# 확률적 언어 모델을 위한 자료 기반 어휘 구축

류성호[0] 김진형
한국과학기술원 전자전산학과 전산학전공
{shryu, jkim}@ai.kaist.ac.kr

# A data-driven approach for lexicon selection for probabilistic language model

Sungho Ryu[0] Jin-Hyung Kim
Div. of Computer Science, KAIST

요 약

한국어를 대상으로 하는 확률적 언어 모델에서는 대부분의 경우 형태소를 기본 어휘로서 사용하고 있다. 그러나, 이러한 모델들은 학습 및 검증을 위하여 사람에 의하여 형태소 분석이 이루어진 말뭉치를 필요로 한다. 또한, 형태소의 자동 분석은 현재 표준말을 중심으로 이루어져 있어 그 적용 분야에도 한계가 있다. 본 논문에서는 한국어의 특징을 고려하여 확률적 언어 모델의 구축에 적합한 어휘의 선택 기준에 대하여 고찰하고, 통계적인 기준을 통하여 확률적 언어 모델의 어휘를 구축하는 방법을 제안한다.

## 1. 서론

확률적 언어 모델은 임의의 문자열이 주어진 언어에서 사용되는 빈도를 표현하는 확률 분포 함수이다. 확률적 언어 모델은 대량의 말뭉치로부터 얻어진 사용빈도의 통계 정보를 바탕으로 생성된다. 현재 확률적 언어 모델은 음성인식, 형태소 분석, 문자인식, 철자교정등 다양한 분야에서 사용되고 있다.

현재, 한국어에서는 형태소를 기본 어휘로 하는 언어 모델이 주로 사용되고 있다. 교착어라는 특성상 한국어에서는 여러 개의 형태소가 조합되어 하나의 어절을 형성하지만, 형태소의 굴절현상이나 불규칙 현상이 다양하여 어절을 기본 어휘로 사용하기에는 무리가 있기 때문이다.

그러나, 형태소 기반 모델에는 크게 두 가지 문제점이 존재한다. 첫째로, 형태소 기반 모델을 훈련시키기 위해서는 형태소 분석이 이루어진 말뭉치가 필요하다는 점이다. 형태소 분석이 이루어진 말뭉치는 사람에의하여 직접 생성되어야 하기 때문에 원시 말뭉치에 비하여 상대적으로 적다. 뿐만아니라 형태소의 정의가 그 적용분야에 따라 주관적으로 이루어질 수 있기 때문에[11], 형태소 분석 말뭉치 사이의 호환성에도 문제가 있다. 둘째로, 형태소 분석이 문법에 맞는 표준말을 대상으로 이루어진다는 점이다. 휴대폰이나 PDA등과 같이 메모나 문자 메시지 전송이 주를 이루는 환경에서는 표준말보다는 구어체나 통신체등을 잘 표현하는 언어 모델이 보다 적합하다.

본 논문에서는 형태소의 대안으로써 대량의 원시 말뭉치로부터 확률적 언어 모델을 위한 어휘를 학습시킬 수 있는 방법을 제안한다. 제안된 방법은 어절들간에 공통적으로 빈번히 사용되는 문자열들을 추출하여 한국어를 위한 확률적 언어 모델의 기본 어휘로서 형태소 대신 사용될 수 있게 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 확률적 언어 모델의 설계시 고려 사항 및 한국어에서의 문제점에 관하여 설명하고, 3장에서는 확률적 언어 모델을 위한 어휘의 선택 기준에 관하여 고찰한다. 4장에서는 원시말뭉치로부터 어휘를 구축하는 방법에대하여 설명한다. 5장에서는 제안된 방법을 실험을 통하여 원시 말뭉치로부터 학습가능한 다른 방법들과 비교하며, 6장에서 전체적인 결과를 정리한다.

## 2. 확률적 언어 모델의 설계

확률적 언어 모델은 어휘와 확률 모델의 두 가지 구성요소로서 정의할 수 있다.

어휘는 주어진 언어를 표현하기 위해 모델에서 사용하는 기본 단위들의 집합이다. 문법에서의 형태소나 단어와 유사한 역할을 수행하며, 언어 모델이 적용되는 분야 및 주어진 제약 조건에 따라 다르게 정의될 수 있다.

확률 모델은 정의된 어휘의 활용법을 확률적으로 표현하는 모델이다. 일반적으로 언어를 어휘열의 마르코프 과정으로 표현하는 모델을 지칭한다.

마르코프 과정을 기반으로 하는 모델은 그 크기에 의하여 훈련 및 구현에 사실상 제약을 받는다. 확률적 언어 모델의 경우, 어휘와 확률 모델에서 표현하는 상관관계의 길이가 모델의 크기를 결정한다. 즉, 어휘를 L이라하고, 모델에서 표현하는 상관관계의 길이를 n이라 할 때, 전체 모델이 표현하는 공간의 크기는 $|L|^n$에 비례한다. 따라서, 가급적 적은 크기의 모델을 사용하여 주어진 자료의 특성을 표현할 수 있도록 어휘와 모델 구조를 선택하는 것이 확률적 언어 모델 설계의 주안점이라고 할 수 있다.

영어의 경우 어절을 기본 어휘로 사용하는 모델이 사실상 표준으로 자리잡고 있다. 영어에서의 어절은 대부분 하나의 형태소로 구성되어 있으며 형태 변이가 제약되어 있어 어휘 사전을 구축하기 용이하다. 또한, 간단한 과정을 거쳐 원시 텍스트로부터 추출해 낼 수 있어 대량의 통계정보를 얻을 수 있다.

반면, 한국어의 경우에는 어절이 어휘의 기본 단위로 사용되기에 부적합하다. 한국어는 교착어라는 특성상, 하나의 어절이 여러 개의 형태소로 구성되어 있으며, 다양한 종류의 굴절 현상 및 불규칙 현상이 존재한다. 따라서, 유사한 크기의 형태소 사전으로부터 조합 가능한 어절의 수가 영어에 비하여 훨씬 많으며, 이로 인하여 훈련시 자료 부족으로 인한 성능 저하가 발생한다.

이로 인하여 한국어의 확률적 언어 모델에서는 주로 형태소, 혹은 그 유사 단위들이 어휘의 기본 단위로써 사용되어왔다[5,9,10]. 어휘 선택에 사용되는 구체적인 기준은 그 적용분야에 따라 다르지만, 대부분 형태소 분석기에 의하여 생성된 자동 분석 결과를 기반으로 어휘 선택 작업이 이루어진다. 따라서, 현재 대부분의 한국어의 확률적 언어 모델은 형태소 자동 분석과 밀접한 연관을 가지고 있다고 볼 수 있다.

그러나, 한국어 형태소 자동 분석은 아직 오류 발생의 여지가 있고, 적용 가능한 분야의 제약이 존재하는 등의 문제점이 있다. 마찬가지로 형태소를 기반으로하는 어휘를 사용하는 확률적 언어 모델들 역시 동일한 문제점으로 인하여 제약을 받는다. 따라서, 형태소의 자동 분석이 적합하지 않은 분야에 대해서는 확률적 언어 모델의 구축을 위한 별도의 어휘 선택 기준이 필요하다. 형태소와 같이 어휘 선택을 위한 문법적 기준이 주어지지 않을 경우, 주어진 문제와 그 적용 분야를 중심으로 어휘선택에 대한 충분한 고찰이 이루어져야 한다.

## 3. 확률적 언어 모델을 위한 어휘 선택

자연어에서 사용되는 어휘는 고정되어 있지 않다. 시간이 지남에따라 새로운 개념의 도입과 더불어 새로운 어휘가 추가되는 한편, 사용되지 않는 어휘는 사라지기도 한다. 그러나, 어휘들의 사용빈도를 조사할 경우 제한된 수의 어휘가 실제 언어에서 사용되는 빈도의 대부분을 차지한다는 사실이 알려져 있다.

확률적 언어 모델에서 어휘를 선택할 경우에도 주로 사용되는 어휘들을 파악하는 것이 중요하다. 확률적 언어 모델의 학습 과정에는 자료에서 관측되지 않았던 사례들을 표현할 수 있도록 하기 위하여 다양한 스무딩 기법들이 적용된다. 그러나, 이러한 방법들은 본질적으로 모델이 등록된 어휘들의 응용력을 가질 수 있도록 하는 것이 주목적으로써, 미등록어의 사용빈도 표현에는 근본적으로 한계가 있다. 따라서, 실제로 모델 사용시 미등록어의 출현 빈도가 높을 경우 성능 저하가 불가피하다. 그러므로, 확률적 언어 모델에서 사용하는 어휘는 미등록어 출현 빈도의 기대값이 낮을수록 바람직하다.

그러나, 단지 미등록어의 출현 빈도를 낮추기 위하여 낱자처럼 단순한 어휘를 사용하는 데에도 문제가 있다. 이런 경우 언어에서 나타나는 대부분의 내용을 확률 모델을 이용하여 표현해야 한다. 따라서, 모델에서 표현해야 하는 상관관계의 길이가 충분히 길어져야 하지만, 모델이 복잡해짐에따라 훈련자료의 부족 및 구현 문제등과 같은 실질적인 제약조건에 부딪히게 된다.

만일 형태소 분석이 항상 완벽하게 이루어지는 것을 가정하면, 형태소는 확률적 언어 모델의 기본 어휘로서 사용되기에 적합하다. 형태소는 언어를 구성하는 최소 기본 단위로서, 가장 작은 크기의 어휘를 사용하여 언어를 표현하는 것이 가능하다. 또한, 대다수의 신조어들은 기존의 형태소의 조합에 의하여 생성되기 때문에, 적절한 형태소 분석이 이루어질 경우 미등록된 형태소가 출현할 확률도 낮다.

또한, 확률적 언어 모델의 기본 구조상, 형태소를 단위로 하는 마르코프 모델이 성능적인 측면에서도 바람직하다. 일반적으로 마르코프 가정은 모델을 단순화하여 계산상의 복잡도를 줄이기 위하여 적용된다. 그러나, 이러한 가정이 실제로 적용

되는 부분의 확률 분포가 상호 독립적인 분포와 거리가.있을 경우, 마르코프 가정을 적용함으로써 발생하는 정보량의 손실로 인하여 모델의 성능이 저하되게 된다.

실제로 말뭉치를 기반으로 문장을 분석해보면 형태소 경계 부분에서의 문자열의 변이가 형태소 내부에서 보다 다양하게 나타난다. 이는 형태소의 경계 부분에서의 문자열간의 상호 정보량이 형태소 내부에서의 문자열간의 상호 정보량보다 낮다는 사실을 의미한다. 따라서, 형태소의 경계 부분에 마르코프 가정을 적용할 경우, 보다 적은 정보량의 손실로 모델을 단순화하는 것이 가능하다.

그러나, 현재 형태소의 자동 분석은 아직 오류의 여지가 존재하며, 미등록어절 분석의 경우 더욱 오류 발생의 확률이 높다. 때문에, 형태소를 기반으로 하는 확률적 언어 모델은 형태소 분석단계에서의 오류로부터 영향을 받게 된다.

또한, 형태소 기반 모델의 훈련 및 검증에 필요한 자료가 상대적으로 부족하다는 점도 문제가 된다. 형태소 기반 모델의 훈련 및 검증을 위해서는 사람에 의하여 분석이 이루어진 말뭉치가 필요하다. 그러나, '의미를 가지는 가장 작은 언어 단위'라는 형태소의 정의에는 주관적 판단이 개입될 여지가 있다. 이로 인하여 동일한 어절에 대해서도 서로 다른 형태소 분석결과가 나타날 수 있으며, 이는 형태소 분석 말뭉치의 상호 호환성에 문제가 발생할 수 있다는 점을 의미한다.

마지막으로, 형태소 분석 작업이 주로 표준말을 대상으로 이루어지고 있다는 점 역시 문제가 될 수 있다. 필기체 문자 인식이나 음성 인식과 같은 분야에서 확률적 언어 모델의 역할은 입력이 이루어진 그대로 인식될 수 있도록 인식기의 오류를 보정하는 것이다. 현재 표준말 이외에도 구어체, 통신체등 다양한 표현들이 일상적으로 널리 사용되고 있다. 따라서, 오직 표준말만을 표현하는 언어 모델은 실용적인 응용분야에의 적용에 제약을 받게된다.

확률적 언어 모델이 말뭉치로부터 추출된 통계정보를 바탕으로 구축되는 사실을 고려할 때, 가능한 많은 양의 자료를 학습에 사용하는 것이 바람직하다. 따라서, 원시 말뭉치를 바탕으로 학습할 수 있으면서도 경계에서의 정보량 변화특성이 형태소와 유사한 어휘를 선택하는 것이 형태소를 기반으로 하는 어휘 구축 방법의 대안으로서 적합하다고 볼 수 있다.

## 4. 확률적 언어 모델을 위한 자료 기반 어휘 구축

형태소 경계에서의 문자열간의 결합확률분포가 낮은 상호정보량을 가진다는 사실을 역이용하면, 문자열간의 결합확률분포를 바탕으로 문자열을 세부 문자열로 분할하는 것이 가능하다. 이와 같은 방법으로 문자열을 분할할 경우, 형태소와 유사한 특징을 갖는 문자열들의 집합을 구성할 수 있다 [1,2].

문자간의 결합 확률 분포를 바탕으로 문자열을 분할하는 방법은 문자열간의 정보량을 판단하는 기준에 따라 다양한 방법이 존재한다. 그중 낱자 n-gram을 사용하는 방법들은 문자열의 일부분만으로 분할 여부를 판단할 수 있기 때문에, 띄어쓰기나 맞춤법 오류의 발견을 위하여 널리 사용되고 있다. 그러나, '대학'과 '대학생'의 경우에서 볼 수 있듯이 형태소간에는 부분적으로 중복 및 포함관계가 형성되므로 이러한 문자열 분할 기법들에는 근본적인 한계가 존재한다. 따라서, 형태소 분할 결과와 보다 유사한 효과를 얻기 위해서는 어절 전체를 분석하는 것이 바람직하다.

### 4.1 토큰 모델

본 논문에서는 한국어 어절 내에서 사용될 수 있는 임의의 길이를 가지는 문자열들을 토큰이라고 정의하고, 이를 한국어를 위한 확률적 언어 모델의 어휘로서 제안한다. 분할 및 조합의 단순화를 위하여 토큰은 낱자 단위로 경계를 구분하도록 정의된다.

토큰 모델에서 임의의 어절은 토큰들의 확률분포 함수 $p_T$를 바탕으로하는 하나의 표본 관측값으로 간주된다. 즉, $p_T$를 기반으로 표본 추출된 연속된 토큰들이 연결되어 어절을 형성한다.

그러나, 반대로 어절로부터 토큰들을 복원할 경우 다양한 토큰열들이 서로 동일한 어절을 생성할 수 있는 모호성이 존재한다. 구체적으로, 길이 $n$인 어절을 표현할 수 있는 토큰열들은 모두 $2^{n-1}$ 가지가 존재한다. 따라서, 이러한 모호성을 해결하기 위하여 각 어절은 해당 어절을 조합할 수 있는 모든 토큰열 중 가장 높은 확률 값을 가지는 것에 의하여 생성된다고 정의된다.

즉, 임의의 어절의 확률 분포 함수를 $p_W$라 할 때, 임의의 어절 $W$를 생성가능한 모든 토큰열의 집합을 $S$라 하면

$$p_w(W) = \max_{s \in S} p_T(s)$$

이며,

$$s_{best} = \arg\max_{s \in S} p_T(s)$$

와 같이 정의된다.

예)
'달맞이꽃'이라는 어절을 조합가능한 토큰열은
달+맞+이+꽃
달+맞+이꽃
달+맞이+꽃
달+맞이꽃
달맞+이+꽃
달맞+이꽃

달맞이+꽃
달맞이꽃
의 8가지가 존재한다.
따라서, 이웃한 토큰들이 상호 독립이라고 가정할 경우, $p_W$('달맞이꽃')은
$p_T$(달)×$p_T$(맞)×$p_T$(이)×$p_T$(꽃)
$p_T$(달)×$p_T$(맞)×$p_T$(이꽃)
$p_T$(달)×$p_T$(맞이)×$p_T$(꽃)
$p_T$(달)×$p_T$(맞이꽃)
$p_T$(달맞)×$p_T$(이)×$p_T$(꽃)
$p_T$(달맞)×$p_T$(이꽃)
$p_T$(달맞이)×$p_T$(꽃)
$p_T$(달맞이꽃)
중 가장 높은 확률을 갖는 토큰열의 확률로서 정의된다.

$p_T$는 토큰열의 마르코프 모델로서 정의된다. 따라서, 동일한 어절을 생성하는 토큰열들 중에서 가급적 적은 수의 토큰으로 구성된 토큰열이 일반적으로 보다 높은 확률값을 가진다. 또한, 동일한 수의 분할이 이루어질 경우 그 중 가장 높은 확률을 가지는 분할 결과가 선택된다. 때문에, 토큰모델은 빈번히 관측된 어절은 해당 어절 전체로써 표현하며, 그렇지 못한 어절의 경우에도 가급적 적은 수의 분할과 적은 양의 확률 감소를 야기하도록 어절을 표현하는 방법을 선택한다고 볼 수 있다. 이는 주어진 어휘 및 그 확률 분포에 의하여 해당 어절에서 최적의 낱자 단위 분할 경계를 탐색할 수 있게하며, 동시에 미등록어의 확률 추정을 위한 back-off효과도 나타낸다고 볼 수 있다.

### 4.2 토큰 모델의 학습

토큰의 확률 분포 함수 $p_T$는 원시 말뭉치로부터 얻어진 토큰들의 사용빈도를 바탕으로 계산된다. 그러나, 원시 말뭉치의 어절에는 토큰들의 경계가 구분되어있지 않기 때문에 어절로부터 토큰들을 분할하여야 하며, 이 과정에서 $p_T$가 요구되는 문제점이 있다.

따라서, 본 논문에서는 EM알고리즘을 기반으로 반복적인 과정을 통하여 $p_T$를 수렴시켜가는 학습방법을 사용한다. 즉, 임의의 초기 확률 분포를 상정하고, 이를 바탕으로 주어진 원시 말뭉치를 토큰으로 분할한다. 이 분할 결과에서 얻어진 통계정보를 바탕으로 새로운 $p_T$를 계산하며, 다시 이를 바탕으로 $p_T$가 수렴할 때까지 위의 과정을 반복한다.

토큰의 초기 확률 분포를 상정하기 위해서는 우선 어떤 문자열들을 어휘로 사용할 것인지와 해당 문자열들이 출현하는 빈도를 어떻게 계산할 것인지가 결정되어야 한다.

어휘의 초기 구성요소들은 원시 말뭉치에서 사용된 낱자의 positional n-gram들을 바탕으로 구축한다. Positional n-gram들은 조사나 어미와 같이 어절의 뒷부분에서만 나타나는 문자열들을 고려하여 어절의 시작부분과 끝부분 양쪽에서 조사한다. 이들 어휘는 그 출현 빈도를 바탕으로 확률값을 계산한다. 원시 말뭉치에도 입력 오류등이 존재할 수 있으므로 초기 후보의 선정시에는 일정 빈도이상 관측된 어절들만을 어휘 구축 및 빈도 조사에 사용한다.

예)
'달맞이' 로부터 추출가능한 positional n-gram들은 다음과 같다.
'달'
'달맞'
'달맞이'
'맞이'
'이'

EM알고리즘이 진행됨에따라 원시 말뭉치의 분할에서 한 번도 사용되지 않은 문자열들은 모델에서 제외된다. 따라서, 초기 어휘의 선택에 일부 문제점이 있더라도, 수렴이 이루어진 뒤 모델에서 사용하는 어휘는 실제로 어절의 분할에 사용되는 문자열들로 구성되게 된다.

## 5. 실험 및 결과

제안된 모델의 성능을 비교하기 위하여 원시말뭉치로부터 생성할 수 있는 확률적 언어 모델들을 중심으로 실험을 수행하였다.

실험에는 어절, 낱자 n-gram, 토큰을 기본어휘로 하는 모델들을 사용하였다. 해당 모델들이 어절의 사용 빈도를 표현하는 성능을 중점적으로 보기 위하여, 각 모델들은 어절간에는 상호 독립을 가정하고 어절 내부에서의 확률 분포만을 표현하도록 하였다.

실험에는 KAIST 원시 말뭉치를 사용하였다. 훈련에는 97년도 원시 말뭉치를 사용하였으며, 이 말뭉치는 약 1550만개의 어절로 이루어져 있다. 훈련시 이 말뭉치는 4개의 균형잡힌 그룹으로 분리하여 훈련 및 검증에 사용하였다. 테스트에는 96년도 원시 말뭉치를 사용하였으며, 이 말뭉치는 약 1580만개의 어절로 이루어져 있다.

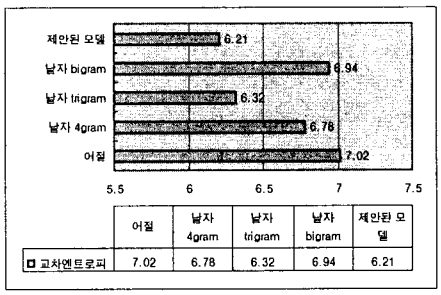
어절 기반 모델은 Good Turing 추정방법을 사용하여 미등록어의 확률을 추정하였다. 낱자 n-gram모델의 경우 Katz's back-off, Jelinek-Mercer 보간기법, Witten-Bell 보간기법, Kneser-Ney back-off등의 다양한 스무딩기법을 적용해 본 결과, Witten-Bell 보간 기법이 일관적으로 가장 좋은 성능을 나타내었다. 토큰 기반 모델에서는 이웃한 토큰들이 상호 독립적인 분포를 가진다고 가정하였다.

학습이 종료된 시점에서 각 모델의 어휘 수는 다음과 같다.

| 모델 | 어휘 수 |
|---|---|
| 제안된 모델 | 157,210 |
| 낱자 bigram | 187,963 |
| 낱자 trigram | 1,110,699 |
| 낱자 4gram | 2,170,348 |
| 어절 | 147,065 |

어절 기반 모델, 낱자 bigram 모델, 토큰 기반 모델이 서로 비슷한 단위의 어휘 수를 가지며, 낱자 trigram이상의 경우 어휘의 수가 상대적으로 많아짐을 알 수 있다.

각 모델을 바탕으로 낱자별 교차 엔트로피를 계산한 결과는 다음과 같다.




| | 어절 | 낱자 4gram | 낱자 trigram | 낱자 bigram | 제안된 모델 |
|---|---|---|---|---|---|
| 교차엔트로피 | 7.02 | 6.78 | 6.32 | 6.94 | 6.21 |

결과를 보면, 토큰 기반 모델, 낱자 n-gram기반 모델, 어절 기반 모델의 순서로 교차엔트로피 값이 낮게 나타남을 알 수 있다. 낱자 n-gram모델에서는 낱자 trigram이 가장 낮은 교차엔트로피를 나타냈으며 보다 긴 문자열을 사용한 낱자 4gram모델에서는 오히려 교차엔트로피가 증가하였다.

본 논문에서는 원시 말뭉치를 훈련 및 실험에 사용함으로써 동일한 조건에서 형태소 기반 모델과 비교 할 수 없었다. 참고로, 김길연의 논문에 완전한 형태소 분석이 이루어진 말뭉치를 바탕으로 형태소 기반 모델을 실험한 결과가 보고되어 있다[12]. 이 논문에서는 형태소 trigram 모델에 Kneser-Ney 스무딩 방법을 적용한 결과 약 6.61의 형태소간 교차 엔트로피를 보였다고 한다. 그러나, 본 논문에서는 이웃한 어절들간에는 상호 독립이라고 가정한 반면, 위의 논문에서는 이웃한 어절의 형태소들의 상관관계도 표현하였으며 완전히 형태소 분석이 이루어진 자료를 바탕으로 실험한 결과이므로 두 수치를 직접적으로 비교하기에는 무리가 있다.

## 6. 결론

본 논문에서는 한국어를 위한 확률적 언어 모델의 구축에 있어서 어휘의 선택에 관하여 고찰하고, 통계적 정보를 바탕으로 형태소와 유사한 특징을 가지는 문자열을 추출함으로써 확률적 언어 모델에 적용하는 방법을 제안하였다.

기존의 형태소 기반 모델들과는 달리, 제안된 모델은 별도의 형태소 분석기 및 형태소 분석이 이루어진 말뭉치를 필요로 하지 않는다. 제안된 모델에서는 원시 말뭉치로부터 직접 어휘를 추출함으로써 사람의 주관적 판단이 개입될 여지를 최소화한다. 또한, 문법적 지식이나 단어의 의미와 무관하게 통계적 정보만을 바탕으로 어휘를 추출하기 때문에, 구어체나 통신어체등 문법적 분석 자료가 충분하지 않은 분야에 대해서도 적용가능하다.

실험 결과, 제안된 토큰 기반 모델은 동일한 조건하에서 낱자 n-gram이나 어절 기반 모델들보다 더 낮은 교차 엔트로피를 보였다. 본 논문에서 수행된 실험은 어절의 확률 분포 추정 성능을 보기 위하여 어절간에 상호 독립 가정을 적용한 것을 고려하면, 어절간 어휘들의 상관 관계를 표현할 경우 모델의 교차 엔트로피는 더욱 낮아질 수 있으리라 기대된다.

## 7. 참고 문헌


[1]. Sabine Deligne and Frederic Bimbot, "Language modeling by variable length sequences : theoretical formulation and evaluation of multigrams", in Proc. of ICASSP, 1995

[2]. Michael R. Brent, "An efficient, probabilistically sound algorithm for segmentation and word discovery", Machine Learning Journal, vol 34, pp.71~106, 1999

[3]. Stanley F. Chen, "Building probabilistic models for natural language", Ph.D. Thesis, Harvard University, 1996

[4]. Stanley F. Chen and Joshua Goodman, "An empirical study of smoothing techniques for language modeling", TR-10-98, Harvard University, 1998

[5]. Oh-Wook Kwon, K.Hwang and J.Park, "Korean large vocabulary continuous speech recognition using pseudomorpheme units", in Proc. EUROSPEECH, pp. 483-486, 1999

[6]. Reihard Kneser and Hermann Ney, "Improved backing-off for m-gram language modeling", in Proc. of ICASSP, 1995

[7]. Slava M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer", IEEE Trans. On Acoustics, Speech and Signal processing, vol.35, no.3, pp.300-401, 1987

[8]. Ian H. Witten and Timothy C. Bell, "The zero-frequency problem : Estimating the probabilities of novel events in adaptive text compression", IEEE Trans. on Information Theory, vol.37, no.4, pp.1085-1094, 1991

[9]. 이경님, 정민화, "의사 형태소 단위의 음성 언어 형태소 해석", 제10회 한글 및 한국어 정보 처리 학술대회 발표 논문집, pp396-404, 1998
[10]. 박영희, 정민화, "대어휘 연속음성 인식을 위한 결합형태소 자동생성", 한국음향학회지, 제 21권, 제4호, pp.407-414, 2002
[11]. 황화상, 시정곤, "형태소 분석을 위한 한국어 어절의 구성 양상 연구", 제 13회 한글 및 한국어 정보처리 학회 발표 논문집, pp.25-32,2001
[12]. 김길연, 최기선, "단어와 클래스 기반의 한국어 언어 모델링", 제 13회 한글 및 한국어 정보처리 학회 발표 논문집, pp.221-225, 2001
[13]. 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회 논문지(B) 제23권 제9호, pp.991-1000, 1996
[14]. Christopher D. Manning and Hinrich Schutze, "Foundations of Statistical Natural Language Processing", MIT Press, 1999
[15]. Daniel Jurafsky and James H. Martin, "Speech and Language Processing", Prentice-Hall, 2000