김일성종합대학학보
정보과학

주체109(2020)년
제66권 제3호

토의 모방에 기초한 조선어자연발화 음성언어코퍼스의 구축방법

리혁철, 김철

경애하는 최고령도자 **김정은**동지께서는 다음과 같이 말씀하시였다.

《과학기술과 경제의 일체화를 다그치고 나라의 경제를 현대화, 정보화하는데서 과학기술부문이 주도적인 역할을 하도록 하여야 합니다.》

선행연구[1, 4]들에서는 자연발화음성언어코퍼스문제를 해결하기 위하여 글말과 입말코퍼스들을 결합하여 리용하거나 웨브자료로부터 입말과 류사한 본문들을 추출하여 막힘(filler)과 휴지(pause)들을 인위적으로 추가하는 방법들을 제안하였다. 그러나 일반적으로 입말코퍼스의 크기가 제한되여있고 막힘이나 휴지현상들은 자연발화음성언어의 구별적특징들중에서 일부에 지나지 않는다.

론문에서는 글말과 구별되는 입말체토들의 문법적기능과 발음특성을 통계적으로 분석하고 대표적인 토들을 모방한 언어코퍼스생성방법을 제안하였다.

1. 문 제 설 정

자연발화음성은 랑독이나 방송보도음성과 구별되는 음향 및 언어적특성들을 가진다. 특히 빠른 발성속도, 불명확한 발음, 발음변동, 장황한 표현, 문법이 맞지 않는 문장, 막힘과 교정과 같은 굴절현상들이 흔히 나타난다.

자연발화음성을 정확히 표기하자면 음성인식체계에서 이러한 음향 및 언어적현상들을 모형화하여야 한다.

대규모의 코퍼스들을 수집하는것은 일반적으로 수동적인 표기화의 비용문제로 하여 비현실적이다.

그러므로 신문과 같이 특정한 과제의 특징들을 나타내는 코퍼스들을 자연발화음성의 특징들을 나타내는 코퍼스와 결합하여 리용할수 있다.

인식체계[1, 2]들에서 이러한 결합방법을 적용하고있지만 화제와 자연발화특성들을 옳게 조화시킬수 없는 문제가 제기된다.

코퍼스의 화제적응성과 자연발화특성문제들을 다같이 해결하기 위하여 선행연구[3, 4]에서는 포괄적인 화제범위의 다량의 웨브페지들을 내리적재하고 그로부터 입말과 류사한 본문들을 선택한 다음 막힘과 휴지와 같은 전형적인 언어현상들을 추가하였다. 그러나 이러한 방법들은 글말과 구별적인 입말어휘들중의 일부로 막힘과 휴지만을 모방하였다.

조선어자연발화음성언어모형을 구축하기 위하여 글말과 구별되는 입말체토들의 문법적기능과 발음특성을 통계적으로 분석하고 대표적인 토들을 모방한 언어코퍼스를 생성하여 리용한다.

2. 입말체토들의 특성분석과 입말체본문코퍼스의 생성

1) 대표적인 입말체토들의 특성분석

일반적으로 조선어글말에서 토들은 자기의 문법적, 기능적특성으로 하여 코퍼스의존성, 화제의존성이 거의나 없다. 그러나 입말체토들은 글말체토와 서로 다른 특성을 가지고 발성된다.

동일한 문법적기능을 수행하는 글말체토와 입말체토사이의 대응관계를 표 1에 보여주었다.

표 1. 동일한 문법적기능을 수행하는 글말체토와 입말체토사이의 대응관계

류형	글말체토	입말체토	실례
줄임형	는	ㄴ	나는 – 난
	를	ㄹ	학교를 – 학괄
바뀜형	고	구	먹고 – 먹구
	로	루	기차로 – 기차루
	으로	으루	집으로 – 집으루
	도	두	동무도 – 동무두
줄임형과 바뀜형의 혼합형	이를	일	영학이를 – 영학일
	에서는	에선	입에서는 – 입에선
	라는	란	쉬라는 – 쉬란
	고서는	구선	싣고서는 – 싣구선
	으로는	으룬	힘으로는 – 힘으룬

줄임형토는 글말체토《는, 를》의 변형이다. 입말에서 이 토들은 마지막글자가 받침이 없는 체언(명사, 대명사)의 뒤에 놓일 때《ㄴ, ㄹ》의 형태로 앞단어에 결합된다. 글말코퍼스에서 마지막글자가 받침이 없는 체언의 뒤에 오는 토《는, 를》들을《ㄴ, ㄹ》로 치환함으로써 줄임형토들을 모방한 입말코퍼스를 생성할수 있다.

바뀜형토는 글말체토《고, 로, 으로, 도》의 변형이다. 입말에서 이 토들의 모음글자《ㅗ》는《ㅜ》의 형태로 바뀌여진다. 줄임형토와 마찬가지로 글말코퍼스에서 토《고, 로, 으로, 도》들을《구, 루, 으루, 두》로 치환함으로써 바뀜형토들을 모방할수 있다.

줄임형토와 바뀜형토의 혼합형토는 글말체토들의 결합에 의하여 형성된 결합토들에서 줄임과 바뀜현상이 동시에 나타나는 입말의 언어적현상이다. 그러므로 최소형태부에서보다 결합형태부들에서 많이 나타난다.

형태부결합처리에 의하여 얻어지는 글말코퍼스의 결합토들에 대해서 줄임형토와 바뀜형토들을 모방하여 혼합형토처리를 진행할수 있다.

2) 토의 모방에 의한 입말체언어코퍼스생성절차

글말코퍼스로부터 입말코퍼스를 자동적으로 생성하는 단계들은 다음과 같다.

① 형태부해석을 통하여 글말본문코퍼스로부터 형태부분할코퍼스를 생성한다. 생성된 형태부코퍼스는 최소형태부단위이다.

② 최소형태부단위의 해당한 토들에 한하여 줄임형토와 바뀜형토로 치환을 진행한다. 결과적으로 얻어지는 코퍼스에는 줄임형토와 바뀜형토들이 생겨나게 된다.

③ 형태부결합처리를 진행한다. 결과적으로 얻어지는 코퍼스에는 여러 품사형태부들

이 결합된 결합형태부들이 존재하게 된다.

④ 결합토들에 한하여 혼합형토처리를 진행한다.

이와 같은 절차에 의하여 글말코퍼스로부터 줄임형토와 바뀜형토, 혼합형토들이 모방된 입말코퍼스가 얻어지게 되며 이 코퍼스에 기초하여 언어모형을 학습시킨다.

3. 평 가 실 험

실험에서 리용된 언어모형들은《로동신문》코퍼스(글말체)와 영화문학 및 극작품집(입말체)들로 구성된 학습자료로부터 구축되였다. 글말체와 입말체본문코퍼스는 형태부단위의 품사표식이 붙은 자료기지로서 크기는 각각 1GB, 5.7MB이고 어휘수는 각각 9만 8천개, 1만 8천개이다.

제안한 방법이 글말체에 주는 영향도 같이 평가하기 위하여 글말체코퍼스의 절반부분만을 입말체코퍼스생성에 리용하고 나머지절반은 그대로 리용하였다. 제안한 방법이 인식성능에 주는 영향을 정확히 평가하기 위하여 어휘밖의(OOV: Out of Vocabulary) 단어가 포함되지 않도록, 학습자료와 중복되지 않도록 검사자료를 선택하였다. 검사자료는 글말체평가용본문 50문장과 입말체평가용본문 50문장으로 구성하였다.

학습자료로부터 3－그람언어모형들을 구축하였다. 이때 어휘사전크기가 65K를 넘지 않도록 낮은빈도단어들은 음절토막화하였다.

기준실험과 제안한 방법에 기초하여 생성된 입말코퍼스를 리용하여 실험을 진행하고 결과분석을 통하여 그 효과성을 검증하였다.

1) 기준실험

기준모형으로 최소 및 결합형태부단위의 전체 글말코퍼스(1GB)와 입말코퍼스(5.7MB)로부터 추정된 3－그람모형들(LM 01, LM 02)을 리용하였다.

최소 및 결합형태부단위의 언어모형(LM: Language Model)들에 대한 기준실험결과를 표 2에 보여주었다.

표 2. 최소 및 결합형태부단위의 언어모형들에 대한 기준실험결과

언어모형	1－그람/수	2－그람/수	3－그람/수	글말 WER/%	입말 WER/%	평균 WER/%
최소형태부단위 언어모형(LM 01)	54.6K	3 153.7K	1 5604.2K	0.96	28.57	14.76
결합형태부단위 언어모형(LM 02)	63.6K	6 265.4K	2 4895.8K	0.34	23.95	12.14

표 2에서는 최소 및 결합형태부단위의 기준언어모형들의 n－그람수와 그것을 리용한 인식실험결과를 보여준다. 여기서 비교실험들에서 리용되는 모방된 입말코퍼스의 그람확장정도를 수자적으로 대비하기 위해 그람수를 주었다.

2) 비교실험

제안한 토의 모방방법으로 생성한 여러가지 입말코퍼스들로부터 학습된 언어모형들의 인식성능비교실험을 진행하였다. 실험은 세가지 단계(준말형토 및 바뀜형토의 모방, 결합형태부의 생성, 혼합형토의 모방)로 진행하였다.

첫번째 단계에서는 최소형태부단위의 글말코퍼스의 절반으로부터 준말형토와 바뀜형토들을 모방한 입말코퍼스를 생성하고 나머지글말코퍼스의 절반과 합쳐서 LM 1을 추정

한다.

두번째 단계에서는 첫번째 단계의 결과코퍼스에 대해 일정한 척도에 준하여 형태부결합을 진행하고 이 결합형태부단위의 코퍼스로부터 LM 2를 추정한다.

세번째 단계에서는 두번째 단계의 결합형태부단위의 코퍼스의 절반에서 혼합형토를 모방하고 나머지코퍼스부분과 함께 LM 3추정에 리용한다.

모방된 입말코퍼스를 리용한 언어모형들에 대한 비교실험결과를 표 3에 보여주었다.

표 3. 모방된 입말코퍼스를 리용한 언어모형들에 대한 비교실험결과

언어모형	1－그람/수	2－그람/수	3－그람/수	글말 WER/%	입말 WER/%	평균 WER/%
LM 1	54.9K	3 294.1K	16 308.4K	1.70	26.43	14.06
LM 2	64.2K	6 527.3K	25 882.2K	1.15	23.44	12.29
LM 3	65.3K	6 545.3K	25 917.9K	1.15	22.98	12.06

표 3에서 보여준것처럼 최소형태부단위의 모형들가운데서 LM 1이 기준모형 LM 01에 비하여 입말검사자료에 한해서는 2.14%, 검사자료전체에 대해서는 0.7%의 WER가 개선되였다는것을 알수 있다. 또한 결합형태부단위의 모형들가운데서 LM 2와 LM 3이 기준모형 LM 02에 비하여 입말검사자료와 검사자료전체에 관하여 각각 0.51%, 0.97%의 WER가 개선되였다.

맺 는 말

조선어자연발화음성언어의 대표적특징인 줄임형토와 바뀜형토들을 모방하여 글말코퍼스로부터 입말코퍼스를 구축하기 위한 방법을 제안하였다. 실험을 통하여 제안한 방법이 조선어자연발화음성의 언어적 및 발성적견지에서 우월하다는것을 확증하였다.

참 고 문 헌

[1] J. Glass et al.; Proc. Eurospeech, 2553～2556, 2007.

[2] M. Cettolo et al.; Proc. ICASSP, 769～772, 2004.

[3] R. Masumura et al.; Proc. Interspeech, 1465～1468, 2011.

[4] X. Hu et al.; Journal of Information Processing, 21, 2, 168, 2013.

주체109(2020)년 5월 5일 원고접수

A Method for Constructing Korean Spontaneous Spoken Language Corpus Based on Imitating *tho*(Particle)

Ri Hyok Chol, *Kim Chol*

In the paper, for constructing the Korean spontaneous spoken language model, we analyze grammatical functions and pronouncing characteristics of spoken vocabulary distinguished from written one, and use a language corpus imitating typical particles.

Keywords: automatic speech recognition, language model, spontaneous speech