K-SLM Toolkit 을 이용한 한국어의 통계적 언어 모델링 비교†

이 진 석, 박 재 득*, 이 근 배
포항공과대학교 컴퓨터공학과 자연언어 처리 연구실
경북 포항시 남구 효자동 산 31 번지
{wolfpack, gblee}@nlp.postech.ac.kr

*한국 전자 통신 연구원 지식 정보 연구부 언어 이해 연구팀
대전시 유성구 가정동 161
jdpark@etri.re.kr

Comparative Analysis of Statistical Language Modeling for Korean using K-SLM Toolkits

Jinseok Lee, JayDuke Park*, Geunbae Lee
Natural Language Processing Lab., Dept. of Computer Science & Engineering, POSTECH
*Language Understanding Lab., Knowledge Technology Research Department, ETRI

요약

통계적 언어 모델은 자연어 처리의 다양한 분야에서 시스템의 정확도를 높이고 수행 시간을 줄여 줄 수 있는 중요한 지식원이므로 언어 모델의 성능은 자연어 처리 시스템, 특히 음성 인식 시스템의 성능에 직접적인 영향을 준다. 본 논문에서는 한국어를 위한 통계적 언어 모델을 구축하기 위한 다양한 언어 모델 실험을 제시하고 각 언어 모델들 간의 성능 비교를 통하여 통계적 언어 모델의 표준을 제시한다. 또한 형태소 및 어절 단위의 고 빈도 어휘만을 범용 언어 모델에 적용할 때의 적용률을 통하여 언어 모델 구축시 어휘 사전 크기 결정을 위한 기초적 자료를 제시한다. 본 연구는 음성 인식용 통계적 언어 모델의 성능을 판단하는 데 앞으로 큰 도움을 줄 수 있을 것이다.

1. 서론

언어 모델링(Language Modeling)은 자연어 안에서 규칙성을 찾아내고 그 규칙성을 이용하기 위한 노력이다. 언어 모델링을 통해 얻어진 언어 모델(Language Model; LM)은 음성 인식이나 기계 번역, 문자 인식, 철자 교정 등 다양한 모습으로 시스템의 정확도를 높이고 수행 시간을 줄이는 데 필수 불가결한 요소이다.

지금까지 언어 모델은 크게 지식 기반 모델(Knowledge-based Model) 과 통계적 모델(Statistical Model)로 나눌 수 있다. 지식 기반 모델은 정규 문법(Regular grammar; RG)나 문맥 자유 문법(Context-free grammar; CFG)을 만들고, 이러한 문법 구조에 어긋난 구조를 탐색 공간(search space)에서 제거함으로써 탐색 범위를 줄이고 인식률을 높이는 방식이다. 그러나 문법 구조를 만들기가 까다롭고 대용량의 어휘를 수용하기 어렵다는 점과 자연어의 비문법성을 고려할 때 다양한 문장의 인식이 어려워 진다는 단점이 있다. 이러한 단점은 범용 언어 모델이나 새로운 영역에 대한 언어 모델을 구성할 때 많은 시간과 노력을 요구하게 된다. 이에 반해 통계적 모델은 대량의 말뭉치(corpus)에서 언어 규칙을 확률로 나타내고 확률값을 통해서 탐색 영역을 제한하는 방법이다. 통계적 모델은 지식 기반 모델보다 더 큰 확장성(extensibility)과 강건함(robustness)으로 특히 음성 인식 쪽에서 좋은 성능을 발휘하였다[1, 8].

본 논문에서는 음성 인식 쪽에서 사용될 수 있는 통계적 한국어 언어 모델을 구축하고 다양한 모델들 사이의 관계를 알아 보았다. 보통 영어에서의 통계적 언어 모델은 단어 단위의 언어 모델이 일반적이지만 한국어 같이 어미나 조사가 발달한 교착어의 특징을 가진 언어에서는 어절뿐 아니라 형태소 단위의 언어 모델이 가능하다. 이를 위해서 CMU-CAM Statistical Language Modeling Toolkit[3, 4]을 한국어에 적합하도록 확장한 K-SLM(Korean-Statistical Language Modeling) Toolkit 을 구현하였다. K-SLM Toolkit 은 상위 언어 처리를 거친 말뭉치를 기반으로 표층 뿐만 아니라 표층에 나타나 있지 않은 부가적인 정보들을 결합하여 다양한 형태의 언어 모델을 만들어 주는 도구이다.

본 연구에서는 K-SLM Toolkit 을 사용하여 형태소 단위의 표층, 주형태, 발음열, 품사(Part-Of-Speech; POS) 태그들을 결합한 언어 모델과 어절 단위의 언어 모델을 비교하였다. 또한 범용 언어 모델 생성을 위해 한국어에서 사용 빈도가 높은 어휘로 어휘 사전(vocabulary)를 구성하고, 어휘 사전의 크기를 변화 시키면서 언어 모델의 성능을 비교하였다.

2. 통계적 언어 모델

음성 인식은 화자(speaker)가 발화한 음성을 텍스트나 상위 언어 처리가 가능한 형태로 정확히 복원하는 작업이다. 즉 통계적 방법을 사용하는 음성 인식 시스템에서 음향 신호열(acoustic signal sequence) A 가 주어 지면 나타날 가능성이 가장 높은 단어열(word sequence) W 를 찾는 것을 의미한다[1]. 이것을 수식으로 나타내면 식 (1)과 같다.

여기에 Bayes rule 을 적용하면 식 (1)은 식 (2)와 같이

† 본 연구는 ETRI 위탁과제(1999. 2 - 1999. 11)연구비 지원으로 이루어진 것임.

나타낼 수 있다.

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{A}) \qquad (1)$$

$$\mathbf{A} \in a_1, a_2, \ldots\ldots, a_i \qquad \mathbf{W} \in w_1, w_2, \ldots\ldots, w_i$$

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W} \mid \mathbf{A}) = \arg\max_{\mathbf{W}} \frac{P(\mathbf{A} \mid \mathbf{W})P(\mathbf{W})}{P(\mathbf{A})} \qquad (2)$$

여기서 P(A|W)는 화자가 단어열 W를 말했을 때 음향 신호열 A가 나타날 확률을 의미하는 음향 모델(acoustic model)이고 P(W)가 본 연구에서 다룰 언어 모델이다.

P(W)를 다시 쓰면

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \ldots\ldots, w_{i-1}) \qquad (3)$$

로 나타낼 수 있다. 그러나 어떤 문장에서 i번째 단어 w_i가 나타날 확률을 구하기 위해 w_1에서 w_{i-1}까지 i-1개의 단어열이 나타날 확률을 구하는 것은 많은 노력을 요구하므로 단어의 확률은 이전 단어에 의존적이라는 Markov 가정에 의해 i번째 단어가 나타날 확률을 구하기 위해 이전의 i-1개의 단어에 대한 확률을 전부 구하지 않고 이전의 N-1개의 단어를 보는 N-gram 모델이 일반적이다. 즉, 식(3)은

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \ldots\ldots, w_{i-1}) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-N+1}, \ldots\ldots, w_{i-1}) \qquad (4)$$

식 (4)와 같이 나타낼 수 있다. 여기서 N이 1일 경우 이러한 모델을 bigram 모델이라 하고, N이 2인 경우 trigram 모델이라 한다.

본 연구에서는 데이터 부족 문제 때문에 bigram 모델로 언어 모델을 구축하여 비교해보았다. bigram 모델의 경우 식 (3)은 다음과 같이 된다(단, boundary condition 무시).

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i \mid w_1, w_2, \ldots\ldots, w_{i-1}) \approx \prod_{i=1}^{n} P(w_i \mid w_{i-1}) \qquad (5)$$

한국어는 식 (5)의 w에 해당하는 단위로 형태소 및 어절 수준에서의 다양한 단위를 적용할 수 있다. 즉, 어절 뿐 아니라 형태소를 영어에서의 단어처럼 사용하여 형태소 수준의 N-gram 모델이 가능하다. 본 연구에서는 형태소 및 어절 수준의 정보들을 조합한 각각에 대해 bigram 모델을 구하였다. 그러므로 식 (5)는 다음과 같이 쓸 수 있다.

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(w_i \mid w_{i-1}) = \prod_{i=1}^{n} P(\mathbf{U}_i \mid \mathbf{U}_{i-1}) \qquad (6)$$

$$(\mathbf{U} \subset \mathbf{C}, \mathbf{U} \neq \phi\ ,\ \mathbf{C} \in \{\mathbf{T}, \mathbf{R}, \mathbf{S}, \mathbf{P}\})$$

여기서 T는 품사 태그 세트열을 의미하고, R은 주형태열(lexical sequence)을, S는 표층열(surface sequence)을, 그리고 P는 발음열(phoneme sequence)을 의미한다. 즉, 형태소 수준의 가능한 모든 정보를 가지고 언어 모델을 구성한다고 하면 식 (6)은 식 (7)과 같이 나타낼 수 있다.

$$P(\mathbf{W}) = \prod_{i=1}^{n} P(\mathbf{U}_i \mid \mathbf{U}_{i-1}) = \prod_{i=1}^{n} P(\mathbf{T}_i, \mathbf{R}_i, \mathbf{S}_i, \mathbf{P}_i \mid \mathbf{T}_{i-1}, \mathbf{R}_{i-1}, \mathbf{S}_{i-1}, \mathbf{P}_{i-1}) \qquad (7)$$

bigram을 사용한 이유는 trigram에 비해서 정확도는 떨어 질 수 있으나 일반적으로 trigram보다 적은 양의 훈련(training[1])으로도 측정이 가능하기 때문이다.

3. K-SLM Toolkit

한국어의 최소 의미 단위는 형태소다. 영어에서는 공백(white space)으로 구분된 단어 단위의 언어 모델의 구성이 일반적이다. 그러나 어미나 조사의 활용이 활발한 한국어에서는 영어에서처럼 공백으로 구분된 단위인 어절 뿐 아니라 형태소 단위로 구성된 언어 모델의 생성이 가능하다. 따라서 형태소 단위의 언어 모델을 가능하도록 도와 주는 K-SLM Toolkit을 제작하였다. K-SLM Toolkit은 영어에 적합하게 설계된 CMU-CAM Statistical Language Modeling Toolkit[3, 4]을 한국어에 맞게 확장하였다.

CMU-CAM Statistical Language Modeling Toolkit (이하 SLM-Toolkit)은 미국의 Carnegie Mellon University(CMU)와 영국의 Cambridge University의 음성 언어팀에서 통계적 언어 모델링을 위해서 개발한 Toolkit이다. 이 SLM-Toolkit은 일반적인 텍스트로부터 단어 빈도 리스트(word frequence list)와 사전 구축(vocabulary), 일반적인 bigram과 trigram 모델 구축, 사전에 기반한 bigram, trigram 모델 구축, bigram과 trigram 모델에 관계된 각종 통계치 계산 등을 손쉽게 해주는 통계적 언어 모델링 도구이다. 또한 SLM-Toolkit에서 계산된 언어 모델을 바탕으로 perplexity, Out-Of-Vocabulary(OOV)빈도, bigram과 trigram 적중률등을 계산할 수 있다.

그러나 SLM-Toolkit은 영어의 공백으로 분리된 unique한 단어를 최소 단위로 이것들을 가지고 여러 가지 통계정보 및 언어 모델링을 하기 때문에 곧바로 한

[1] 통계적 언어 모델에서 훈련이란 모델의 파라메터값을 대량의 말뭉치에서 자동으로 결정하는 것을 의미한다[2].

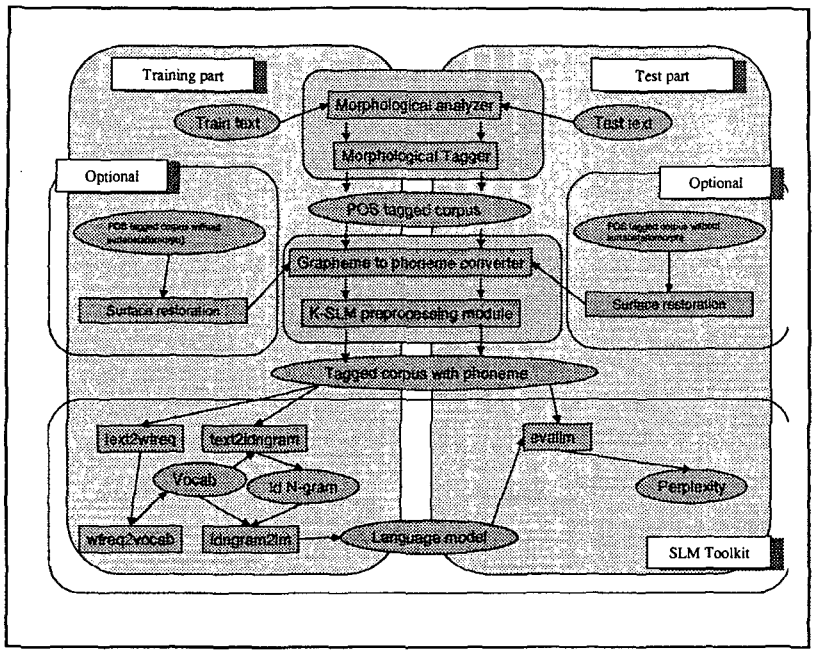

그림 <1> K-SLM Toolkit 의 구성도

국어에 적용하는 것은 무리가 있다. 왜냐하면 영어는 공백으로 분리된 단어 단위가 일반적인 언어 모델의 단위인 반면 앞서 말했듯이 한국어는 형태소 단위의 언어 모델 구성이 가능하기 때문에 한국어에 적합하도록 확장할 필요가 있다. 여기서 사용한 방법은 SLM-Toolkit 의 변경은 최소화로 하고, 전 처리를 통하여 다양한 한국어 언어 모델을 생성할 수 있도록 하였다. 이렇게 구성한 이유는 자칫 Toolkit 이 가지고 있는 성격과 확장성을 떨어 뜨릴 수 있으므로 전 처리를 통하여 다양한 기능을 수용할 수 있도록 하였다.

K-SLM Toolkit 에서 다루는 한국어의 단위는 크게 형태소와 어절 단위이다. 형태소 단위와 어절 단위 각각에서 추출할 수 있는 언어 모델 단위는 표층과 주형태, 발음열과 품사 태그 세트이다.

이를 위해서는 부가적인 여러 가지 작업이 선행되어야 한다. 다음 절에서는 K-SLM Toolkit 의 전반적인 구성 및 특징에 대해서 설명한다.

3.1 K-SLM Toolkit 의 구성과 전처리

그림 <1>은 SLM Toolkit 의 기본적인 사용 흐름과 이를 한국어에 맞게 확장한 K-SLM Toolkit 의 구성도이다.

한국어는 영어와 달리 공백으로 구분된 단위(어절)가 언어 모델의 최소 단위로 부적합할 수 있다. 따라서 한국어에 맞는 word 의 개념을 정립하는 것이 한국어 통계적 언어 모델의 출발이다.

한국어는 형태소 수준의 언어 모델이나 어절 수준의 언어 모델, 기타 다양한 수준의 언어 모델이 가능하다. 또한 한국어가 표음문자에 속하지만 음운현상이 활발하기 때문에 발음 정보도 또한 언어 모델의 일부가 될 수 있다. 그러므로 어절뿐만 아니라 한국어의 의미 최소 단위인 형태소나 형태소에 대응되는 표층정보, 그리고 발음열까지 다양한 단위의 언어 모델이 가능하다.

따라서 이를 위해서는 상위 언어 처리인 형태소 분석과 태깅[2]이 선행되어야 한다. 일단 원시 말뭉치(raw corpus)가 형태소 분석과 태깅을 거쳐야만 비로소 형태소 단위의 언어 모델을 구성할 수 있기 때문이다. 또한 형태소 분석을 마친, 태그가 부착된 말뭉치(tagged corpus)에 나타난 표층정보는 자소열-음소열 변환기(Grapheme To Phoneme Converter;G2P[3])의 정보로 사용되어 한국어의 음운 현상을 반영한 발음열을 출력하게 된다. 한국어는 형태소 내의 음운 변화와 형태소와 형태소 간, 어절과 어절간의 음운 변화가 활발하므로 주형태소나 표층 정보 뿐 아니라 발음열도 언어 모델의 구성단위가 될 수 있다.

3.2 K-SLM Toolkit 의 처리 과정

그림 <1>에서 같이 입력된 문장은 일단 형태소 분석과 태거를 거치고, 그 결과가 자소열-음소열 변환기(이하 G2P)를 거치면서 태그가 부착된 말뭉치에 발음열까지 추가 하게 된다. 그리고 SLM-Toolkit 의 입력형태로 적당한 전처리를 거쳐서 SLM-Toolkit 의 모듈로 입력하게 된다. 전처리를 하는 이유는 형태소 분석과 태깅, G2P 를 거쳐서 나온 결과를 각 단위별 (*e.g.*,주형태소, 표층, 발음) 뿐만 아니라 여러 단위를 하나의 단위로 조합한 단위를 언어 모델의 기본 단위로 만들 수 있어야 하기

[2] 본 연구에서 사용한 형태소 분석기 및 태거는 포항공대 SKOPE 내에 사용된 형태소 분석기 및 태거를 사용하였다[5].

[3] 본 연구에서 사용한 자소열-음소열 변환기(Grapheme To Phoneme Converter)는 포항공대 SKOPE 내에 사용된 자소열-음소열 변환기를 사용하였다[6].

때문이다. 이를 위해서는 위의 형태소 분석과 G2P 를 거친 원시 말뭉치는 실제 언어 모델을 구축하기 위한 전단계로 먼저 언어 모델의 기본 단위별 구분을 공백으로 하고, 각 문장 경계를 표시하는 context cue[4](<s>)가 문장의 시작과 끝에 붙어야 한다. 또한 주형태소, 표층, 발음의 모든 정보 중 사용자가 원하는 것만 선택해서 언어 모델을 만들 수 있어야 한다.

3.3 K-SLM Toolkit 의 처리 예

그림 <2>은 K-SLM Toolkit 의 입력과 전처리를 거친 예를 보여준다.

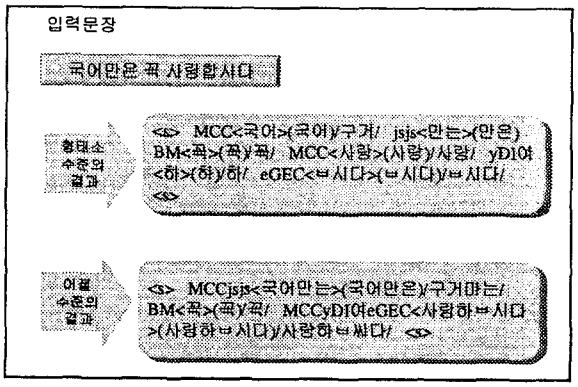

그림 <2> K-SLM Toolkit 의 입출력 예

K-SLM Toolkit 은 원시 말뭉치를 입력으로 받아 형태소 분석, 태깅, 그리고 G2P 를 거치면서 각 품사 관련 정보 및 발음열까지 부착된 말뭉치를 그림 <2>처럼 형태소 수준과 어절 수준으로 출력하게 된다. 형태소 수준에서 결과 중 축약이 있거나, 복합 어미, 복합 조사나 보조 용언과 어미는 그림 <2>에서의 'jsjs'처럼 하나의 형태소 단위로 취급한다. 그러한 이유는 상위 언어 처리를 손쉽게 하고, G2P 에서 처리를 원활하게 하기 위함이다[5, 6].

4. 모델 실험 및 결과

실험은 크게 두 가지로 나눌 수 있다. 각각의 실험 단위는 형태소와 어절이다. 두 가지로 단위로 나눈 이유는 형태소는 한국어의 최소의 의미 단위이고 따라서 언어 처리의 가장 기본이고 어절은 문장 성분의 최소의 단위라 볼 때 문장 단위의 처리에서 최소 단위일 수 있기 때문이다. 그러므로 두 단위의 비교는 한국어의 특징 파악에 중요한 요소이다.

실험 1 은 한국어에 자주 나타나는 어휘로 사전을 구성하고, 새로운 말뭉치에 대한 어휘 사전의 적용률(coverage)을 알아보았다. 이것은 범용 언어 모델을 만들 때 어휘 사전 크기의 판단 기준이 될 수 있으며, 한국어의 일반적 어휘사용 범위를 밝힐 수 있다.

실험 2 는 실험 1 에서 얻은 어휘 사전으로 형태소 및 어절 단위의 다양한 언어 모델을 만들고 각각의 언어 모델의 perplexity[5]를 구하였다[2]. 언어 모델에서 perplexity 의 대략적 의미는 한 단어 뒤에 올 수 있는 랜덤 변수(random variable)의 개수라고 말할 수 있다. perplexity 를 구하는 식은 식 (8)과 같다.

$$PP_M(T) = 2^{-\sum_x P_T(x)\cdot\log P_M(x)} \qquad (8)$$

식 (8)에서 $P_T(x)$, $P_M(T)$는 각각 테스트 말뭉치와 모델의 확률 함수(probability fuction)를 의미한다. 따라서 perplexity 는 식 (8)에서 볼 수 있듯이 모델 M 에 의해 예상되어 지는 테스트 말뭉치의 Entropy 값을 2 의 승수로 취한 값이다.

본 연구에서 사용한 말뭉치는 국어 정보 베이스중 일부를 사용하였고, 실험에 사용한 smoothing 방식은 Toolkit 에서 지원하는 Good-Turing Discounting 과 Witten-Bell Discounting 방식이다[1, 4]. 표 1 은 실험에 쓰인 말뭉치의 크기이다. 자세한 것은 각각의 실험에 대한 부분에서 다룬다.

표 1 실험에 쓰인 말뭉치의 크기

추출 단위 / 용도	형태소	어절
어휘 사전 추출용 말뭉치 크기	약 800 만 개	약 200 만 개
언어 모델 훈련 (Training) 용 말뭉치 크기	약 200 만 개	약 100 만 개
언어 모델 테스트용 말뭉치 크기	약 40 만 개	약 20 만 개

4.1 어휘 사전의 적용률

이 실험은 고 빈도의 한국어 형태소 및 어절을 추출하고 추출한 어휘 사전이 어느 정도의 적용률을 가지는지에 대한 실험이다. 이 실험을 위해 국어 정보 베이스에서 800 만개의 형태소와 200 만개의 어절을 추출하고 추출한 것에서 가장 많이 쓰이는 형태소 및 어절을 각각 15000, 20000, 25000, 30000 어휘를 추출하였다. 여기서 추출된 어휘 사전은 다음 실험인 언어 모델의 어휘 사전으로 이용된다. 따라서 형태소 및 어절 어휘 사전은 언어 모델을 기준으로 다양하게 만들어 졌다. 예를 들어 표층과 품사 태그를 가지고 언어 모델을 구성하기 위한 어휘 사전은 {품사, 표층}으로 이루어 진다. 실험 2 에서 구성한 언어 모델은 표 2 와 같다.

표 2 에서와 같이 형태소 및 어절에 대해 각각 15 개의 언어 모델 구축을 위한 각각 15 개의 어휘 사전을 표제어의 개수를 바꿔가면 구성하였다. 그리고 각 어휘 사전을 기준으로 언어 모델을 구성하고 테스트 말뭉치에 대한 어휘 사전의 적용률을 알아 보았다.

[4] 문장 시작과 끝 또는 문단 등을 다른 단어와 다르게 취급하기 위한 기호를 의미한다[3, 4].

[5] 여기서 말하는 perplexity 는 cross-perplexity 를 의미한다.

표 2 각 언어 모델의 구성 및 예 (형태소)

언어 모델	어휘 사전 표제어 예
R	<국어>
TR	MCC<국어>
S	(국어)
TS	MCC(국어)
P	/구거/
TP	MCC/구거/
RS	MCC(국어)
TRS	MCC<국어>(국어)
SP	(국어)/구거/
TSP	MCC(국어)/구거/
RP	<국어>/구거/
TRP	MCC<국어>/구거/
RSP	<국어>(국어)/구거/
TRSP	MCC<국어>(국어)/구거/
T	MCC

대체적으로 거의 모든 언어 모델에서 비슷한 결과를 나타냈다. 이것은 고 빈도의 어휘에서는 각 정보들이 의존적임을 알 수 있다. 즉, 고 빈도 어휘에서는 하나의 정보를 알 수 있다면 나머지 다른 정보들을 결정할 수 있다는 의미이다.

약간의 차이를 보이는 언어 모델을 살펴 보면, 형태소에서 가장 적용률이 높은 것은 주형태({R})로 이루어진 언어 모델이고 가장 많은 정보를 포함하는 태그, 주형태, 표층형태, 발음열({T, R, S, P})로 이루어진 언어 모델에서 적용률이 가장 떨어 진다. 어절에서는 발음열({P})로 이루어진 언어 모델의 적용률이 가장 높았으며 형태소와 같이 모든 정보({T, R, S, P})로 이루어진 언어 모델이 가장 적용률이 떨어 졌다. 대체적으로 한가지 정보로만 이루어진 언어 모델의 적용률이 올라가는 일반적인 현상을 보였다. 그러나 형태소에서는 발음열을 포함하는 언어 모델들의 적용률이 대체적으로 낮게 나타났고 어절에서는 오히려 발음열만으로 이루어진 언어 모델의 적용률이 가장 좋게 나왔다. 그러한 이유는 한국어의 발음이 형태소 단위에서 표층이나 주형태보다 변화가 활발해서 OOV(Out-of-vocabulary) rate 가 올라 가는 한편 오히려 어절 단위에서는 활용이 다른 것에 비해 떨어져서 적용률이 높아짐을 의미한다.

그림 <3>과 그림 <4>은 형태소 및 어절 단위의 각 언어 모델에 따른 어휘 사전의 적용률을 보여 준다.

4.2 각 언어 모델의 비교

언어 모델의 실험은 표 2에서 제시한 15 가지의 언어 모델을 기준으로 형태소와 어절로 나누어서 실험을 하였다. 언어 모델에 사용되는 어휘 사전(vocabulary)은 실험 1 에서 사용했던 어휘 사전을 사용하였다. 따라서 각각의 어휘 사전 크기에 따라 15 가지의 언어 모델을 형태소 및 어절 단위로 만들었다.

그림 <5>와 그림 <6>은 형태소 수준과 어절 수준의 언어 모델 중 어휘 사전의 크기가 20000 인것으로 가장 대표적인 것만을 뽑은 결과이다. 막대 그래프는 각 언어 모델의 perplexity 를 나타내는 것이고, 꺽은선 그래프

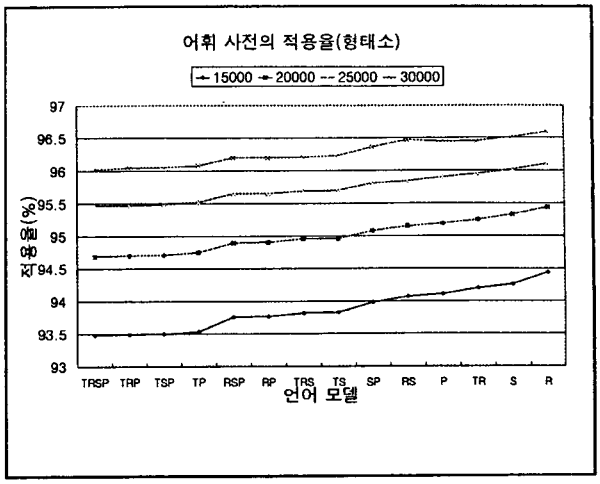

그림 <3> 언어 모델에 따른 어휘 사전 적용률 (형태소)

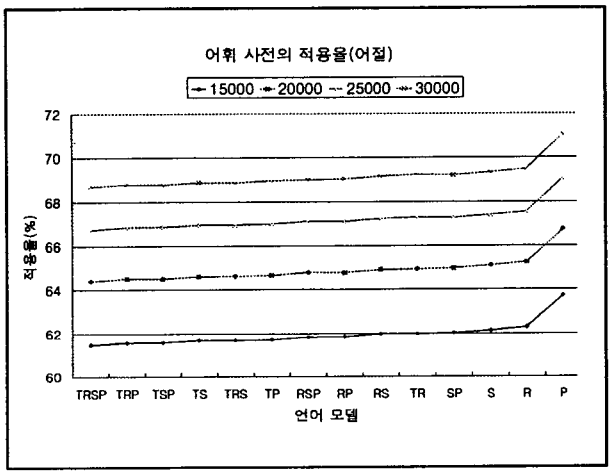

그림 <3> 언어 모델에 따른 어휘 사전 적용률 (어절)

는 그림 <3>과 그림 <4> 에서 보았던 2000 어휘에 대한 적용률을 나타낸 것이다.

그림에서 알 수 있듯이 형태소 단위의 언어 모델보다 어절 단위의 언어 모델의 perplexity 가 10 배 이상 차이가 났다. 이것을 통해 어절을 이용한 범용 한국어 언어 모델을 구축한다는 것은 한계가 있음을 알 수 있다. 따라서 한국어의 최소 의미 단위가 형태소이기 때문에 형태소 수준의 언어 모델을 구축하고, 상위의 언어 처리를 통해 형태소 수준에서 처리하기 힘든 부분을 처리하는 것이 바람직하다고 볼 수 있다.

형태소 수준에서 가장 우수[6]하게 나타난 모델은 perplexity 와 적용률을 모두 고려해 볼 때 태그와 주형태({T,R})로 이루어진 모델이고 가장 뒤떨어지는 모델은 모든 정보를 가지고 만든 언어 모델이었다. 어절 수준에서는 형태소에서 가장 우수했던 태그와 주형태({T,R})

[6] perplexity 가 작은 값일수록 우수한 모델이 된다.

로 이루어진 언어 모델도 대체적으로 우수하게 나타났지만 적용률도 같이 고려해 볼 때 발음열({P})로 이루어진 언어 모델이 가장 우수했으며, 모든 정보를 가지고 만든 언어 모델이 역시 가장 나쁜 perplexity를 가졌다. 대체로 많은 정보를 가지고 만든 언어 모델의 perplexity가 나쁜 것으로 나타났는데, 이것은 한 정보와 다른 정보의 조합으로 나타날 수 있는 경우의 수가 늘어나기 때문이다.

그러나 실험 1의 적용률 실험처럼 실험 2에서도 거의 전 모델의 perplexity가 그렇게 현저한 차이는 보이지 않았다. 따라서 응용 프로그램의 필요에 따라 모든 정보를 가지고 만든 언어 모델도 적용할 수 있을 것이다. 그러한 이유는 모델이 가진 정보가 많을수록 상위 언어 처리의 부담을 줄일 수 있기 때문이다.

형태소와 어절 사이의 비교에서 특징은 적용률과 마찬가지로 발음열로만 이루어진 언어 모델이 형태소 수준에서는 큰 perplexity 값을 가지는 반면 어절에서는 작아짐을 볼 수 있었다.

일반적으로 언어 모델의 정량적 비교 기준이 될 수 있는 것은 어휘 사전이 같고, 훈련 말뭉치(training corpus)

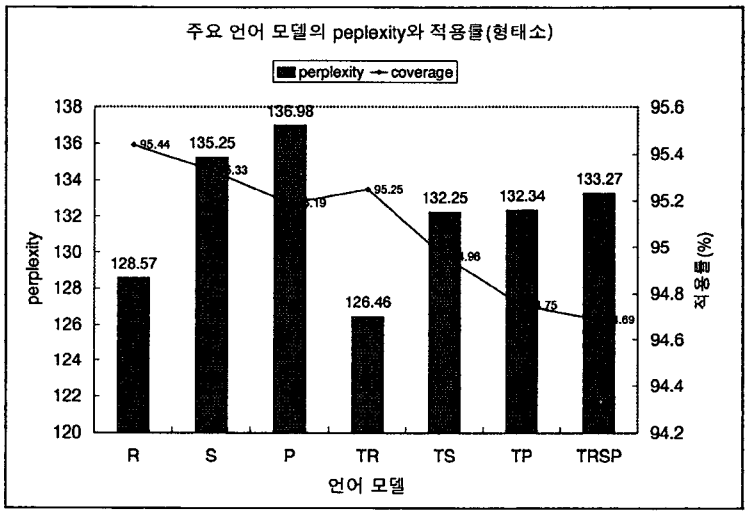

그림 <5> 주요 언어 모델의 perplexity와 적용률(형태소)

(적용률은 그림 <3>의 2000 어휘와 동일)

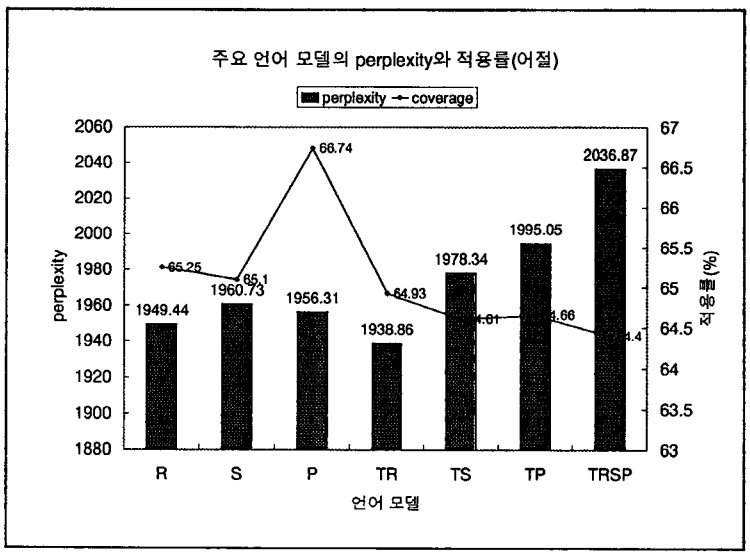

그림 <6> 주요 언어 모델의 Perplexity와 적용률(어절)

(적용률은 그림 <4>의 2000 어휘와 동일)

가 동일하며, 같은 테스트 말뭉치(test corpus)를 가지고 측정한 값이 비교의 대상이 될 수 있다[2, 7]. 그러나 본 연구에서 수행한 모델들은 각각의 특성상 이러한 기준을 전부 지키기는 힘드므로 동일한 량의 말뭉치에서 구성할 수 있는 언어 모델이라는 기준으로 각 모델을 비교하였다.

5. 결과 및 토의

본문에서는 범용 한국어 언어 모델 구축을 위한 기초적인 작업 중 하나인 고 빈도 어휘들의 적용률과 다양한 언어 모델을 통한 한국어 특징, 전형적으로 구성할 수 있는 범용 한국어 언어 모델들 사이의 관계를 형태소와 어절 단위로 알아 보았다. 이를 위해 K-SLM Toolkit을 제작하고 그 특징 및 구성을 살펴보았다. K-SLM Toolkit은 CMU-CAM Statistical Language Modeling Toolkit을 한국어에 맞게 확장한 Toolkit이며 여기서는 어절 및 형태소 단위의 다양한 레벨의 언어 모델 구축을 도와 주는 Toolkit이다.

결과로 형태소 단위에서 고 빈도 어휘 사전 적용률이 가장 높은 것은 주형태({R})로 이루어진 언어 모델인 반면에 어절에서는 발음열({P})로 이루어진 언어 모델이 가장 우수함을 알 수 있었고, 적용률과 perplexity를 같이 고려한 비교에서는 형태소 수준은 태그, 주형태짝({T,R})이 가장 우수했고, 어절 수준에서는 발음열({P})로 이루어진 언어 모델이 가장 우수했다.

여기서 수행한 연구는 상위 자연어 처리를 고려하지 않은 기본적인 언어 모델의 차이와 특징을 밝히려했다. 그러나 전술했듯이 언어 모델의 정량적 비교는 본 연구에서 수행한 언어 모델들 간에는 적용할 수 없을 수도 있다. 그러나 본 연구에서 사용한 언어 모델은 동일한 말뭉치(corpus)에서 추출할 수 있는 언어 모델이라는 점에서 각 모델들마다 어휘 사전 적용률과 perplexity의 차이를 상위 언어 처리가 요구하는 수준에서 비교한다면 우수한 언어 모델을 선택할 수 있을 것이다.

또한 K-SLM Toolkit을 좀 더 보완하고 확장한다면 다양한 확률 학습에 사용될 수 있고, 본 연구에서 수행한 방법을 기초로 다른 응용을 추가 한다면 더 좋은 언어 모델을 구축할 수 있을 것이다[8, 9].

6. 참고 문헌

[1] F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Massachusetts, 1998.

[2] R. Rosenfeld. *Adaptive Statistical Language Modeling: A Maxium Entropy Approach.* PhD thesis, School of Computer Science, Carnegie Mellon University, April 1994.

[3] R. Rosenfeld. The CMU Statistical Language Modeling Toolkit, and its use in the 1994 ARPA CSR Evaluation. In *ARPA Spoken Language Technology Workshop, Austin, TX*, January 1995.

[4] P. Clakson, R. Rosenfeld. Statistical Language Modeling Using The CMU-CAMBRIDGE Toolkit. In *Proceedings of the Eurospeech'97, vol.5*, 1997.

[5] 이원일. 단일화 기반 범주 문법에 기반한 음성 한국어 처리, 박사 학위 논문, 포항공과대학 대학원, 1998.

[6] B. Kim, G. Lee, J. Lee. Unlimited Vocabulary Grapheme to Phoneme Conversion with Probabilistic Phrase Break Detection. In *Proceedings of the 8th International Conference on Computer Processing of Oriental Languages*, March 1999

[7] D. Jurafsky, J. H. Martin. *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing.* Internet Draft, June 1999. (http://www.cs.colorado.edu/~martin/slp.html)

[8] B. Srinivas. "Almost parsing" techniques for language modeling. In *Proceedins of the 4th Internationtal Conference on Spoken Langugae Processing(ICSLP-96)* 1996.

[9] P. A. Heeman, J. F. Allen. Incorporation POS Tagging into Language Modeling. In *proceedings of the Eurospeech'97, vol.5*, 1997.