김일성종합대학학보
(정보과학)
주체107(2018)년
제64권 제1호

합성명사분해에 기초한 조선어발음자료기지구축의 한가지 방법

리명일, 김동수

조선어발음자료는 자모별발음자료, 주제별발음자료, 특수단어발음자료로 구성되여있다.

자모별발음은 조선어단어의 첫 글자의 자모별순서에 따라 정돈한 발음이며 주제별발음은 이야기의 주제로 설정될수 있는 단어와 그것과 관계되는 단어들에 대한 발음이다. 실례로《기계》,《학용품》,《음식》 등을 주제로 설정할 때 이 주제에 속하는 단어들에 대한 발음이다. 특수단어발음은 발음하기 힘든 단어와 성구, 속담들에 리용된 단어들에 대한 발음이다.

론문에서는 주제별발음자료기지에 추가되는 합성명사에 대한 분할에 기초하여 조선어발음자료기지를 구축하는 한가지 방법을 제기하였다.

발음자료기지는 품사에 따르는 홑단어 혹은 합성명사단어들로 구축되여있다. 품사에 따르는 홑단어자료는 단어사전을 리용하여 쉽게 구축할수 있으며 특수단어발음자료는 발음하기 힘든 단어들과 속담, 성구에 리용된 단어들을 선택하여 구축할수 있다. 합성명사단어는 어떤 의미를 가지는 단어인가에 따라 대응되는 주제별발음자료기지에 구축할수 있다.[1, 2]

선행연구[3]에서는 규칙을 리용한 합성명사분해에 대하여 고찰하였다.

합성명사는 어근과 접사의 결합으로 구성되였으며 일반적형태는 다음과 같다.

$$N=a_1\,N_1[\cdots N_k\,a_i\,N_{k+1}\cdots Na][a_p]$$
$$N=[a_1]N_1[\cdots N_k\,a_i\,N_{k+1}\cdots Na]\,a_p$$
$$N=[a_1]\,N_1[\cdots N_k\,a_i\,N_{k+1}\cdots Na]\,[a_p]$$

여기서 a_i는 접사, N_k는 어근(단일명사), [ ]는 생략할수 있는 항목에 대한 표시기호, a_1은 접두사 또는 1개 문자명사, a_p는 접미사 또는 1개 문자명사이다.

규칙을 리용한 합성명사의 분해는 앞뒤최장법에 기초하여 진행한다.

합성명사를 이루는 문자렬을 $c_1c_2\cdots c_n$으로 표시하고 규칙을 리용한 합성명사의 분해과정을 보기로 하자.

합성명사의 분해는 접사처리단계와 분해단계를 거쳐 진행된다.

① 접사처리단계

ㄱ) c_1을 선택한다.

ㄴ) $c_1 \in T$ 인가를 판정한다.($T=\{t_i \mid i=\overline{1,n}\}$, n: 접사의 개수)

$c_1 \in T \Rightarrow TF=1 \Rightarrow$ ㄷ)

$c_1 \notin T \Rightarrow$ ㄷ)

ㄷ) $c_1c_2 \in$ WD 인가를 판정한다.(WD: 단어사전)

$c_1 \in T$ 이고 $c_1c_2 \notin \mathrm{WD} \Rightarrow c_1$: 접사로 판정한다.

우에서 T는 n개로 이루어진 접사(t_i)들의 모임으로서 접사사전을 의미하며 TF는 접사로 판정되였음을 알리는 기발등록기이다.

접사의 처리에서는 먼저 첫 글자를 선택하고 접사인가를 판정한 다음 명사에 대한 분해를 진행한다.

② 분해단계

ㄱ) $c_2 \in \mathrm{WD}, c_2c_3 \in \mathrm{WD} \Rightarrow c_2$, c_2c_3을 후보단어로 설정한다.

ㄴ) $c_2 \notin \mathrm{WD}, c_2c_3 \in \mathrm{WD} \Rightarrow c_2c_3$을 단어로 설정한다.

ㄷ) 문자렬 $c_4c_5\cdots c_n$에 대하여 ㄱ), ㄴ)의 처리를 진행한다.

여기서 WD=$\{w_i \mid i=\overline{1, n}\}$로서 단어($w_i$)들의 모임이다.

규칙에 기초한 합성명사의 처리에서는 접사사전과 단어사전을 리용하므로 사전에 속하는 요소들(단어와 접사)의 수가 고정되면 등록되지 않은 요소에 대하여서는 판정하지 못하며 따라서 합성명사분해률이 낮은 제한성을 가진다.

1. 조선어발음자료기지구축

조선어단일명사들은 대부분 2개 문자로 구성되여있다.

15만개의 합성명사모임으로부터 단일명사들을 분리해본 결과 10 623개의 단일명사중 2개 문자명사는 8 905개로서 83.8%, 3개 문자, 4개 문자명사는 1 361개로서 12.8%, 1개 문자명사는 78개로서 0.7%를 차지하였다. 이것은 2개 문자명사가 다른 길이의 명사에 비해 압도적비중을 차지한다는것을 알수 있다.

이러한 사실을 고려하여 합성명사형태를 다음과 같이 결정하였다.

$$c_1c_2/\cdots/c_{i-2}\ c_{i-1}/c_i/c_{i+1}\ c_{i+2}/\cdots/c_{n-1}c_n$$

$$c_1/\ c_2\ c_3/\cdots/c_{i-1}\ c_i/\cdots/\ c_{n-1}\ c_{n-2}/\ c_n$$

$$c_1c_2/\cdots/c_{i-2}\ c_{i-1}/c_i/\ c_{i+1}\ /c_{i+2}\ c_{i+3}/\cdots/c_{n-1}c_n$$

일반적으로 통계적분해는 문자렬들의 출현빈도특성을 반영하여 분해하는 방법으로서 합성명사가 분해되는 류형은 다양하다.

따라서 합성명사를 분해하기 위하여서는 목적하는 부분에서의 생성확률이 가장 큰 단어렬을 추출하는 모형 즉 합성명사의 분해에서 최대생성확률을 얻기 위한 확률적언어모형이 필요하다.

문자렬을 $C=c_1c_2\ldots c_n$ (n: 합성명사의 문자수)이라고 할 때 생성가능한 단어렬들의 모임 W는 다음과 같다.

$$W=\{W_i \mid i=\overline{1, r}\}$$

여기서 $W_i=\{W_{i1}, W_{i2}, \cdots, W_{im}\}(1 \le m < n)$, m은 단어렬의 수이다.

문자렬 C로부터 생성가능한 매개 단어 $W_{ij}(j=\overline{1, m})$의 생성확률을 $P(W_{ij})$라고 하고 그것들의 적을

$$P(W_i)=\prod_{j=1}^{m} P(W_{ij})$$

라고 하면 합성명사분해는 $P(W_i)$가 최대로 되는 단어렬 $\hat{W}$을 구하는 문제에 귀착된다. 즉 $\hat{W}=\arg\max P(W_i)$이다.

이 언어모형에 기초하여 $P(W_i)$가 최대로 되는 단어렬 $\hat{W}$을 구하는 문제를 2－gram 통계량에 기초하여 고찰하자.

문자렬 $C=c_1c_2\ldots c_n$으로부터 생성된 2－gram문자렬들의 모임을 A라고 하면 $A=\{c_ic_{i-1}\}$, $i=\overline{1,\ n-1}$이다.

문자렬 c_ic_{i+1}의 출현확률은 다음의 식으로 계산된다.

$$P(c_ic_{i+1})=f(c_ic_{i+1})/N$$

여기서 $f(c_ic_{i+1})$는 문자렬 c_ic_{i+1}의 출현빈도, N은 시험모임에서 합성명사의 총수이다.

합성명사분해에 기초한 조선어발음자료기지구축과정은 다음과 같다.

① $P(c_ic_{i+1})=f(c_ic_{i+1})/N$을 구한다.

② $P(c_1c_2)$와 $P(c_2c_3)$의 크기관계를 비교하여 접사와 명사를 분리한다.

$P(c_1c_2)<P(c_2c_3)$이면 $c_1\in T$, $c_2c_3\in$WD에 대한 판정을 진행한다. 즉 접사사전과 단어사전을 리용하여 존재상태를 판정한다. 등록되여있지 않으면 접사사전과 단어사전에 등록한다.

$P(c_1c_2)>P(c_2c_3)$이면 $c_1c_2\in$WD, $c_2c_3\in$WD에 대한 판정을 진행한다.

③ 문자렬 c_ic_{i+1}에 대하여 ①, ②를 반복진행한다.

④ 분해된 합성명사에서 주제단어를 결정한다.

⑤ 주제단어와 합성명사를 조선어주제별발음자료기지에 등록한다.

우의 과정은 이웃한 두 문자렬들의 출현확률을 구하고 확률이 큰 문자렬을 분해후보로 선정하면서 생성해나가는 과정이다.

실례 $C=$《초강도행군》

$P(c_1c_2)<P(c_2c_3)$이므로 c_2c_3이 단일명사, c_1은 접두사로 될 가능성이 크다. 따라서 c_1, c_2c_3에 대하여 각각 접사사전과 단어사전을 참고한다. 즉《초》는 접두사이며《강도》는 단어라는것을 알수 있다. 다음《행군》에 대하여 단어사전에 존재하는가를 판정하고 다음과 같은 분해결과를 얻는다.

{초/강도/행군}

오유처리는 다음과 같이 진행한다.

분해오유를 처리하는 문제는 합성명사분해효률을 높이기 위하여 나서는 중요한 문제이다.

합성명사는 단일명사들의 병렬결합으로 구성되는데 그 결합부분사이의 문자렬들은 합성명사모임에서 출현빈도가 커지는 특성을 나타내는 경우가 있다. 즉 각이한 단일명사들이 다양한 형식으로 서로 결합되므로 이러한 특성이 더 빈번해진다. 실례로《전동식자동차》,《가동식다리》,《진동식선별기》,《동식물기름》등의 합성명사들에서 부분문자렬《동식》은 접미사《식》과 어근과의 결합, 어근들사이의 결합에 의해 생성되는데 이러한 결합에 의해 생성되는 부분문자렬들의 출현빈도는 어근의 출현빈도보다 커질수도 있다.

분해결과는 $\hat{W}=${이, 동식, 건물}로 되는데 이때 문자《이》는 1개 문자명사이므로 정확한 분해로 잘못 출력된다. 이것은 의미없는 문자렬《동식》이 합성명사모임내에서 출현빈도가 크기때문에 생긴 오유이다. 이러한 분해오유를 해소하기 위하여《동식》과 같은 문자렬을 례외자료모임에 등록하고 식 $P(W_i)=\prod_{j=1}^{m}P(W_{ij})$에서 확률 $P(W_{ij})$를 가장 작은 값으로 설정한다. 그러면 작은 값을 가지는 생성확률 $P(W_{ij})$가 계산되므로 분해오유를 줄일 수 있다.

2. 실험 및 평가

규칙과 통계적특성에 기초하여 합성명사를 분해한 결과를 다음의 표에 보여주었다.

표. 규칙과 통계적방법에 의한 분해률

명사의 개수/개	규칙/%	통계/%
5	72	86
6	69	78
7	66	77
8	66	76
9	63	71

표에서 보는바와 같이 규칙을 리용하였을 때 분해률은 평균 66.2%, 통계적특성을 리용하였을 때 분해률은 평균 73.8%로서 통계적특성을 리용하면 합성명사가 보다 정확히 분해된다는것을 알수 있다.

맺 는 말

통계적방법에 리용되는 합성명사의 형태를 결정하였으며 합성명사분해에 기초한 조선어발음자료기지구축방법을 실현하고 분해률을 평가하였다.

참 고 문 헌

[1] A. C. Buck et al.; Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, 25, 2014.

[2] Q. N. Rockiman; Natural Language Process, 2, 32, 2015.

[3] Paul Piwek; Natural Language Process, 2, 12, 2010.

주체106(2017)년 11월 5일 원고접수

A Method of Korean Pronunciation Database Construction based on the Complex Noun Division

Ri Myong Il, Kim Tong Su

The Korean pronunciation data is composed of alphabet, subject and special words.

In this paper we suggested a method of the Korean pronunciation database construction based on the complex noun division database using statistical property.

Key words: database, language processing, information retrieval