대어휘 음성인식 언어모델링을 위한 텍스트 코퍼스 구축

김정세*, 윤애선**, 권혁철***
*한국전자통신연구원, 네트워크연구소, 음성정보연구센타,
**부산대학교 인지과학협동과정, ***부산대학교 전자전기정보컴퓨터공학부

Text Corpus Construction for Language Model

Jeong-se Kim*, Aesun Yoon**, Hyuk-Chul Kwon***
*Speech Technology Research Center, Network Laboratory, ETRI
**Pusan National University, Department of cognitive science
***Pusan National University, School of Electrical & Computer Engineering
E-mail : jungskim @etri.re.kr, asyoon@pusan.ac.kr, hckwon@pusan.ac.kr

요 약

본 논문은 음성정보연구센터에서 추진하고 있는 대용량 텍스트 코퍼스 구축에 관하여 기술한다. 총 3 년 동안 약 3 억~5 억 어절 수집을 목표로 하고 있으며, 주 목적은 대어휘 음성인식용 언어모델링을 위한 통계정보 추출용으로 활용할 예정이다. 1 차년도인 2002 년에 수집할 텍스트의 양은 약 6 천만 어절로 주요 일간지와 방송뉴스를 대상으로 하고 있다. 이 중 2 천만 어절은 띄어쓰기, 철자오류 수정 등을 수동으로 수행하고, 나머지 어절은 자동 검증 툴을 사용하여 오류를 수정하고자 한다. 본 논문에서는 공동 이용 가능한 텍스트 코퍼스의 구축 방안과 구축 시의 고려해야 할 사항들을 제시하고자 한다.

1. 서론

지금까지 한국어의 음성 및 텍스트 데이터베이스는 각 연구자가 필요에 따라 만들어 왔다. 음성 연구가 진보되어 감에 따라 처리 가능한 데이터의 수가 많아져 가고, 따라서 준비해야 할 데이터의 양도 대폭적으로 증가되어왔다.

여기에 대어휘 음성인식의 연구가 활발히 진행됨에 따라 언어모델링을 위한 대용량의 텍스트 데이터의 필요성이 절실하다.

본 논문에서는 대용량 텍스트 코퍼스 구축 방안과 구축 시 고려해야 할 전사규칙, 띄어쓰기 규칙, 숫자 처리, 외국어 처리, 기호 처리 등을 일간지와 방송뉴스에서 나올 수 있는 다양한 표현을 정리하고 이들을 처리하기 위한 방안을 제시하고자 한다.

2. 텍스트 코퍼스 수집

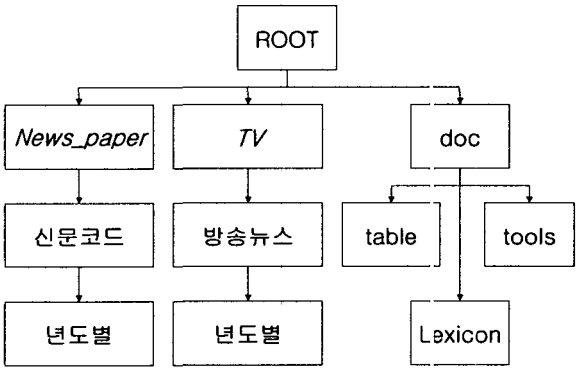
2.1. 텍스트 코퍼스 수집 내역

1 차년도인 2002 년에 수집할 텍스트의 양은 약 6 천만 어절로 2001 년도 일간신문 기사와 방송뉴스를 수집한다. 이 중 2 천만 어절은 수동 수정을 하고, 나머지는 자동 툴을 이용하여 수정한다.

2.2. 디렉토리의 구조

(그림 1)은 각 데이터가 들어갈 디렉토리의 구조다. 여기에는 크게 신문, 방송뉴스, 그리고 각 정보들을 입력할 DOC 디렉토리를 가지고 있다.

DOC 디렉토리의 TABLE 은 업계정보, 텍스트 작업 참여자와 각종 코드 리스트들을 기술한다. TOOLS 디렉토리는 DB 들을 작성하면서 생성된 TOOL 들과 그에 대한 설명서들을 기술한다. LEXICON 디렉토리는 축약이나 고유명사, 외국어 리스트들을 기술한다.



(그림 1) 디렉토리 구조

2.3. 파일내의 데이터 구조 및 전사규칙

본 연구에서는 데이터 저장 및 재사용성을 극대화 하기 위해 XML 에 기반한 정보 구조를 정의한다. 각 요소(element) 정보는 풍부한 언어구조 확장성을 고려하여 요소 기반 접근법을 활용하였다. [2]

[표 2] 문자 저장을 위한 XML 태그 정의

태그명	설명
Section	하나의 기사를 의미하는 root node. Id 속성을 부여하여 다른 기사와 구분한다.
Para	하나의 문단(Paragraph) 단위. 만약 하나의 문단이 부제인 경우 sub 속성값을 true 로 한다. sub 속성을 넣지 않거나, false 로 하면 일반 문장으로 처리된다.
P	문장 (Sentence)
Vc	한글 이외의 정보(한자, 숫자, 영문자)를 음성으로 발음하는 경우. Src 는 원래 텍스트 문장으로 기록된 정보를 저장하며, case 는 소리내어 읽는 한글 정보를 저장한다.
Src	만약 src 의 정보가 고유명사인 경우 prop 속성값을 true 로 지정한다.
Case	Vc 태그에서 한 가지 이상으로 발음할 수 있는 경우, 해당하는 범주를 넣어준다.
Title	신문사명
author	기자명
Date	기사 작성일자
Topic	분류체계 (신문사별로 상이할 수 있다.)
Head	기사의 제목

본 정의에서 VC element 를 별도로 정의한 이유는 문서 정보에 기록되어 있는 사항과 발음 정보를 모두 남겨둘 경우 향후 정보 처리 및 검색, 자료의 재가공 방법에 있어 높은 효율성을 가져올 수 있기 때문이다. 아래는 위의 태그에 대한 예문이다.

```
- <section id="274748">
    <title>중앙일보</title>
    <author>김준술 기자</author>
    <date>2000.09.04</date>
    <topic>정보통신</topic>
    <head>미국 …</head>
  - <body>
    - <para sub="true">
        <p>…</p>
      </para>
    - <para>
        …
      - <p>
          전력을 사용하면서 최대
        - <vc>
            <src>1 GHz</src>
            <case>일 기가헤르쯔</case>
```

```
    </vc>
    의 처리 속도를 낸다.
  </p>
...
```

3. 세부규격

기본 원칙은 표준문법에 맞게 표기하나, 관례로 사용하는 것들은 수용한다. 그리고 기호, 숫자, 단위, 외래어, 외국어 등은 " //" 를 이용하여 표준 한글 표기를 덧붙여 준다.

3.1. 띄어쓰기

띄어쓰기는 우리말 맞춤법을 따르되 맞춤법에서 해석이 모호한 경우나 신문 따위에서 관례로 쓰는 형태는 그대로 수용한다. 즉, 신문이나 언론사의 자료에는 띄어쓰기가 틀린 경우가 많다. 주로 지면편집과 일반적인 관례를 따르다 보니 띄어쓰기오류가 발생하지만, 일부는 맞춤법을 잘못 이해하여 생겼다. 예를 들어 '3 만 5 천여명'('3 만 5 천여 명'이 바름)처럼 쓰는 것은 오히려 글을 읽기 쉽게 할 수도 있고, 그냥 관례로 이렇게 써왔으므로 이렇게 쓰는 것으로 판단한다. 또 '에베레스트산'은 '에베레스트 산'으로 써야 하지만 '한라산'은 '산'을 붙여써야 하는 맞춤법 규칙 (외래어에서는 '산', '강' 따위를 고유명사와 띄어써야 한다)을 따르지 않고 관례로 이렇게 쓰는 것 같다. 다른 예로는 '김 씨' '김 사장'를 '김씨', '김사장'으로 붙여쓰는 오류도 자주 보인다. 이런 신문사나 언론사의 띄어쓰기 오류는 허용한다. 또한 원문을 최대한 살린다는 의미에서 많은 띄어쓰기 규칙들이 관례에 따라 완화된다.

3.2. 복합어 처리

복합어는 원시 데이터에 띄어쓰지 않은 형태로 나타난 복합어 중 띄어 쓸 수 있는 곳에 " _" 를 삽입하여 이것이 복합어임을 구분한다. 예를 들어 '전국대회'는 '전국_대회'로 표시하며, 원문에 '전국 대회'로 표기된 것은 그대로 둔다. 그러나 다음의 경우, 띄어 쓸 수 있는 경우라도 처리를 하지 않는다

(1) 복합어가 사전에 등재된 경우는 분리하지 않는다.

(2) 접두어 또는 접미어가 붙어 형성된 복합어

(3) 사이 시옷이 들어 있는 복합어

(4) '1 음절짜리' 단어로 구성된 복합어로, 띄어쓰기를 하는 경우 중의성이 증가되어 한 단어로 처리되는 것이 더욱 타당한 용어

A. 스포츠, 바둑 등 용어: 타수차, 타수, 동타, 호투, 결승골, 볼투입, 골밑슛, 골차, 골세례, 골문, 눈가림수, 주경기장

(5) 수 부류사 '개' + "1 음절짜리 명사'

A. 20 개사, 20 개교, 20 개국

(6) 회사명, 브랜드명과 같은 고유명사는 원문을 그대로 유지한다.

3.3. 숫자처리

수를 적을 때는 만 단위로 띄어쓴다. 원래 십진법에 따라 띄어쓰던 것을 '만' 단위로 개정하여 '만, 억, 조' 및 '경, 해, 자' 단위로 띄어쓰는 것이다. 십진법에 의하여 띄어쓰면, 합리적이긴 하지만 너무 작게 갈라놓아 오히려 의미 파악이 어려워진다. 일반적으로 신문에서는 이 규칙을 완화하여 쓴다. 따라서 되도록 이 부분은 신문의 관례를 따른다. 다음의 경우가 그 예이다.

(1) 한글로 표시된 단위 등은 변환 범위에 포함하지 않는다.

A. 5 분 => {{오}}//{{5}}분

B. 라이브 1 관 => 라이브{{일}}//{{1}}관

(2) 한자어, 영문, 기호로 표시된 단위 등은 따로

변환한다.

A. 3m => {{상}}//{{3}}{{미터}}//{{m}}

(3) 단, 고유명사나 연어(collocation)으로 분류될 수 있는 것은 동일 변환 범위이 포함한다.

A. 샤넬 No.5 => 샤넬 {{넘버 파이브}}//{{No.5}}

(4) 숫자의 읽는 방법은 번호독식과 봉독식이 있는데 두 가지 다 읽어도 무방한 것은 둘 다 표현한다.

A. 20 개 => {{스무, 이십}}//{{20}}개

(5) 표기예외규칙

A. 40~50 대 => {{사오십 대}}//{{40~50 대}}

3.4 외국어 처리

소스데이터에 나온 모든 외국어는 한글로 변환하되, 표기 기준은 국립국어 연구원의 '외래어 한글 표기법'을 기준으로 한다. 단, 아래와 같은 경우는 변환이 굳이 필요 없거나 변환의 어려움으로 인해 변환하지 않는다.

(1) '한글 음사' 다음 괄호 안에 적은 영문자

A. 빅맥(big-mac), 팅(ⓣing)

(2) 한국어 다음 의미의 명확성을 위해 적은 영문자

A. 생명윤리(bioethics)

(3) 2 단어 이상의 영문 고유 영사 영어 구문 이상의 단위

(4) 이메일 주소, 이메일 ID, 웹 url, 컴퓨터 경로명, 파일명

그리고, 외국어, 수, 기호 등이 결합된 고유명사는 전체를 한 단위로 처리한다.

3.5 기호 처리

기본적으로 발성에 포함되지 않는 기호들은 삭제한다. 그러나 아래와 같은 것은 문맥상 별표가 발성되어야 하기 때문에 표기한다.

원제 Back to the Future3. 1990 년작. ★★★☆

=> {{별 세 개}}//{{★★★☆}}

" -" 은 다시, 에서, 부터 등으로 다양하게 발성이 되나 이것은 문맥에 따라서 결정한다.

4. 결론

공동 이용 가능한 텍스트 코퍼스의 구축 방안과 구축 시의 고려해야 할 사항들을 살펴보았다. 구축된 텍스트 코퍼스는 대학이나 관련 연구기관에서 활용할 수 있도록 배포할 예정이며, 지속적으로 수정 보완 및 확장해 나갈 예정이다.

Acknowledgement

본 연구는 정보통신부 출연 " 음성정보처리기반기술" 과제의 일환으로 수행되었습니다.

참고문헌

[1] 이용주, 김봉완 외, " 국어공학센터의 한국어 음성 DB 구축계획" , 제 12 회, 음성통신 및 신호처리 워크샵 논문집, pp 276-279, 1995.6.

[2] Laurant, St., XML primer 3rd edition, John willy and sons, 2001.