김일성종합대학학보
(자연과학)
주체106(2017)년 제63권 제10호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 63 No. 10 JUCHE106(2017).

조선어련속음성인식체계의 인식단위결정에서 음절의 리용방법

리 혁 철

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《나라의 과학기술을 세계적수준에 올려세우자면 발전된 과학기술을 받아들이는것과 함께 새로운 과학기술분야를 개척하고 그 성과를 인민경제에 적극 받아들여야 합니다.》

(《**김정일**선집》 증보판 제11권 138～139페지)

대어휘련속음성인식에서 리용되는 인식단위는 언어에 따라 서로 다르다. 영어를 비롯한 유럽계언어들은 굴절어의 특성을 가지는것으로 하여 단어 그자체를 인식단위로 리용할수 있다.

조선어련속음성인식에서는 조선어의 교착어적특성으로 하여 단어 그자체를 인식단위로 리용하는것이 불합리하므로 형태부를 기본인식단위로 리용한다. 그러나 앞붙이와 뒤붙이와 같이 길이가 짧은 형태부들은 형태부들사이의 경계에서 발생하는 발성전이현상을 정확히 반영하기 어려우며 이것은 음성인식에서 삽입, 탈락오유의 기본원인으로 된다.

형태부들을 결합하여 합성단어들을 생성하고 그것들을 인식단위로 취하면 형태부들사이 경계에서 일어나는 발성전이현상을 반영할수 있을뿐아니라 언어모형측면에서 국부적으로 고차의 n–gram적용효과를 얻을수 있다.[1] 그러나 많은 분야들을 대상으로 하는 대어휘련속음성인식의 응용실천에서 형태부와 합성단어들을 그대로 전부 리용하는것은 어휘수증대를 가져오며 그에 따라 단어식별자의 자료형이 커져 모형의 비대화를 초래하게 된다. 물론 여러가지 모형압축과 축소화기술들이 제안되여있지만 그것들은 성능저하를 동반하게 된다. 또한 국부적으로 저빈도단어들을 무시하는 방법도 있지만 그것은 어휘의 피복률을 저하시킨다.

다음으로 최대우도분할방식으로 학습코퍼스를 가변길이음절렬들로 표현하고 모형파라메터들을 추정하는 방법이 제안되였다.[2] 이 방법에서는 분할과 재추정의 반복으로 가변길이음절렬들을 생성하고 수렴조건에 의하여 생성개수를 조절할수 있으며 모형의 분기수에서도 개선이 있었다. 그러나 이 방법은 순수 통계적으로 결합된것으로 하여 많은 음절렬들이 언어적인 의미정보를 류실하며 결국 음성인식성능에 부정적인 영향을 주게 된다.

이로부터 론문에서는 높은 인식성능을 보장하면서도 어휘수증가를 막기 위한 방도로서 합성단어와 부분적인 음절토막화를 결합하여 리용하는 방법을 제안하고 실험을 통하여 그 효과성을 검증하였다.

1. 형태론적통계에 기초한 국부적인 음절토막화

합성단어들은 학습코퍼스에서 자주 출현하는 형태부쌍들로 선택한다. 이때 낮은 빈도를 가지는 형태부쌍들을 합성단어로서 어휘에 추가하면 복호화과정에 그것과 비슷한 다른

단어들과 음향학적인 혼돈을 가져올수 있다.

결과적으로 학습코퍼스에는 합성단어들과 형태부들이 존재하게 되는데 품사정보에 따르는 통계적분석자료는 표 1과 같다.

표 1에서 구체적인 수값들은 학습코퍼스에 따라 차이날수 있지만 전반적인 통계는 일반성을 잃지 않는다. 그리고 론문에서는 어휘사전에서 많은 비중을 차지하는 대표적인 품사들에 대하여서만 서술하고 코퍼스의존성이 거의 없는 앞붙이, 뒤붙이, 감동사와 같은 품사들은 략하였다. 표 1에서 알수 있는바와 같이 외래어를 포함한 고유명사와 일반명사들이 대략 전체 어휘의 75%를 차지하고있으며 그것들사이에 빈도차이는 매우 심하다.

고유명사와 일반명사사전의 빈도특성을 표 2에 주었다.

표 1. 학습코퍼스의 대표적인 품사별통계

품사	백분률/%
합성단어	10.25
일반명사	21.59
고유명사	53.10
형용사	1.66
동사	2.42
부사	2.82
결합토	2.71

표 2. 고유명사 및 일반명사사전의 빈도특성

빈도	고유명사사전안에서의 백분률/%	일반명사사전안에서의 백분률/%
1이하	41.50	11.42
5이하	69.75	25.59
10이하	78.48	33.36
20이하	85.18	41.45
50이하	91.83	52.76
100이하	95.23	61.22
200이하	97.39	69.76

표 2에서 알수 있는바와 같이 고유명사들은 일정한 범위내에서는 포화특성을 가지지만 일반명사들은 거의 선형성에 가까운 특성을 보여주고있다.

고유명사에는 외래어들이 대다수를 차지하고있으며 전반적으로 빈도수가 작은것으로 하여 이러한 단어들이 개별적으로 체계의 인식성능에 미치는 영향은 매우 작다. 그러므로 빈도수가 작은 고유명사들은 어휘사전에서 배제될 가능성이 크지만 배제되는 단어수가 전체 어휘수의 상당한 비중(거의 50%)을 차지하므로 어휘의 피복률과 체계의 전반적성능의 견지에서는 커다란 손실을 주게 된다. 이러한 현상은 일반명사에 대해서도 마찬가지이며 단지 빈도분포특성이 고유명사와 다를뿐이다.

또한 일반적으로 빈도가 작은 단어들은 일반 사용자들에게 잘 알려지지 않은 단어들이므로 조선어인 경우 그러한 단어들에 대해서는 음절단위로 발성할 가능성이 많다.

이러한 통계적인 해석과 조선어의 어음학적특성에 기초하여 우리는 합성단어와 형태부들을 그대로 리용하는것과 함께 전체 어휘사전의 많은 비중을 차지하고 코퍼스의존성이 강한 고유명사 및 일반명사사전에 대해서 저빈도단어들을 음절단위로 토막화하고 음절 n-gram으로 표현함으로써 어휘의 피복률과 어휘수, 인식성능문제들을 다같이 해결하였다.

2. 실험 및 결과분석

실험에서 리용된 언어모형들은《로동신문》과 상식, 경제, 군사, 철학, 력사, 법률을 비롯한 사회정치문화분야의 본문코퍼스들로부터 구축되였다. 본문코퍼스는 형태부단위의 품사표식이 붙은 자료기지로서 크기는 2.3GB이고 어휘수는 14만 6천개이다.

검사자료로서 학습에 참가한 상식코퍼스의 1 176개 문장을 선택하였다.

14만 6천개의 전체 어휘를 다 포함하고있는 단어 3-gram모형을 기준모형 LM0으로, 64 000개 어휘규모의 가변길이음절 3-gram모형을 비교모형 LM1로 정하였다.

또한 14만 6천개의 전체 어휘중에서 고유명사사전과 일반명사사전에서 빈도가 각각 50, 20이하인 단어들을 배제한 62 000개 어휘규모의 단어 3-gram모형을 비교모형 LM2로 선정하였다.

모형 LM2에서 배제된 단어들을 음절토막화하여 학습시킨 단어 및 음절 3-gram모형을 LM3(제안모형)으로 하였다.

LM2를 제외한 3개 모형들은 어휘의 피복률이 100%이며 LM0을 제외한 3개 모형들은 어휘수가 65 000개 미만으로서 단어식별자의 자료형은 2B이다.

우의 4가지 언어모형들에 대한 단어오유률평가를 조선어련속음성인식체계《룡남산》에서 진행하였으며 그 결과를 표 3에 제시하였다.

표 3. 각이한 인식단위 3-gram모형들의 성능평가

모형	인식단위	어휘수/개	단어오유률/%
LM0	합성단어, 형태부	146 000	4.18
LM1	가변길이음절렬	64 000	4.65
LM2	합성단어, 형태부	62 000	4.68
LM3	합성단어, 형태부, 음절	64 000	4.27

표 3에서 보는바와 같이 제안모형(LM3)의 단어오유률이 어휘수와 피복률이 거의 류사한 선행모형(LM1)에 비하여 상대적으로 0.4%나 개선되였다는것을 알수 있다.

또한 LM3은 LM2에서 배제된 어휘들을 음절 3-gram들로 표현함으로써 LM2에 비하여 피복률이 높아지고 결국 인식성능이 개선되게 되였다.(대략 0.4%) 그리고 LM0은 합성단어와 형태부단위의 어휘들을 모두 포함하고있는것으로 하여 인식성능은 상대적으로 제일 높지만 모형이 비대해지는데(대략 2.7배로 증가) 비해볼 때 제안모형에 비한 상대적성능개선(대략 0.1%)은 작다고 볼수 있다.

합성단어개수와 음절토막화하려는 단어선택빈도턱값은 인식체계의 규모와 대상에 따라 실험적으로 조절할수 있다.

맺 는 말

조선어대어휘음성인식체계의 응용실천에서 제기되는 인식단위와 어휘수문제를 해결하기 위하여 합성단어와 함께 품사별통계에 기초한 국부적인 음절토막화를 결합하여 리용하는 방법을 제안하였다.

실험을 통하여 제안한 방법이 실천적견지에서 이전의 방법들에 비해 우월하다는것을 확인하였다. 제안한 방법은 이전의 방법들에 비하여 단어오유률을 대략 0.4% 개선하였다.

참 고 문 헌

[1] **김일성**종합대학학보(자연과학), 52, 12, 26, 주체95(2006).

[2] 리혁철; 콤퓨터와 프로그람기술, 3, 24, 주체104(2015).

주체106(2017)년 6월 5일 원고접수

Study on Usage of Syllable for Determining a Recognition Unit in Korean Continuous Speech Recognition System

Ri Hyok Chol

In large vocabulary continuous speech recognition, it is a matter of significance for improvement of its performance to determine the recognition unit reasonably.

In this paper, we proposed the local segmentation into a syllable to resolve a mismatch between the recognition unit and vocabulary size that occurred frequently in Korean large vocabulary speech recognition applications, and then evaluated the performance.

Key words: recognition unit, n-gram model, morpheme, syllable