한국어 텍스트에 사용된 이음표의 자동 전사

윤애선, 권혁철*†
부산대학교

Aesun Yoon and Hyuk-Chul Kwon. 2003. Automatic Transcription of the Union Symbols in Korean Texts. *Language and Information 7.1*, 23–39. In this paper, we have proposed Auto-TUS, an automatic transcription module of three union symbols—hyphen, dash and tilde ('-', '—', '~')—using their linguistic contexts. Few previous studies have discussed the problems of ambiguities in transcribing symbols into Korean alphabetic letters. We have classified six different reading formulae of the union symbols, analyzed the left and right contexts of the symbols, and investigated selection rules and distributions between the symbols and their contexts. Based on these linguistic features, 86 stereotyped patterns, 78 rules and 8 heuristics determining the types of reading formulae are suggested for Auto-TUS. This module works modularly in three steps. The pilot test was conducted with three test suites, which contains respectively 418, 987 and 1,014 clusters of words containing a union symbol. Encouraging results of 97.36%, 98.48%, 96.55% accuracy were obtained for three test suites. Our next phases are to develop a guessing routine for unknown contexts of the union symbols by using statistical information; to refine the proper nouns and terminology detecting module; and to apply Auto-TUS on a larger scale. (**Pusan National University**)

Key words: TTS, 현대한국어(Modern Korean), 줄표(dash), 붙임표(hyphen), 물결표(tilde)

1. 서 론

빠르게 진화하는 정보화 사회에서 사용되는 언어정보 중 문자보다는 음성의 필요성이 크게 증대되고 있다. 이는 문자보다는 음성이 더 오랫동안 인간 언어의 속성을 구성해 왔다는 역사성보다는, 정보의 입출력 단계에서 나타나는 음성의

* 609-735 부산 금정구 장전동 산 30 부산대학교 인문대학 불어불문학과 / 공과대학 전자전기정보컴퓨터학부, E-mail: {asyoon, hckwon}@pusan.ac.kr

† 본 논문은 과학기술부의 국가지정연구실사업의 지원을 받아 이루어졌음을 밝힌다 (과제명: 언어 중심의 지능적 정보처리를 위한 단계적(scalable) 우리말 분석기술의 개발 (M10203000028-02J0000-01510)). 또한, 초고의 오류와 미처 생각하지 못했던 예를 상세히 지적하여 준 세 분 심사위원에게 감사의 뜻을 전한다.

효율성에서 그 원인을 찾을 수 있을 것이다.[1] 문자를 매개로 한 입출력은 노트북, PDA처럼 입력을 위한 자판이나 펜과 출력을 위한 화면이 필요하므로 정보 저장 장치의 크기를 제한하고 따라서 이동성이 떨어진다.[2] 따라서 음성정보를 문자정보만큼 효과적으로 저장, 검색, 가공하기 위한 음성공학 연구가 90년대부터 활발히 진행되고 있다. 음성공학(speech processing)은 언어학의 제 분야, 음향학, 전자공학을 포괄하는 학제간 연구분야로, 음성정보의 처리 방향에 따라 생성에 관여하는 '음성합성(TTS: Text-To-Speech)'과 분석에 필요한 '음성인식(Speech Recognition)'으로 크게 구분할 수 있다. 두 분야는 밀접한 상호 관계를 맺고 있으나, 중의성이나 잡음도가 높은 음성인식보다는 음소나 변이체를 분절·추출하여 이를 조합하는 음성합성의 난이도가 좀 더 낮으며, 더욱 빠르게 실용화 단계에 이르렀다.

이전에는 증권정보, 전화번호 안내, 게임 스코어 안내 등 비교적 제한된 전문 분야에만 사용되던 음성합성 시스템도, 최근에는 신문 기사와 같이 다양한 전문 분야를 포함하는 문자정보를 실시간에 음성으로 변환하는 데 사용되고 있다.[3] 현재 개발된 음성합성 시스템의 음절조합 성능과 개별 어휘의 생성 정확도는 매우 높다. 하지만, 억양 단위와 휴지부(pause)의 설정, 복합어 끊어 읽기, 동철이의어(homograph)나 다의어(polysemy)에 따른 음운변화의 처리는 아직 미숙하다. 또한 이러한 시스템은 텍스트에 사용되는 다양한 기호나 부호와 같은 중의성을 가진 비-문자(non-literal) 정보를 정확한 음성정보로 변환하지 못한다. 따라서 보다 개선된 시스템의 개발을 위해서는 정교한 언어분석이 선행되어야 한다.[4] 이 중 휴지부 및 복합어 분석은 자연언어처리(Natural Language Processing)의 성과를, 동철이의어나 다의어에 따른 음운변화는 음운론과 음성학의 성과를 부분적으로 이용할 수 있다. 하지만 텍스트에 사용되는 중의적 기호나 부호의 읽기 규칙에 관해서는 이론 언어학, 자연언어처리, 음성공학 어느 분야에서도 체계적으로 연구된 바가 없다.[5]

문자정보에 비해, 아라비안 숫자, 문장부호나 기호 등 비-문자 기호의 사용은 텍스트의 가독성(readability)을 높여주므로, 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, 이는 좌우 문맥에 따라 다른 분포를 갖는다. 본 논문의 3장에서 자세히 소개 되는 것처럼, 문장부호인 이음표('-', '—', '~')도 형태적 동인성에 기인하여 좌우 문맥에 따라 '에서, 대, 의/에, 빼기,

[1] 매체의 효율성은 정보의 속성과 사용 목적에 따라 다르다. 예를 들어 편지글이나 이야기 등은 음성으로 더욱 빠르게 전달될 수 있으나, 복잡한 표나 대차대조표 등은 시각을 사용하는 문자 텍스트가 음성에 비해 훨씬 효율적이며 정확도도 높다.

[2] 휴대폰의 경우, 음성을 주로 이용하지만 정보처리의 관점에서 볼 때 가공할 수 있는 분절된 언어 정보 형태로 입출력이 이루어지는 것이 아니고, '음'이라는 물리적인 뭉치 자료가 전달될 따름이다.

[3] 국내 2개 일간지에서 TTS시스템을 이용한 일반 음성기사가 제공되고 있고, 1개 일간지에서는 시각장애자용 점자신문의 형태로 서비스된다. 상용 음성합성 시스템으로는 대표적인 시스템이 4개가 있는데, 이메일이나 웹 문서 음성합성 등에 이용되고 있다. 2003년 4월 현재 3개 데모 버전이 자사의 웹페이지에 제공되고 있다.

[4] 음성합성의 핵심 기술은 크게 ① 형태소 분석, 미등록어 처리, 품사 태깅 및 음성 단위의 최적 구문 분석을 하는 언어처리부, ② 음의 피치, 지속시간, 에너지를 예측할 수 있는 운율생성부, ③ 음운환경을 고려한 단위음 데이터베이스 구축과 이를 기반으로 명료한 음성을 합성할 수 있는 신호합성부를 포함한다. 지금까지 음성공학 연구는 상대적으로 ②, ③에 치중하였다.

[5] 아라비안 숫자의 읽기 규칙 및 좌우 문맥을 고려한 규칙의 선택에 관해서는 김상준(1986, 1992), 유재원(1997, 1999), 윤애선 외(2003), 이영직(2000), 정영임 외(2002), 채완(1983)에서 부분적으로 연구된 바 있다.

마이너스' 및 영형태(zero morpheme)인 6가지 방식으로 문자화된다.[6] 또한, 좌우 문맥이 아라비아 수인 경우, 수의 읽기에도 영향을 미친다.[7] 하지만 국내에서 개발하여 사용하고 있는 기존의 TTS시스템에서 이음표 읽기의 정확도가 낮아, 정보 전달력이 약화된다.

본 연구에서는 음성합성의 전처리(preprocessing) 단계의 일부로 이음표를 문자로 자동전사[8]하기 위한 모듈 구현을 목표로 한다.[9] 이를 위해 2장에서는 기존 음성합성 시스템에 나타난 오류를 분석함으로써 선행 연구의 문제점을 알아보고, 3장에서는 본 연구에 사용된 분석 말뭉치 및 평가 말뭉치의 구성과 연구 대상의 특성을 검토하고, 연구 범위 및 방법론을 살펴본다. 4장에서는 문맥의 분석을 통해 이음표의 문자화 규칙과 휴리스틱스를 설정하여, 이음표의 자동 전사 모듈을 구현한다. 5장에서는 평가 말뭉치를 이용하여 구현된 자동 전사 모듈을 실험하여 그 정확도를 알아보고, 오류 유형을 밝히고, 6장에서는 모듈의 성능 향상 방안과 후속 과제를 제시한다.

2. 선행연구 및 문제점

문장부호나 기호에 관한 선행 연구나 관련 기술로는 의미·문체적 관점에서 그 사용법을 규정한 것이 대부분을 차지한다. 한글 맞춤법 관련 규정에서 이음표를 '줄표(—)'와 '붙임표(-)' 및 '물결표(~)'로 구분하고, 다음과 같은 사용법을 제시한다.[10]

종류	용법
줄표	① 문장 중간에 앞의 내용에 대해 부연하는 말이 끼어들 때 쓴다. ② 앞의 말을 정정 또는 변명하는 말이 이어질 때 쓴다.
붙임표	① 사전, 논문 등에서 합성어를 나타낼 적에, 또는 접사나 어미임을 나타낼 적에 쓴다. ② 외래어와 고유어 또는 한자어가 결합되는 경우에 쓴다.
물결표	① "내지"라는 뜻에 쓴다. ② 어떤 말의 앞이나 뒤에 들어갈 말 대신 쓴다.

[표 1] 한글맞춤법 규정의 이음표 용법

한글 맞춤법에서 선의 길이에 의해 구분되는 줄표와 붙임표는 실제 언어 자료에서는 형태적 변별성이 유지되지 않으므로,[11] 이에 따른 의미적 구분을 하지

[6] '문자화'는 비-문자 기호를 한글로 변환하는 것을 지칭하고, '읽기'는 문자정보 사이에 나타날 수 있는 음성적 변환까지 포함한다는 의미에서 후자가 좀 더 광범위한 용어다. 하지만 문자정보의 음성적 변환은 본 연구의 대상에서 벗어나므로 '문자화'와 '읽기'를 구별하지 않고 사용한다.

[7] 쌍점(:), 온점(.), 빗금(/)의 문자화도 매우 중의적이나, 이에 대한 분석은 향후 연구를 기약한다.

[8] '전사(transcription)'은 ① 음을 문자로 변환하거나, ② 한 언어나 기호를 다른 언어로 바꾸어 적으며 그 음성정보를 유지하는 것을 지칭하며 좁은 의미의 '문자화 변환(transliteration)'와 동의 관계를 갖는다. (http://www.wikipedia.org/wiki/ Transcription+(linguistics)) 본 연구에서 '전사'는 후자의 정의를 따른다.

[9] 본 연구의 최종 목표는 모든 비-문자 기호의 문자화를 대상으로 하며, 본 고의 내용은 이 최종 목표의 일부분을 구성한다.

[10] 국어 문장 부호 사용법은 1988년 개정된 한글 맞춤법에 부록으로 규정된 것을 근간으로 한다.

[11] 유니코드에서는 줄표와 붙임표는 '2010~2015'와 '8208~8213'으로 표현되며 숫자가 높아질수록 줄의 길이가 길어진다. 하지만 사용자 인터페이스 관점에서 볼 때 붙임표는 101 표준 한글키보드에서 한번에 입력할 수 있으나, 줄표의 경우 특수 기호 넣기 기능을 이용해야 하므로 적어도

않는다. 이런 사용 규정 외에 실제로 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.

[표 2]의 예문 (1)~(9)는 기존 TTS시스템에서 찾아 볼 수 있는 읽기 오류이다.[12]

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

[표 2] 기존 음성합성 시스템의 이음표 읽기 오류

기존 TTS 시스템에서는 [표 3]처럼 이음표의 읽기를 2~3개 정도로만 단순 변환하므로 다양한 읽기 방식을 모두 제시하지 못한다. 시스템에 따라서 (1)이나 (6)과 같이 음수를 표시하는 '마이너스'로 읽지 못하거나, (5)에서처럼 범위의 출발점을 나타내는 '에서'로 변환하지 못한다.

출처	'-'의 읽기 방식	'~'의 읽기 방식
D	'대', 영형태	'에서'
M	'마이너스', 영형태	'에서', '틸드', 영형태
V	'에', '마이너스', 영형태	'에서', 영형태

[표 3] 기존 TTS시스템의 이음표 읽기 방식

또한, 기존 TTS 시스템의 이음표 읽기 규칙이 정교하지 않아 발생하는 오류도 많다. 이음표 '-'가 각 종 미사일·무기류명이나 병균명과 같이 고유명사로 나타날 때는 이음표가 문자와 숫자 간 구분자로 사용되는데, (2)에서와 같이 이를 '대'로 읽거나, (3)과 같이 '마이너스'로 처리하는 오류를 범한다. (4)처럼 대회명에 붙는 연도표시 숫자를 전화번호로 인식한다. (8), (5), (9)에서 볼 수 있듯이

3번의 키나 마우스 포인팅이 필요하다. 따라서 대부분의 디지털텍스트에서 이 두 부호의 구분을 하지 않고 사용한다.

[12] 예문 (1)~(9)의 출처인 D, M, V는 각각 음성기사 서비스를 제공하고 있는 동아일보, 매일경제와 데모프로그램 중 실시간 음성합성이 가능한 VoiceText의 약자이다. 본 연구의 기존 시스템 평가는 2003년 3월과 4월에 이루어졌다.

범위수의 읽기에 오류를 보일 뿐 아니라, 이 세 예와 부정수를 나타내는 (7)과 같은 경우를 구분하지 못하고 동시에 분류사와 아라비안 숫자와의 선택 제약을 반영하지 못한다.[13]

3. 연구 내용과 범위

이 장에서는 이음표의 문자 전사화 연구를 위한 말뭉치 구성과 이음표 사용의 특성을 살펴봄으로써 연구 내용과 범위를 기술한다.

3.1 말뭉치의 구성

2장에서 살펴본 바와 같이 이음표의 문자 전사화나 읽기 규칙에 관한 선행 연구가 거의 없으므로, 가능한한 다양한 경우가 포함될 수 있는 말뭉치를 구성하는 것이 매우 중요하다. 이를 위해 본 연구에서는 전자화된 텍스트 중 다양한 전문 분야가 포함되고 가독성을 높이기 위해 기호 사용이 빈번한 신문 기사를 이용하여 연구대상 말뭉치를 구성하였다.[14]

이를 위해 학습말뭉치는 1개 신문 2년치(2000년 1월~2001년 12월) 모든 분야의 기사로부터 이음표를 사용한 어절 19,767개에서 실험말뭉치3을 제외한 18,756개로 구성하였다. 실험말뭉치는 분석대상 말뭉치와 다른 시기(2002.1.1~ 2002.12.31)에 발행된 동일 신문에서 크기와 표본추출 방식에 따라 3개를 구성하였다. 실험말뭉치1과 2는 분석대상 말뭉치의 2%와 5%에 해당되는 비율을 무작위 추출(random sampling) 방식으로 선택하였다. 이에 따라 14일과 37일에 발행된 모든 분야의 기사에서 이음표 사용 어절 418개와 987개로 구성하였다. 무작위 추출 방식으로 구성된 말뭉치는 모든 경우의 이음표 읽기 규칙을 포함하지 못할 수 있다. 따라서 실험 말뭉치3은 분석대상 말뭉치에 나타난 이음표 읽기 규칙의 출현 빈도에 비례하여 전체 중 대표성을 갖는 약 5% 비율에 해당하는 경우를 추출하여 1,014개 어절로 구성하였다.[15]

3.2 연구 대상의 특성

이 절에서는 이음표의 사용 환경과 용법, 문맥에 따른 다양한 읽기 방식을 살펴봄으로써 연구 대상의 특성을 알아본다.

3.2.1 이음표의 용법. 학습말뭉치에서 실제로 관찰할 수 있는 줄표 및 붙임표[16]와 물결표의 용법은 [표 1]의 한글맞춤법 규정에서 제시한 용법인 예문 (10-a~d)의 출현 빈도는 극히 낮다. 오히려 규정에 비해 예문 (11)처럼 매우 다양하게 사용

[13] 부정(不定)수는 범위수의 부분집합이며, '1-2, 4-5, 20-30, 500-600, 7000-8000' 등 '일, 십, 백, 천, 만, ...'과 같이 각 자릿수를 기준으로 인접 수를 나타내는 경우를 지칭한다. 범위수의 경우, 이음표는 '에서'로 전사되고, 이 부호 앞뒤의 아라비안 수의 읽기는 이음표가 사용되지 않은 단일 수사열의 읽기 규칙이 적용된다. 부정수와 범위수의 읽기 차이에 대해서는 본고 3.2.2를 참조하라.

[14] 말뭉치는 사용 목적에 따라 다르게 구성된다. 글말(written language)을 입말(spoken language)로 변환하는 TTS시스템의 정확성 향상을 위해 비-문자 기호의 읽기 규칙을 규명하기 위해, ① 글말의 특성을 가지며, ② 다양한 분야의 용어를 포함하며, ③ 현대 한국어의 특성을 잘 반영하는 신문기사는 본 연구의 목적에 가장 적합한 분석대상이라고 판단한다. 물론 신문 기사가 전자화되어 있으며, 누락된 날짜나 부분이 없다는 편의적인(opportunistic) 특성도 연구대상 말뭉치 선정에 영향을 미친 요소이다.

[15] 실험말뭉치3은 대표성을 가진 균형말뭉치(representative & balanced corpus)의 특성을 갖는다.

[16] 본 연구의 분석말뭉치에서 줄표와 붙임표는 구분없이 사용되었다. 따라서 본 고의 3장부터 줄표와 붙임표를 구분없이 사용한다.

구분	표본추출방식	텍스트 생성 시기	이음표 포함 어절수
학습말뭉치		2000.1.1-2001.12.31 638일 중 실험말뭉치3 제외	18,756
실험말뭉치1	무작위 추출	2002.1.1-2002.12.31 중 14일	418
실험말뭉치2	무작위 추출	2002.1.1-2002.12.31 중 37일	987
실험말뭉치3	대표적, 균형적 추출	2000.1.1-2001.12.31에서 학습말뭉치의 비율과 종류에 따라 선정된 5%	1,014

[표 4] 말뭉치 구성 및 특성

된다. 예를 들어 (11-a)은 인용 또는 출처를, (11-b, c)는 구(phrase)나 어휘 단위의 구분 또는 대조를 , (11-d)는 '빼기'로 읽어야 하며, (11-e, f, g, h)은 공간, 시간, 수의 구간을 나타내며, (11-i)은 시간적 순서를 나타낸다.

(10) (a) 만의 하나 -절대 그럴 리야 없겠지만- 의혹이 생긴다면

(b) ① 크로이츠펠트-야코프씨병 ② 역사가의 '에고(ego)-역사' ③윈-윈 전략 ④ http://www.chung-gu.pusan.kr

(c) 300~500여 명

(d) 중국은 웃고, 한국은 울고-

(11) (a) "한국 국가채무 우려 안 해." - 주한 미 상의회장

(b) 개각의 특징은 '거시경제팀은 유임-실무 집행부서는 물갈이'로 요약된다.

(c) ① 김도훈-양현정 콤비 ② 아시아-유럽 프레스 포럼 ③ 현대-북한 협상 ④ 김옥두-하순봉 영수회담 ⑤ 미-러 전략핵 감축협상

(d) 매출액에서 순이자비용(이자비용-이자수익)이 차지하는 비중

(e) ① 부곡-용산 간 화물열차 ② 구리~판교 간 고속도로,

(f) 금강~설악~두타~청옥산을 지나

(g) 패밀리 레스토랑 베니건스는 주중(월-금요일) 점심시간에

(h) 매물은 여러 차례에 걸쳐 수천 주-수만 주씩 쏟아져 나왔다.

(i) ① 군사령관의 서울시 방문계획 통보-연기-취소로 이어져,
② 퇴직금 누진제를 폐지하면 성과급제-연봉제로 이어지는 구조조정

이와 함께 예문 (12)-(14)에서 볼 수 있듯이 특정한 문맥에서 이음표가 사용되는 비율이 아주 높고 때로는 줄표 및 붙임표와 물결표 간 구분이 없다. (12-a~h)는 축구 전술, 전화번호, 통장번호, 주민등록번호, 주소, 버스번호, 고유명사, 웹주소와 같이 특정한 유형으로 분류할 수 있는 패턴화된 구조이고, (13-a~e)는 수사와 함께 사용되면서 문맥에 의해 문자 전사화가 다양하게 형성되는 경우, (14-a~c)는 구어의 장음화 현상을 반영하는 기호로 사용하는 다양한 경우를 보여준다.

(12) (a) 서독이 우승하면서 썼던 3-5-2 포메이션

(b) ① 문의는 전화 016-316-3494 ② 시환경보호과 031-828-2821~5

(c) 한빛은행 108-05-001401

(d) 700119-1068189

(e) 부산광역시 사하구 신평동 산 41-10 부산 대동고등학교

(f) 상암동과 종로구 평창동을 오가는 135-2번 노선

(g) 정부가 6-1광구 고래 구조

(h) CDMA2000-1x 기능을 갖춘 PDA

(i) sookmyung.ac.kr/~mtherapy

(13) (a) 동원증권이 -4.8%

(b) 월드컵 D-100일

(c) 대한항공과의 여자 단체전 결승전에서 3-1 역전승

(d) ① 0.5-0.75%포인트의 콜금리 인상 ② 1994~96년에 샀던 쏘나타II ③ 5~10년짜리 장기 국내 채권

(e) ① 2-3만 원대, ② 2~3백 명 정도

(14) (a) ① 위-잉 하는 모터소리 ② '삐~'하는 신호음

(b) ① '선영아 사랑해-.' ② "다음 역은 아우라지역입니다.~"

(c) ① 긴장을 전혀 늦추지 않는-다. ② 연분홍 치~마가

3.2.2 이음표의 읽기 방식. 예문 (11)-(14)에서 본 것처럼 용법이 다양한 것만큼 이음표의 읽기도 다양한 방식으로 구현된다. 이때, 읽기 방식에 어떤 화계(話階)를 기준으로 삼느냐는 것도 중요하게 고려할 대상이다. 본 연구의 말뭉치를 신문 기사로 구성한 만큼 한글맞춤법 표준어규정의 표준발음법이 제시하는 규범에 따르기로 한다.[17] 또한 패턴화된 (12), 구어의 장음을 전사한 (14)의 경우와 다

[17] 일상 구어체와 표준발음법에 근간한 표준방송어법의 규범 간에는 순화용어의 채택 여부나 분류사와의 선택관계 등의 차이가 있다. 예를 들어, 표준방송어법에서는 줄표나 붙임표를 영어 'dash'의 일본어 발음인 '다시'로 읽는 것과 같은 비-순화용어 사용이 허용되지 않는다. 분류사 '명, 건' 등과 같이 한국어 고유어로 읽는 수사를 일상 구어체에서는 한자어 수사로 읽는 것이 허용되나 표준방송어법에서는 허용되지 않는다.(김상준 1986, 1992; 박갑수 1996) 또한, 수의 크기가 커지면 일상 구어체에서는 한자어 수사의 사용이 자연스러우나,(유재원 1997, 1999; 이영직, 2000; 채완 1983) 표준방송어법에서는 수사와 분류사 간 결합 규칙의 적용을 받는다.(김상준 1986, 윤애선 외 2003, 정영임 외 2002)

른 예를 구분해야 한다. (12)에서도 (h)와 같은 고유명사를 제외하고는 각 패턴은 특정한 읽기 규칙을 갖는다.[18]

(12′) (a) 3-5-2 [삼ø오ø이][19]

(b) ① 016-316-3494 [공일육ø삼일륙ø삼사구사]
② 031-828-2821~5 [공삼일ø팔이팔{에/ø} 팔이일에서 오]

(c) 108-05-001401 [일공팔ø공오ø공공일사공일]

(d) 700119-1086154 [칠공공일일구ø일공팔륙일오사]

(e) 산 41-10 [사십일의 십]

(f) 135-2번 [백삼십오의 이]

(g) 6-1광구 [육의 일]

(i) sookmyung.ac.kr/~mtherapy [sookmyung.ac.kr/물결표mtherapy]

(11′)-(13′) 예의 이음표 읽기에서 볼 수 있듯이 줄표 및 붙임표와 물결표는 각각 '에, 에서, 마이너스, 대, 빼기'와 '에서, 물결표'로 전사되거나 영형태([ø])로 전사된다.

(11′) (d) 순이자비용(이자비용-[빼기]이자수익)이 차지하는 비중

(e) ① 부곡-[ø]용산 간 화물열차

(g) 주중(월-[에서]금요일) 점심시간에

(h) 수천 주-[에서] 수만 주씩 쏟아져 나왔다.

(13′) (a) 동원증권이 -[마이너스]4.8%

(b) 월드컵 D-[마이너스]100일

(c) 대한항공과의 여자 단체전 결승전에서 3-[대]1 역전승

(d) ① 0.5-[에서]0.75%포인트의 콜금리 인상
② 1994~[에서]96년에 샀던 쏘나타II

(e) ① 2-[ø]3만 원대, ② 2~[ø]3백 명 정도

예문 (14)와 같이 이음표가 장음을 나타내는 경우, 첫째, (14-c-①)처럼 자음으로 끝나면 해당 한 음절을 '초성+중성'과 '중성[20]+종성'으로 두 음절로 변환하고, 둘째, (14-a-①)처럼 복합 모음이면 두 번째 모음을 반복하고, 셋째, 나머지 경우처럼 이음표 앞의 음절이 단순 모음으로 끝나면 해당 단순 모음을 반복한다.

[18] 패턴화된 구조의 수사도 특정한 방식의 읽기 규칙이 적용된다. 아라비안 숫자의 읽기 방식과 패턴화된 구조에 적용되는 규칙에 관해서는 김상준(1986, 1992)를 참조하라.

[19] (12′) 이하 예문에서 문자로 전사된 내용은 [ ]안에, ( )은 수의적 요소를 표시하고, { }은 2개 이상의 전사 결과가 있는 경우를 나타낸다. 또한 { } 안의 /는 2개 이상의 문자화가 가능한 경우를 표시하며, 이음표를 읽지 않은 경우는 ø로 나타낸다.

[20] 이때 중성이 단순 모음이냐, 복합 모음이냐에 따라 둘째, 셋째 규칙이 적용된다.

(14′) (a) ① 위-[이]잉 하는 모터소리 ② '삐~[이]'하는 신호음

(b) ① '선영아 사랑해-[애].' ② "다음 역은 아우라지역입니다.~[아]"

(c) ① 긴장을 전혀 늦추지 않는-[느은]다. ② 연분홍 치~[이]마가

이음표의 읽기가 좌우 문맥의 영향을 받을 뿐 아니라, 다음과 같이 좌우 문맥이 아라비안 숫자일 때, 숫자의 읽기에 역으로 영향을 주기도 한다.

(15) (a) 4-7세 [사에서 **칠** 세]

(b) 4-7살 [**넷**에서 **일곱** 살]

(16) (a) 3-4세 [**삼사** 세/?삼에서 사 세]

(b) 3-4살 [**서너** 살/*세네 살/?셋에서 네 살]

(c) 30-40세 [**삼사십** 세/삼십에서 사십 세]

(d) 30-40살 [**삼사십** 살/*서른마흔 살/서른에서 마흔 살]

한자어 수관형사와 결합하는 분류사 '세(歲)'와 한국어 고유어 수관형사와 결합하는 분류사 '살'이 이음표와 함께 범위수 및 부정수와 함께 쓰이는 예문 (15), (16)을 살펴 보자. 한자어 수사처럼 수명사와 수관형사가 동일한 형태소를 갖는 (15-a)에서는 이음표의 사용이 수 읽기에 영향을 미치지 않는다. 하지만 한국어 고유어처럼 수명사와 수관형사의 형태가 다른 (15-b)와 같은 경우, 이음표 다음에 출현하는 수사열은 수관형사형을 유지하나 이음표 앞의 수사열은 분류사의 선택 규칙의 제약을 받으나 수명사형으로 변환된다. (16-a)처럼 한자어를 선택하는 부정수의 경우 이음표를 읽지 않는 것이 더욱 자연스러우나, 고유어와 결합하는 (16-b)는 수관형사 기본형의 결합인 [*세네]가 아닌 변이형 [서너]로 사용해야 한다. (16-c, d)에서처럼 십, 백, 천, 만 자리의 부정수 표현은 (15)와 같이 범위수처럼 읽기와 (16-a, b)와 같이 부정수처럼 읽기가 모두 가능하다.[21] 후자의 경우, 고유어를 선택하는 분류사라도 십 단위를 넘는 부정수는 한자어 수사와 결합한다.

4. 이음표의 자동전사

3장에서 살펴본 바와 같이 이음표의 읽기 방식은 다양하고 문맥에 의해 결정될 뿐 아니라, 역으로 이음표의 유무가 좌우 문맥에 출현하는 수사의 읽기에 영향을 미친다. 수사 읽기 방식은 다음과 같이 품사, 어원, 수의 형태, 기·서수, 기본형 여부에 따라 다음과 같이 구분한다. (윤애선 2003, 정영임 외 2002)

4장에서는 이음표가 포함된 어절의 패턴과 좌우 문맥 정보에 따라 이음표가 문자화되는 규칙과 휴리스틱스를 제시하고, 이를 바탕으로 이음표 자동전사 시스템을 설계한다.

[21] 십 단위 이상에서 부정수를 '2-30, 3-400'이라고 쓰지 않고 '20-30, 300-400'라고 표기하는 가장 큰 이유는 전자가 범위수와 부정수를 모두 표시할 수 있는 중의성을 갖기 때문이다.

어원	특성				약어	예
고유어	명사	기수	定수		Kca_n	*셋, 넷*
			不定수, 범위수		Kca_ni	*서넛, 셋에서 넷*
		서수			Kor_b	*셋째, 넷째*
	형용사	기수	定수	기본형	Kca_b	*세, 네*
				이형태	Kca_v	*서/석, 너/넉*
			不定수, 범위수		Kca_i	*서너*
		서수	定수		Kor_b	*셋째, 넷째*
			不定수, 범위수		Kor_i	*서너째*
영어					Brn	*쓰리, 포*
한자어	기수	整수	+자릿수	定수 기본형	Cca_b [+U]	*육, 십, 이백 십 오*
				定수 이형태	Cca_v [+U]	*유, 시*
				不定수, 범위수	Cca_i	*삼사, 삼사십*
			-자릿수		Cca_b [-U]	*이일오*
		小수	+자릿수		Cca_d [+U]	*이 점 일오*
			-자릿수		Cca_d [-U]	*이일오*
		分수			Cca_f	*이와 삼분의 일*
	서수	정수			Cor_b	*(제)삼*
		부정수, 범위수			Cor_i	*(제)삼사*

[표 5] 수사의 읽기 방식

4.1 이음표의 문자화를 위한 규칙과 휴리스틱스

이음표 사용 문맥은 크게 특정화된 형식을 갖는 패턴화된 구조와 그렇지 않은 것으로 구분할 수 있다. 전자를 검색하기 위해서는 패턴 구성 규칙을 추출해야 하고, 후자의 경우 일반화할 수 있는 규칙이나 휴리스틱스를 규정해야 한다.[22]

4.1.1 패턴 검색 규칙. 통장번호, 주민등록번호, 우편번호, 사업자등록번호, 웹주소, 운동경기 전술 등은 수사열의 개수가 고정되어 있거나, 특정한 형식을 갖는다. 하지만, 전화번호와 같은 경우는 필수적 요소가 정규화되어 있지 않고 임의적 요소의 수와 종류가 다양하여 검색할 패턴의 추출이 용이하지 않다. 전화번호 검색 루틴을 개발하기 위해 다음과 같은 단계의 분석이 필요하다.

1.전화번호를 구성할 수 있는 모든 요소를 추출한다.

2.각 요소의 읽기 방식을 규정한다.[23]

3.이 중 필수적 요소와 임의적 요소를 구분한다. 전화번호의 경우 필수적 요소는 국번과 사용자 번호로 구성된다.

[22] 실제 한국어 텍스트에서 띄어쓰기 규칙은 그리 잘 지켜지고 있지는 않다. 특히 전문용어, 신조어, 복합어의 사용 비율이 높은 신문의 경우 띄어쓰기 오류는 매우 높다. 또한 비-문자 기호의 사용은 사용자 주관이 개입된 경우가 많다. 따라서 4.1절에서 규칙이나 휴리스틱스를 규정하는 데 있어 이 점을 고려하여야 한다.

[23] 전화번호의 국번과 사용자 번호를 읽는 방식은 Cca_b [-U]를 사용하는 경우와 Cca_b [+U]를 사용하는 두 가지가 모두 허용된다. 전화번호 검색이 이루어지면 어느 경우라도 문자전사의 난이도는 같다. 본 고에서는 편의적으로 전자 방식을 구현한다.

4.전화번호와 유사한 수사열 표현과의 중의성을 해결할 수 있도록 각 요소의 특성을 분석한다. [표 6]처럼 수사열의 크기,[24] 특정 목록, 수의 형태, 다른 기호 사용의 가능성을 고려한다. 예를 들어 ⓓ의 '정수'라는 특성은 ⓖ의 추가번호 구분자로 온점을 쓴 경우인 3098-4125.6과 소수로 된 범위수 표현인 309.8-4125.6를 구분하는 데 중요한 특성이 된다.

5.하지만 3098-4125.6와 같은 경우가 범위수일 가능성은 여전히 높다. 이와 같이 전화번호 패턴으로 잘못 분류될 수 있는 경우로는 연도 범위 표현을 들 수 있다. 이러한 중의성을 낮추기 위해 ⓐ~ⓓ의 좌문맥이나 ⓕ~ⓖ의 우문맥의 특성을 추출할 수 있다. 예를 들어 ⓐ~ⓓ의 좌문맥에 '안내, 문의, 전화, 팩스, 신청, 참가, 연락처' 등이 포함되어 있거나 ⓕ~ⓖ의 우문맥에 '번'이 포함되면 해당 수사열을 전화번호 패턴으로 인식하고, ⓕ~ⓖ의 우문맥에 '번'을 제외한 다른 분류사사 출현하면 전화번호가 아닌 것으로 인식하여 ⓓ와 ⓕ를 Cca_b [+U]로 읽고 그 사이에 나타나는 이음표는 '에서'로 전사한다.[25]

	ⓐ	ⓑ	ⓒ	ⓓ	ⓔ	ⓕ	ⓖ
구분	사업자	국가	지역	국번	-	사용자 번호	추가번호
필수성	-	-	-	+	+	+	-
수사 읽기방식	Cca_b [-U]	Cca_b [-U]	Cca_b [-U]	Cca_b [-U]	에	Cca_b [-U]	Cca_b [-U]
수사열의 크기	3	1-3	2-4	3-4		4	1-4
특성	특정목록	특정목록	특정목록	특수국번의 목록			온점, 쉼표, 물결표, 붙임표로 시작될 가능성
	'0'으로 시작	여는 괄호로 시작 가능성	닫는 괄호로 끝날 가능성				

[표 6] 전화번호 패턴 검색 결정 조건

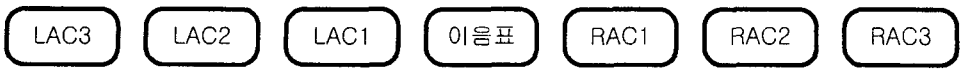
본 연구에서는 이와 같은 패턴화된 구조를 86개 추출하였다.

4.1.2 문맥 정보를 이용한 규칙과 휴리스틱스. 이음표를 포함하거나 인접한 어절에서 패턴화된 구조를 검색하지 못하는 경우가 더 빈번하다. 이 경우의 대부분은 이음표 좌우 문맥이 이음표 및 이음표 양 옆의 수사 읽기 방식을 결정한다. 이음표를 중심으로 띄어쓰기를 할 수 있는 좌우 문맥을 각각 LAC(Left-Associated Constituent)와 RAC(Right-Associated Constituent)로 구별하고, 인접 거리에 따라 LAC1, LAC2, LAC3 등으로 구분한다.[26]

[24] 567-8012을 5678-012와 같이 표시하는 것과 같이 국번과 사용자번호의 수사열 크기를 사용자 임의대로 분절하여 사용하는 비정규적 표현은 고려하지 않는다.

[25] 이 때 검색할 좌문맥과 우문맥은 검색에 소요되는 시간을 고려하여 ⓐ~ⓖ로 추정되는 수사열의 좌우 첫 어절만으로 제한한다.

[26] LAC1, 이음표, RAC1 간에는 띄어쓰기 유무를 고려하지 않는다. 문맥의 크기는 검색속도와 직



[그림 1] 이음표와 좌우 문맥의 구분

이상과 같이 구분된 문맥을 규칙 또는 휴리스틱스로 구성하기 위해서는 다음과 같은 분포 분석이 필요하다.[27]

1.LAC1과 RAC1이 동일한 성질을 갖는 자료(한글/한문, 영문자, 아라비안 숫자)인가 아닌가?[28]

2.LAC1과 RAC1에 모두 아라비안 숫자가 포함되어 있다면, RAC1 또는 RAC2에 자릿수 표현이나 분류사(classifier)가 출현하는가? 출현한다면 그 분류사는 어떤 숫자 읽기 방식을 선택하는가?[29]

3.LAC1과 RAC1에 모두 아라비안 숫자가 포함되어 있을 때, LAC1와 RAC1의 수는 부정수나 범위수에 속하는가?[30]

4.인접 3어절 내에 다른 이음표가 사용되는가?

5.LAC2, LAC3나 RAC2, RAC3에 특정 어휘나 어휘군이 사용되는가?

예를 들어 [표 7]은 LAC1과 RAC1에 아라비안 숫자(NA)가 포함되어 있으며, NA를 모두 Cca_b[+U]로 읽는 경우다. 가령 ①처럼 LAC2에 '전반/후반/스코어/점수/전적'이라는 어휘나 팀이름이 나타나거나, RAC1에 '(으)로/을/를/에게'이 포함되거나, RAC2에 '이기-/지-/완승-/완패-'라는 어휘나 팀이름이 출현하

접적인 관련이 있다. 본 연구에서 문맥의 크기를 좌우 3개 어절로 제한한 것은 검색 속도를 유지하면서도 문맥을 고려할 수 있을 것이라는 경험적 결정이며, 이음표 연구에서 문맥의 크기와 검색 효율성 및 속도와의 관계는 향후 연구를 통해 검증해야 할 대상이다.

27 여기에서는 분석 과정 중 LAC1과 RAC1에 아라비안 숫자가 모두 나타나는 경우의 일부만을 제시한다.

28 텍스트 정보가 동일 성격의 자료인지 판단하는 기준은 다음과 같이 사용하는 코드의 영역으로 쉽게 구분할 수 있다.

종류		유니코드 표현 영역
숫자		0030~0039
한글		AC00~D7AF
알파벳	대문자	0041~005A
	소문자	0061~007A
	서유럽어 확장	00C0~00FF
	동유럽어 확장	0100~024F
한자	CJK기본한자	4E00~9FFF
	CJK호환성한자	F900~FAFF

29 본 고의 예문 (15), (16)을 나타난 것과 같다. 분류사, 분류사 전치어, 분류사 후치어, 숫자 전치어와 숫자 읽기 방식의 선택에 관해서는 유재원(1999), 윤애선 외(2003), 정영임 외(2002)를 참조하라.

30 범위수는 LAC1<RAC1의 조건을 만족시켜야 한다. 부정수는 이 조건과 함께 'RAC1-LAC1=1 or 10 or 100 or 1000 or 10000'을 만족하여야 한다.

면 이음표를 '대'로 읽는다. 하지만 LAC1과 RAC1에 아라비안 숫자(NA)가 부정수가 아닌 범위수의 조건을 만족하고, ②와 같이 RAC1에 아라비안 숫자의 한자어 읽기 방식을 선택하는 분류사가 나타나거나, ③처럼 LAC1에 '지수/주가'라는 어휘가 나타나거나 RAC1에 '선' 또는 RAC2에 '사이'라는 어휘가 출현하거나, ④처럼 LAC1에 인명/왕호/대회명 등이 나타나면, 이음표를 '에서'로 읽는다.[31]

	LAC2	LAC1	이음표[읽기]	RAC1	RAC2
①	'전반/후반/스코어/점수/전적'	NA	-/~ [대]	NA+'(으)로/을/를/에게'	'이기-/지-/완승-/완패-'
	팀이름	NA			팀이름
②		NA	-/~ [에서]	NA+Cca분류사	
③	'지수'	NA		NA+'선'	'사이'
④	인명/왕호/대회명	NA		NA	

[표 7] 이음표 문맥 분석

위와 같이 기술할 수 있는 78개의 문맥 제약과는 달리 출현 빈도가 매우 낮거나, 불규칙적인 분포를 갖는 경우를 위해 6개의 휴리스틱스를 구성하였다. 예를 들어 예문 (14)와 같은 경우 휴리스틱스는 ⓐ LAC1이 종결어미인 경우, ⓑ 이음표 앞뒤에 출현하는 LAC1과 RAC1을 결합하여 의성어 및 의태어를 구성하거나, ⓒ 한 어절을 구성하면 본 고 3.2.2에 기술한 것과 같이 모음을 반복하거나 음절을 분리한다.

4.1.3 이음표 읽기의 디폴트 값. 패턴, 문맥을 이용한 규칙 또는 휴리스틱스가 적용되지 못하는 경우 모듈이 제시할 디폴트 값은 다음과 같은 방식으로 정하였다. 학습말뭉치에서 LAC1의 마지막 음절(LAC1-LAST)과 RAC1의 시작 음절(RAC1-FIRST)을 자료 특성에 따라 아라비안숫자(NA), 라틴알파벳(LA), 한글(KA), 한자(CA), 문장부호(SYM), 빈칸(NULL)로 구분하고, 실제로 나타나는 조합쌍 72개[32] 중 실제로 예를 찾을 수 있는 40쌍에 대해 각각 읽기방식의 출현 비율을 살펴보았다. 그 일부를 살펴보면 [표 8]과 같다.

LAC1-LAST	RAC1-FIRST	이음표 종류	읽기방식의 수					계
			[ø]	[마이너스]	[대]	[에서]	[의]	
NA	NA	-	5660	0	780	1428	95	7963
KA	NA	-	151	44	2	3	5	205
KA	KA	-	3945	0	0	6	0	3951
NA	NA	~	2553	0	0	4048	0	6601
KA	NA	~	12	0	0	1460	0	1472
KA	KA	~	70	0	0	1123	0	1193

[표 8] 이음표의 좌우음절 종류에 따른 읽기방식의 출현 빈도(발췌)

위와 같은 자료를 근거로 [표 9]와 같은 디폴트 값을 정할 수 있었다.

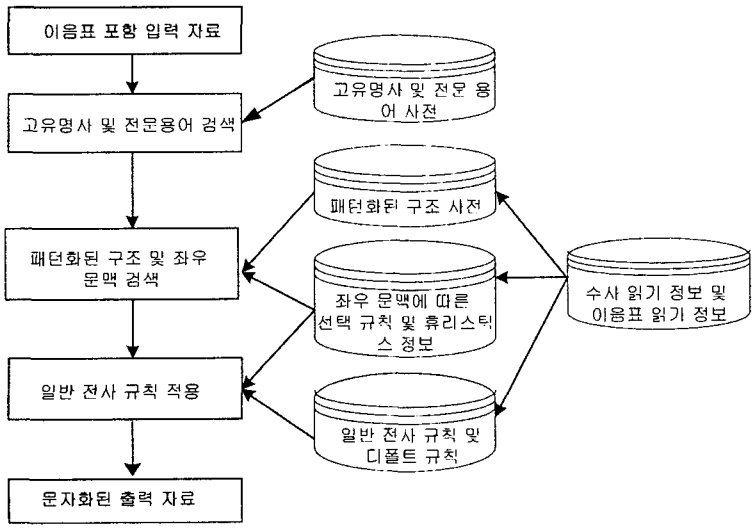
[31] 인명, 왕호, 대회명 등의 목록은 부산대학교 한국어정보처리연구실의 분류 목록을 사용한다.
[32] 6종류(LAC1의 자료 종류)×2개(이음표 종류)×6종류(RAC1의 자료 종류)

	LAC1-LAST	RAC1-FIRST	이음표 종류	디폴트 읽기방식
①	SYM	NA	-	[마이너스]
②	①을 제외한 나머지 쌍		-	[ø]
③	KA	SYM	~	[ø]
	LA	LA		
④	③을 제외한 쌍		~	[에서]

[표 9] 이음표의 디폴트 값

4.2 이음표 자동 전사 모듈의 구조

4.1절에서 제시한 이음표의 패턴화된 구조 86개와 좌우 문맥 정보 84개를 이용하여 <그림 2>와 같은 이음표 자동 전사 모듈을 구현하였다. 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 사전과 문맥에 따른 선택 규칙과 휴리스틱스 정보를 이용해 86개의 구조와 부합되는 구조가 처리된다. 다음 단계로 ③ 좌우 문맥 정보를 이용하여 중의성있는 구조를 검색하고 일반 전사 규칙 및 디폴트 규칙을 적용하여 이음표를 한글로 전사한 결과값을 갖는다.



[그림 2] 이음표 자동 전사 모듈의 구조

5. 실험 및 평가

이상과 같이 구현한 자동전사 시스템의 성능평가를 위해 3.1절에서 소개한 3개의 실험말뭉치를 사용하였다. 그 평가 결과는 [표 10]과 같다. 양과 균형도가 다른 실험말뭉치 모두에서 96% 이상의 정확도를 얻을 수 있으므로, 본 연구에서

구현한 이음표 자동전사 시스템이 실용적임을 확인할 수 있었다. 무작위 추출된 실험말뭉치1과 2의 정확도가 균형적 특성을 가진 실험말뭉치3보다 더 높은 것은 일반적인 읽기 규칙의 적용범위가 비교적 넓게 분포되어 있기 때문이다. 대상어절의 수가 더 많은 실험말뭉치2의 정확도가 실험말뭉치1보다 더 높다는 점은 본 시스템이 특정한 유형의 이음표 읽기를 잘 처리할 수 있다는 점을 보여준다. 특히 기존에 개발된 V음성합성 시스템과 비교해 봤을 때, 3개의 실험말뭉치 모두에서 훨씬 더 높은 정확도를 얻을 수 있었다. 특히 이음표 용법의 모든 경우가 실제 출현 비율에 따라 제시된 실험말뭉치 3의 경우 기존 음성합성 시스템의 오류율이 매우 높은 것과는 달리 본 연구에서 개발한 시스템의 오류율 증가는 소폭에 그친다.

	정확도(맞은 어절/전체 어절)	
	본 연구 시스템	V 음성합성 시스템
실험말뭉치1	97.36% (407/418)	85.16%(356/418)
실험말뭉치2	98.48% (972/987)	90.78%(896/987)
실험말뭉치3	96.55% (987/1014)	81.16%(823/1014)

[표 10] 이음표 자동전사 시스템의 평가

(17) 북위 37-00-30 [*삼십칠-공공-삼십/삼십칠 도 영 분 삼십 초]

(18) (a) 영명(472-[*에서/의]4027)

(b) 588-[*에서/에]2299/http://www.hanwha.co.kr

(19) 이 연수원(옹진군 덕적면 북리 124-[*ø/의]2.032[*이 점 영삼이/이. 공삼이]-[ø]834-[*ø/의]2117-[*ø/에서]9)은

(20) (a) 보행-[*ø/에서]주행신호로 바뀔 때 사고 운전자에

(b) 역으로 일본어-[*ø/에서]한국어로 번역되는 과정에서

(21) (a) "근~[*근/그은]배야,"

(b) '프리~[*ø/이]덤(자유)'을

(22) (a) 수도인 말레는 한국-[*ø/마이너스]3시간

(b) 반대로 하늘에서 떨어질 때 -2[*ø/마이너스]에서

(23) 대우차 희망센터 김경운 팀장 -[*마이너스/ø]1700명)이며

(24) (a) 명지대에 3-0(25-22,2 5-22,2 팔 대 이십육) 으로

(b) 본부 신고접수 담당자는 "최근에는 '011-9×××'

본 연구에서 개발한 이음표 자동전사 시스템의 오류 비율이 높지는 않으나, 오류문 (17)~(23)은 시스템 개선할 수 있는 (17)~(20)과 그렇지 않은 경우로 구분할 수 있다. (17)은 실험말뭉치에서 처음 발견된 패턴으로 방위표현과 함께 쓰인

오류유형	실험말뭉치1	실험말뭉치2	실험말뭉치3
(17)	0	1	0
(18)	3	4	9
(19)	1	2	3
(20)	2	2	5
(21)	0	1	3
(22)	1	2	5
(23)	1	1	3
(24)	3	4	7
계	11	17	35

[표 11] 유형별 오류 수

것이며, 패턴화된 구조에 추가할 수 있다. (18)은 전화번호 패턴 검색 오류이고, (19)는 원문에서 전화번호를 잘못 기록한 것이다. 이 경우 사업자 번호나 지역번호와 같은 특정 번호, '세자리수+이음표+네자리수'와 같은 특정 숫자열에 가중치를 높여줌으로써 검색율을 높일 수 있을 것이다. (20)은 '문자+이음표+문자' 구조에서 이음표를 '에서'로 읽는 문맥을 확장할 필요를 보여준다. 본 시스템에서는 LAC1과 RAC1이 시간 명사나 수사열인 경우에 한해, 이음표를 '에서'로 전사하였으나, 이에 덧붙여 RAC2, 3에 '바뀌다, 번역하다' 등과 같은 전환동사의 출현 등의 조건을 추가할 수 있을 것이다. 하지만 (21)~(23)과 같은 예는 시스템 개선에 사용하기 힘들다. (21)은 장음 표시로서 '근배'와 같은 인명, '프리덤'과 같은 외국어는 '이음표 양쪽의 LAC1과 RAC1이 한 어절을 형성하는지 판단할 수 있는 사전 내에 발견되지 못한 경우다. (22)나 (23)은 문장 또는 문단 수준의 의미 분석이 있어야 해결할 수 있다. (24-a)는 원문 자체에 오류가 있는 경우이며, (24-b)는 숫자 대신 다른 기호가 사용되어 전화번호로 인식하지 못한 것이다.

6. 결론 및 향후 연구

이상과 같이 신문 텍스트에서 이음표('-', '~')의 자동 문자화 방식 및 이음표가 좌우 문맥 숫자에 미치는 영향을 분석하여 규칙과 휴리스틱스를 찾아내고 이에 기반하여 이음표의 자동 전사 모듈을 구현하였다. 실험말뭉치를 이용한 평가 결과 평균 97% 정도의 정확도를 얻을 수 있었다. 오류율의 30% 정도는 전화번호 패턴 검색에 있으며 이는 패턴 정보에 가중치를 부여함으로써 해결할 수 있을 것이다. 앞 서 언급하였듯이 기호 사용이나 띄어쓰기는 사용자에 따라 편차가 있다.[33] 따라서 다른 텍스트를 대상으로 성능 평가를 할 필요가 있다. 이음표 이외에 텍스트에 사용되면서 읽기의 중의성을 가진 기호로는 온점(.), 쌍점(:), 빗금(/)이 있다. 이에 대한 분석은 향후 연구에서 이루어질 것이며, 이음표 자동전

[33] 현대 한국어에서는 예문 (12-i)처럼 웹주소와 이메일이 점차 많이 쓰인다. 이 두 표현에는 이음표 사용이 빈번한데, 줄표와 물결표를 읽는 방식의 차이가 있다. 물결표의 경우 영어의 '틸드'보다는 '물결표'로 읽나, www.chung-gu.pusan.co.kr에 나타나는 줄표의 경우 '줄표'보다는 비표준어인 '다시'(영어 dash에서 비롯된 외래어)를 더 빈번하게 사용한다. 본 연구의 목적이 표준 발음을 어떤 기준으로 설정하는지보다 비문자 기호와 그 읽기 방식 간의 분포를 알아보는데 있으므로 언중이 자주 사용하지 않는 '줄표'라는 전사를 제외하였다. 차후 웹주소와 이메일 등에 사용된 줄표의 표준읽기 방식이 결정되었을 때, 본고 4.1.1에 제시한 바와 같은 패턴 매칭 방법으로 줄표를 자동전사할 수 있다.

사 모듈, 아라비안 숫자 읽기 모듈과 함께, 비-문자 기호의 자동전사 시스템을 구성하게 될 것이다.[34]

<참고문헌>

김병주. 2000. 정보인식을 위한 고유명사 및 수사 추출. 영남대학교 대학원 석사학위 청구논문.

김상준. 1986. 방송언어와 수의 표현. KBS 표준방송언어, 125-151. 한국방송공사.

김상준. 1992. 방송언어연구: 한국어 음성표현의 이론과 실제. 도서출판 홍원.

김수연. 2002. Numeral Quantifiers in Small Clauses. 언어 27.4: 557-579.

김영희. 1976. 한국어 수량화 구문의 분석. 언어 12: 89-112.

김영희. 1983. 한국어 셈술화 구문의 통사론. 탑출판사.

박갑수. 1996. 한국 방송언어론. 집문당.

시정곤. 2000. 국어 수량사구의 통사구조. 언어 25.1: 73-101.

유동준. 1989. 국어 분류사와 수량화. 국어국문학 24: 53-72.

유재원. 1997. 자연어 처리를 위한 의존명사의 하위 범주 분류. 제9회 한글 및 한국어 정보처리 학술대회 학술발표 논문집, 136-142.

유재원. 1999. 자연어 처리를 위한 수사의 하위 범주 분류. 언어와 언어학 24: 103-110.

윤애선, 권혁철. 2003. 음성 합성을 위한 아라비안 숫자의 자동 전사. 2003 한국언어학회 동계 학술논문 발표집, 109-110.

이영직. 2000. 방송 뉴스 전사 문장의 수사 및 단위의 발성 방식. 제17회 음성통신 및 신호처리 학술대회 학술발표논문집, 285-288.

이은정. 1993. 최신 표준어·맞춤법 사전. 백산출판사.

이희승, 안병희. 2001. 새로 고친 한글 맞춤법 강의 (2판 2쇄). 신구문화사.

임홍빈. 1991. 국어 분류사의 성격에 대하여. 국어문법의 심층 3, 235-262. 태학사.

전재연. 2002. 한국어 분류사와 수. 불어불문학 연구 51: 305-325.

정영임, 김정세, 김상훈, 이영직, 윤애선. 2002. 현대 한국어에서 아라비안 숫자의 읽기 규칙 연구. 제14회 한글 및 한국어 정보처리 학술대회 학술발표 논문집, 16-23.

채완. 1983. 국어 수사 및 수량사구의 유형적 고찰. 어학연구 19.1: 19-34.

참고 웹사이트

디지털 동아: www.donga.com

매일 경제: www.mk.co.kr

한국일보 소리신문 반디: www.hankooki.com/event/bandi/bandi-1.htm

VoiceText: www.voiceware.co.kr/demo/demo_text.html

[34] 본 연구의 방법에 대해 전적으로 수동 구축이며, 지식구축의 자동화에 대한 고려가 없다는 비판이 있었다. 물론 학습말뭉치를 기초로 한 지식 구축의 자동화를 배제하지는 않는다. 하지만 초기의 지식 구축에서 정확도를 높이려면, 충분히 대표성을 가진 예가 포함된 학습말뭉치의 분석을 통해 규칙과 휴리스틱스 등이 설정되어야 한다. 본 논문을 통해 소개한 이음표의 자동전사 모듈과 앞으로 구현할 문장 부호 및 아라비안숫자 자동전사 모듈은 '공통 음성DB 구축'을 위한 텍스트 전처리에 사용될 예정이다. 이 사업 1단계의 목표 어절의 수는 약 3억 개다. 본 논문의 원시말뭉치인 중앙지 1개의 2년치 기사의 어절수가 약 1,880만 개이고 그중 19,767개 어절에서 이음표가 나타나므로, 목표 어절 전체에서 나타날 이음표의 수는 약 30만 개 정도로 추정된다. 즉 본 연구에 사용된 학습말뭉치의 크기는 이음표 자동전사 모듈을 적용할 대상의 6.7%정도에 지나지 않는다. 이 크기는 수동 분석이 가능할 정도이며, 다양한 읽기 방식을 보여주는 자료가 포함되어 있었다.

VoiceTopia: www.slworld.co.kr/voicetopia/voicetopia3.htm
CoreVoice: www.corevoice.com/demo/demo_1.html

접수일자: 2003년 5월 17일
게재결정: 2003년 6월 11일