아라비안 숫자를 동반한 중의적 기호의 자동전사: 온점, 쌍점, 빗금을 중심으로

윤애선, 정영임, 권혁철*†
부산대학교

Aesun Yoon, Youngim Jung and Hyuk-Chul Kwon. 2004. Automatic Transcription of Three Ambiguous Symbols Used with Arabic Numerals: Period, Colon and Slash. *Language and Information 8.1*, 117–136. In this paper, we have proposed Auto-TSS, an automatic transcription module of three ambiguous symbols—period (.), colon (:) and slash (/)—using their linguistic contexts. Few previous studies have discussed the problems of ambiguities in reading those symbols into Korean alphabetic letters in order to improve the current Korean TTS (Text-To-Speech) systems. We have classified 9 different reading formulae of the three symbols, analyzed their left and right contexts, and investigated selection rules and distributions between the symbols and their contexts. Based on these linguistic features, 30 stereotyped patterns, 53 rules and 5 heuristics determining the types of reading formulae are investigated for Auto-TSS. This module works modularly in 4 steps. The pilot test was conducted with three test suites, which contain respectively 6,979, 3,491 and 2,450 morpheme clusters containing at least one of three ambiguous symbols and Arabic numeral(s). Encouraging results of 94.3%, 93.0%, 94.2% accuracy were obtained for the 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 de-

* 609-735 부산시 금정구 장전동 산 30 부산대학교 인문대학 불어불문학과/공과대학 전자전기컴퓨터 공학부, E-mail: {asyoon, acorn, hckwon}@psuan.ac.kr

† 본 논문은 과학기술부의 국가지정 연구실 사업(과제명: '언어 중심의 지능적 정보처리를 위한 단계적(scalable) 우리말 분석 기술의 개발' (M10203000028-02J0000-01510))의 지원을 받아 이루어졌음을 밝힌다. 또한 초고의 오류와 미처 생각하지 못한 부분에 대해 상세한 지적을 해주신 심사위원에게 지면을 빌어 감사의 뜻을 전한다. 물론 논문의 모든 오류는 전적으로 저자의 책임임을 다시 한번 밝힌다.

본 논문의 내용은 한국인지과학회 2003년 춘계학술대회에서 '한국어 TTS 시스템의 성능향상을 위한 문장기호 전사시스템의 구현'으로 발표한 내용을 보완하고, 성능 평가의 범위를 확장한 것이다. 여기에서 사용하는 '전사(transcription)'라는 용어는 ① 음을 문자로 변환하거나, ② 한 언어나 기호를 다른 언어로 바꾸어 적으며 그 음성정보를 유지하는 것을 지칭하며 좁은 의미의 '문자화 변환(transliteration)'과 동의 관계를 갖는다 (http://www.wikipedia.org/wiki/Transcription+(linguistics)). 본고에서 '전사'는 후자의 정의를 따른다.

© 2004 한국언어정보학회

tecting module; and to apply Auto-TSS on a larger scale. (**Pusan National University**)

Key words: 음성합성 전처리(preprosessing for TTS), 아라비안 숫자(Arabic numeral), 온점(period), 쌍점(colon), 빗금(slash)

1. 서론

컴퓨터 대용량 저장 장치의 발전에 힘입어 음성합성 기술이 간단한 단어 또는 숫자의 조합으로 문장을 합성하여 증권정보, 114 안내, 게임 스코어 안내 등 제한적 분야에 사용되던 단계에서 어휘나 문장 수에 제한 없이 임의의 문장을 합성하는 무제한 음성합성(unlimited TTS: Text-To-Speech) 방식의 기술로 발전하였다. 현재 국내의 TTS 기술은 상당 부분 발전하여 음성 기사 서비스, E-mail 읽기 시스템, 언어 학습 시스템, 언어 장애자를 위한 발성 대용 및 훈련 시스템 등 다양한 분야에서 그 기술이 응용되고 있고 국내 기업 및 연구소에 의해 상용화를 위한 TTS 시스템이 개발되었다. 실용화된 TTS 시스템이 개발되려면 음성 합성 단위의 자연스러운 음성적, 운율적 연결이 구현되어야 함은 물론, 입력 텍스트의 정확한 정보전달을 위해 텍스트에 사용된 비-한글 문자 기호의 전처리(preprocessing)에 관한 정보도 필요하다[권철홍, 2004; 김형순, 2001; 이용주, 1995]. 음성공학을 중심으로 전자에 관한 연구는 활발히 이루어져 왔으나, 후자에 관해서는 자연 언어 처리의 관점에서 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, 이음표의 경우 9가지 의미를 가지고 6가지로 읽을 수 있다. 온점(.), 쌍점(:), 빗금(/)도 형태적 동인성에 기인하여 문장부호로 사용될 뿐만 아니라 중의성을 가지는 기호로 사용된다. 특히, 아라비안 숫자와 함께 나타나면 온점의 경우, '점, 닷, 년·월·일' 및 영형태(zero morpheme)로 읽히고, 쌍점은 '대, 장·절, 년·월·일' 및 영형태, 빗금은 '나누기, 분의, 년·월·일' 및 영형태 등으로 읽힌다. 또한, 온점, 쌍점, 빗금은 이들 기호와 함께 사용된 아라비안 숫자의 읽기에도 영향을 미쳐 좌·우 문맥에 따라 아라비안 숫자가 부정수, 단위가 있는 한자어 및 단위가 없는 한자어로 다양하게 읽힌다. 하지만 기존의 TTS 시스템에서 이러한 세 기호 읽기의 정확도가 낮아 정보 전달력이 떨어진다[윤애선/권혁철, 2003; 정영임 외, 2003].

본 연구에서는 한국어 TTS 시스템의 성능 향상을 위해[1] 신문 텍스트에서 출현율

[1] 이음표의 자동전사에 관련된 선행 논문과 본고는 '한국어 텍스트에 사용된 비-문자 기호의 문자화 시스템 개발'이라는 최종 목표의 한 단계를 구성한다.

이 높은[2] 동시에 중의성을 띤 '아라비안 숫자 동반 온점, 쌍점, 빗금'의 문자화[3] 시스템을 구현한다. 2장에서 선행 연구의 문제점 및 기존 TTS 시스템에서 나타나는 세 기호의 읽기 오류 유형을 살펴본다. 3장에서는 본 연구에 사용된 분석 말뭉치 및 평가 말뭉치의 구성과 연구 대상의 특성을 검토하고, 연구 범위 및 방법론을 살펴본다. 4장에서는 온점, 쌍점, 빗금의 읽기 방식에 따라 문자화 규칙을 설정하고 이들 기호에 의해 영향을 받는 숫자 읽기를 고려하여, 세 기호의 자동 문자화 시스템을 구현한다. 5장에서는 구현된 자동 전사 시스템을 실험하여 그 정확도를 알아보고, 오류 유형을 밝히며, 6장에서는 시스템 성능 향상 방안과 향후 연구방향을 제시한다.

2. 선행연구 및 문제점

온점, 쌍점, 빗금과 관련된 어문규정으로는 한글 맞춤법의 부록에서 문장부호로서의 정의와 용례가 나타나는데, 특히 아라비안 숫자와 함께 쓰인 세 기호의 사용법은 [표 1]과 같이 제시되었다[이은정, 1993; 이희승/안병희, 2001, 임동훈 2002]. 그러나 현대 한국어에서 온점은 반점이나 가운뎃점을 대신하여 여러 단위를 열거할 때, 기념일을 나타낼 때 사용되며 소수점으로도 사용된다. 빗금은 온점과 마찬가지로 연월일을 표시할 때나 수학식에서 '÷'를 대신하여 나누기 기호로 사용된다. 또한, 온점, 쌍점, 빗금은 웹 주소나 이메일 주소에 사용되고 있으며, 의미적 차이를 가지는 두 개 이상의 기호 포함 숫자열 간 구분자로도 기능한다. 하지만 어문규정에서는 실제 언어 자료에 나타나는 다양한 용례를 다 포함하여 기술하지 못하였다[채완, 2002].

종류	용법
온점	① 아라비안 숫자만으로 연월일을 표시한다. ② 표시 문자[4] 다음에 사용한다.
쌍점	① 시·분, 장·절을 구별할 때 사용한다. ② 둘 이상을 대비할 때 사용한다.
빗금	① 대응, 대립되거나 대등한 것을 함께 보이는 단어와 구, 절 사이에 쓴다. ② 분수를 나타낼 때 사용한다.

[표 1] 한글 맞춤법 규정에 제시된 온점, 쌍점, 빗금의 용법

[2] 본 연구의 원시 말뭉치인 1개 신문 2년치 모든 분야의 기사에서 전체 어절 대비 '아라비안 숫자를 동반한 온점, 쌍점, 빗금'의 출현율은 10.08%으로, 이음표(줄표, 붙임표, 물결표)의 출현율 4.72%보다 높다. 문자와 함께 나타나는 온점의 용법 구분은 자연언어처리에서 문장을 구분하는 표지임과 동시에 TTS 시스템에서 억양을 결정하는 중요한 요소이나, 본 연구의 초점에서 벗어나므로 다음 연구를 기약한다.

[3] '문자화'는 비-문자 기호를 한글로 변환하는 것을 지칭하고, '읽기'는 문자 정보 사이에 나타날 수 있는 음성적 변환까지 포함하므로, 후자가 좀더 광범위하다. 하지만 문자 정보의 음성적 변환은 본 연구 다음 단계에 수행되어야 하므로, 본고에서는 '문자화'와 '읽기'를 구분하지 않고 사용한다.

[4] 표시 문자는 '1.', '가.' 등 (11-b)에서처럼 항목의 표시에 사용한다.

온점, 쌍점, 빗금은 다양한 방식으로 문자화된다. 그러나 전산언어학이나 음성공학 분야에서는 아직 온점, 쌍점 및 빗금이 다양하게 읽히는 방식에 대한 연구가 거의 시도되지 않았고, 이에 따라 현재 제공되고 있는 TTS시스템[5]에서 온점, 쌍점, 빗금의 읽기 방식에서 [표 2]와 같이 많은 오류를 보인다.

온점, 쌍점, 빗금은 숫자와 함께 여러 가지 패턴화된 구조로 사용되며, 좌·우 문맥에 따라 읽기 방식이 결정된다. 그러나 이러한 패턴화된 구조나 문맥에 따른 온점, 쌍점, 빗금의 문자화 규칙 설정에 대한 기존의 연구가 거의 없고, 현재 제공되고 있는 TTS시스템에서 온점, 쌍점, 빗금의 문자화 처리의 정확도가 매우 떨어진다.

실제 TTS시스템을 이용하여 음성기사 서비스를 제공하고 있는 D, M신문에서는 온점, 쌍점 및 빗금의 문자화를 일괄적으로 처리하고 있으며, V음성합성 시스템과 C시스템의 경우 좌·우 문맥을 고려하지 않고 규칙이 적용되어 [표 2]의 (1)~(10)과 같은 오류가 나타난다.

기존의 TTS시스템에서는 [표 3]처럼 세 기호의 읽기를 2~4개 정도로 단순 변환하므로 읽기 오류율이 높다. 온점이 반점(,)이나 가운뎃점(·)을 대신하여 둘 이상의 대등한 것을 나열하거나 기념일 표현에 사용되는데,[6] 이 때 온점의 읽기가 영형태이나 오류 예문 (1), (2)와 같이 기존 TTS 시스템 대부분이 오류를 범하고 있다. 쌍점은 둘 이상을 대비할 때 '대'로 문자화되어 사용되는 경우나 장·절의 구별하는 사용법이 문자화 규칙에 포함되지 않아 (5)~(7)과 같은 오류가 발생한다. 빗금 역시 연월일과 분수 표현을 나타내는데, (9)~(10)처럼 이러한 읽기를 제공하지 못한다.
또한, C와 V는 연월일과 소수를 표현하는데 사용할 수 없는 수사열임에도 불구하고 (3)과 같이 처리하고 있어 온점의 문자화 규칙이 정교하지 못하다. 오류 예문 (4)와 같이 전화번호의 마지막 고유번호 사이의 구분자로서의 온점을 인식하지 못해, C는 온점 포함 수사열을 정확하게 읽지 못하고, V는 온점 이하의 숫자를 별개의 토큰으로 인식한다. 쌍점의 경우, '비율', '창(세기)'와 같이 문맥이 뚜렷한 단어가 연접하였지만 C는 (6)과 같이 쌍점을 '시·분'으로 문자화하거나, (7)처럼 '장·절'을 '대'로 문자화 처리하는 오류를 보였다. 구분자로 사용되어 영형태로 문자화되는 빗금의 경우도 (8)과 같이 전화번호의 마지막 번호의 구분 표지를 잘못 처리한다.

[5] 국내 3개 일간지에서 음성합성 시스템을 이용한 음성기사가 제공되고 있고, 국내 기업 및 연구소 5군데에서 상용 TTS시스템을 개발하였다. 그 중 성능이 우수하다고 평가되는 2개의 상용 시스템과 2개 일간지 음성기사 서비스의 온점, 쌍점, 빗금 읽기 방식과 문자화 처리되는 유형을 분석하였다. 본 논문에서는 상용 시스템인 보이스웨어의 음성합성 시스템을 V, 코어보이스를 C로 나타내며 음성기사 서비스를 제공하는 동아일보를 D, 매일경제신문을 M으로 약칭한다.

[6] 유니코드에서는 온점과 가운뎃점(middle dot)은 각각 '002E와 '00B7'로 표현된다[The Unicode Consortium, 1996]. 가운뎃점은 기본 라틴 문자에 속하지 않으므로 워드프로세서에서 입력이 불편하다. 예를 들어 가운뎃점을 입력하려면 특수기호 넣기 기능을 이용해야 하므로 적어도 3번의 키나 마우스 포인팅이 필요하다. 또한 가운뎃점과 유사한 형태를 가진 기호가 5개 종류가 더 있어 각 신문사마다 총 6개의 가운뎃점 리스트 중 임의로 하나를 선택하므로 가운뎃점을 다른 코드로 전환했을 때 일관성이 없고 쉽게 깨진다. 따라서 실제 신문과 같은 한국어 텍스트에서 반점(,)과 가운뎃점을 대신하는 온점의 사용이 빈번하다.

예	읽기 오류	바른 읽기	출처
(1) 2.11.13.14호	*이점 일일점 일삼점 일사	이 십일 십삼 십사	C,V
(2) 9.11 테러	*구 점 일일	구일일	D,M, C,V
(3) 업체 016.018.019로	*십육년 십팔월 십구일	공일육 공일팔 공일구	C
	*공일육점 공일팔점 공일구		V
(4) 02-2188-6041. 028	*공이 다시 이일팔팔 다시	공이에 이이팔팔에 육공사일 육공이팔	C
	*공이에 이이팔팔에 육공사일 육천이십팔		V
(5) 중반 이후 7:3의 일방적	*칠 삼	칠 대 삼	D,V
(6) 무게비율 10:20	*열시 이십분	십 대 이십	C
	*십 이십		V
(7) 창13:13	*십삼 대 십삼	십삼 장 십삼 절	C
	*십삼 십삼		V
(8) 문의 02-6242-5305/6.	*공이에 육이사이에 오삼공오 나누기 육	공이에 육이사이에 오삼공오 또는 육	V
(9) 2차 코스 1/17~1/29로	*일 나누기 십칠에서 일 나누기 이십구	일 월 십칠 일에서 일 월 이십구 일	C,V
(10) 3/1000mm	*삼 나누기 천	천분의 삼	C,V

[표 2] 기존 음성합성 시스템의 온점, 쌍점, 빗금 읽기 오류

	'.'의 읽기	':'의 읽기	'/'의 읽기
D	'점', 영형태	영형태	영형태
M	'점', 영형태	-	-
V	'점', 영형태	영형태	'나누기', 영형태
C	'점', '연월일', '조', 영형태	'대', '시분초', 영형태	'나누기', 영형태

[표 3] 기존 TTS시스템의 온점, 쌍점, 빗금 읽기 형태

3. 연구 내용과 범위

이 장에서는 온점, 쌍점, 빗금의 문자 전사화 연구를 위한 말뭉치 구성과 온점, 쌍점, 빗금 사용의 특성을 알아보고 연구 내용과 범위를 기술한다.

3.1 말뭉치의 구성

앞 장에서 살펴본 바와 같이, 온점, 쌍점, 빗금의 문자 전사화나 읽기 규칙에 관한 선행 연구가 거의 없으므로, 가능한 한 다양한 경우가 포함될 수 있는 말뭉치를 구성하

는 것이 매우 중요하다. 이를 위해 본 연구에서는 전자화된 텍스트 중 다양한 전문 분야가 포함되고 가독성을 높이기 위해 기호 사용이 빈번한 신문 기사를 이용하여 연구 대상 말뭉치를 구성하였다.

이를 위해 사용된 말뭉치는 1개 신문 2년치(2000년 1월~2001년 12월)[7] 모든 분야의 기사에서 아라비안 숫자와 함께 사용된 온점, 쌍점, 빗금을 포함한 69,000여 개 어절 중, 20%인 13,959개 어절을 임의 추출하여 분석하였다. 이 분석 결과를 바탕으로 숫자와 함께 사용된 온점, 쌍점, 빗금의 패턴화된 구조를 찾아내며 좌·우 문맥 정보를 이용하여 이들 기호의 문자화 규칙을 설정한다. 또한, 이들 기호에 의해 영향을 받는 아라비안 숫자의 읽기 방식을 고려하여 이들 기호의 문자화 규칙을 적용하여 온점, 쌍점, 빗금 기호의 전사 시스템을 구현한다. 시스템의 성능을 평가하기 위해 전체 말뭉치의 10%, 5% 크기에 해당하는 평가 말뭉치 2개 세트를 만들었으며, 각 기호의 다양한 문자화 방식을 모두 포함한 3.5% 크기의 대표성을 가진 균형 말뭉치(balanced corpus) 1개 세트를 구성하여 모두 3개 세트의 평가 말뭉치를 실험에 사용하였다.

구분	표본 추출 방식	말뭉치 생성	어절 수	원시말뭉치 대비 비율
원시 말뭉치	무작위 추출	2000.1.1-2001.12.31 638일 기사 중 아라비안 숫자를 동반한 기호(온점, 쌍점, 빗금) 포함 어절 전체 추출	69,795개	100%
분석 말뭉치	무작위 추출	비정렬된 원시말뭉치의 20%	13,959개	20%
실험 말뭉치1	무작위 추출	(원시말뭉치-분석말뭉치)의 12.5%	6,979개	10%
실험 말뭉치2	무작위 추출	(원시말뭉치-분석말뭉치-실험말뭉치1)의 7.8%	3,491개	5%
실험 말뭉치3	대표적, 균형적 추출	(원시말뭉치-분석말뭉치-실험말뭉치1-실험말뭉치2)에서 분석말뭉치의 비율과 종류에 따라 선정된 5.4%	2,450개	3.5%

[표 4] 말뭉치 구성 및 특성

3.2 연구 대상의 특성

이 절에서는 온점, 쌍점, 빗금의 사용 환경과 용법, 문맥에 따른 다양한 읽기 방식 등 연구 대상의 특성을 알아본다.

3.2.1 온점, 쌍점, 빗금의 용법. 학습말뭉치에서 실제로 관찰할 수 있는 온점, 쌍점, 빗금의 용법은 [표 1]의 한글맞춤법 규정에서 제시한 용법인 예문 (11-a~f)의 출

[7] 본 연구에 사용된 말뭉치는 윤애선/권혁철(2003)의 이음표 전사 시스템 개발에 사용한 것과 동일하다.

현 빈도는 극히 낮다. 오히려 이러한 규정에 제시되지 않은 예문 (12)와 같은 사용이 더욱 빈번하다.

(11) a. ① **2004.4.15.**
② **2004.4.**
③ **4.15.**

b. **1.** 마침표

c. ① 오후 **3:16**
② 요한복음 **3:16**

d. 현대가 삼성을 **65:60**으로 이기고

e. 백이십오 원/**125**원

f. ① **3/4**분기
② **3/20**

(12) a. 주가가 **4.15**% 상승

b. **4.15**총선에서

c. 이동통신업체 중 **016.018.019**로

d. IP주소는 **164.125.8.5**이고

e. 이번 100m 기록은 **10:05**에 도달하였다

f. 2차 모집기간은 **4/15**~**4/18**로

g. **24/5 = 4.8**

온점은 흔히 (12-a)처럼 소수점 표시와 (12-b)처럼 기념일 표시로 사용한다. 또한 (12-c)처럼 반점 또는 가운뎃점 대신 분리 기호로 이용되며, (12-d)처럼 특정한 패턴에서 출현한다. 쌍점은 함께 사용되는 아라비안 숫자의 크기와 문맥에 따라 (12-e)처럼 읽는 방식이 결정된다. 빗금은 (12-f)에서와 같이 온점처럼 연월일을 나타내기도 하고, (12-g)처럼 수식표현에 사용된다.

숫자와 동반한 세 기호의 용법을 의미에 따라 구분하면 아래 [표 5]와 같다.

3.2.1 온점, 쌍점, 빗금의 읽기 방식. 이상에서 본 것처럼 온점, 쌍점, 빗금의 용법이 다양한 것만큼 읽기 방식도 다양하다. (13-a~h)처럼 연월일 표현, 열거, 웹주소, 스포츠 경기 점수 표현, 고유명사와 같이 특정한 유형으로 분류할 수 있는 패턴화된 구조이고 (14-a~h)는 아라비안 숫자와 함께 사용되면서 문맥에 의해 기호 및 숫자의 전사가 다른 방식으로 구현되는 예이다.

의미 분류 / 기호	시간	수량	순서	대비	구분자	번호 (패턴)	명칭
·	기념일 연월일	소수점 부정수	순서 열거		분리 기호	IP, 웹 주소	고유명사
:	연월일 시분초			비율 점수	분리 기호	장·절 조·항 IP, 웹 주소	고유명사
/	연월일	분수 나누기	순서 열거		분리 기호	IP, 웹 주소	고유명사

[표 5] 온점, 쌍점, 빗금의 의미 분류

(13) a. ① **1998. 12. 3.***[천구백구십팔 년 십이 월 삼 일]*
② **1998.12.***[천구백구십팔 년 십이 월]*
③ **12.3.**[십이 월 삼 일]

b. ① **3.** 이 부분의*[삼0]*
② **3.** This part*[삼0]*

c. ① http://**203.456.12.8***[이공삼 점 사오육 점 일이 점 팔]*
② ftp://**203.456.12.8***[이공삼 점 사오육 점 일이 점 팔]*

d. ① 커뮤니케이션의 하위노드: **5***[오]*
② subcategories of communication: **5***[오]*

e. **삼성 3:2 한화***[삼성 대 한화 삼 대 이]*

f. **1/4**분기*[일사]*

g. ① **1998/12/3***[천구백구십팔 년 십이 월 삼 일]*
② **1998/12***[천구백구십팔 년 십이 월]*
③ **12/3***[십이 월 삼 일]*

h. **IEEE802.11b***[아이트리플이 팔공이 점 일일비]*

(14) a. ① 담배를 하루에 **2**대 이상 피우면*[두|*이]*
② 담배를 하루 평균 **2.3**대*[이 점 삼|*두 점 세]*

b. **60.70**년대에는*[육칠십]*

c. **8.15**m*[팔 점 일오]*

d. ① **8.15**기념행사*[팔일오]*
② **12.12**사태*[십이십이]*[8]

e. 오전 **3:3***[세 시 삼 분]*

f. ① 창세기 **3:3***[삼 장 삼 절]*
② 창**3:3***[삼 장 삼 절]*

g. 한화와 LG가 **3:3**으로 동점을*[삼 대 삼]*

h. **2/5-3/16**까지 세일을*[이월 오일에서 삼월 십육일]*

온점의 경우, 예문 (14-a)의 ①과 ②를 비교하면 알 수 있듯이 온점이 소수점으로 사용될 때, 고유어 수사와 결합하는 분류사 '대'라도 함께 사용된 소수는 한자어 수사로 읽는다[윤애선, 2002]. (14-a)와 같은 소수표현과 (14-b)처럼 범위수를 나타내는 경우를 구분해야 하며, (14-c)와 (14-d)에서 볼 수 있는 것처럼 소수와 기념일은 다르게 읽는다. 쌍점의 경우에는 (14-e)~(14-g)에서 볼 수 있듯이 문맥에 따라 쌍점의 읽기와 동반된 아라비안 숫자의 읽기 방식이 다르다. 빗금의 경우도 마찬가지로 숫자의 크기, 문맥에 따라 (14-h)에서처럼 다르게 전사된다. 문맥에 따라 세 기호는 [표 6]과 같이 문자화된다.

의미 분류 / 기호	시간	수량	순서	대비	구분자	번호 (패턴)	명칭
.	[Ø]	[점]	[Ø]		[Ø]	[점]	다양
	[년월일]	[Ø]				[닷]	
:	[년월일]			[대]	[Ø]	[장절]	다양
	[시분초]					[조항]	
						[콜론]	
/	[년월일]	[분의]	[Ø]		[Ø]	[슬래쉬]	다양
		[나누기]					

[표 6] 의미 분류에 따른 온점, 쌍점, 빗금의 문자화

4. 온점, 쌍점, 빗금의 자동전사

2장과 3장에서 살펴본 바와 같이, 기존 기사 음성 서비스에서는 실제 언어 자료에서 나타나는 온점, 쌍점, 빗금의 다양한 용법이 모두 제시되지 못하였고, 문맥에 따라 다르게 읽히는 온점, 쌍점, 빗금의 다양한 문자화 규칙이 정교하게 설정되지 못하였다.

[8] 예문 (14-d②)는 이영직(2000)에서 가져왔다.

4장에서는 실제 신문텍스트에서 임의 추출된 학습말뭉치 13,959개 어절의 좌우 문맥을 분석하여, 아라비안 숫자 동반 기호의 자동전사 시스템을 구현한다.

4.1 기호의 문자화를 위한 규칙과 휴리스틱스

온점, 쌍점, 빗금의 문맥은 크게 특정화된 형식을 갖는 패턴화된 구조와 그렇지 않은 것으로 구분할 수 있다. 전자를 검색하기 위해서는 패턴화된 구조의 검색 조건을 구성해야 하고, 후자의 경우는 중의성을 해소할 수 있는 문맥 정보 및 기호의 문자화를 일반화할 수 있는 규칙이 필요하다. 그리고 규칙이 상충할 경우, 규칙의 우선 적용 순서를 정하기 위한 휴리스틱스를 설정해야 한다.

4.1.1 패턴화된 구조의 검색 조건. 예 (13)에서 볼 수 있는 패턴 중 웹주소, 열거, 스포츠 경기 점수 표현, 분기 표현 등은 특정한 형식을 가지므로 패턴 검색하기 용이하다. 하지만 (4), (8)과 같은 전화번호나,[9] (13-a, g)의 와 같은 연월일 표현의 검색은 특정한 조건을 만족해야 한다.

1. 기호 포함 수사열이 N1.N2.N3.[10] 또는 N1/N2/N3인가?
2. 연월일 표현과 유사한 수사열 표현과의 중의성을 해결할 수 있도록, [표 7]과 같이 수사열을 구성요소로 분석하여, 각 요소의 특성을 알아보고 각 요소의 읽기 방식을 규정한다.

	N1	./	N2	./	N3	./∅
구분	연도 표현		월 표현		일 표현	
필수성	+	+	+	+	+	-
수사열의 크기	2 또는 4		1 또는 2		1 또는 2	
특성	수사열의 크기=4일 때, N1≤2999			N2≤12	N3≤31	
읽기방식	Cca_b[+U][11]	년	Cca_b[+U]	월	Cca_b[+U]	일

[표 7] 연월일 표현의 패턴 검색 결정 조건

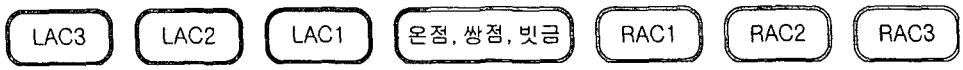
[9] 전화번호 패턴 검색에 대해서는 윤애선/권혁철(2003:32-33)을 참조하라.

[10] 실제 한국어 텍스트에서 띄어쓰기 규칙은 그리 잘 지켜지고 있지는 않다. 특히 비-문자 기호 사용에서 띄어쓰기는 사용자 주관이 개입된 경우가 많다. 따라서 규칙이나 휴리스틱스를 설정하는데 있어 이 점을 고려하여야 한다. 이 경우 수사열 N1.N2.N3.은 'N1. N2. N3.', 'N1. N2.N3.', 'N1.N2. N3.' 등의 변이 형태를 모두 포함한다.

[11] 온점, 쌍점, 빗금 기호와 동반하는 수사읽기 방식의 구분은 어원, 기·서수 여부, 자릿수 유무 및 형태 등의 특성에 따라 20가지로 나뉜다. 본 논문에 인용된 'Cca_b[+U]'는 '1234[*천이백삼십사*]'와 같이 숫자를 자릿수 포함 한자어 기수 기본형으로 읽는 방식을 의미하고, 'Cca_b[-U]'는 '1234[*일이삼사*]'와 같이 자릿수를 포함하지 않은 한자어 기수 읽기 방식을 의미한다. 수사 읽기방식의 자세한 구분은 윤애선/권혁철(1003:32)을 참조하라.

본 연구에서는 이러한 패턴화된 구조 및 패턴화된 구조 검색 조건을 '전화번호', '웹 주소', '우편번호', '시분초' 표현 등을 포함하여 30개 추출하였다.

4.1.2 문맥정보를 이용한 중의성 해소와 문자화 일반 규칙. 온점, 쌍점, 빗금을 포함한 수사열이 패턴화된 구조로 사용되지 않는 경우가 더 빈번하다. 이 경우의 대부분은 수사열의 좌우 문맥이 기호 및 아라비안 숫자의 읽기 방식을 결정한다. 세 기호를 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.[12]



[그림 1] 온점, 쌍점, 빗금의 좌·우 문맥 구분

[그림 1]에 나타난 문맥의 구성 단위는 [표 5]에서 구분한 온점, 쌍점, 빗금의 용법 및 문자화에 따라 8개 의미 범주[13]로 나뉘어 진다. 분석 코퍼스를 통해 수집된 기호의 LAC, RAC를 8개 범주로 분류하고, 이들 문맥을 통해 중의성이 해소된 규칙으로 구성하기 위해서는 다음과 같은 분포 분석이 필요하다.[14]

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

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?[15]

③ 인접 3어절 내에 동일 기호나 상이한 기호가 사용되는가?

④ LAC2, LAC3나 RAC2, RAC3가 한글/한문이라면 8개의 의미 범주 중 어느 범주에 속하는가?[16]

[12] 주 10에서 밝힌 것처럼 LAC1, 중의적 기호, RAC1 간에는 띄어쓰기 유무를 고려하지 않는다. 문맥의 크기는 검색속도와 직접적인 관련이 있는데, 선행 연구와 본 연구에서 문맥의 크기를 좌우 3개 어절로 제한한 것은 검색 속도를 유지하면서도 문맥을 고려할 수 있을 것이라는 경험적 결정이며, 중의적 기호 연구에서 문맥의 크기와 검색 효율성 및 속도와의 관계는 향후 연구를 통해 검증해야 할 대상이다.

[13] [표 5]에서 분석된 세 기호의 의미 범주는 모두 7개이나, 실제 데이터에서 이 7개 범주에 속하지 않는 구성 성분을 구분하기 위한 1개 범주를 더 두어, 기호의 LAC 및 RAC가 형성할 수 있는 문맥의 범주는 각각 8개씩이다.

[14] 이 분석 과정은 아라비안 숫자 동반 이음표의 분석 준거와 유사하다.

[15] 본고의 예문 (15), (16)에 나타난 것과 같다. 분류사, 분류사 전치어, 분류사 후치어, 숫자 전치어와 숫자 읽기 방식의 선택에 관해서는 김상준(1986, 1992), 유재원(1999), 윤애선/권혁철(2003), 정영임 외(2002), 채완(1983)을 참조하라.

[16] LAC, RAC가 '한글/한자/한자+한글'로 된 자료라면 형태소 분석기를 통해 형태소 분석이 이루어진

이상과 같이 분석된 자료의 특성과 문맥의 구성 단위 의미 범주를 이용하여 [표 8]과 같이 기호 포함 수사열의 중의성을 해소하고 기호의 의미 분류에 따른 문자화 규칙을 설정할 수 있다.

	LAC2	LAC1	기호 [읽기]	RAC1	RAC2
①	평균|+|-|타율|지수	N1 Cca_b[+U]	. [점]	N2+(분류사)+(의) Cca_b[-U]	범위| 비율
				N2+(분류사)+(을|를) Cca_b[-U]	기록하다|밑돌다|앞지르다
				N2+(분류사)+(으)(로) Cca_b[-U]	끝나다|높아지다|출발하다
②		N1 Cca_b[+U]	. [∅]	N2 Cca_b[-U]	선거|의거|특별법|기념
③				N2+을|를 Cca_b[-U]	맞이하다|앞두다
④				N2+에 Cca_b[-U]	즈음하다
⑤	전반|후반|점수|전적|스코어	N1 Cca_b[+U]	: [대]	N2+(으)로|을|를|에게 Cca_b[+U]	이기다|지다|완승|완패

[표 8] 온점과 쌍점의 문맥을 통한 중의성 해소(발췌)

예를 들어 [표 8]은 LAC1과 RAC1에 아라비안 숫자(N1, N2)가 포함되어 ① 처럼 좌문맥에 '평균|+|-|타율|지수'가 나타나거나, 우문맥에 분류사가 나타나거나 '의 + 범위| 비율', '을|를 + 기록하다|밑돌다|앞지르다', '(으)로 + 끝나다|높아지다|출발하다'와 같이 '수량'의 문맥을 형성되면 온점은 [점]으로 읽고 N1과 N2는 각각 Cca_b[+U]와 Cca_b[-U] 방식을 선택한다. ②, ③, ④와 같이 'N1+.+N2'가 '기념일'을 나타내는 경우에는 기념일에 관련한 특정 어휘가 우문맥으로 나타난다. 이때 N1과 N2는 각각 Cca_b[+U]과 Cca_b[-U] 방식으로 읽히며,[17] 온점은 읽지 않는다. ⑤처

다. 따라서 조사, 어미 등의 기능어가 분리되므로 분석 코퍼스를 통해 수집되고 8개로 의미 분류된 어휘군과 문자열 비교를 통한 매칭이 가능하다.

[17] '11.5부동산대책*[십일오]*, 8.15광복*[팔일오]*, 12.6사태*[십이육]*, 5.18민주화운동*[오일팔]*' 등에서 볼 수 있듯이 기념일의 읽기는 'N1(Cca_b[+U])+∅+N2(Cca_b[-U])'의 규칙성을 가진다. 하지만 '6.10만세*[육십]*, 12.12사태*[십이십이]*'는 예외적으로 'N1(Cca_b[+U])+∅+N2(Cca_b[+U])'의 읽기 방식을 채택하고, '10.10절*[쌍십]*'은 특수한 경우이므로, 이 세 가지는 [그림 2]에서 소개된 Auto-TSS에서는 고유명사와 동일하게 처리한다.

럼 LAC2에 '전반|후반|점수|전적|스코어'이라는 어휘가 출현하거나, RAC1에 '(으)로/을/를/에게'이 포함되거나, RAC2에 '이기다|지다|완승|완패'라는 어휘가 나타나면 쌍점이 포함된 수사열은 '대비, 비율'을 의미하며 이때 쌍점은 [대]로 읽는다.

이상과 같이 구분된 규칙이 서로 상충되지 않고 적용되기 위해서 다음과 같이 규칙의 우선 적용 순서를 정할 수 있는 휴리스틱스가 필요하다.

① 아라비안 숫자와 온점, 쌍점, 빗금의 결합 패턴이 4.1.1에서 결정된 30개의 패턴화된 구조 중 특정 패턴화된 구조의 모든 검색 조건을 만족한다면 패턴화된 구조에 따른 기호의 문자화가 우선 적용된다.

② LAC에 아라비안 숫자와 이음표가 포함되어 있을 때, 아라비안 숫자가 '0'으로 시작되면 온점, 빗금은 다른 문맥 규칙에 관계없이 전화번호에 사용된 구분자로 인식되어 읽기 방식이 정해진다.

③ LAC1과 RAC1에 모두 아라비안 숫자가 포함되어 있을 때, LAC1와 RAC1의 수가 부정수에 속하는지[18]를 판단하여 기호와 부정수 숫자 읽기 방식을 택한 후, 다음 LAC2, LAC3나 RAC2, RAC3에 따라 결정되는 기호의 읽기 방식을 적용한다.

④ 한 기호의 LAC와 RAC가 서로 다른 문맥을 형성하는 경우, RAC가 선택하는 기호의 읽기 방식을 우선 적용한다.

⑤ 인접할수록 기호의 읽기에 영향을 미치는 정도가 크므로 우문맥, 좌문맥의 각각 적용되는 순서는 RAC1, RAC2, RAC3와 LAC1, LAC2, LAC3의 순서로 정한다.

4.2 온점, 쌍점, 빗금의 당연(default) 값

패턴, 문맥을 이용한 규칙 또는 휴리스틱스가 적용되지 못하는 경우, 모듈이 제시할 당연값은 다음과 같은 방식으로 정하였다. 분석 말뭉치에서 LAC1의 마지막 음절(LAC1-LAST)과 RAC1의 시작 음절(RAC1-FIRST)을 자료 특성에 따라 아라비안 숫자(NA), 라틴알파벳(LA), 한글(KA), 한자(CA), 문장부호(SYM), 빈칸(NULL)로 구분하고, 실제로 나타나는 조합 78개[19] 중 실제로 예를 찾을 수 있는 53개 쌍에 대해 각각 읽기 방식의 출현 비율을 살펴보았다. 그 일부를 살펴보면 [표 9]와 같다.

위와 같은 자료를 근거로 [표 10]과 같은 기호 읽기의 당연값을 정하였다.

---

[18] '3.5.7호선'과 같이 열거한 수는 LAC1<RAC1의 조건을 만족시켜야 한다. 부정수는 이 조건과 함께 'RAC1-LAC1=10^n,n=정수, n≥0'을 만족하여야 한다.

[19] 숫자와 동반하는 기호 읽기방식의 당연값을 얻기 위해 고려하는 기호와 좌우 음절 자료의 조합 쌍은 다음과 같이 정해진다. 1종류(LAC1 중 NA)×3개(기호 종류)×6종류+5종류(LAC1 중 NA를 제외한 나머지)×3개(기호 종류)×4종류(RAC1 중 NA, LA, SYM, NULL)

[20] [표 7]에서 NA+'/'+NA의 조합 쌍에서의 기호 읽기방식의 출현 빈도는 4가지 방식에 고르게 분포되어 있어 당연값을 구하기가 힘들다. '/' 읽기의 4가지 방식 중 웹 주소에 나타났을 때의 읽기 방식인 [슬래쉬]가 최대 빈도 검색이 용이한 웹 주소에서의 '/' 읽기를 규칙으로 처리하고, NA+'/'+NA 조합 쌍에서 '/' 읽기의 당연값은 [∅]로 결정하였다.

LAC1-LAST	RAC1-FIRST	기호 종류	읽기방식의 수							계
			[∅]	[점]	[대]	[슬래쉬]	[년월일]	[장절]	[나누기]	
NA	NA	.	1,249	9,307	0	0	134	0	0	10,690
NA	KA	.	35	0	0	0	5	0	0	40
NA	NA	:	0	0	32	0	0	9	0	41
NULL	NULL	:	413	0	5	0	0	0	0	418
NA	NULL	:	1	0	12	0	0	0	0	13
NA	NA	/	17	0	0	18	9	0	15	59
SYM	LA	/	57	0	0	84	0	0	0	141

[표 9] 기호의 좌우음절 종류에 따른 읽기방식의 출현 빈도(발췌)

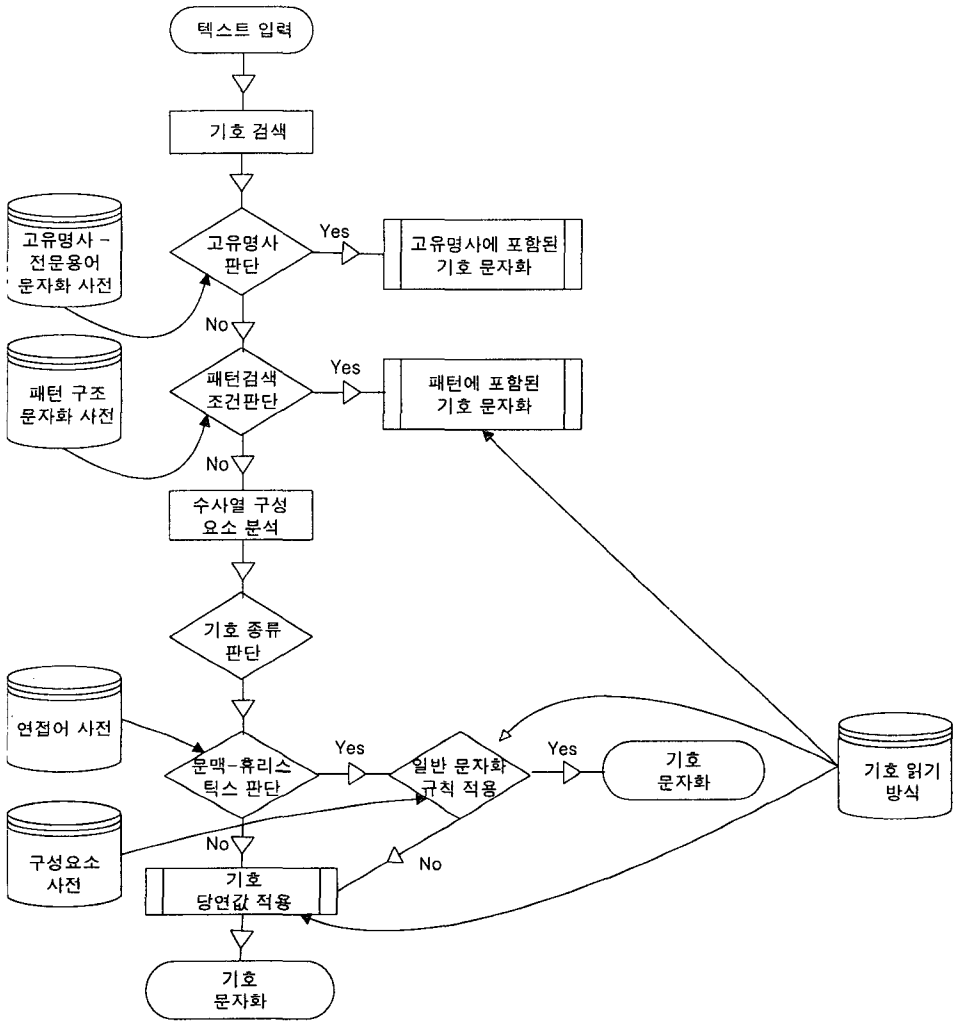
	LAC1-LAST	RAC1-FIRST	기호 종류	당연값 읽기방식
ⓐ	NA	NA	.	[점]
ⓑ	ⓐ를 제외한 조합 쌍		.	[∅]
ⓒ	NA	NA	:	[대]
ⓓ	NA	NULL	:	[대]
ⓔ	ⓒ, ⓓ를 제외한 조합 쌍		:	[∅]
ⓕ	NA	NA	/	[∅][20]
ⓖ	ⓕ를 제외한 조합 쌍		/	[슬래쉬]

[표 10] 기호 읽기의 당연값

4.3 온점, 쌍점, 빗금의 자동전사 시스템

4.1절에서 제시한 온점, 쌍점, 빗금의 패턴화된 구조 30개와 좌우 문맥 정보를 이용한 규칙 53개를 바탕으로 온점, 쌍점, 빗금의 일반 문자화 규칙 및 휴리스틱스를 처리할 수 있는 알고리즘을 만들고, 4.2절에서 밝힌 온점, 쌍점, 빗금의 문자화 당연값을 포함하여 [그림 2]와 같이 자동전사 시스템을 구현하였다.

온점, 쌍점, 빗금을 포함한 텍스트가 입력되면 ① 고유명사 및 전문 용어 사전을 통해 대회명, 지명 등에 고유명사 정보를 준다. 이 외 'IEEE802.11b'와 같이 패턴화되기 어려운 구조를 가진 고유명사에 포함된 기호를 전사한다. 다음으로, ② 패턴화된 구조 사전을 검색해 30개의 구조와 매칭되는 구조가 처리된다. 다음 단계로 ③ 좌·우 문맥 정보를 이용하여 중의성있는 구조를 검색하여 ④ 일반 전사 규칙을 적용하며, 규칙 간 적용 순서는 휴리스틱스를 따른다. ⑤ 패턴, 문맥 규칙 및 휴리스틱스 적용을 통해 전사되지 못하는 온점, 쌍점, 빗금은 각각 모듈이 제시하는 당연값을 적용하여 한글로 전사한 결과값을 갖는다.



[그림 2] 온점, 쌍점, 빗금의 자동 전사 시스템

5. 실험 및 평가

본 연구에서는 동일 신문 2년치(2000년 1월~2001년 12월) 기사에서 숫자와 함께 사용된 온점, 쌍점, 빗금을 포함한 어절 69,000여 건 중, 전체의 10%와 5% 크기의 말뭉치 세트 2개를 임의추출하고, 3장에서 제시된 온점, 쌍점, 빗금의 다양한 문자화 방식을 모두 포함하여 전체 말뭉치의 3.5%에 해당하는 크기의 균형 말뭉치를 구성하여 총 3개 세트를 실험하였다. 실험은 3개 세트를 자동 전사 시스템으로 자동 처리한 결과와 직접 분석한 결과를 비교 분석하여 자동 처리 결과의 정확도와 오류율을 밝히고,

균형 말뭉치를 대상으로 적용된 개별 처리 단계의 정확도를 밝혔으며, 자동 처리 결과 나타난 오류 유형을 분석하였다. 우선, 본 연구에서 개발된 자동 전사 시스템의 성능을 기존 TTS 시스템과 비교하기 위해, 동일한 실험 말뭉치로 C음성합성시스템, V음성합성시스템 및 본 연구에서 개발한 Auto-TSS로 실험한 결과 [표 11]과 같은 정확도를 보였다.

TTS 시스템의 종류	비율	평가 말뭉치1		평가 말뭉치2		균형 말뭉치	
		어절 수	비율	어절 수	비율	어절 수	비율
C시스템	정확	6,220개	89.1%	3,105개	88.9%	2,028개	82.7%
	오류	759개	10.9%	386개	11.1%	422개	17.3%
V시스템	정확	6,042개	86.6%	3,042개	87.1%	2,141개	87.4%
	오류	937개	13.4%	449개	12.9%	309개	12.6%
Auto-TSS	정확	6,583개	94.3%	3,248개	93.0%	2,307개	94.2%
	오류	396개	5.7%	243개	7.0%	143개	5.8%
계		6,979개	100%	3,491개	100%	2,450개	100%

[표 11] TTS 시스템의 평가

Auto-TSS의 성능은 기존 음성합성 시스템의 성능에 비해 정확도가 평균 5.2%~7.7%가 높았고, 균형 말뭉치를 실험한 결과, 본 자동전사 시스템의 정확도가 6.8%~11.5%만큼 높았다.

본 자동전사 시스템의 성능을 보다 구체적으로 분석하기 위해 [표 12]와 같이 균형 말뭉치를 대상으로 한 시스템의 개별 처리 단계를 통해 적용되는 기호 전사의 정확도를 분석하였다.

정확도 / 처리단계	각 단계에서 문자화된 어절 수	바르게 문자화된 어절 수	비율 (%)
고유명사	4	3	75
패턴화된 구조	22	13	59.09
문맥 정보	171	159	92.98
일반 문자화 규칙	2,029	1,982	97.68
당연값	224	150	66.96

[표 12] 전사 시스템 개별 처리 단계의 정확도

[표 12]에서 보는 바와 같이 중의성을 해결하기 위한 문맥 정보 규칙과 일반 전사 규칙의 정확도가 높았다. 규칙을 통해 대부분의 기호가 정확하게 전사되었다. 단, 패턴화된 구조 검색을 통한 처리 단계의 정확도가 낮은데 이에 대한 분석은 아래 오류 유형 분석에서 예문의 설명과 함께 기술될 것이다. 패턴화된 구조, 문자화 규칙 및 휴리스틱스로 처리되지 못한 기호는 전체 데이터의 9%(224/2,450)에 해당하며 당연값

을 적용하여 67%의 정확도로 전사되었다.

[표 13]은 Auto-TSS를 균형 말뭉치에 적용하는데 나타난 오류 유형을 보여준다.

정확도			어절 수
오류유형	자동 처리 결과	바른 읽기	
①	당연값 처리	[점], [∅]	15
②	[점]	[∅]	20
③	당연값 처리	[∅]	57
④	미처리, [점]	[∅]	6
⑤	[점], [∅]	[년월일], [∅]	19
⑥	[점], 당연값 처리	[∅], [년월일]	15
⑦	[∅]	[대]	2
⑧	고유명사		9
계			143

[표 13] Auto-TSS의 오류 유형

오류는 예문(15)와 같이 분석 데이터에 없는 패턴화된 구조나 좌·우 문맥정보를 가진 어절을 처리하지 못하는 것이 대부분이다.

(15) a. **8,815.00**루피아를 기록했다*[*8, 팔백십오 점 영영/팔천팔백십오점 영영]*

b. **5.16**의 평가가 이러하므로*[*오 점 일육/오∅일육]*

c. http://208.234.31.**19.02**-785-4411~3*[*일구 점 공이/일구∅공이]*

d. **2-②**. 이 결과 내리꽂는 공*[미처리/이의 이∅]*

오류 유형 ①은 소수 표현 뒤의 미분석된 기호를 포함한 어절을 처리하지 못하는 오류이다. 예문 (15-a)와 같이 숫자의 자리수 표현을 하는 반점(,)에 대한 규칙이 설정되어 있지 않으므로 온점을 중심으로 분석되는 좌우 숫자열의 당연값을 적용한다. 예문 (15-b)에서와 같이 오류 유형 ②는 '평가'가 기념일의 좌·우 문맥 정보로 분석되지 않아 발생하는 오류이다. 오류 유형 ③은 '주소·전화번호', '상호명·전화번호·웹주소'와 같이 숫자·영문자·문자열·기호가 띄어쓰지 않고 붙어서 나올 때 온점과 빗금이 구분자로 사용되나 전체 어절이 패턴화된 구조로 인식하기 어려운 경우이다. 따라서 예문 (15-c)와 같이 '.'의 읽기를 당연값으로 처리하여 발생한 오류이다. 예문 (15-d)에서 보는 것처럼 오류 유형 ④는 글의 제목 등의 순서를 표시하는데 사용되는 'N-원문자'와 같이 미분석된 패턴화된 구조를 처리하지 못한 오류이다. 당연값 적용에 있어서도 원문자는 기호의 좌우 음절 출현 자료 유형으로 포함시키지 못했기 때문에 당연값 적용도 실패한 경우이다.

①~④의 오류들은 규칙으로 일반화시킬 수 있는 것을 자동 전사 시스템에 추가로 적용하여 시스템의 성능을 향상시킬 수 있다. 그러나 예문 (16)의 경우, 새로운 규칙, 패턴 및 기호 좌우 음절 유형의 추가만으로 기호 읽기 방식을 결정하기는 어렵다.

(16) a. **1.5**군으로 브라질의 정예*[*일Ø오/일 점 오]*

b. 모델인 **2.5**, 월 마트에서 인기*[*이월 오일/이 점 오]*

c. **2000.1** 대만 염소 구제역*[*이천 점 일/이천년 일월]*

d. 한나라당 **7:** 무소속 후보 3*[*칠Ø/칠 대]*

e. '**3.3** 법칙'이 적용된다는*[*삼 점 삼/삼Ø삼]*

오류 ⑤는 시스템이 규칙을 잘못 적용하는 경우 발생한다. 예문 (16-a)에서는 기념일 규칙으로 주어진 문맥 정보 '군'이 오적용되어 'N.N'의 숫자·기호의 결합을 소수 표현으로 인식하지 못하고 기념일로 처리하였다. 반대로 예문 (16-b)는 소수 표현인 'N.N'에 대해 다음에 오는 '월, 일'을 기호를 [년월일]로 읽는 문맥 정보로 잘못 적용하였다. 오류 ⑥은 예문 (16-c)에서처럼 숫자와 기호의 결합이 'N.N'이거나 'N/N'과 같이 중의적인 구조이지만 좌·우 문맥 정보가 부족하여 잘못 처리되는 오류이다. 오류 ⑦은 예문 (16-d)에서 숫자와 기호의 결합이 'N:N'의 형태가 아니며 패턴화된 구조로 설정하기 어려워 쌍점을 [대]로 처리하지 못한 오류이다. 이러한 오류는 텍스트의 의미 분석이 되어야 해결될 수 있는 오류이다. 예문 (16-e)와 같은 미분석된 고유명사는 당연값이 적용되거나 규칙이 잘못 적용되어 기호의 읽기 방식의 결정에 오류가 빈번하다.

6. 결론 및 향후 연구

이상 본 연구에서는 신문 텍스트에서 숫자와 함께 나타나는 온점('.'), 쌍점(':'), 빗금('/')의 자동 문자화를 패턴화된 구조, 좌우문맥의 의미 분류 및 규칙에 기반하여 이들 기호의 자동 전사 시스템을 구현해 보았다. 실제 신문 텍스트 코퍼스를 시스템을 통해 자동 처리한 결과 정확도가 93.0~94.3%로 기존 TTS 시스템의 성능에 비해 5.2~7.7% 더 향상되었고, 균형 말뭉치의 실험 결과로는 기존 시스템보다 6.8~11.5%정도 정확도가 높았다. 본 자동전사 시스템은 균형 말뭉치 실험을 통해 나타난 오류율은 5.8%로, 실험 결과를 통해 나타난 오류 유형을 분석한 결과, 미분석된 좌·우 문맥 정보나 패턴화된 구조를 추가 적용시키고, 전화번호, 주소와 함께 나타나는 웹 주소 표현의 분리를 통해 4%의 오류를 교정할 수 있다. 따라서 문맥 부족에 의한 오류 및 고유 명사를 포함하더라도 실제 오류율은 1.8%로 아주 낮은 수치일 것으로 예상된다. 단, 새로운 문맥 정보와 패턴화된 구조를 추가할 때 시스템이 데이터에 나타난 정보를 자동으로 추출하여 학습할 수 있도록 기호 전사 학습 모델에 관한 연구가 계속 되어야 하며 현재 진행 중에 있다. 또한 전체 오류 중, 오류 유형 ⑤, ⑥과 같

이 패턴화된 구조가 중의적이거나 텍스트의 좌우 문맥 정보가 충분하지 못한 경우 나타나는 오류는 개별 규칙 적용의 정확도에 대한 통계적 정보를 통합한 방식으로 해결해야 할 것이다. 이러한 통합적 방식에 대한 연구는 향후 과제로 남긴다. 나아가 신문 텍스트에서 중의성을 가지는 온점, 쌍점, 빗금 기호 및 이음표 기호와 아라비안 숫자를 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.

<참고문헌>

권철홍. 2004. 음성합성 기술. *음성신호처리 기술 및 응용* (IDEC 공개강좌 자료집). 부산대학교 반도체설계교육지역센터. pp. 85–121.

김상준. 1986. 방송과 수의 표현. *KBS 표준방송언어*. 한국방송공사. pp. 125-151.

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

김형순. 2001. 음성공학과 언어학의 협동연구방향. *언어와 언어학* 27.1, 157–176.

유재원. 1999. 자연어 처리를 위한 수사의 하위 범주 분류. *제9회 한글 및 한국어 정보처리 학술대회 학술발표 논문집*, 136–142.

윤애선. 2002. 아라비안 숫자의 자동 전사를 위한 연어관계 연구. *2002년도 한국불어불문학회 동계학술대회 논문집*, 57–69.

윤애선, 권혁철. 2003. 한국어 텍스트에 사용된 이음표의 문자 전사. *언어와 정보* 7.1, 23-40.

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

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

이용주. 1995. 한국어 음성의 규칙합성을 위한 한글 자동 읽기 규칙의 구현. *원대논문집* 29.2, 225–239.

이은정. 1993. *최신 표준어·맞춤법 사전*. 백산출판사.

이정철. 2003. 음성합성(Text-to-Speech) 기술. *음성합성 기술 세미나 발표 자료*. 부산대학교 항공관 (2003년 10월 24일).

임동훈. 2002. 현행 문장 부호의 보완과 세칙안. *'문장 부호 세칙안' 마련을 위한 토론회 자료*. 국립국어연구원. pp. 25–32.

정영임, 김정세, 김상훈, 이영직, 윤애선. 2002. 현대 한국어에서 아라비안 숫자의 읽기 규칙 연구. *제14회 한글 및 한국어 정보처리 학술대회 학술발표논문집*, 16–23.

정영임, 강미영, 윤애선, 권혁철. 2003. 한국어 TTS시스템의 성능향상을 위한 문장 기호전사 시스템의 구현. *한국인지과학회 2003년도 춘계학술대회*, 195–200.

채완. 1983. 국어 수사 및 수량사구의 유형적 고찰. *어학연구* 19-1, 19–34.

채완. 2002. 세칙안의 마침표와 쉼표. *'문장 부호 세칙안' 마련을 위한 토론회 자료*. 국립국어연구원. pp. 42–49.

The Unicode Consortium. 1996. *The Unicode Standard, Version 2.0*, Addison Wesley Developers Press.

보이스웨어 TTS 시스템. http://www.voiceware.co.kr
코아보이스 TTS 시스템. http://www.corevoice.com/
동아일보 보이스 뉴스. http://www.donga.com/
매일경제 음성 기사 서비스. http://www.mk.co.kr/

접수 일자: 2004년 5월 6일
게재 결정: 2004년 6월 14일