한국어 발음열 자동 생성을 위한 형태소 태그 정보 기반의 텍스트 전처리기

이경님° 정민화
서강대학교 컴퓨터학과
{knlee, mchung}@sogang.ac.kr

Text Preprocessor for Generating Korean Automatic Pronunciation Variants Using Morpheme-tag Information

Kyong-Nim Lee° Minhwa Chung
Dept. of Computer Science, Sogang University

요 약

일반적으로 발음열 자동 생성기는 음성 인식 및 음성 합성에 사용되며, 그 주된 역할은 입력된 한글 철자에 대해 발음 나는 데로 표기된 음소열로 출력하는 것이다. 그러나 실제 입력되는 문장에는 특수 기호 및 알파벳, 아라비아 숫자, 영어 단어, 알파벳과 숫자가 혼용된 약어, 기호 단위 명사 등이 포함되어 있다. 게다가 아라비아 숫자의 경우 단위명사의 종류에 따라서 뿐만 아니라, 문맥에 따라 숫자를 읽는 방식이 달라지게 된다. 이러한 모든 현상들을 발음열 생성기 내부에서 처리하게 되면 선행 작업이 상대적으로 크게 되어 과부하 문제가 발생된다. 또한 어절 내의 문맥 정보만으로 정확한 변환 결과를 얻기 힘들기 때문에 형태소 분석 수행 결과 및 예외처리를 위한 루틴을 포함하여 한글 자소 단위의 입력 형식으로 변환하는 전처리 시스템을 구성하였다.

1. 서론

일반적으로 발음열 자동 생성기의 역할은 음성인식 측면에서 보면 음성 데이터에 대한 학습용 발음열 생성이고, 합성 측면에서는 입력된 텍스트가 합성음으로 변환될 발음 표기의 생성이다. 두 영역 모두 입력된 한글 철자에 대해 발음 나는 데로 표기된 발음열을 출력하는 것이다. 이러한 발음 표기 변환은 입력된 한글 자소 문맥과 형태소 태그 정보에 따라 해당 음운 변동 규칙을 적용하여 올바른 발음열을 생성한다.

그러나 실제 입력되는 문장에는 제어문자, 특수기호, 알파벳, 아라비아 숫자, 영어 단어, 알파벳 및 숫자가 혼용된 약어, 기호 단위 명사 등이 포함되어 있다. 게다가 아라비아 숫자의 경우 문맥에 따라 숫자를 읽는 방식이 달라지기 때문에 발음열 생성기 내부에서 이 모든 현상을 수용하여 처리하는 것은 주된 목적 수행을 위한 선행 작업이 상대적으로 크게 되는 문제가 발생된다.

이러한 문제를 해결하기 위해 이 논문에서는 시스템 생성에 맞는 입력 형식으로 비한글 문자를 한글로 변환하는 작업을 수행하였다. 일반적으로 어절 내의 문맥 정보만으로는 예측 가능한 결과가 많을 뿐만 아니라, 정확한 변환 결과를 얻기 힘들기 때문에 발음열 생성기에서 사용되는 형태소 태그 정보를 사용함으로써 좀 더 정확한 결과를 얻도록 텍스트 전처리 시스템을 구성하였다.

2. 관련연구

주로 음성합성 시스템을 위한 전처리 과정 및 문서 분석기에 대한 연구들이 많이 수행되어져 왔다. [1]의 연구에서는 음성학적 전처리 과정으로 문자열 정형화부, 문장 구조 추출부, 음운변동 처리부로 세분화 하였으며, [4]의 연구에서는 전처리 모듈, 형태소 태거, 발음 표기 변환 모듈, 구문 분석기 등으로 구성하였다.

이 과정에서 생성되어지는 발음열 표기 결과는 음성 인식에서도 기본적으로 같은 처리 과정을 거쳐 생성되어지며 공통으로 사용되어진다.

기존 연구에서는 형태음운론적 분석에 기반하여 문자열을 자동으로 발음열로 변환하는 한국어 자동 발음열 생성 시스템을 구축하였다[2]. 한국어의 특성상 음소열로 변환하기 위해서는 형태소 분석 및 태깅 작업이 수행되어야 올바른 발음열을 유도할 수 있다. 이 시스템에서는 형태소 분석을 선행한 후, 한국어에서 빈번하게 발생하는 음운변화 현상의 분석을 통해 정의된 음소변동 규칙과 변이음 규칙을 다단계로 적용하여 형태소, 어절, 언절 또는 문장 등의 다양한 형태의 입력에 대해 가능한 모든 발음열을 생성하였다.

본 논문에서 제안한 텍스트 전처리 시스템은 앞서 작성된 발음열 생성 시스템의 올바른 동작을 수행하기 위한 입력 형식을 작성하기 위한 작업으로 입력 문서에서 사용된 비한글 문자들을 모두 한글로 바꾸는 작업을 수행하였다.

3. 텍스트 전처리기

앞서 설명한 바와 같이 발음열 자동 생성기의 역할은 입력된 한글 철자에 대해 발음 나는 데로 표기된 음소열로 출력하는 것이다. 그러나 실제 입력되는 문장에는 비한글 문자인 특수기호 및 알파벳, 그리고 숫자 등이 포함 되어 있으며, 문맥 정보에 따라 읽는 방식이 달라지기 때문에 이에 대한 고려사항도 필요하다. 따라서 시스템 생성에 맞는 입력 형식으로 변환하는 작업을 수행하기 위해 다음과 같은 단계를 거쳐 전처리기를 수행하였다.

크게 1) 특수 문자 및 2바이트 코드 변환 루틴, 2) 태그 정보를 위한 형태소 분석 루틴, 3) 기호 단위명사 변환 루틴, 4) 외국 어 및 비한글 고유명사 변환 루틴, 5) 숫자 처리 변환 루틴, 6) 예외 발음사전 탐색 및 음운 변동 규칙 처리를 위한 재접속 규칙 루틴 등으로 구성되어진다.

3.1 특수 문자 및 2byte 코드 변환

문장에서 제거 가능한 코드문자는 제거하고, 「」「」와 같은 괄호문자와 ℃, m²와 같은 기호 단위, 2바이트 로마자 숫자 Ⅱ, Ⅲ 등은 아라비아 숫자로, 그 외 2바이트 코드는 해당 1바이트 아스키 코드로 변환작업을 수행한다.

3.2 형태소 태깅

기존에 구축한 전처리 시스템은 앞뒤 문맥 정보만을 보고 변환 작업을 수행하도록 구축하여 형태소 분석 과정을 거치지 않았으나, 앞뒤에 고려되는 가지수가 너무 많아지는 단점을 지니고 있다. 발음열 생성 시스템 내부에서 형태소 분석 정보를 사용하므로 전단계에서 형태소 태깅 작업을 수행하게 된다. 여기서 수행된 형태소 태그를 제약 정보로 사용하여 전처리기에서 좀 더 정확한 분석 결과를 얻고자 하였다.

일반적으로 문자기반의 형태소 분석기의 경우, 원형을 복원하는 루틴이 포함되어 용언의 불규칙 처리나 생략된 조사와 축약된 형태까지 복원하게 된다. 그러나 음성 인식이나 합성에서는 소리값을 유지하여야 하기 때문에 문자기반의 형태소 태거[3]를 수정하여 원형 복원을 하지 않도록 태깅을 수행하였고, 태그 기호는 표준 품사 태그 표준안[6]에 따랐다.

3.3 기호 및 단위 명사 변환

su(단위기호), nbu(단위성 의존명사)로 태깅된 경우, 알파벳 기반 단위명사는 발화되는 한글 단위명사로 변환한다. 예를 들어 기호 'kg'은 '킬로그램'으로 변환 작업을 수행한다. 일반적으로 대문자, 소문자, 대소문자 혼용에 관계없이 변환 작업을 수행하였다.

그러나, 'M 과 m', 'G와 g'와 같은 일부 단위명사의 경우 대소문자 구분에 따라 서로 다른 발성 방식을 갖는 경우가 발생한다. 실생활에서 혼용이 되어 쓰이기는 하지만, 입력 방식을 고정하여 [표 1]에서와 같이 대부분 사용되는 발성 방식으로 변환 작업을 수행하였다.

단위기호	변환	단위기호	변환
%	퍼센트	bps	비피에스
cc	씨씨	db	데시벨
km	킬로미터	ha	헥타아르
kw	킬로와트	ppm	피피엠
kg	킬로그램	Mbit	메가비트
kv	킬로볼트	Mbyte	메가바이트
mg	밀리그램	M	메가
mm	밀리미터	m	미터
cm	센치미터	g	그램
t	톤	G	기가

[표 1] 기호 변환 예제

기타 발화되지 않는 구분 기호로 어절 경계에 변화가 없는 경우, 기호와 태그만 제거하고 좌우의 단어를 결합하여 하나의 어절로 구성하였다. sp(pause symbol)의 경우는 쉼표에 해당되는데, 일반적으로 쉼표 바로 뒤는 띄어쓰기가 실현되나, 대용량 텍스트 처리시 띄붙오류가 많이 생기므로 전처리시 sp 태그를 제거하고 어절 경계를 주었다. 마지막으로 sd(dash symbol)는 이음표로, 일반적인 '-' 기호의 경우는 sp와 같은 처리를 하였고, '~'의 경우는 숫자와 연결시 '에서'로 변환하고 어절 구분을 하였다.

3.4 외국어 및 고유명사 변환

특정 날짜를 표기하는 숫자 고유명사와 알파벳과 숫자가 섞인 고유명사 처리 문제로 대부분 외래어나 외국어, 제품명과 같은 고유명사에서 많이 발생한다.

기본적으로 코퍼스 내에서 발생되는 변환 리스트를 map 파일 형식으로 작성하여 변환 루틴을 생성하였다. 기본적으로 알파벳과 숫자는 한 문자 단위로 일대일 매칭하고, 그 외의 경우는 [표 2]와 같이 예외 처리로 변환하였다.

특히 영단어의 경우는 작성된 리스트에 대해 변환 작업을 수행하지만, 앞으로 공개된 영어사전을 사용하여 변환하는 루틴에 대한 추가 작업이 필요하다.

분류	입력	태그	변환
특정 날자	3.1 12.12	고유명사 고유명사	삼일 십이십이
숫자·영문	펜티엄III 3COM MP3	일반명사 고유명사 일반명사	펜티엄쓰리 쓰리콤 엠피쓰리
영문 약어	FIFA	고유명사	피파
영단어	Internet Center Solution	일반명사 or 외국어	인터넷 센터 솔루션

[표 2] 고유명사 변환 예제

3.5 숫자처리 및 변환

아라비아 숫자의 경우는 뒤에 오는 단위명사에 따라 '일이' 식이나 '한두' 방식으로 발화하게 되므로 단위명사 분류별로 처리하였다. '한두' 방식으로 읽게 되는 단위명사 목록으로는 대표적으로 [표 3]과 같으며 100 단위 이하의 경우에만 적용하였다.

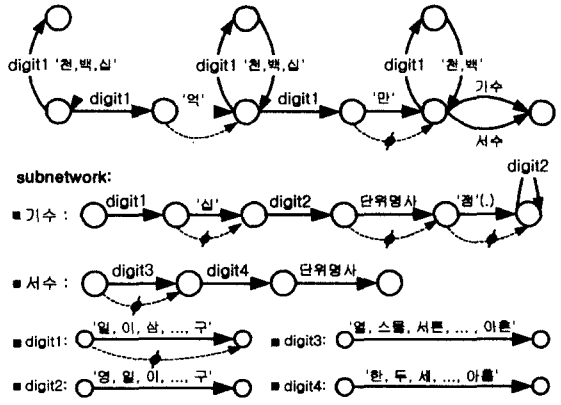
단위 기호 명사 외에 '바늘, 집' 등과 같은 보통명사 앞에서도 숫자가 '한두' 방식으로 발성되는 경우가 발생하므로 텍스트 분석을 통하여 문맥에 따른 변화도 고려하였다.

분류된 단위명사의 경우 화자의 습관에 따라 '일이' 식으로 혼용되어 읽는 경우도 발생하여 음성과 일치하지 않는 오류가 전처리기에서 발생할 수도 있다. 오류가 발생하는 경우에는 입력 자체를 한글 수사로 변환하여 혼동을 줄이도록 하였다. 또한 수사 및 단위발성에 대한 기존 연구[5] 결과를 참고하여 대상으로 삼은 코퍼스에서 발생하는 이외의 단위 명사 발성 방식에 대해서도 모델링 하였다.

방식	단위명사 목록 예제 (일부)
한두	가마, 가지, 개, 곡, 골, 곳, 군데, 건, 경기, 나라, 달, 돌, 마리, 명, 바퀴, 발, 배, 번째, 병, 부분, 살, 상자, 수, 시, 시간, 쌍, 알, 잔, 장, 차례, 척, 칸, 토막, 포기, 품목, 팅
일이	도, 년, 미터, 원, 월, 일, 퍼센트
혼합	동, 단계, 등급, 순위, 점, 차원, 차선, 층

[표 3] 후위 단위 명사에 따른 숫자 읽기 방식

그 외에 소수점 및 일반적인 숫자 읽기 단위를 처리하도록 구성하였고, 날짜, 시간, 점수, 수식, 전화번호 등의 패턴을 분석하여 해당 발음으로 변환하도록 하였다. 전체적인 숫자 처리 방식에 대한 오토마타는 [그림 1]과 같다.



[그림 1] 숫자 처리 루틴 오토마타

3.6 예외 발음 처리를 위한 루틴

예외 발음사전 탐색 및 발음 변환 현상이 규칙에 의하여 처리되지 않는 경우를 위해 형태소 재접속 결합규칙에 따라 자동 재태깅 작업을 수행하였다.

복합어의 경우, '솜이불[솜니불]'과 같이 음절 경계에서 발생하는 'ㄴ-첨가' 규칙을 적용하기 위해 복합어 리스트를 기록한 복합 명사 사전을 구축하고, 사전 검색을 통해 '솜이불/cn2'과 같이 복합 명사 태그와 경계 위치를 갖도록 구성하였다.

일부 합성어 및 파생어 특히 '~적, ~성, ~권' 등과 같은 접미사가 붙은 한자어의 경우 비규칙적인 경음화 현상이 발생한다. 이는 국어학적으로도 도출 가능한 규칙이 정해져 있지 않기 때문에 대부분 예외처리를 할 수 밖에 없다. 예외사전을 이용한 예외처리를 수행하기 위해 해당 접미사의 경우 일반명사로 결합하는 규칙을 추가하였다.

4. 실험 및 평가

텍스트 전처리기 구성 및 성능을 평가하기 위하여 신문기사 및 서적으로부터 추출한 낭독체 45,000문장을 사용하였다. 이 문장들은 음향모델을 생성하기 위해 균형잡힌 트라이폰 목록을 포함하도록 구성되어졌으며, 한 화자당 100문장씩 발성하였다. 낭독을 위한 스크립트는 숫자 및 기호 등이 많이 포함되어 있다. 특수 문자 및 2byte 코드 변환 단계를 거쳐 형태소 분석을 수행한 뒤, 형태소 분석기의 오류를 줄이기 위해 수작업으로 수정 작업을 거친 후 전처리기를 수행하였다. 다음 [표 4]는 실제 전처리기를 거쳐 변환된 대표적인 결과이다.

입력	전처리 과정을 거친 후
24/nnn+명/nbu+의/jcm	스물/nnc+네/nnc+명/nbu+의/jcm
2002/nnn+년/nbu	이천/nnc+이/nnc+년/nbu
10.2/nnn+%/su	십/nnc+쩜/xsn+이/nnc+퍼센트/nbu
40/nnn+kg/nbu+당/xsn	사십/nnn+킬로그램/nbu+당/xsn
3.1/nq+운동/ncn	삼일/nq+운동/ncn
NASA/f+는/jxc	나사/nq+는/jxc
CF/f+요정/ncn	씨에프/ncn+요정/ncn

[표 4] 텍스트 전처리 수행 결과

텍스트 문장 분석 결과, '-적'으로 끝나는 명사의 종류가 599개, '-성'의 경우 107개, '-권'의 경우 89개의 단어가 발생하였다. 이 중 경음화 현상이 반영되지 않은 단어 중 예외적으로 경음화 현상이 나타나는 단어의 수를 카운트한 결과, '-적'의 경우 481개 중 9개가, '-성'의 경우 80개, '-권'의 경우 74개의 단어가 경음화 현상이 발생하였다.

예외적으로 경음화 현상이 발생하는 경우에 한해 현재 구축된 예외사전에 없는 단어들만 추가하였다. 분석 결과 '-성, -권' 접미사가 붙는 경우 대부분 경음화 현상이 일어나므로 태그와 문맥 정보 뿐만 아니라 음절 정보까지 반영하여 경음화 처리 문제를 해결하도록 할 예정이다.

대상으로 삼은 코퍼스에서 발생된 예외처리 고유명사는 약 40여 개이고, 복합명사 리스트는 총 67개이다.

발음열 생성기에 적용된 음운 규칙으로 적용되지 않는 현상들을 반영하기 위해 구축된 예외 사전의 크기는 약 10K 정도로 불필요한 검색을 최소화하여 구축하였다.

전체적으로 테스트 결과 발생하는 오류의 경우, 분석을 통하여 수정 및 안정화 작업을 수행하였다.

5. 결론 및 향후 연구과제

본 논문에서는 한국어 자동 발음열 생성기를 수행하기 위한 정확한 입력을 생성하기 위해 텍스트 전처리 과정을 세분화하여 시스템을 구축하고 안정화 작업을 수행하였다.

이로서 생성기 내부에서 처리할 수 없었던 비한글 문자 변환 작업과 문맥 정보 및 형태소 태그 정보를 활용하여 발성 방식의 모호성을 최대한으로 줄임으로써 생성기를 유용하게 이용할 수 있도록 하였다.

앞으로 형태소 태그 정보가 정확하지 않은 경우에도 어절 내의 단어 및 문맥 정보만으로도 올바른 변환이 가능하도록 전처리기 시스템을 재구성하고 있다.

6. 참고 문헌

[1] 강용범, 김진영, "무제한 음성합성시스템을 위한 전처리과정", 제 11 회 음성통신 및 신호처리 워크샵 논문집, pp.334-337, 1994.

[2] 이경님, 전재훈, 정민화, "한국어 연속음성 인식을 위한 발음열 자동 생성", 한국음향학회지, 제 20 권, 제 2 호, pp. 35-43, 2001.

[3] 이상호, 서정연, 오영환, "KTS: 미등록어를 고려한 한국어 품사 태깅 시스템", 제 12 회 음성통신 및 신호처리 워크샵 논문집, pp. 195-199, 1995.

[4] 이상호, 오영환, 서정연, "한국어 문서 음성 변환 시스템을 위한 문서 분석기", 한국음향학회지 제 15 권 3 호, pp. 50-59, 1996.

[5] 이영직, "방송 뉴스 전사문장의 수사 및 단위의 발성 방식", 제 17회 음성통신 및 신호처리 워크샵 논문집, pp. 285-288, 2000.

[6] 최기선, 남영준, 김진규, 한영균, 박석문, 김진수, 이춘택, 김덕봉, 김재훈, 최경진, "한국어 정보 베이스를 위한 형태·통사 태그 표준에 관한 연구" 한국인지과학회 논문지, pp.43-61, 1996.