다.

$$P(w_i|w_{i-1}) = \begin{cases} P(w_i|w_{i-1}) & ,\text{if } C(w_{i-1}w_i) > 0 \\ \alpha \times P(w_i) & ,\text{if } C(w_i) > 0 \\ \beta & ,otherwise \end{cases} \quad (8)$$

3.3 단어 격자구조 생성

그림 2의 (3)에서는 생성된 전이, 방출 확률을 이용하여 단어 격자구조를 생성한다. 단어 격자구조는 정점(vertex)과 간선(edge)으로 구성된다. 정점은 생성후보 단어를 의미하고 간선은 이전 단어와 다음 단어의 관계를 나타낸다. 간선의 가중치(weight)는 전이확률과 방출확률의 곱으로 설정한다. 일반적으로는 두 단어 사이의 완전이분그래프로 구성하지만, 본 논문에서는 다양한 문장의 생성을 위해 무작위 추출법을 활용한다. 격자구조 생성시 무작위 난수를 생성하여 미리 설정한 ϵ(epsilon) 값 이상일시 간선의 가중치를 최대화한다.

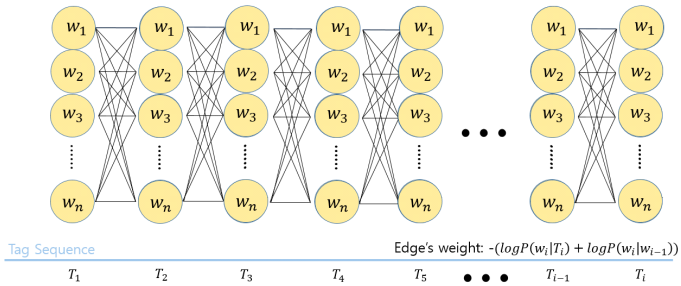

그림 3. 단어격자구조 구성도

3.4 최단거리 경로 선정

그림 2의 (4)에서는 3.3절에서 설명한 단어 격자구조의 최단경로를 Bellman-Ford 최단경로 알고리즘[21]을 이용해 탐색하게 된다. 이렇게 탐색된 최단경로 정점들을 합쳐서 최적의 문장을 생성한다.

3.5 혼잡도에 의한 유효한 문장 선정

그림 2의 (5)에서는 (3),(4) 과정을 거쳐 생성된 문장의 평가를 위해 혼잡도(Perplexity)를 사용한다. 혼잡도란 문장을 생성할 때 다음 단어로 예측할 수 있는 단어의 개수이다[22]. 혼잡도가 낮을수록 모델의 성능이 우수하다고 볼 수 있다. 문장의 혼잡도를 구하는 수식은 식 (7)과 같다.

$$PPL(S) = \left(\prod_{1}^{n} P(w_i|w_{i-1})\right)^{-\frac{1}{n}} \quad (7)$$

$$S = w_1, w_2, \ldots, w_n$$

3.5에서는 생성된 문장의 혼잡도가 말뭉치 전체 문장의 혼잡도 평균보다 높으면 출력으로 사용하지 않고 (3)~(4)의 과정을 다시 반복한다.

4. 실험 및 결과

본 논문에서는 엑소브레인 말뭉치[1]와 자체 제작한 한국어 개체명 말뭉치를 합쳐 총 24,086개의 문장을 사용하였다. 문장 생성 시 무작위성을 부여하는 ϵ값은 0.9로

1) http://aiopen.etri.re.kr/service_dataset.php

설정하였으며 캐이츠 평활화에 사용된 α와 β는 각각 10^{-4}과 10^{-5}으로 설정하였다. 사용한 말뭉치의 총 단어 수는 46,964개이며 기존의 BIO표지의 O표지를 품사 표지로 대체한 193종의 BIT표지를 사용하였다. 이를 통해 생성한 문장의 예는 그림 4와 같다.

지난 4일 오후 방송된 JTBC '뉴스룸'에는 영화 '많은 사제들'의 주인공 강동원이 출연 했습니다.

신세경 악플러 고소 인신공격적이어 유명인 댓글에 분노

자유 한국당은 조건부 해체가 크게 됐지만, 정직한 성과도 공개질의에 답변하지 않았다.

공자 역시 이렇게 말했다.

그림 4. 생성된 문장의 예시

생성된 문장을 분석한 결과 그림 4와 같이 비교적 자연스러운 문장이 생성되기도 했지만, 부적절한 동사 및 명사로 인해 부자연스러운 문장이 생성되기도 하였다. 부적절한 문장의 예시는 그림 5와 같다.

배우 강동원이 명예훼손 및 모욕죄로 악플러 들을 고소해 관심이 *출연* 했습니다.

배우 신세경이 명예훼손 및 모욕죄로 악플러 들을 고소해 관심이 집중 됐다.

배우 신세경이 명예훼손 및 모욕죄로 악플러 들을 *거쳐 팀이 집중* 됐다.

배우 신세경이 명예훼손 및 모욕죄로 악플러 들을 고소해 관심이 *중단* 됐다.

그림 5. 부적절한 문장 생성

본 논문에서는 문장을 생성할 때 말뭉치 전체의 평균 혼잡도(12.80)를 생성 기준으로 활용한다. 생성된 문장의 평균 혼잡도는 15.24였다. 또한, 3,286개의 문장을 생성하였을 때 버려진 문장의 수는 1,753개의 문장으로 손실률은 53.44%였다.

생성된 문장을 평가하기 위해 기계번역에서 번역된 문장의 성능을 평가하는 지표인 BLEU 점수를 이용하였다. BLUE 점수는 모델이 번역한 문장과 정답 문장 간에 일치하는 n-gram의 수를 기반으로 계산하며 높을수록 잘 번역되었음을 의미한다. 본 논문에서는 말뭉치의 문장을 정답문장으로 사용하였으며, 이를 생성한 문장과 비교하여 BLUE 점수를 계산하였다. 계산 결과는 표 1과 같다.

표 1. 생성된 문장의 BLEU 점수

	BLEU-2	BLEU-3	BLEU-4	BLEU
전체 생성 문장	0.3989	0.2860	0.2107	0.2886
30 어절 이하	0.4221	0.3019	0.2238	0.3439
30 어절 이상 60 어절 이하	0.3637	0.2629	0.1899	0.2628
60 어절 이상	0.3590	0.2635	0.1983	0.2657

전체적으로 생성된 문장의 길이가 짧을수록 BLEU 점수가 높은 것을 알 수 있다. 어절의 개수가 30개 이하인 경우가 0.3439로 가장 높았다. 전체 문장에 대한 평균은 0.2886의 BLEU 점수를 보였다. BLEU 점수는 높을수록 출력된 문장과 정답 문장이 일치한다는 의미이므로, 제안한 시스템이 생성한 문장은 본래 말뭉치의 문장과는 약 70% 정도 다른 새로운 문장을 생성할 수 있음을 알 수 있다.

5. 결론

본 논문에서는 은닉 마르코프 모델 기반의 언어모델을 이용하여 표지 정보를 추가한 문장 생성 시스템을 제안하였다. 생성된 3,286 문장에 대한 평균 혼잡도는 15.24이고 BLEU 점수는 0.2886 으로 측정되었다. 기존 말뭉치에 등장하지 않는 약 70% 다른 새로운 문장을 생성하였다. 향후 연구로는 장기 의존성 문제를 해결한 것으로 알려진 순환신경망을 기반으로 한 생성적 적대 신경망 모델에 표지 정보를 추가한 연구를 진행하고자 한다.

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] P. Sun, X. Yang, X. Zhao, and Z. Wang, An Overview of Named Entity Recognition, International Conference on Asian Language Processing, pp.273-278, 2018.

[1] E. Reiter and R. Dale, Building Natural Language Generation Systems, Cambridge University Press, 2000.

[2] 최희열, 신경망 기반 기계번역 모델의 이해, 정보과학회지, Vol. 37, No. 2, pp. 16-24,2019

[3] 전재원, 황현선, 이창기, 추출요약과 생성 요약을 결합한 2단계 문서 요약, 한국정보과학회 학술발표 논문집, pp.581-583, 2019.

[4] 배장성, 이창기, 딥러닝을 이용한 한국어 이미지 캡션 생성, 한국정보과학회 학술발표 논문집, pp.488-490, 2016.

[5] 김재훈, 중간언어방식을 이용한 기계번역에서의 한국어 격조사 생성을 위한 한국어 격틀 설정, 석사학위논문, 한국과학기술원, 1988.

[6] 안동언, Corpus를 기반으로 하는 한국어 술어의 양상 생성, 박사학위논문, 한국과학기술원, 1995.

[7] 권일재, 송만석, 표현기술언어를 이용한 한국어 생성에 관한 연구, 언어정보과학회 학술발표 논문집, pp. 117-120, 1995.

[8] 이강천, 이상호, 서정연, 의미 중심어 주도 방식에 기반한 한국어 생성 시스템, 언어정보과학회 학술발

표 논문집, Vol.23, No. 1A, pp. 949-952, 1996.
[9] 박영진, 박인철, 안동언, 이용석, 개념 그래프에 기반한 한국어 문장 생성 방법, 한국정보과학회 학술발표 논문집, Vol.22, No.2A, pp.635-638, 1995.
[10] 서영애, 이종혁, 이근배, 개념 그래프에서의 한국어 문장 생성, 한국정보과학회 학술발표 논문집, pp.255-258, 1997.
[11] 김양훈, 황용근, 강태관, 정교민, LSTM 언어모델 기반 한국어 문장 생성, 한국통신학회 논문지, Vol.41, No. 5, pp.592-601, 2016
[12] 허윤석, 김주애, 박영민, 강상우, 서정연, LSTM 언어모델을 이용한 한국어 자연어 생성모델, 한국정보과학회 학술발표 논문집, pp.853-855, 2017
[13] 최세목, 박정희, LSTM을 이용한 한글 문장 생성 모델, 한국정보과학회 학술발표 논문집, pp. 827-829, 2018.
[14] D. Jurafsky. and J. H. Martin. Speech and Language Processing, 2nd edition, Prentice Hall Publishers. 2009
[15] R. Richard, Mathematical Statistics: An Introduction to Likelihood Based Inference, Wiley & Sons, 2018.
[16] B. Alison D. Guthrie, and L. Guthrie, “Another look at the data sparsity problem”, Proceedings of International Conference on Text, Speech, and Dialogue, pp 327-334, 2006.
[17] S. F. Chen and J. Goodman, “An empirical study of smoothing techniques for language model”, Journal of Computer Speech & Language, Vol. 13, No. 4, pp. 359-394, 1999.
[18] 윤호, 김창현, 천민아, 박호민, 남궁영, 최민석, 김재균, 김재훈, BIT 표기법을 활용한 개체명 인식, 제 31회 한글 및 한국어 정보처리 학술대회, 2019.
[19] S. M. Katz, “Estimation of probabilities from sparse data for language model component of a speech recognizer”. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 35, No. 3, pp. 400-401, 1987.
[20] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, second edition, MIT Press, 2001.
[21] S. F. Chen, D. Beeferman, and R. Rosenfeld, Evaluation metrics for language models, Proceedings of DARPA Broadcast News Transcription and Understanding Workshop, pp. 275-280, 1998.