한국어 음성 합성을 위한 '이음표'의 문자 전사

정영임[O] 정휘웅 윤애선 권혁철
부산대학교 인지과학 협동과정, 전자전기정보컴퓨터공학부, 한국어 정보처리 연구실
{acorn[O], hwjeong, asyoon, hckwon}@pusan.ac.kr

Transcribing Some Text Symbols for Improving Korean TTS System

Youngim Jung[0], Hwiwoong Jeong, Aesun Yoon, Hyukchul Kwon
Korean Language Processing Lab.
Dept. of Cognitive Science, Dept. of Computer Science,
Pusan National University

요 약

최근 신문기사의 음성 서비스 등 음성합성 연구가 실용단계로 접어 들고 있으나, 텍스트의 비-문자 처리에는 오류율이 높다. 본 연구는 신문 텍스트에 나타나는 비-문자 중 중의성이 높은 이음표의 문자화 유형을 6가지로 제시하고, 이음표를 포함한 어절의 패턴화된 구조 및 좌우 문맥 정보를 이용하여 이음표의 문자화 규칙을 알아본다. 제시된 이음표의 문자화 규칙과 이음표가 좌우 문맥 숫자의 문자화에 미치는 영향에 따른 숫자 읽기 방식을 포함하여 이음표 포함된 텍스트의 문자화 전사 시스템을 구현하였고, 2년치 J신문 텍스트를 코퍼스로 하여 이음표 문자화 시스템의 정확도를 측정하였다. 아울러 실험 결과에서 오류로 나타난 유형을 분석하여 정확도를 향상시킬 수 있는 방안을 제시하였다.

1. 서 론

정보기기의 소형화 및 경량화에 따라 언어정보 중 음성정보의 필요성이 크게 증대되고 있다. 이전에는 증권정보, 전화번호 안내, 게임 스코어 안내 등 제한적 분야에서 사용되던 TTS(Text-To-Speech) 시스템도, 최근에는 신문과 같이 다양한 전문 분야를 포함하는 문자정보를 실시간에 음성합성하는 데 사용되고 있다.[1] 실용화된 TTS시스템이 개발되려면 음운규칙, 음성규칙, 억양단위, 휴지부 등에 관한 정보는 물론, 정보전달의 효율성을 위해 텍스트에 사용된 기호 및 부호의 전처리에 관한 정보도 필요하다. 음성공학을 중심으로 전자에 관한 연구는 활발히 이루어져 왔으나, 후자에 관해서는 최근에야 비로소 주목을 받고 있다.

문자정보에 비해, 아라비안 숫자, 문장부호나 기호는, 텍스트의 가독성(readability)을 높여주고, 정보 전달력을 향상할 뿐 아니라, 공간 효율적이다. 하지만, 역으로 이들의 문자 정보화는 중의성이라는 문제점을 안고 있다. 예를 들어, 현대 한국어에는 어원·수의 종류·수 단위의 유무 등에 따라 20가지 아라비아 수 읽기 방식이 존재하며[1, 2, 3], 이는 좌우 문맥에 따라 다른 분포를 갖는다[3]. 문장부호인 이음표('-', '~')도 형태적 동인성에 기인하여 '에서, 대, 의/에, 빼기, 마이너스' 및 영형태(zero morpheme)인 6가지 방식으로 문자화된다. 또한, 좌우 문맥이 아라비아 수인 경우, 수의 읽기에도 영향을 미친다.

본 연구에서는 신문 텍스트에서 규칙에 기반한 이음표의 문자화 전처리를 위한 자동 전사 시스템을 구현한다. 이를 위해 사용된 말뭉치는 1개 신문 2년치(2000년 1월~2001년 12월) 기사에서 이음표를 포함한 어절 37,000여 건을 추출하여 분석하였고, 동일 신문 2달치와 4달치의 기사에서 이음표를 포함한 어절 418건과 569건을 추출해 평가용 말뭉치로 사용하였다.

2장에서 선행 연구의 문제점 및 기존 TTS 시스템에서 나타나는 이음표 전사 오류를 살펴본다. 3장에서 이음표의 읽기에 따라 6가지 규칙을 설정하고 이음표에 의해 영향을 받는 숫자 읽기를 고려하여, 4장에서는 이음표의 자동 전사 시스템을 구현한다. 5장에서는 구현된 자동 전사 시스템을 실험하여 그 정확도를 알아보고, 오류 유형을 밝히고, 6장에서는 시스템 성능 향상 방안과 후속 과제를 제시한다.

1) 국내 2개 일간지에서 TTS시스템을 이용한 음성기사가 제공되고 있고 상용 음성합성 시스템으로는 대표적인 시스템이 2개가 있는데, 이메일 음성합성 등에 이용되고 있다.

2. 선행연구 및 문제점

문장부호 및 기호에 관해서는 맞춤법 관련 연구에서는 이음표를 '줄표(—)'와 '붙임표(-)' 및 '물결표(~)'로 구분하고, 줄표는 '이미 말한 내용을 부연하거나 보충하기 위해 사용되며', 붙임표는 '접사, 어미, 복합어 결합 관계를 표시하는 데 사용한다.' 고 밝히는 의미적 구분만을 하고 있으며 '~'는 '내지'라는 뜻으로 쓰이거나 '어떤 말의 앞, 뒤로 들어갈 말 대신'에 사용된다고 밝히고 있다[4]. 하지만, 실제 언어 자료에서는 줄표와 붙임표의 이러한 형태적, 의미적 구분을 하지 않는다. 이 밖에 이음표는 범위 표지, 구분자, 수학 기호 등으로 광범위하게 사용되며 문맥에 따라 문자화되는 형태도 다양하다. 그러나 전산 언어학이나 음성 공학 분야에서는 아직 이음표의 다양한 문자화에 대한 연구가 시도되지 않았고, 이에 따라 현재 제공되고 있는 TTS시스템의 정확성이 매우 떨어진다.

실제 TTS시스템을 이용하여 음성기사 서비스를 제공하고 있는 D신문, M신문과 V음성합성 시스템은 이음표를 정확히 줄표와 붙임표로 구분하지 않고 있으며, 이음표의 읽기를 <표 1>과 같이 매우 간략한 규칙으로 처리하고 있으나 (1)~(10)과 같은 오류가 나타난다.

<표 1 기존 TTS시스템의 이음표 읽기 형태>

	'-'의 읽기	'~'의 읽기
D신문	'대', 영형태	'에서'
M신문	'마이너스', 영형태	'에서', '틸드', 영형태
V시스템	'에', '마이너스', 영형태	'에서', 영형태

(1) **-0.24%** [*영점 이사/마이너스 영점 이사](D)
(2) **T-50** [*티 대 오십/*티 마이너스 오십/티 오십](D), (V)
(3) 미그-19기 [*미그 마이너스 십구/미그 십구](M), (V)
(4) **011-9XX** [*공일일 대 구엑스엑스/공일일 구엑스엑스](D)
(5) **2000-2001** [*이공공공(에) 이공공일/이천 이천일](D), (V)
(6) **14-16일** [*십사 마이너스 십육/*일사에 일육 일/십사에서 십육](M), (V)
(7) 신용등급 **A-** [*에이/에이 마이너스](M), (V)
(8) **3~4개** [*삼에서 사/서너](D), (V)
(9) 3.15~3.50달러 [*삼 점 일오 틸드 삼 점 오공/삼 점 일오에서 삼점 오공](M)
(10) 3억~5억 원[*삼억오억/삼억에서 오억](M), (V)

기존 TTS 시스템에서는 이음표의 다양한 읽기를 모두 반영

하지 않아 D신문의 경우, 이음표 '-'가 주가나 매매가의 변동률이 음수임을 나타낼 때는 '-'를 [마이너스]로 읽어야 하는데 오류 예문 (1)과 같이 읽으며, M신문, V시스템 역시 범위수를 나타내는 '-'를 '에서'로 처리하지 못한다(6). D신문과 V시스템은 숫자와 함께 나타나는 이음표 '~'에 대해서 '에서'로 일괄처리하여 (8)에서처럼 부정수로 나타나는 수 표현을 정확하게 표현하지 못한다.

또한, 기존 TTS 시스템에서는 이음표 문자화 규칙의 비정교성으로 인한 오류도 많이 발생한다. D신문에서는 이음표 '-'가 각 종 미사일·무기류명이나 병균명과 같이 고유명사로 나타날 때는 이음표가 문자와 숫자 간 구분자로 사용되는데, (2)에서와 같이 이를 '대'로 읽는 오류를 보이며, M신문과 V시스템은 '마이너스'로 처리하는 오류를 범한다(2), (3). D신문은 전화번호에 사용되는 구분자를 잘못 인식하였고(4), D신문과 V시스템은 대회명에 붙는 연도표시 숫자를 전화번호로 인식하여 (5)와 같이 잘못 읽는다. M신문과 V시스템은 '-'를 마이너스로 읽는 규칙에 대해서도 오류 예문 (6), (7)에서 보듯이 이음표의 문자화 규칙이 정교하지 않다. 이음표 '~' 역시 '에서'뿐만 아니라 (9), (10)에서 보는 것과 같이 '틸드', '영형태'으로 일정한 규칙이나 일관성이 없이 읽고 있다.

3. 이음표의 문자화

2장에서 살펴본 바와 같이, 기존 기사 음성 서비스는 문맥에 따른 이음표의 다양한 문자화가 정교하게 이루어지지 못하였다. 3장에서는 이음표가 포함된 어절의 패턴과 좌우 문맥 정보에 따라 이음표가 문자화되는 규칙을 제시하고 역으로 이음표가 좌우 문맥 숫자의 문자화에 미치는 영향을 알아본다.

3.1 패턴과 좌우 문맥에 따른 이음표의 문자화 구분

본 연구에서는 이음표가 포함된 어절의 패턴과 좌우 문맥에 기반하여 이음표가 문자화되는 규칙을 이음표의 읽기에 따라 아래의 6가지로 구분할 수 있다.[2)]

<표 2 이음표 문자화 규칙 1>

LAC2	LAC1	PATTERN	RAC1	RAC2	읽기
		CST-NA CST-ALP_NA ALP-NA ALP-ALP_NA			_(null)
문의/전화/팩스/예약/신청/연락처/안내/참가		NA-NA NA-NA-NA NA)NA-NA (NA-NA-NA)	번		_(null)
우편번호		NA-NA			_(null)
계좌/은행		NA-NA{-NA-NA}			_(null)
		N-N-N N-N-N-N	전술/시스템/전형/포메이션/형태		_(null)

<표 3 이음표 문자화 규칙 2>

LAC2	LAC1	PATTERN	RAC1	RAC2	읽기
전반/후반/스코어/점수/전적	을	NA-NA	으로/로/의	이기-/지-/완승-/완패-/	_대_
팀이름		NA-NA	팀이름		_대_

<표 4 이음표 문자화 규칙 3>

LAC2	LAC1	PATTERN	RAC1	RAC2	읽기
		NA-NA NA~NA	분류사		에서_
지수		NA-NA NA~NA	선/사이		에서_
	인명/왕호/대회명	(NA-NA) (NA~NA)			에서_
		CST~CST			에서_

<표 5 이음표 문자화 규칙 4>

LAC2	LAC1	PATTERN	RAC1	RAC2	읽기
		NA-1 NA-2 NA-3	번	좌석버스/마을버스/노선/버스	의_
주소표현	산	NA-NA	번지	주소표현	의_
리포트/취재/기획/시리즈	[제목]/<제목>/"제목"	NA-NA CST-NA			의_

<표 6 이음표 문자화 규칙 5>

LAC2	LAC1	PATTERN	RAC1	RAC2	읽기
번동률/성장률/오름세/하락률/예상치		-NA (-NA	% %p %포인트	)	_마이너스_
기업명		(-NA	화폐도량형	)	_마이너스_

<표 7 이음표 문자화 규칙 6>

LAC2	LAC1	PATTERN	RAC1	RAC2	읽기
	수학 연산자	NA-NA	수학연산자/는/은		_빼기_

3.2 이음표가 좌우 문맥 숫자의 문자화에 미치는 영향

이음표의 문자화가 좌우 문맥의 영향을 받을 뿐 아니라, 다음과 같이 좌우 문맥 숫자의 읽기에 역으로 영향을 주기도 한다.

이음표가 9이하의 두 수 사이에서 부정수를 표현할 때, '3~4[서너]', '4-5[네댓]'과 같이 고정화된 표현으로 읽으므로 이음표 앞 뒤 숫자의 문자화에 영향을 미친다. 10이상의 부정수에서는 숫자 다음에 오는 분류사가 고유어 수관형사를 취하는 것이라도 이음표 앞 뒤의 숫자를 '20-30명[이삼십 명]', '30-40그루[삼사십 그루]'와 같이 한자어 수사로 발음한다.

이음표가 두 수 사이의 범위를 나타내어 '에서'로 문자화될 때 이음표 앞의 숫자는 명사형 수사로, 이음표 뒤의 숫자는 수관형사형으로 문자화된다.(예: 1-3개[하나에서 세 개])

또한, 전화번호, 계좌번호, 우편번호 및 스포츠의 공격 전형을 나타내는 표현과 같이 패턴화된 표현에서 '-'가 나타날 때는 각 숫자를 수 단위를 붙이지 않고 한자어로 읽는다.(예: 051-510-3463[공오일 오일공 삼사육삼])

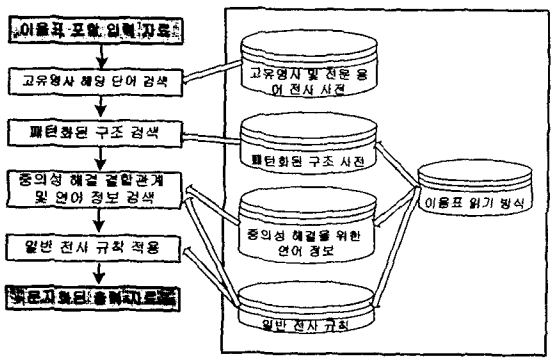
4. 이음표의 자동 문자화를 이용한 자동 전사 시스템

3.1절에서 제시한 이음표의 패턴화된 구조 86개와 좌우 문맥 정보 84개를 바탕으로 이음표의 자동 문자화 규칙 및 휴리스틱스를 처리할 수 있는 알고리즘을 만들고, 3.2절에서 밝힌 이음표에 의해 영향을 받는 좌우 문맥 숫자의 문자화 규칙을 포함하여[3)] <그림 1>과 같이 자동 전사 시스템을 구현하였다.

2) 표에 사용된 약자의 의미는 아래와 같다. LAC: 좌연접어, RAC: 우연접어, PATTERN: 이음표가 포함된 패턴화된 표현, CST:한글 문자열, NA: 숫자열, ALP: 영문자열, N: 단일 숫자. 표 안에 사용된 {}는 중괄호 안 요소가 선택적임을 나타내고, 나머지 소괄호, 각괄호, 대괄호 및 따옴표는 패턴에 포함된 구성 요소이다.

3) 이번 연구에 포함된 이음표에 의해 영향을 받는 숫자 읽기는 2-3개[두세 개], 1-2년[일이 년]과 같이 중의성 없이 숫자의 읽기가 결정되는 경우 숫자의 문자화가 맞게 제시될 수 있다. 중의성이 있는 분류사와 함께 오는 경우

이음표를 포함한 텍스트가 입력되면 ① 고유명사 및 전문 용어 사전을 통해 인명, 대회명, 팀이름 등에 고유명사 정보를 준다. 이 외 '콕사키 바이러스-B6'와 같이 패턴화되기 어려운 구조를 가진 고유명사를 전사한다. 다음으로 ② 패턴화된 구조 사전을 검색해 86개의 구조와 매칭되는 구조가 처리된다. 다음 단계로 ③ 좌우 문맥 정보를 이용하여 중의성있는 구조를 검색하여 ④ 일반 전사 규칙을 적용하면 이음표를 한글로 전사한 결과값을 갖는다.



<그림 1. 이음표의 자동 전사 시스템>

5. 실험 및 평가

실험은 동일 신문의 2달치(2002.2~3), 4달치(2002.4~7) 기사에서 이음표를 포함한 418건, 569건의 어절을 자동 전사 시스템으로 자동 처리한 결과와 동일한 코퍼스를 직접 분석한 결과를 비교 분석하여 자동 처리 결과의 정확도의 오류율을 밝히고 여기서 나타난 오류 유형을 아래와 같이 분석하였다.

<표 8. 이음표의 자동 전사 시스템 실험 결과>

	평가 말뭉치1		평가 말뭉치2	
	어절 수	비율(%)	어절 수	비율(%)
정확도	395	94.5	547	96.13
오류율	23	5.5	22	3.87
계	418	100	569	100

<표 9. 이음표 자동 전사 시스템의 오류 유형>

정확도(숫자는 문자화 규칙)			어절 수
오류유형	자동처리 결과	직접 분석 결과	
①	1	4	3
②	미처리	1	7
③	미처리	2	4
④	웹 주소 및 이메일 주소 표현		28
⑤	문맥 부족		1
⑥	원문 오류		2
계			45

오류 유형 ①은 'CST-CST' 구조에서 문자열이 지명이나 시간을 나타내는 말은 아니나 이음표가 '에서'의 의미를 가지는 유형으로 의미 분석이 필요한 오류이다.(예: '보행-주행신호 바뀔때 사고 운전자에 형사', '역으로 일본어-한국어로 번역되는 과정에서') ②는 전화번호가 N-N의 구조이며 미등록된 고유명사인 기관명이나 상점명이 좌연접어로 오므로 전화번호 구조로 인식하지 못하는 오류이다.(예: 영명식당(472-4027)) 또한 분석되지 않은 기호의 사용으로 전화번호 구조로 인식하지 못하는 경우나(예: 본부 신고접수 담당자는 "최근에는 '011-9××××')

웹주소와 붙어서 나오는 전화번호 인식의 오류이다.(예: 한화콘도(588-2299/http://www.hanwha.co.kr)) ③은 '-N'의 구조로 다양한 우연접어가 붙어서 나오며 패턴화된 구조나 좌우 문맥 규칙을 결정하기 어려운 경우이다. (예: '수도인 말레는 한국-3시간', '반대로 하늘에서 떨어질 때 -2에서') ④는 웹주소나 이메일 주소에 포함된 '~'나 '-' 읽기 문제로 발생한 오류이다. (예: '김동호 기자<e-news@joongang.co.kr>', 'http://dong-gu.incheon.kr')

⑤는 추출한 어절을 통해 충분한 문맥 정보를 얻을 수 없는 경우(대우차 희망센터 김경운 팀장 -1700명)이며, ⑥은 원문 텍스트 자체에 글자가 깨지거나 표기를 잘못 한 경우 생기는 오류이다.(예: '02-7744~')

6. 결론 및 향후 연구

이상 본 연구에서는 신문 텍스트에서 이음표('-', '~')의 자동 문자화 방식 및 이음표가 좌우 문맥 숫자에 미치는 영향을 바탕으로 규칙 기반하여 이음표의 자동 전사 시스템을 구현해 보았다. 실제 신문 텍스트 코퍼스를 시스템을 통해 자동 처리한 결과 정확도가 95.5%였다. 시스템의 오류율이 약 4.5%로 높은 편이지만, 실험 결과를 통해 나타난 오류 유형을 분석한 결과, 전화번호와 함께 나타나는 웹 주소 표현의 분리를 통한 전처리나 웹 주소나 이메일 주소에 나타나는 이음표의 문자화 규칙 설정을 통해 약 2.9%의 오류를 교정할 수 있다. 따라서 원문 오류, 문맥 부족에 의한 오류 및 고유 명사를 포함하더라도 실제 오류율은 1.6%로 아주 낮은 수치일 것으로 예상된다. 단, 패턴화된 구조나 좌우 문맥 정보로 이음표의 문자화 규칙을 만들기 어려운 경우 개별 규칙 적용의 정확도에 대한 통계적 정보를 통합한 방식으로 해결해야하는데 이러한 통합적 방식에 대한 연구는 향후 과제로 남긴다. 나아가 신문 텍스트에서 이음표 외 기호로 중의성을 가지는 '.', ':' 및 '/' 기호 등에 관한 연구도 계속되어 아라비안 숫자와 여러 가지 기호를 하나의 모듈로 구성하여 음성합성 전처리에 이용하면 현 TTS 시스템의 성능을 크게 향상시킬 수 있을 것이라 예상된다. 이에 대한 연구 또한 향후 과제로 진행 중에 있다.

Acknowledgement

본 논문은 과학기술부의 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적(scalable) 우리말분석기술의 개발 (M10203000028-02J0000-01510))의 지원을 받아 이루어졌음을 밝힌다.

참고 문헌

[1] 채완(1983), "국어 수사 및 수량사구의 유형적 고찰", 『어학연구』 제 19권 제1호, pp.19~34

[2] 유재원(1999), "자연어 처리를 위한 수사의 하위 범주 분류", 제9회 한글 및 한국어 정보처리 학술대회 학술발표 논문집, pp.136~142

[3] 정영임, 김정세, 김상훈, 이영직, 윤애선(2002). "현대 한국어에서 아라비안 숫자의 읽기 규칙 연구", 제14회 한글 및 한국어 정보처리 학술대회, pp.16~23

[4] 이희승, 안병희, 『새로 고친 한글 맞춤법 강의』 신구문화사. 2001.

(1-2대[일이 대/한두 대])나 중의성이 있는 다른 기호에 의해 숫자 읽기가 달라지는 경우(1.1~1.5[일 점 일에서 일 점 오/일 월 일 일에서 일 월 오 일]) 의 문자화는 임의로 [일이 대], [일 점 일에서 일 점 오]와 같이 처리하였다. 중의성이 있는 분류사나 기호 등과 함께 읽는 숫자의 문자화 및 이음표 외 기호의 문자화를 고려하여 아라비안 숫자와 기호를 하나의 모듈로 구성하는 연구는 다음 과제로 남긴다.