음절단위 bigram정보를 이용한 한국어 단어인식모델

신중호, 박혁로
연구개발정보센터

A Statistical Model for Korean Text Segmentation Using Syllable-Level Bigrams

Joong Ho Shin, Hyuk Ro Park
Korea Research and Development Information Center
{jhshin,hrpark}@kordic.re.kr

요 약

일반적으로 한국어는 띄어쓰기 단위인 어절이 형태소 분석의 입력 단위로 쓰이고 있다. 그러나 실제 영역(real domain)에서 사용되는 텍스트에서는 띄어쓰기 오류와 같은 비문법적인 형태도 빈번히 쓰이고 있다. 따라서 형태소 분석 과정에 선행하여 적합한 형태소 분석의 단위를 인식하는 과정이 이루어져야 한다. 본 연구에서는 한국어의 음절 특성을 이용한 형태소 분석을 위한 어절 인식 방법을 제안한다. 제안하는 방법은 사전에 기반하지 않고 원형코퍼스(raw corpus)로부터의 필요한 음절 정보 및 어휘정보를 추출하는 방법을 취하므로 오류가 포함된 문장에 대하여 견고한 분석이 가능하고 많은 시간과 노력이 요구되는 사전구축 및 관리 작업을 필요로 하지 않는다는 장점이 있다. 한국어 어절 인식을 위하여 본 논문에서는 세 가지 확률 모델과 동적 프로그래밍에 기반한 인식 알고리즘을 제안한다. 제안하는 모델들을 띄어쓰기 오류문제와 한국어 복합명사 분석 문제에 적용하여 실험한 결과 82-85% 정도의 인식 정확도를 보였다.

1. 서론

형태소 분석을 수행하기 위해서는 형태소 분석의 입력 단위가 되는 단어 혹은 어절을 문장으로부터 인식하는 작업이 선행되어야 한다. 이 과정은 분석을 대상으로 하는 언어의 특성에 따라서 그 방법과 난이도를 달리 한다. 예를 들어 영어의 경우는 공백(space)으로 구분되는 단어들로 보아도 무방하다. 그러나 중국어 문장이나 일본어와 같이 문장이 공백 등의 구분자가 없이 단어들의 열로 구성된 경우 문장으로부터 단어를 분리하는 과정이 선행되어야 하고 영어에 비해 단어 분리에 대한 문제가 부각되고 있다.

한국어의 경우도 일반적으로 띄어쓰기 단위인 어절이 형태소 분석의 입력 단위로 쓰이고 있다. 그러나 실제 영역(real domain)에서 사용되는 텍스트에서는 띄어쓰기 오류와 같은 비문법적인 형태도 빈번히 쓰이고 있다. 따라서 형태소 분석 과정에 선행하여 적합한 형태소 분석의 단위를 인식하는 과정이 이루어져야 한다. 예를 들어, 신문 데이타나 Homepage 데이타 등에서 쉽게 발견할 수 있는 미등록어를 포함하는 복합 명사의 경우 형태소 해석 단계의 미등록어 추정 만으로는 충분한 성능을 보장할 수 없다. "고성능팬티엄개발"라는 복합명사에서 "팬티엄"이 미등록어인 경우가 그러한 예이다. 특히 본 과제의 적용 대상 영역 중의 하나인 Home page 텍스트의 경우, 문서 형식이나 강조를 위한 표기 방법에 따라 정확한 띄어쓰기 정보를 얻기 힘든 경우도 빈번히 발생한다. 예를 들어 "기 계 번 역 을 위 한 홈 페 이 지"와 같은 표기 형식은 Homepage 제목에서 흔히 볼 수 있는 예이다.

본 연구에서는 한국어의 음절 특성을 이용한 형태소 분석을 위한 어절 인식 방법을 제안하다. 제안하는 방법은 사전에 기반하지 않고 원형코퍼스(raw corpus)로부터의 필요한 음절 정보 및 어휘정보를 추출하는 방법을 취하므로 오류가 포함된 문장에 대하여 견고한 분석이 가능하고 많은 시간과 노력이 요구되는 사전구축 및 관리 작업을 필요로 하지 않는다는 장점이 있다. 그러므로 제안하는 방법은 띄어쓰기 오류가 빈번한 텍스트에 대한 전처리나 방대한 사전 정보를 요구하는 복합명사 분석에 이용될 수 있다.

본 연구에서는 우선 일반적인 관점에서 문장이 단어 단위로 분리되는 과정을 수학적으로 모델링한다. 그 다음으로 수학적으로 모델링된 단어분리 과정에서 음절정보를

이용하여 단어 확률을 정의한다. 마지막으로 문장으로부터 최대의 확률값을 같는 단단위를 분리하는 과정을 전개한다.

2. 단어분리모델

본 절에서는 음절의 열로부터 단어가 분리되는 과정을 전개하고 전개된 과정에서 단어의 확률을 정의한다. 음절의 열에서 단어를 분리하는 과정을 전개함에 있어 Luo에 의해서 제안된 언어모델에 기반하였다. 음절 $u_1, u_2...u_n$로 구성된 음절의 열을 S라고 할 때, 단어 인식과정은 음절의 열 S로부터 단어의 열 $w_1w_2....w_m$을 인식하는 과정으로 볼 수 있다. 이 과정을 다음수식과 같이 전개할 수 있다.

$$\begin{aligned} S &= u_1u_2...u_{n-1}u_n \\ &= (u_1...u_{x_1})(u_{x_1+1}...u_{x_2})...(u_{x_{m-1}+1}...u_{x_m}) \\ &= w_1w_2...w_m \end{aligned}$$

위의 수식에서 x_k는 단어 k_{th}의 마지막 음절을 나타낸다. 예를들어 w_k은 음절의 열 $u_{x_{k-1}+1}...u_{x_k}$로 구성된 단어를 가르킨다.

이때, 음절의 열 S에 대한 여러가지 가능한 단어의 열들의 후보집합을 $G(S)$라고 하면 $G(S)$는 다음과 같이 나타낼 수 있다.

$$G(S) = \{(x_1, ...x_m)|1 \leq x_1 \leq ... \leq x_m, m \leq n\}$$

위에서 정의된 개념들을 이용하여 음절의 열으로부터 단어를 분리하는 과정은 음절의 열 S로부터 분리 가능한 단어열의 후보들 중 최대의 확률값을 가지는 g^*를 찾는 문제로 정의하고 그 과정은 다음과 같다.

$$g^* = \arg \max_{g \in G(S)} P_g(w_1, ..., w_m)$$

위의 수식에서 $P_g(w_1, ..., w_m)$ 단어의 열 $w_1, ..., w_m$로 분리된 후보열 g의 확률값을 나타낸다. 위의 수식은 사슬규칙(chain rule)에 의하여, 단어 확률을 이전 단어에 대한 조건부 확률로 전개함으로써 다음과 같이 나타낼 수 있다.

$$\begin{aligned} P(w_1, ...w_m) &= p(w_1)p(w_2|w_1)p(w_3|w_2, w_1) \\ &\quad ...p(w_m|w_{m-1}, ..w_1) \\ &= \prod_i p(w_i|h_i) \end{aligned}$$

위의 수식에서 h_i는 w_i에 선행하는 단어열 $w_1, .., w_{i-1}$, 즉 w_i의 history를 나타낸다.. h_i를 적용하는 범위에 따라, trigram 모델은 이전 두 단어 w_{i-1}, w_{i-2}만을 고려하는 모델이 되고 bigram 모델은 w_{i-1}만을 history 고려하는 모델이다.

모델링하는 과정에서 많은 history 정보를 고려하는 것이 보다 정교한 단어 인식 모델을 설계하는 직접적인 방식이지만, 이 경우 모델을 학습시키기 위해서는 고려하는 history의 범위에 비례하여 많은 양의 데이타가 필요하다는 단점이 있다. 실제적으로, 대부분의 연구들이 단어들의 unigram 이나 bigram 만을 history 정보로 고려하는 이유도 trigram 이상의 정보를 history로 고려하는 경우 학습 데이타 양과 학습 속도가 현실적인 면에서 수용할 수 없기 때문이다. 심지어 단어의 unigram 정보 만을 고려하는 모델도 실제 영역(real domain)에 적용하는 경우 쓰이는 단어 수가 광범위하기 때문에 학습에 필요한 충분한 정보를 얻지 못하는 자료희귀(data sparseness) 문제를 겪게 된다.

단어에 기반하는 모델의 자료희귀(data sparseness) 문제 등을 극복하고, 띄어쓰기 오류가 빈번한 텍스트에 대해서도 견고한 분석을 수행하는 것을 목표로 본 연구에서는 음절단위 정보를 이용한 단어인식 전처리 모델을 제안한다. 제안하는 모델은 크게 세가지 단계로 전개되었다. 첫째로 단어의 경계 부분의 음절 bigram에 기반한 기본 모델을 제안하고 기본 모델에 단어를 구성하는 음절 bigram 정보를 추가하여 기본 모델을 확장한다. 마지막으로 단어 bigram 정보를 고려한 모델과 통합하여 제안하는 모델을 완성한다.

2.1 기본모델

음절 정보만을 이용한 모델이 단어인식을 올바르게 수행하기 위해서는 음절 정보가 단어 분리의 유용한 판단 기준이 될 수 있다는 것이 입증되어야 한다. 제안하는 모델은 단어와 단어가 인접하는 경계에 나타나는 음절 형태와 단어 내부에 쓰이는 음절의분포가 다르다는 것을 기본 가정으로 한다. 이 가정에 따라서 단어 분리를 위한 단어 확률은 음절bigram을 이용하여 정의할 수 있다. 실제 단어 경계 부분에 나타난 음절 bigram과 단어 내부에 쓰이는 음절 bigram의 차이의 정도를 측정하기 위하여 다음과 같은 두 함수 간의 거리 측정 함수를 이용하였다.

$$|p(\beta|u_i, u_j) - p(\omega|u_i, u_j)|$$

위의 수식에서 $p(\beta|u_i, u_{i+1})$는 음절 bigram u_i, u_{i+1}이 단어와 단어의 경계에 쓰일 확률을 나타내고 $p(\omega|u_i, u_j)$는 음절 bigram u_i, u_{i+1}이 단어 내부에 나타날 확률을 가르킨다.

즉 위의 수식은 두 함수 분포 사이의 거리의 절대값을 구하는 방법이다. 다시말해, 1에 가까울 수록 변별력이 강한 분포를 가지는 경우이고 0에 가까울 수록 유사한 분포를 가짐을 나타낸다. 위의 수식을 실험 1에서 사용한 학습 데이타에 적용하였을 때 차이값이 0.86으로 측정되었다. 이 값은 음절 bigram의 분포를 고려한 함수가 단어의 경계와 단어의 내부를 구분하는 함수로써 충분한 변별력을 가짐을 나타낸다. 따라서, 음절 bigram은 단어의 경계를 결정하는 중요한 자질정보(feature)가 될 수 있다.

그림1은 임의로 추출한 10개의 음절 bigram들 (0.시들 1.스팡 2.세탁 3.를좋 4.어녕 5.신전 6.몽고 7.운도 8.리처

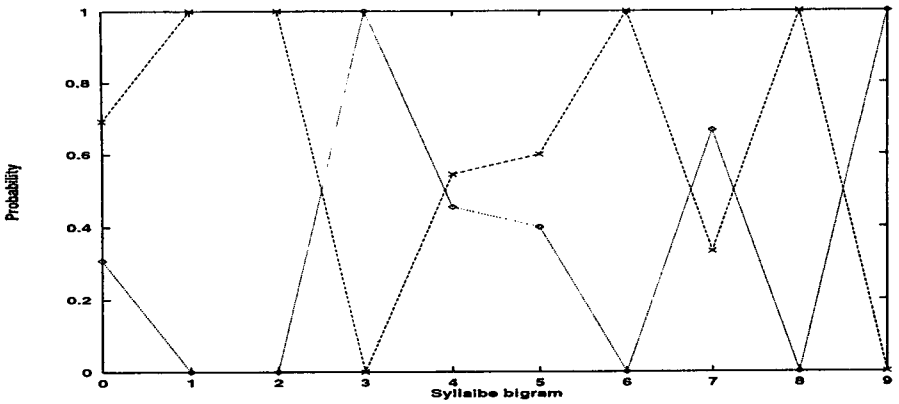

그림 1: 임의로 선택한 10개의 음절 bigram에 대한 사용빈도 분표의 예

9.게없)에 대한 단어 내부에 나타나는 빈도수와 단어 간의 경계 부분에 나타난 빈도수의 확률값을 나타낸다.

기본 모델에서는 단어의 확률 $p(w_i)$을 음절 bigram 패턴들의 단어 경계 부분에 쓰일 확률을 이용하여 다음과 같이 정의한다.

$$p(w_i) = \varsigma_1 p(\beta|u_{x_{i-1}}, u_{x_{i-1}+1}) p(\beta|u_{x_i}, u_{x_i+1}) \quad (1)$$

위의 수식에서 ς_1 단어들의 확률값 $p(w)$의 총합을 1로 만들기 위한 상수값이고, $p(\beta|u_i, u_{i+1})$은 음절 bigram u_i, u_{i+1}이 단어와 단어의 경계에 쓰일 확률값이다. 확률값 $p(\beta|u_i, u_{i+1})$은 학습 데이타로부터 다음과 같은 유도식을 통하여 얻을 수 있다.

$$p(\beta|u_i, u_{i+1}) = \frac{c(\beta, u_i, u_{i+1})}{c(\beta, u_i, u_{i+1}) + c(\omega, u_i, u_{i+1})} \quad (2)$$

$$p(\omega|u_i, u_{i+1}) = \frac{c(\omega, u_i, u_{i+1})}{c(\beta, u_i, u_{i+1}) + c(\omega, u_i, u_{i+1})} \quad (3)$$

위의 식에서 $c(\beta, u_i, u_{i+1})$는 학습 데이타에서 음절 bigram $u_i u_{i+1}$이 단어 간의 경계에 쓰이는 경우의 수이고, $c(\omega, u_i, u_{i+1})$는 음절 bigram $u_i u_{i+1}$ 학습 데이타에서 단어 내부에 나타나는 경우의 수이다.

2.2 Model 2

기본 모델에서는 음절 bigram이 단어의 경계에 쓰일 확률 정보만을 이용하였다. 두번째 모델에서는 단어 확률을 나타내기 위하여 음절 bigram이 단어의 경계에 쓰일 경우와 단어의 내부에 쓰이는 경향의 차이를 아래의 수식과 같이 모두 반영하였다.

$$p(w_i) = \varsigma_2 p(\beta|u_{x_{i-1}+1}, u_{x_{i-1}}) p(\beta|u_{x_i+1}, u_{x_i}) \prod_{j=x_{i-1}+1}^{x_i-1} p(\omega|u_{j+1}, u_j)$$

위의 수식에서 ς_2는 단어들의 확률 $p(w)$의 총합을 1로 정규화시키기 위한 상수이다. 예를 들어, "정보검색시스템"가 입력 음절열일 때 "검색"이 단어가 될 확률 p(검색)은 $p(\beta|$보검$)p(\beta|$색시$)p(\omega|$검색$)$로 정의된다.

2.3 Model 3

모델 3은 단어 bigram 정보를 반영하는 모델이다. 단어 단위의 정보를 이용함으로 수반되는 자료희귀(data sparseness)문제를 경감시키기 위하여 모델 2에서 제안된 단어의 확률값 $p(w_i)$를 보정함수(smoothing function)로 이용하였다.. 다음은 단어 bigram 정보와 음절 bigram을 이용하여 정의된 제안하는 모델이다.

$$p(w_i|h_i) = (1-\lambda) p(w_i|w_{i-1}) + \lambda p(w_i) \quad (4)$$

위의 수식에서 λ는 단어 bigram 함수 $p(w_i|w_{i-1})$와 본 연구에서 제안된 단어 확률 함수 $p(w_i)$의 값을 통합하기 위한 보정상수(interpolation coefficient)이다.

3. 단어 인식 알고리듬

입력 음절열부터 단어를 분리하는 과정은 앞 절에서 정의된 단어 확률값 $P(w)$과 동적 프로그래밍(dynamic programming)을 이용하여 아래와 같은 알고리듬을 이용하여 나타낼 수 있다.

아래의 알고리즘에서, 단어 w_{ij}는 i번째 음절부터 $(i+j)$번째 음절까지의 음절들로 구성된 단어를 나타낸다. 그리고 $L(i,j)$는 첫번째 단어에서 단어 $w_{i,j}$까지로 구성된 단어열의 최대 확률값을 나타낸다. 음절의 열에서 왼쪽부터 첫번째 단어가 되는 후보들을 초기화한 후에 $L(i,j)$은 아래와 같은 방법으로 재귀적으로(recursively) 정의된다. 각 $L(i,j)$를 구하는 과정이 완료되면 최대의 확률값을 같은 단어들의 열, 즉 구하고자 하는 해가 분절되는 위치를 지정하는 $b(i,j)$를 통하여 지정되게 된다.

Initialization :

$$L(0,j) = \log P(w_{0j})$$

Recursion :

From $i = 1$ to $len(w), j = 1$ to i, for all i, j

$$L(i,j) = \max_{1 \le k \le i-j} [L(j,k) + \log P(w_{ij}|w_{jk}) + \log P(w_{ij})]$$

$$b(i,j) = \arg \max_{1 \le k \le i-j} [L(j,k) + \log P(w_{ij}|w_{jk}) + \log P(w_{ij})]$$

Path backtracking:

$$i = len(w) - \max_{1 \le k \le len(w)} [L(len(w)-k), k]$$

$j = len(w) - i$

repeat

select w_{ij}

$j = i - b(i,j)$

$i = b(i,j)$

until $i > 0$

4. 실험

본 절에서는 소규모 실험을 통해서 제안하는 단어 인식방법이 실제 응용에 대하여 적절한 전처리를 수행함을 보인다. 본 실험에서는 실험대상 분야로 붙여쓰기가 된 문장에 대한 띄어쓰기 오류복원과 복합명사 분석 두가지 분야에 대하여 실험하였다.

4.1 실험 1 (띄어쓰기 오류복원)

띄어쓰기 오류복원은 잘못 붙여쓰어진 어절들을 올바르게 복원하는 과정이다. 앞 절에서 제안하는 단어인식 과정을 적용하여 한국어 문장에서 한국어 어절을 인식하는 문제에 응용할 수 있다. 예를들어 다음과 같이 띄어쓰기가 되어있지 못한 문장이 입력되었을 경우

"우리는정보검색시스템을개발하였다"

어절인식 문제는 다음과 같이 올바른 어절단위로 분리하는 과정이다.

"우리는 정보검색시스템을 개발하였다"

본 실험에서는 띄어쓰기가 바르게 수행된 130,103 어절을 이용하여 다음과 같은 세가지 부분으로 나누어 학습과 실험을 수행하였다.

- 띄어쓰기가 바르게 수행된 110,103 어절을 이용하여 모델을 학습하였다.
- 10,000 어절을 단어 bigram 정보와 음절 bigram 정보를 통합하는데 쓰인 보간상수(interpolation coefficient)을 구하는 데 이용하였다.
- 띄어쓰기가 바르게 수행된 10,000 어절 전부 붙여쓰기를 한 후 실험에 이용하였다.

본 연구에서 제안하는 방법들의 성능을 측정하기 위하여 우선 띄어쓰기가 바르게 수행된 10,000 어절 전부 붙여쓰기를 하였다. 붙여쓰기가 된 어절들을 본 연구에서 제안하는 어절인식 프로그램으로 복원한 후, 복원된 결과를 원래 데이타와 비교하여 어절인식 정확도를 측정하였다. 실제로 한국어에서는 여러 형태의 띄어쓰기 규칙이 문법적으로 허용이 된다. 예를 들어, "정보검색시스템을"와 "정보검색 시스템을"은 모두 문법적으로 올바르게 띄어쓰기된 경우이다. 본 실험에서 사용된 정확도 측정 방법은 위의 경우 원래 어절이 "정보검색시스템을" 이고 구현된 프로그램이 내 준 결과가 "정보검색 시스템을"인 경우분석 오류로 간주하게 된다. 그러므로 실제 어절 인식 정확도는 본 실험에서 사용된 원래 데이타와의 절대적인 차이 비교 방법에 비하여 높다고 할 수 있다.

4.2 Experiment 2 (복합명사 분석)

복합명사 분석 문제도 앞 절에서 제안한 단어 분리 모델을 이용하여 모델링 할 수 있다. 이 경우 입력 분석 대상이 되는 복합명사를 입력어절로 그리고 분석된 단순 명사를 인식된 단어에 대응된다. 예를 들어, 입력 음절의 열이 "정보검색시스템" 인 경우 단어 분리 모델은 "정보 검색 시스템" 단순 명사들의 열의 형태로 분석한다. 본 실험에서는 학습을 위하여 수동 분석된 12,305 개의 복합 명사를 이용하였다. 12,305 개의 복합명사는 다음과 같이 세부분으로 나뉘어 실험에 이용되었다.

- 10,305개의 수동 분석된 복합명사를 모델학습을 위한 학습 데이타로 이용하였다.
- 10,000개의 복합명사를 단어 bigram 정보와 음절 bigram 정보를 통합하는데 쓰인 보간상수(interpolation coefficient)을 구하는 데 이용하였다.
- 1,000개의 복합명사를 제안하는 방법의 성능평가를 위하여 사용하였다.

아래의 그림은 학습 데이타 증가에 따른 분석 정확도에 대한 실험 결과이다.

본 연구에서 제안한 방법을 띄어 오류 복원과 복합명사 분석에 응용한 실험 결과에서 주목할 만한 현상은 띄어쓰기 오류복원에서는 제안하는 모델2가 모델3과 유사한 성능을 보인데 반해 복합명사 분석의 경우에는 모델 3이 모델 2에 비하여 높은 성능 향상이 가능했다는 점이다. 이것은 실험 1에서 분석대상이 되는 문장 단위가 실험 2에서 분석 대상이 되는 복합명사에 비하여 음절의 길이가 일반적으로 길다는 사실에서 기인한다. 즉 분석 대상이 되는 음절의 길이가 길 경우 많은 양의 음절 bigram이 반영되므로 음절 bigram 정보 만을 이용한 모델 2만으로도 충분한 정확도를 보인다. 또한 모델 3의 경우 단어 bigram 정보를 고려하므로 음절의 길이가 긴 경우, 즉 고려해야할 단어 수가 많은 경우 자료희귀 문제로 인하여 충분한 성능을 발휘

표 1: 띄어쓰기 오류복원의 정확도

Model	학습데이타 크기(어절수)			
	15,028	30,012	60,004	110,103
Model 1	60.01	71.38	72.93	74.06
Model 2	68.38	80.03	81.69	84.36
Model 3	67.98	79.92	81.88	85.53

모델 1: (단어 경계 부분의 음절 bigram 정보를 고려한 모델)
모델 2: (단어 경계와 단어 내부의 음절 bigram을 고려한 모델)
모델 3: (모델 2와 단어 bigrams을 고려한 모델)

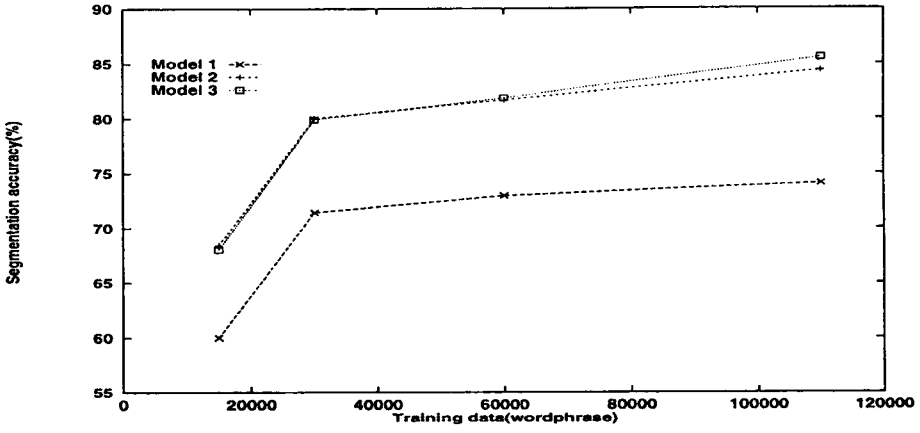

그림 2: 학습 데이타 증가에 따른 제안하는 방법들의 정확도 증가비교

표 2: 복합명사 분석의 정확도

Model	학습데이타 크기			
	12,520	25,012	5,151	10,305
Model 1	77.79	79.12	79.93	79.82
Model 1	77.54	79.04	79.93	80.93
Model 1	79.68	80.71	82.01	82.34

모델 1: (단어 경계 부분의 음절 bigram 정보를 고려한 모델)
모델 2: (단어 경계와 단어 내부의 음절 bigram을 고려한 모델)
모델 3: (모델 2와 단어 bigrams을 고려한 모델)

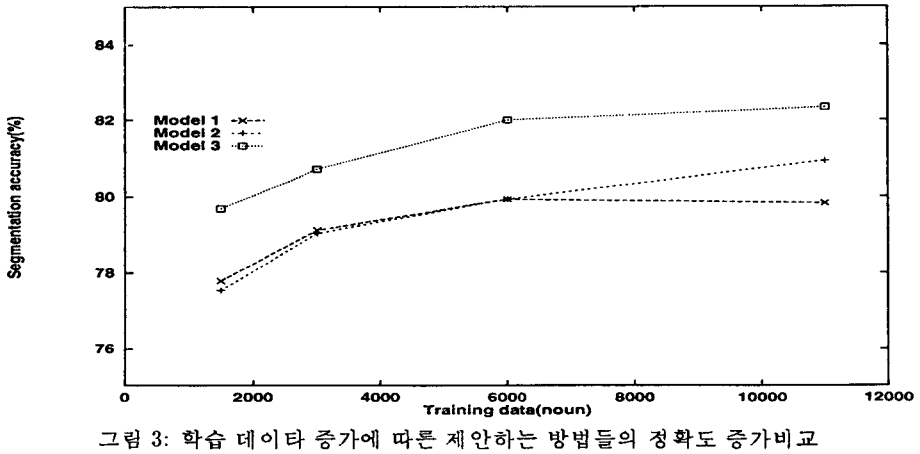

그림 3: 학습 데이타 증가에 따른 제안하는 방법들의 정확도 증가비교

하지 못 하였다.

5. 결론

본 장에서는 음절 bigram에 기반한 통계적 단어 분리 방법을 제안하였다. 제안된 방법은 소규모 실험을 통해 한국어에 있어 제안하는 모델이 띄어쓰기 오류복원과 복합명사 인식에서 유용하게 사용될 수 있음을 보였다. 본 연구에서는 적용대상을 한국어의 두 문제에만 국한시켰지만 제안하는 방법은 특정 언어에 제한적인 언어학적 지식은 배제하고 통계적인 정보만을 고려하였으므로 다른 언어나 유사한 문제에 대하여도 직접적으로 응용될 수 있다. 특히 본 연구에서 이용한 음절 단위의 정보는 기존의 단어 단위의 정보를 이용한 접근 방식에 비해 대용량의 사전관리의 부담이 없고, 소규모의 데이타만으로 모델의 충분한 학습이 가능하다는 장점이 있다.

참고 문헌

[1] Luo, Xiaogiang, Salim Roukos. 1994. "An Iterative Algorithm to Build Chinese Language Models" *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics*, 139-143.

[2] Richard Sproat, Chilin Shih, William Gale and Nancy Chang. 1996. "A Stochastic Finite-State Word-Segmentation Algorithm for Chinese" *Computational Linguistics*, 22(2): 377-403.

[3] Kim-Teng Lua and Kok-Wee Gan, 1994. "An Application of Information Theory in Chinese Word Segmentation" *Computer Processing of Chinese and Oriental Languages*, 8(1): 115-123.

[4] Jung H. Shin, Young S. Han, Young C. Park, Key-Sun. Choi. 1995. A HMM Part-of-Speech Tagger for Korean with wordpharsal Relations. In *Proceedings of Recent Advances in Natural Language Processing.*

[5] Hyouk R. Park, Young S. Han, Key-Sun Choi and Kang H.Lee. 1996. "A Probabilistic Approach to Compound Noun Inexing in Korean Texts" *Proceeding of 16th International Conference on Computational Linguistics*, 514-518.

[6] Joon Ho Lee and Jeong Soo Ahn. 1996. "Using n-Grams for Korean Text Retrieval" *Proceeding of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 216-224.

[7] Shiho Nobesawa, Junya Tsutsumi, Tomaki Nitta, Kotaro Ono, Sun Da Jiang, Masakzu Nakanishi. 1994. "Segmenting A Sentence Into Morphemes Using Statistic Information Between Words" *Proceeding of 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 216-224.