한국음향학회지 제27권 제8호 pp. 435~442 (2008)

한국어의 어순 구조를 고려한 Two-Path 언어모델링

Two-Path Language Modeling Considering Word Order Structure of Korean

신 중 휘*, 박 재 현*, 이 정 태**, 임 해 창**
(Joong-Hwi Shin*, Jae-Hyun Park*, Jung-Tae Lee**, Hae-Chang Rim**)

*고려대학교 컴퓨터학과, **고려대학교 컴퓨터·전파통신공학과
(접수일자: 2008년 9월 24일; 수정일자: 2008년 10월 20일; 채택일자: 2008년 10월 21일)

n-gram 모델은 영어와 같이 어순이 문법적으로 제약을 받는 언어에 적합하다. 그러나 어순이 비교적 자유로운 한국어에는 적합하지 않다. 기존 연구는 어절 간 어순의 고려가 어려운 한국어의 특성을 반영한 twoply HMM을 제안했으나, 인접 어절 간 어순 구조를 반영하지 못하였다. 본 논문에서는 용언형태소 사이에 나타나는 인접 어절 간에 어순 특성을 반영하기 위해 두 어절을 결합하는 세그먼트 단위를 정의하고, 제안한 세그먼트 단위에서 문맥에 따라 확률을 달리 추정하는 two-path 언어모델을 제안한다. 그 결과 기존 한국어 언어모델에 비해 제안하는 two-path 언어모델은 기존 연구보다 25.68% 혼잡도를 줄였으며, 어절 간에 결합이 일어나는 경계인 용언형태소에서는 94.03%의 혼잡도를 줄였다.
핵심용어: 한국어, 언어모델링, 어순, 용언형태소, 세그먼트 단위
투고분야: 음성처리 분야 (2.7)

The n-gram model is appropriate for languages, such as English, in which the word-order is grammatically rigid. However, it is not suitable for Korean in which the word-order is relatively free. Previous work proposed a twoply HMM that reflected the characteristics of Korean but failed to reflect word-order structures among words. In this paper, we define a new segment unit which combines two words in order to reflect the characteristic of word-order among adjacent words that appear in verbal morphemes. Moreover, we propose a two-path language model that estimates probabilities depending on the context based on the proposed segment unit. Experimental results show that the proposed two-path language model yields 25.68% perplexity improvement compared to the previous Korean language models and reduces 94.03% perplexity for the prediction of verbal morphemes where words are combined.
Keywords: Korean, Language modeling, Verbal morpheme, Word order, Segment unit
ASK subject classification: Speech Signal Processing (2.7)

I. 서 론

언어모델은 문장이 나타날 확률로 다음 단어를 예측하거나 중의성을 해소하는 등의 역할을 한다. 따라서 언어모델은 음성인식 [1], 기계번역 [2] 등 다양한 자연어처리 분야에서 이용된다. 언어모델링의 대표적인 방법은 n-gram 모델 [3]이 있다. 어절의 어순을 이용하는 n-gram 모델은 간단하고 좋은 성능으로 많은 연구에서 활용되는 방법이다. 하지만 어순이 문법적으로 제약을 받는 영어와는 다르게 한국어는 비교적 자유로운 어순 구조를 갖는 언어이므로 n-gram 모델을 직접 적용하는 것은 어렵다. 따라서 한국어의 어순 구조를 고려하는 한국어 언어모델링은 중요한 의미를 지닌다.

한국어의 어순 구조의 특징을 고려하는 기존 연구는, 어절 간 어순의 고려가 어려운 한국어의 특성을 반영하여 문맥에 따라 다른 확률을 이용하는 연구 [4]가 있다. 이 연구는 형태소 단위에서 어절의 첫 형태소의 확률을 추정하는 어절 생성 확률과 어절 내의 나머지 형태소의 확률을 추정하는 어절 내 형태소 전이 확률의 twoply HMM (Hidden Markov Model)을 이용한다. 하지만 용언형태소를 포함한 어절과 그 이전 어절 간 어순 구조를 고려하지 않아서 해당 어절 간에 발생하는 어절의 생성 확률을 올바로 추정하지 못한다. 또 어절 내 형태소 전이

책임저자: 임 해 창 (rim@nlp.korea.ac.kr)
136-713 서울시 성북구 안암동 5-1 고려대학교 컴퓨터·전파통신공학과
(전화: 02-924-2054; 팩스: 02-929-7914)

확률 추정 시 HMM [5]의 특성상 왼쪽 한 문맥 이상을 이용하지 않는 문제가 있다.

문맥에 따라 다른 확률을 추정하여 모델링을 하는 교착어의 연구로는 어절의 첫 형태소를 예측할 때 원거리 문맥 정보를 이용하고, 어절 내 형태소 확률을 예측할 때 어절의 경계는 제외한 채 전체 문맥을 이용하는 언어모델링이 있다. 이는 터키어 [6]와 일본어 [7]를 대상으로 한 것이다. 하지만 빈번한 어순 도치가 있는 한국어에서 원거리 문맥 정보는 확률 분포를 분산시키는 문제가 있고, 어절 단위의 경계를 이용하지 않는 어절 내 형태소 확률 추정은 불필요한 이전 어절의 정보를 이용하는 문제가 있다.

본 논문에서는 용언형태소를 포함하는 어절과 그 이전 어절 간에 어순 구조가 있음을 밝히고, 두 어절의 결합을 통해 해당 인접 어절 간의 어순 구조를 고려하는 새로운 단위의 세그먼트를 정의한다. 또한 이렇게 정의된 세그먼트 단위에 적합한 문맥 정보를 이용하는 언어모델링을 제안한다. 이는 문맥에 따라서 다음과 같은 두 경로의 확률을 추정하는 문맥 기반 two-path 언어모델이다. 먼저 세그먼트 생성 확률 추정에서는 twoply HMM과 같이 근거리 문맥만을 이용하여 확률을 추정함으로써 확률분포의 잘못된 분산을 막는다. 그리고 세그먼트 내 형태소 확률 추정에서는 twoply HMM의 한계인 HMM을 확장하여 세그먼트 내 문맥을 효과적으로 보는 세그먼트 내 형태소 단위 n-gram 모델을 제안한다. 마지막으로, 제안하는 세그먼트 단위와 언어모델을 어절 단위의 twoply HMM 및 기존 연구 [6]과 [7]의 언어모델과 비교실험을 하였다. 그 결과 제안하는 세그먼트와 언어모델이 기존 연구보다 한국어에서 더 적합함을 밝힐 수 있었다.

이후, 본 논문의 구성은 다음과 같다. 2장에서는 기존 연구를 논하고, 3장에서는 본 논문에서 제안하는 새로운 세그먼트 단위에 대해 기술한다. 4장에서는 본 연구에서 제안하는 two-path 언어모델링을 설명하고, 5장에서는 실험 및 실험 결과에 관하여 기술한다. 마지막으로 결론 및 향후 연구에 대하여 언급한다.

II. 관련 연구

본 논문에서 사용하는 형태소 단위의 언어모델은 교착어의 형태소 교착에 의한 어절 생성으로 어절의 종류가 급격히 증가하는 문제를 완화하기 위해 만들어진 것이다. 따라서 많은 교착어에서 형태소 단위 언어모델 [8]을 이용하고 있고, 한국어 [9]에서도 이미 적용되고 있다.

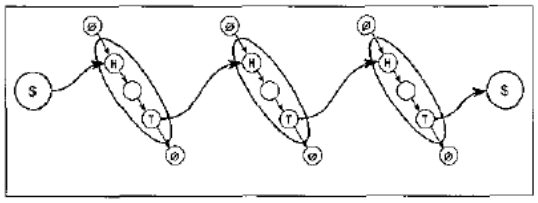

그림 1. 어절 생성 확률과 어절 내 형태소 전이 확률 경로 [4]
Fig. 1. Paths of word generation probabilities and intra-word morpheme transition probabilities.

형태소 단위 언어모델의 대표적인 경우는 형태소 단위 n-gram 모델 [8][9]이다. 하지만 이 모델은 한국어 어절의 어순 변화에 취약한 문제를 가지고 있다.

한국어에서 어절의 자유로운 도치 현상을 고려하는 기존 연구 방법은 twoply HMM [4]이 있다. 이 모델은 형태소 단위 모델에서는 고려할 수 없는 어절 단위 문맥 정보를 이용하는 모델로, HMM을 이용하여 품사 부착을 하는 모델이다. 해당 연구의 모델은 문맥에 따라 어절 생성 확률과 어절 내 전이 확률이라는 두 경로에 HMM을 적용한 방법으로 문맥 기반 two-path 모델이다. 각 경로에서 어절 생성 확률은 어절의 첫 번째 형태소가 나타날 확률로 추정하고, 어절 내 전이 확률은 어절 내 형태소를 HMM을 이용하여 추정한다. 그림 1은 twoply HMM이 고려하는 두 경로를 보여준다. 이 모델에서는 어절 생성 확률을 추정할 때 이전 어절의 마지막 형태소만을 고려한다. 이렇게 근거리 문맥 정보만을 이용함으로써 이전 어절의 변화로 인한 확률 분포의 분산을 최소화하는 언어모델을 만들 수 있다.

표 1. 한국어 어순 변화의 예
Table 1. Example of transformation in word sequence of Korean.

(ㄱ)	밥+을 # 먹+으러 # 식당+에 # 가+ㄹ+것+이+다
(ㄴ)	식당+에 # 밥+을 # 먹+으러 # 가+ㄹ+것+이+다

표 1의 예제에서 "밥"의 확률에 대해 형태소 단위 n-gram 모델[1]은 (ㄱ)에서는 이전 문맥이 없으므로 P(밥)의 확률을 취한다면, (ㄴ)에서는 P(밥|식당,에)의 확률을 갖는다. 반면에 twoply HMM은 (ㄱ)에서는 P(밥)으로 동일하지만, (ㄴ)에서는 P(밥|에)로 이전 문맥이 축약된다. 이는 이전 어절에 장소가 있었는지의 여부만으로 확률을 나눌 뿐, 그 이상의 문맥을 고려하지 않음으로써 어절의 도치에 의한 이전 어절의 심한 변화에도 확

1) 형태소 단위 trigram 모델에 back-off를 이용한 예제

률 분포가 크게 분산되지 않는 효과를 얻을 수 있다.

그러나 기존 연구 [4]의 twoply HMM은 어절의 자유 어순 구조만을 고려하여 용언형태소를 포함한 어절과 그 이전 어절 간 어순 구조를 반영하지 못하는 문제가 있다. 예를 들어 "먹"을 예측할 때, twoply HMM은 $P(\text{먹}|\text{을})$를 이용한다. 이는 "밥"이라는 행위의 대상이 되는 문맥을 이용하지 못하기 때문에 $P(\text{먹}|\text{밥},\text{을})$의 확률보다 부정확한 확률 추정이 발생한다. twoply HMM의 또 다른 문제는 어절 내 형태소 전이 확률을 추정할 때 HMM의 특성에 의해 왼쪽 한 문맥 이상을 고려하지 않아 충분한 문맥 정보를 이용하지 못한다는 것이다.

기존의 twoply HMM과 비슷한 종류의 다른 교착어 연구는 터키어 [6]와 일본어 [7]에서도 있었다. 두 언어모델 모두 어절의 첫 형태소와 어절 내 형태소 열의 추정을 달리하는 언어모델이다. [6]과 [7]의 연구는 어절의 첫 형태소를 예측할 때 이전 두 어절의 첫 형태소를 이용하여 추정하는 방법을 사용하였다. 그러나 예제에서도 볼 수 있듯이 어절의 어순이 바뀌는 교착어의 자유 어순 구조에서는 이전 두 어절의 첫 형태소 열이 좋은 문맥 정보가 될 수 없다. 따라서 기존 연구 [6]과 [7] 모두, 앞선 (ㄱ)과 (ㄴ) 예제에서 "식당+에"의 어순 변화에 대해 올바른 확률을 추정 할 수 없다. 반면에 어절 내 형태소 확률 추정에서는 n-gram 모델을 이용하여 twoply HMM보다 문맥 정보를 많이 이용한다. 하지만 어절의 경계를 이용하지 않고 형태소 단위 n-gram 모델을 적용하기 때문에 어절의 어순 도치에 적절하지 않다. 그 이유는 n-gram 모델의 문맥 n이 늘어감에 따라 이전 어절의 불필요한 형태소 열 문맥 정보가 반영되기 때문이다.

Ⅲ. 인접 어절의 어순 구조를 고려한 세그먼트 단위

3.1. 한국어 어절의 구성

한국어는 형태소의 열로 구성된 어절 단위를 갖는다. 한국어 어절은 그림 2와 같이 하나의 실질형태소와 여러 개의 기능형태소로 구성되어 있는 특성[2]을 가지고 있다. 실질형태소는 단어의 의미를 표현하며 주로 명사나 동사

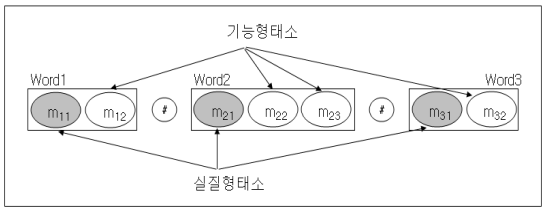

그림 2. 한국어 어절의 구성
Fig. 2. Formation of words in Korean

등으로 구성되고, 기능형태소는 실질형태소의 문법적 관계를 설명하며 조사나 어미 등으로 구성된다.

어절 내의 형태소 열은 어두에 하나의 실질형태소가 나타나고 그 뒤에 기능형태소 열이 나타난다. 각 기능형태소는 그 역할에 맞게 문법적으로 정해진 순서에 따라 나타난다. 예를 들면 "식사+하+시+었+다"의 어절에서 "식사"는 어절의 의미를 나타내는 실질형태소이다. 그 뒤에 나타나는 형태소 열은 모두 기능형태소로, "하"는 행동을 "시"는 높임을 "었"은 과거를 "다"는 종결을 나타내는 역할을 한다. 이렇게 각각의 형태소 열은 모두 그 위치가 정해져서 나타나게 된다.

3.2. 한국어 어절의 어순 구조

한국어는 기능형태소에 의해 문법적 역할이 결정되므로 어절이 자유로운 어순을 갖는다. 하지만 모든 어절이 자유롭게 배치되는 것은 아니다. 일반적으로 한국어 문장에서 주어는 문장의 처음에 나타나고 서술어는 문장의 마지막에 나타나는 등 어절의 순서가 존재한다. 따라서 한국어는 '부분적으로 어순을 갖는 언어'라고 할 수 있다. 그러나 주어의 생략이 매우 빈번하고 [10], 서술어가 나타날 시점을 예측하기 어렵기 때문에 이러한 특성을 반영하여 언어모델링을 하는 것은 매우 어려운 일이다.

본 연구에서는 언어모델링이 용이하면서도 한국어의 어순 구조를 반영할 수 있도록 두 인접 어절 간에 어순 구조가 존재하는 경우를 살펴 볼 것이다. 그 이유는 한국어의 자유로운 어순 구조로 인해 문장 내 어절 열을 이용하여 다음 어절을 예측하는 것은 어려운 일이지만, 인접한 두 어절 간에 어순 구조가 존재하면 두 어절 중 처음 어절을 이용하여 다음 어절을 예측하기가 용이하기 때문이다.

3.3. 용언을 포함한 어절과 그 이전 어절 간의 관계

한국어에서 인접한 두 어절 간에 어순을 갖는 경우는 용언을 포함하는 어절과 그 이전 어절 사이에서 나타난다 [10]. 용언형태소 이전에 나타나는 어절은 용언의 목적어와 부사이다. 각각은 모두 용언을 포함하는 어절과 밀접

2) 한 어절에서 실질형태소가 어두 (語頭)에 나타나지 않는 경우는 접두사가 존재할 때가 있다. 또한 실질형태소가 여러 개 나타나는 경우는 복합명사인 경우이다. 본 논문에서는 이들 경우에 대하여 각 형태소를 결합하여 하나의 실질형태소로 고려하였다.

표 2. 용언형태소와 이전 어절 간 어순 구조의 예
Table 2. Examples of word-order structures between verbal morphemes and preceding words.

용언의 목적어	(부사)	다음 어절의 후보 용언
문+을	(세게 · · ·)	열/VV, 닫/VV · · ·
옷+을	(빨리 · · ·)	입/VV, 벗/VV · · ·
밖+을	(멀리 · · ·)	보/VV, 응시/VV · · ·
꽃+이	(매우 · · ·)	아름답/VX, 예쁘/VX · · ·
오류+가		아니/VCN, 있/VX, · · ·

한 관계가 있다.

표 2는 용언과 이전 어절 간에 나타나는 어순 관계의 예이다. 일반적으로 용언의 목적어와 용언이 연속으로 나타나고 부사가 그 사이에 포함될 수 있다. 그 특성으로, 용언을 포함한 어절과 용언의 목적어 사이에는 어순의 도치나 생략이 일어나지 않는다. 그리고 의미적으로는 용언의 목적어 다음에 나타나는 용언형태소의 후보수가 매우 적다. 예를 들어 "문을"이라는 대상 어절 다음에는 "열/VV"이나 "닫/VV"이라는 용언형태소가 후보가 될 수 있고, "입/VV"나 "아니/VCN"와 같은 용언형태소는 나타나지 않는다. 또한 부사가 나타나는 경우에도 "문+을 # 세게 # 닫+다"와 같은 문장은 존재하지만 "문+을 # 매우 # 닫+다"와 같은 문장은 존재하지 않는다. 이 같은 특성에 따라 본 논문에서는 용언을 포함하는 어절의 이전 어절을 용언의 대상 어절로 정의하도록 한다.

3.4. 용언의 대상 어절을 이용한 용언형태소 예측 효과

용언의 대상 어절이 갖는 형태소 열을 통해 실질형태소인 용언을 예측하는 것은 적절한 방법이 될 수 있다. 그 이유는 용언을 예측하기 위해 이전 어절 형태소 정보를 모두 이용하면 $P(\text{열}/VV\text{문},\text{을})$, $P(\text{입}/VV\text{옷},\text{을})$, $P(\text{보}/VV\text{밖},\text{을})$의 확률로 각 용언형태소의 확률을 추정할 수 있기 때문이다. 이 경우 "열/VV"을 예측할 때 "옷"과 "밖"은 나타날 수 없으므로 효과적으로 예측 후보수를 줄이면서 정확한 확률 예측이 가능하다. 이는 부사가 나타나는 예인 $P(\text{닫}/VV\text{세게})$에서도 나타난다. 이 경우 역시 "매우"나 "상당히"는 "닫/VV"을 예측할 때 나타나지 않기 때문이다. 그 결과 II장에서 밝힌 twoply HMM [4]에서 인접 어절의 어순을 고려하지 않아 발생하는 확률 추정 문제를 해결할 수 있는 방법이 된다.

3.5. 인접 어절의 어순을 고려한 세그먼트 단위

용언의 대상 어절과 용언을 포함한 어절 사이의 어순을 고려하여 언어모델링을 하는 방법 중에 하나는 용언을 포함한 어절과 용언의 대상 어절을 결합하여 언어모델을 적용하는 것이다. 이때 어절의 결합이 일어나 만들어지는 새로운 어절 단위를 본 논문에서는 세그먼트라고 정의한다. 새롭게 만들어지는 세그먼트 단위로 언어모델을 하는 이유는 3.1절에서 언급했듯이 어절 내 형태소는 기능형태소로 그 순서가 일정하게 나타나며, 이 특징을 이용하여 세그먼트 내 형태소 확률 모델을 구축하기 때문이다. 따라서 어순이 존재하는 두 어절을 결합하여 한 세그먼트로 만듦으로써 해당 두 어절의 어순을 고려하는 방법을 제안할 수 있다.

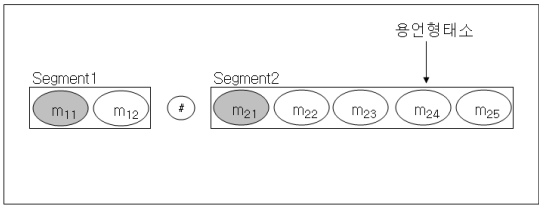

그림 3. 인접 어절의 어순 구조를 고려한 세그먼트 단위
Fig. 3. Segment unit considering the word-order structure of adjacent words.

그림 3은 그림 2에서 3번째 어절 첫 형태소인 m_{31}이 용언형태소일 때 새로운 세그먼트 단위를 보여주는 그림이다. 따라서 그림 3에서는 해당 용언형태소가 2번째 세그먼트가 되므로 m_{24}으로 나타나는 것을 볼 수 있다. 한국어에서는 간단하게 용언형태소 앞의 띄어쓰기만 제거함으로써 용언의 대상 어절과 용언형태소를 포함하는 어절을 통합하는 새로운 세그먼트 단위를 만들 수 있다.

IV. Two-Path 언어모델링

본 논문에서 다루는 언어모델은 twoply HMM과 같이 세그먼트의 첫 형태소와 세그먼트 내 형태소 열의 확률을 문맥에 따라 두 경로를 이용하여 예측하는 문맥 기반 two-path 언어모델이다. 본 논문에서 제안하는 two-path 언어모델을 일반화 하면 다음과 같다.

4.1. 언어모델의 Two-Path 분리

문장이 나타날 확률은 개별 형태소의 생성 확률을 이용하여 다음과 같이 구할 수 있다.

$$P(sentence) = \prod_{i=1}^{N}\prod_{j=1}^{len(i)} P(m_j^i | m_1^1 ... m_{j-1}^i) \quad (\text{단}, m_0^i = m_{len(i-1)}^{i-1}) \tag{1}$$

m_j^i는 문장 내 i번째 세그먼트의 j번째 형태소이다. 위첨자 i는 문장 내 세그먼트의 인덱스이고, 아래첨자 j는 세그먼트 내 형태소의 인덱스이다. N은 문장 내 전체 세그먼트의 개수이며, $len(i)$는 문장 내 i번째 세그먼트의 형태소 개수를 알려주는 함수이다.

수식 (1)은 세그먼트와 형태소 열로 이루어진 문장의 확률을 구하는 식이다. 이 때, 세그먼트의 첫 형태소인 m_1^i의 확률은 문장의 처음 형태소 m_1^1부터 이전 어절의 마지막 형태소 $m_{len(i-1)}^{i-1}$까지 구성되는 형태소 열로 구한다. 세그먼트 내의 나머지 형태소 m_j^i는 m_1^1부터 세그먼트 내의 이전 형태소인 m_{j-1}^i까지 구성되는 형태소 열로 구한다. 세그먼트 단위를 생성하는 방법은 3.5절에서 주어졌기 때문에 각각의 확률은 띄어쓰기 정보를 이용해서 구할 수 있다.

$$P(j=1|\theta)=1, P(j\neq 1|\theta)=0 \quad (\text{단}, \theta: \text{띄어쓰기}) \quad (2)$$

수식 (2)는 띄어쓰기를 이용하여 세그먼트의 첫 형태소와 세그먼트 내 형태소를 분리하는 간단한 방법이다. 따라서 수식 (1)을 세그먼트 생성 확률과 세그먼트 내 형태소 확률로 분리할 수 있다.

$$P(sentence)= \prod_{i=1}^{N}\prod_{j=1}^{len(i)} \begin{Bmatrix} P(m_j^i|m_1^1...m_{len(i-1)}^{i-1}), & \text{if } j=1 \\ P(m_j^i|m_1^1...m_{j-1}^i), & \text{if } j\neq 1 \end{Bmatrix} \quad (3)$$

이렇게 분리한 수식 (3)의 처음 확률은 세그먼트 생성 확률을 모델링 한 것이고, 두 번째 확률은 세그먼트 내 형태소 확률을 모델링 한 것이다. 따라서 각 확률의 유도를 통해 수식 (3)에 두 확률을 적용한다.

4.2. 세그먼트 생성 확률

형태소의 세그먼트 생성 확률은 II장의 기존 연구 분석에서 언급했듯이 자유로운 세그먼트 순서에 의한 확률 분포가 분산되는 문제를 해결하기 위하여 그림 1과 같이 이전 어절의 마지막 형태소로 추정한다. 수식은 다음과 같다.

$$P(m_1^i|m_1^1...m_{len(i-1)}^{i-1}) \approx P(m_1^i|m_{len(i-1)}^{i-1}) \quad (4)$$

따라서 수식 (4)의 세그먼트 생성 확률은 이전 세그먼트의 마지막 형태소와 현재 세그먼트의 첫 형태소의 bigram 확률로 구할 수 있다.

4.3. 세그먼트 내 형태소 확률

세그먼트 내 형태소 확률은 세그먼트 내 형태소 단위 n-gram 모델로 확률을 추정한다. 그 이유는 세그먼트 내 형태소는 형태소 간 순서가 일정하게 나타나고 n-gram 모델은 어순 정보를 이용하는 대표적인 언어모델이기 때문이다. 따라서 수식은 다음과 같다.

$$P(m_j^i|m_1^1...m_{j-1}^i) \approx P(m_j^i|m_{j-n+1}^i...m_{j-1}^i), \text{ if } j\neq 1 \quad (5)$$

수식 (5)는 임의의 i번째 세그먼트 내에서 형태소 단위 n-gram 모델을 적용한 확률이다. 이를 통해 기존의 twoply HMM [4]에 비해 충분한 문맥으로 확률을 추정할 수 있다. 또한 세그먼트의 경계를 이용하지 않는 기존 연구 [6][7]보다 세그먼트 순서의 도치에 적합한 세그먼트 내 형태소 확률을 구성할 수 있다.

4.4. Two-Path 확률 통합 최종 언어모델

수식 (4)와 수식 (5)에서 유도한 세그먼트 생성 확률과 세그먼트 내 형태소 확률을 수식 (3)에 적용하면 다음과 같은 two-path 언어모델의 최종수식이 나온다.

$$P(sentence)= \prod_{i=1}^{N}\prod_{j=1}^{len(i)} \begin{Bmatrix} P(m_j^i|m_{len(i-1)}^{i-1}), & \text{if } j=1 \\ P(m_j^i|m_{j-n+1}^i...m_{j-1}^i), & \text{if } j\neq 1 \end{Bmatrix} \quad (6)$$

V. 실험 및 평가

5.1. 실험 환경

본 연구에서는 21세기 세종 프로젝트[3]의 형태소 품사 부착 말뭉치를 정제하여 이용한다. 이 말뭉치는 뉴스, 소설, 수필, 가사 등 다양한 종류의 문서를 포함한다. 말뭉치는 약 100만 문장과 1,200만 어절 및 2,500만 형태소로 구성되어 있다. 말뭉치는 80%의 학습 말뭉치와 10%의 개발 말뭉치, 10%의 실험 말뭉치를 임의로 분할하였다. 개발 말뭉치는 cut-off를 통한 형태소 종류의 크기를 결정하는데 이용[4]하였으며, 실제 실험 결과는 실험 말뭉치에

3) http://www.sejong.or.kr/

4) cut-off는 10개 이하 빈도의 형태소로 하였고 cut-off 비율에 따라 혼잡도와 언어모델 크기는 trade-off 관계가 있다.

서 얻었다. 그 결과 형태소의 종류는 37,753개이고 OOV (Out-Of-Vocabulary) rate는 2.58%이다.

평탄화 (smoothing) 기법은 Katz back-off [11]를 사용하였고, 언어모델은 SRILM toolkit [12]으로 생성하였다.

성능평가는 언어모델에서 이용하는 평가척도인 혼잡도 (perplexity; ppl) [3]를 이용하여 혼잡도가 줄어들수록 성능이 높아지고 혼잡도가 증가할수록 성능이 하락한다.

5.2. 실험 결과

실험 결과는 다음과 같은 순서를 따른다. 5.2.1절에서는 제안하는 세그먼트 단위와 어절단위를 비교하고, twoply HMM [4]과 제안하는 언어모델을 비교한다. 5.2.2절에서는 터키어 [6]와 일본어 [7]에서 제안한 세그먼트 생성 확률 추정 및 세그먼트 내 형태소 확률 추정을 제안하는 언어모델과 비교한다. 이러한 비교실험을 통해 제안하는 모델이 한국어에 가장 적합한 언어모델임을 밝힌다.

5.2.1. Twoply HMM과 제안하는 Two-Path 언어모델 비교 실험

본 논문에서 제안하는 세그먼트 단위와 어절 단위 각각에서의 twoply HMM의 성능은 다음과 같다.

표 3. 어절 단위와 제안하는 세그먼트 단위의 혼잡도 비교 실험 결과
Table 3. Perplexity comparison between word unit and the proposed segment unit.

Twoply HMM	ppl (증감률%)
어절 단위	97.96
제안하는 세그먼트 단위	86.80 (-11.39%)

표 3은 twoply HMM을 어절 단위와 제안하는 세그먼트 단위에 적용한 실험 결과이다. 실험 결과에서는 제안하는 세그먼트 단위에서 혼잡도가 떨어지는 것을 볼 수 있다. 이는 제안하는 세그먼트 단위가 twoply HMM을 이용할 때 더 효과적임을 나타낸다.

표 4. 어절 단위와 제안하는 세그먼트 단위에서 용언형태소의 혼잡도 비교 결과
Table 4. Comparison of verbal morpheme perplexity between word unit and the proposed segment unit.

Twoply HMM	용언형태소 ppl (증감률%)
어절 단위	683.02
제안하는 세그먼트 단위	81.37 (-88.08%)

표 4는 어절과 세그먼트 간의 차이가 되는 용언형태소의 혼잡도를 어절과 세그먼트의 단위에서 각각 보여준다. 실험 결과에 따르면 용언형태소를 예측할 때에는 용언의 대상 어절인 이전 어절의 형태소를 이용하여 언어모델링 하는 것이 바람직함을 보여주고 있다. 제안된 세그먼트 단위에서 혼잡도가 낮은 이유는, 용언형태소는 다른 실질형태소와 다르게 기능형태소와 비슷한 속성을 보이기 때문이다. 이는 기능형태소 (9.01%)와 용언형태소 (3.04%)의 종류가 한정되어 있고 새로운 종류의 형태소가 나타날 확률이 적으며, 출현빈도가 높은 특성을 갖는 것과 밀접한 관계를 가지고 있다. 따라서 용언형태소를 기능형태소와 함께 세그먼트 내 형태소 확률 모델로 언어모델링을 함으로써 높은 성능 향상을 얻을 수 있었다.

표 5. 제안하는 세그먼트 내 형태소 단위 n-gram을 적용한 two-path 언어모델의 비교 실험 결과
Table 5. Comparison of two-path language models based on intra-segment morpheme unit n-grams.

Two-Path 언어모델	ppl (증감률%)
세그먼트 내 형태소 bigram	86.80
세그먼트 내 형태소 trigram	72.80 (-16.12%)
세그먼트 내 형태소 fourgram	71.34 (-17.81%)
세그먼트 내 형태소 fivegram	71.17 (-18.00%)

표 5는 제안하는 세그먼트 내 형태소 확률을 추정할 때 형태소 단위 n-gram 모델을 적용한 결과이다. 먼저 bigram 모델을 적용한 모델은 기존의 twoply HMM과 동일한 모델이며, 이 모델을 기준으로 n-gram 모델의 n을 증가시키며 비교실험 하였다. 그 결과 세그먼트 내 형태소 확률을 구할 때는 적정 수준의 문맥을 고려하는 것이 적합했으며, fourgram 모델에서 혼잡도를 약 17.81% 낮추며 수렴하는 성능을 보여주는 것을 알 수 있다. 따라서 세그먼트 내 형태소 확률 추정은 세그먼트 내 HMM보다 형태소 단위 n-gram 언어모델링이 적합하다.

표 6. 어절 단위 twoply HMM과 제안하는 세그먼트 단위 two-path 언어모델의 혼잡도 비교 실험 결과
Table 6. Perplexity comparison between word unit based twoply HMM and the two-path language model based on the proposed segment unit.

Two-Path 언어모델	ppl (증감률%)	용언형태소 ppl (증감률%)
어절 단위 twoply HMM	97.96	683.02
제안된 세그먼트 단위 세그먼트 내 trigram	72.80 (-25.68%)	40.76 (-94.03%)

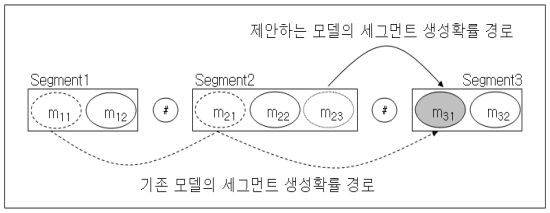

그림 4. 각 언어모델이 형태소 m_{31}을 예측할 때 세그먼트 생성 확률 경로

Fig. 4. Path of segment generation probability in case of predicting m31.

표 6은 어절 단위 twoply HMM [4]과 제안하는 세그먼트 단위 two-path 언어모델과의 최종 성능 비교를 보여준다. 최종 성능은 많은 응용 분야에서 이용하는 trigram 모델을 세그먼트 내 형태소 확률 모델로 이용하였다. 실험 결과 언어모델의 혼잡도는 25.68% 줄었으며 특히 어절과 세그먼트의 차이가 되는 용언형태소에 대해서 94.03%의 혼잡도를 줄이는 결과를 보였다.

5.2.2. 원거리 문맥과 세그먼트 경계를 이용하지 않는 Two-Path 언어모델과 제안하는 언어모델 비교 실험

본 연구는 세그먼트 생성 확률 모델을 이용할 때 이전 세그먼트의 마지막 형태소만을 이용하여 언어모델링을 한다. 반면에 기존 연구 [6]과 [7]는 이전 두 세그먼트의 첫 형태소 열을 이용하여 언어모델링을 한다. 그림 4는 각 언어모델의 세그먼트 생성 확률의 경로[5]이다.

표 7. 세그먼트 생성 확률의 경로에 따르는 two-path 언어모델의 혼잡도 및 파라미터 크기 비교 실험 결과

Table 7. Comparison of perplexity and parameter size of two-path language model depending on segment generation probability paths.

Two-Path 언어모델	ppl (증감률%)	파라미터 크기 (증감률%)
원거리 세그먼트 생성 확률 모델	72.45	3,748,794
제안하는 세그먼트 생성 확률 모델	72.80 (+0.48%)	2,351,768 (-37.26%)

표 7에서 볼 수 있듯이 이전 두 세그먼트의 첫 형태소와 이전 세그먼트의 마지막 형태소와의 혼잡도 차이는 거의 없다. 그 이유는 형태소 종류의 대다수를 차지하는 세그먼트의 첫 형태소 (약 93%) 열로 구성된 이전 두 세그먼트의 첫 형태소 열은 데이터부족 문제가 심하기 때문에 back-off에 의해 확률 값이 결정되기 때문이다. 반면에 파라미터의 크기는 제안하는 모델이 37.26%가 감소한다. 그 이유는 이전 세그먼트의 마지막 형태소는 기능형태소가 대다수이므로, 형태소 간에 나타나는 조합의 개수가 세그먼트의 첫 형태소보다 적기 때문이다. 따라서 two-path 언어모델의 세그먼트 생성 확률은 근거리 문맥 정보인 이전 세그먼트의 마지막 형태소가 적절함을 알 수 있다.

표 8. 세그먼트 경계의 이용 여부에 따르는 세그먼트 내 형태소의 혼잡도 비교 실험 결과

Table 8. Comparison of intra-segment morpheme perplexity depending on the use of segment boundaries.

Two-Path 언어모델	세그먼트 내 형태소 ppl (증감률%)
세그먼트 경계 미 이용	15.64
세그먼트 경계 이용	13.79 (-11.82%)

표 8의 결과는 세그먼트 경계의 이용 여부에 따르는 세그먼트 내 형태소의 혼잡도를 보여준다. 실험결과에 의하면 세그먼트의 경계를 이용함으로써 혼잡도를 11.82%를 줄일 수 있었다. 그 이유는 세그먼트 도치로 인한 이전 세그먼트의 변화에 따라, 이전 형태소 열이 갖는 확률 분포의 분산을 줄일 수 있었기 때문이다.

VI. 결론 및 향후 연구

본 논문에서는 한국어의 인접 어절 간 어순을 고려하여 언어모델링을 하는 방법을 제안하였다. 이를 위하여 먼저 인접 어절 간 어순이 나타나는 용언 대상 어절과 용언을 포함하는 어절 사이의 결합을 통한 세그먼트 단위를 정의하였다. 또한 제안한 세그먼트 단위의 문맥에 적합한 two-path 언어모델을 세그먼트 생성 확률과 세그먼트 내 형태소 확률 각각에 적합한 언어모델링으로 제안하였다. 그 결과 제안하는 세그먼트 단위와 two-path 언어모델은 기존의 어절 단위 언어모델들이 갖는 용언 대상 어절과 용언을 포함하는 어절 간 어순을 고려하지 않는 문제를 해결하였고, 한국어에 더 적합한 언어모델임을 보였다.

본 연구의 향후 연구는 제안하는 모델을 한국어와 비슷한 어순의 구조를 갖는 다른 교착어에 적용하는 것이다. 또한 실제 응용인 기계번역이나 음성인식 등에서 제안하는 언어모델을 이용해 볼 수 있을 것이다.

5) 기존 연구 [6]과 [7]은 원거리 문맥을 고려할 때 보간법을 이용하여 n-gram이나 이전 두 어절의 도치를 고려했으나, 본 논문은 원거리 문맥의 효과만을 보기 위해 그림 4의 경로를 이용하여 비교하였다.

감사의 글

이 논문은 2단계 BK21사업과 2008년도 한국과학재단의 지원을 받아 수행된 연구임 (No. R01-2006-000-11162-0).

참 고 문 헌

1. F. Jelinek, "Self-organized language modeling for speech recognition", Readings in Speech Recognition, A. Waibel and K. F. Lee, eds., Morgan Kaufmann, 450-506, 1990.
2. P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and Roossin, P.S. "A statistical approach to machine translation", Computational Linguistics, **16**(2), 79-85, 1990.
3. D. Jurafsky and J. H. Martin, *Speech and Language Processing* (Prentice Hall, 2007) Chap.4, pp.83-121.
4. 김진동, 임희석, 임해창, "Twoply HMM: 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델", 정보과학회논문지 (B), **24**(12), 1502-1512, 1997.
5. L. Rabiner and B. Juang, "An Introduction to hidden Markov models", ASSP Magazine IEEE Signal Processing, **3**(1), 4-16, 1986.
6. E. Arısoy and M. Sarac¸lar, "Lattice extension and rescoring based approaches for LVCSR of turkish", in INTERSPEECH, 1025-1028, 2006.
7. J. Gao, H. Suzuki, and Y. Wen, "Exploiting headword dependency and predictive clustering for language modeling", in EMNLP-2002, 248-256, 2002.
8. M. Creutz, T. Hirsim¨aki, M. Kurimo, A. Puurula, J. Pylkk¨onen, V. Siivola, M. Varjokallio, E. Arısoy, M. Sarac¸lar, and A. Stolcke, "Morph-based speech recognition and modeling of out-of-vocabulary words across languages", ACM TSLP, **5**(1), 2007.
9. O. Kwon and J. Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units", Speech Communication, **39**(3-4), 287-300, 2003.
10. N. Hideki, *Corpus-based approaches to sentence structures* (John Benjamins Pub Co., 2005), Chap.3, pp.51-76.
11. S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer", IEEE Transactions on Acoustics, Speech and Signal Processing, **35**, 400-401, 1987.
12. A. Stolcke, "SRILM-an extensible language modeling toolkit", in ICSLP-2002, 901-904, 2002.

저자 약력

•신 중 휘 (Joong-Hwi Shin)

1982년 4월 4일생
2006년 : 고려대학교 컴퓨터학과 학사
2006년~현재 : 고려대학교 컴퓨터학과 석사과정

•박 재 현 (Jae-Hyun Park)

1979년 9월 19일생
2005년 : 고려대학교 컴퓨터학과 학사
2007년 : 고려대학교 컴퓨터학과 석사

•이 정 태 (Jung-Tae Lee)

1983년 10월 7일생
2006년 : 고려대학교 컴퓨터학과 학사
2008년 : 고려대학교 컴퓨터학과 석사
2008년~현재 고려대학교 컴퓨터·전파통신공학과 박사과정

•임 해 창 (Hae-Chang Rim)

1953년 2월 26일생
1981년 : Missouri 주립대학 학사
1983년 : Missouri 주립대학 석사
1990년 : Texas 주립대학 박사
1991년~1994년 : 고려대학교 전산과학과 조교수
1994년~1999년 : 고려대학교 컴퓨터학과 부교수
1999년~현재 : 고려대학교 컴퓨터·전파통신공학과 교수