결정 트리를 이용한 '이음표' 문자화의 중의성 해소

정영임° 이동훈 남현숙* 윤애선 권혁철
부산대학교 한국어정보처리연구실, 동서사이버대학교 인터넷학과*
{acorn°, huni77, asyoon, hckwon}@pusan.ac.kr, nosenam@bdu.ac.kr*

Disambiguating in Transliteration of some Text Symbols using Decision tree

Youngim Jung°, Donghun Lee, HyeonSook Nam*, Aesun Yoon, Hyukchul Kwon
Korean Language Processing Lab at Pusan National University
Dept. of Internet contents at Busan Digital University*

요 약

한국어 텍스트 음성합성에서 문장 기호의 문자화에 나타나는 오류는 기호의 중의성에 기인한다. 선행연구에서 규칙에 기반하여 중의성을 해결하는 방안이 제안되었으나 여전히 기호는 다양한 문맥에서 높은 중의성을 가지고 문자화된다. 따라서 본 연구에서는 신문 텍스트에 나타나는 문장 기호 중 이음표의 문자화를 이음표를 포함한 어절의 패턴, 패턴의 좌우에 위치하는 어절 정보 및 휴리스틱스 자질을 학습하여 제시된 이음표의 문자화의 중의성을 해소하는 방안을 제안하였다. 이를 위해 국내 1개 일간지 2년 치 기사에서 이음표를 포함한 어절 49,000여 개를 임의 추출하여 분석하였고, 분석된 자질을 자동추출하여 결정 트리를 구성하였다. 실험 결과, 96.2%~97.7%의 정확도를 보였다.

1. 서 론

한국어 텍스트 음성합성에서 숫자, 문장 기호, 로마자 등의 비-문자 정보의 정확한 한글 변환은 음성합성기의 명료한 합성음 판단의 중요한 선결 조건이 되며 기존 TTS 시스템이 해결해야할 문제이다[1, 2]. 특히, 아라비아 숫자와 결합되어 사용되는 문장 기호의 읽기는 문맥에 따라 3~6가지로 결정되며, '1-3위 [**일에서** 삼 위]', '1-3[일 **대** 삼]으로 지다', '10:30[**열 시** 삼십 **분**]', '1:2[일 **대** 이]의 비율'과 같이 기호의 읽기에 따라 합성음의 의미는 크게 달라진다[2, 3, 4]. 또한 역으로 기호와 결합한 숫자의 읽기가 단독 숫자의 읽기와는 달라 숫자와 결합한 기호에 따라 숫자 읽기 시스템의 정확도에 많은 영향을 미친다[1].

이에 본 연구에서는 신문 텍스트에 나타나는 문장 기호 중 가장 중의적으로 사용되는 이음표의 읽기를 결정하는 자질을 자동추출하고 추출된 자질의 학습을 통해 이음표 문자화의 중의성을 해결한다. 이를 위해 사용된 말뭉치는 1개 신문 2년 치(2000년 1월~2001년 12월) 기사에서 숫자와 결합한 이음표를 포함한 49,000여 개 어절이다. 여기서 추출한 이음표 포함 어절의 좌우연접어, 숫자와 이음표 결합 패턴 및 숫자 크기 등의 자질 및 자질의 적용 순서를 학습하여 모델을 구축한다. 구축된 모델의 평가는 10-fold cross-validation의 측정과 동일 신문에서 이음표를 포함한 어절 1,000개를 임의추출한 평가용 말뭉치의 실험을 통해 이루어진다.

2장에서는 기존 TTS 시스템에서 나타나는 이음표 전사 오류와 선행 연구의 문제점을 살펴본다. 3장에서 이음표의 7가지 읽기의 분류에 결정적인 영향을 미치는 자질과 자질값을 설정하고 설정된 자질의 적용 순서를 결정 트리를 이용하여 구현한다. 4장에서는 구현된 모델의 정확도를 알아보고, 5장에서는 시스템 성능 향상 방안과 후속 과제를 제시한다.

2. 선행연구 및 문제점

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 국어학적 의미 구분만을 하고 있으며 '~'는 '내지'라는 뜻으로 쓰이거나 '어떤 말의 앞, 뒤로 돌어갈 말 대신'에 사용된다고 밝히고 있다[5]. 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 다양하나 이음표의 용법과 문자화에 대한 문법 기술이 충분하지 않다. 또한 전산 언어학 분야에서는 이루어진 이음표의 다양한 문자화에 대한 연구가 아직 실제적으로 적용이 되지 않아 현재 제공되고 있는 TTS시스템의 정확도가 매우 떨어진다.

실제 TTS시스템을 이용하여 음성기사 서비스를 제공하고 있는 D신문, M신문과 V음성합성 시스템은 이음표의 읽기를 <표 1>과 같이 매우 간략한 규칙으로 처리하고 있어 (1)~(10)과 같은 오류가 나타난다[4].

<표 1 기존 TTS시스템의 이음표 읽기>

	'-'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) .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)

1) 부산대학교 한국어정보처리연구실에서 구축한 규칙 기반 숫자 읽기 시스템(Auto-TAN)을 이용해 J일보 2년 치 기사에 포함된 숫자 표현 1,209,598개 어절의 문자화 정확도를 측정하였다. 그 결과, 오류로 나타나는 28,952개 어절 중 전체 오류의 31.4%에 해당하는 9,104개 어절이 기호를 포함한 숫자 표현으로 나타나 숫자 읽기 시스템의 정확도에 큰 영향을 미침을 알 수 있었다.

본 연구의 선행 연구로 이루어진 [3, 4]에서는 이음표의 다양한 문자화를 결정하는 규칙을 설정하여 규칙에 기반한 문자화 시스템을 구현하였다. 구현된 시스템의 정확도는 95.5%로 높은 정확도를 보였다.

그러나 규칙에 기반한 방법은 미등록 단어가 좌우 연접어로 오거나 2byte 기호, 웹주소 등이 바로 연결되어 기호 포함 어절의 패턴이 기존 규칙에 기술되지 않은 패턴으로 인식될 경우 그에 대한 처리를 할 수 없고 새로운 규칙을 계속해서 수작업으로 추가해야 하므로 비용과 노력이 상당하게 든다. 또한 좌우문맥이 불충분하여 규칙을 적용할 수 없는 경우나 규칙이 서로 상충되어 데이터에 적용시키기 어려운 경우에도 규칙에 기반한 방법은 한계를 가진다.

3. 이음표 문자화 학습

3장에서는 이음표 문자화 학습을 위해 이음표의 문자화에 영향을 미치는 요인을 분석하고 그 자질을 자동 추출하여 결정트리를 이용해 추출된 자질의 적용 순서를 구현한다.

3.1 자질의 추출

이음표의 문맥에 따른 문자화는 <표 2>와 같이 7가지로 분류된다[2].

<표 2 이음표 문자화>

종류	읽기	SRF	예
'-, ~'	영형태	SRF1	'3-4-3전형', 'F-16'
		SRF2	'3~4개', '20-30세'
	'에서'	SRF3	'9-11월', '3~6명'
	'영형태, 에서'	SRF4	'02-3422-1500~3'
'–'	'마이너스'	SRF5	'-34%'
	'대'	SRF6	'3-1로 이기다.'
	'의'	SRF7	'200-1번 버스'

숫자와 결합하여 다양한 의미를 가지고 사용되는 이음표는 문맥 내의 연접어와의 의미적 공기 관계에 의해 중의성이 해결된다는 특성 외에도 기호의 특성상 숫자와의 결합 패턴, 숫자의 자리 수, 숫자의 크기 등과 같은 자질을 통해 이음표 문자화의 중의성이 해결될 수 있다. 이에 따라 이음표 문자화를 결정할 수 있는 자질은 <표 3>과 같이 분석할 수 있다.

<표 3 자질과 자질값>

Feature			value
LAC	Null	1	0: Null, 1: Space
	Alphabets	2	0: a~z, 2: D
	Symbols	3	0: (,), [, 1: ', ', ", ", 2: 2byte 기호, 3: 문장기호(이음표 제외)
	한글, 한자	4	0: default, 1: 인명, 2: 국가명, 3: 지역명, 4~30: 시간, 번호, 수량표현 등 의미 영역에 따른 세부 범주
RAC	Null	5	0: Null, 1: Space
	Alphabets	6	0: a~z, 1: 도량형 기호로 사용된 영문자(열)
	Symbols	7	0:),], 1: ', ", 2: 2byte 기호, 3: 문장기호(이음표 제외), 4: 도량형 기호
	한글, 한자	8	0: default, 1~24: 시간, 스포츠, 수량 표현 등 의미 영역에 따른 세부 범주
Pattern	이음표 종류	9	0: '~', 1: '-', 2: '-, ~'
	이음표 개수	10	0: default, 1: 1, 2: 2, 3: 3
	숫자의 개수	11	0: default, 1: 1, 2: 2, 3: 3, 4: 4
Heuristics	숫자의 크기	12	0: default, 1: 0<x<61, 2: 0<x<25, 3: 0<x<32, 4: 0<x<13, 5: 1950<x<2010
	두 숫자의 차	13	0: default, 1: $(y-x)10^n=1$, 2: y-x>0
	숫자의 자리수	14	0: default, 1: 2자리, 2: 3자리, 3: 4자리
	처음 숫자	15	0: default, 1: 0
			(n, x, y: 0≤n, x, y 인 정수)

<표 3>에서 분석된 LAC와 RAC 자질 중 '4, 8'은 이음표를 포함한 어절의 좌우에 나타나는 단어를 의미별로 범주화하여 각각 27개, 24개의 자질값을 부여하였다. default[3]를 제외하고 각 자질값에 해당하는 단어는 목록화하여 각 목록에 있는 단어가 검색되면 해당 자질의 값이 추출된다. 그 외의 자질은 명세된 항목과 비교하여 자질값이 자동으로 추출된다. 분석된 자질은 자질값에 따라 이음표의 문자화가 SRF1~SRF7 중 한 가지로 결정되어야 중의성 해소에 적합하다고 할 수 있다. <표 4>는 각 자질에 따른 이음표의 문자화를 통계 수치로 나타내었다.

<표 4 자질과 이음표 문자화>

이음표 문자화	자질과 자질값								
	15	9			13		11		
	1	0	1	2	1	2	0	1	2
SRF1	100	0	3	0	0	0	0	1.2	59.9
SRF2	0	0	0	0	43.7	0	0	0	0
SRF3	0	85.4	0	0	0	56.3	0	7.7	0
SRF4	0	0	0.3	11.3	0	0	0	0	0
SRF5	0	0	0	0	0	0	6.5	0	0
SRF6	0	0	0	0	0	0	0	21.8	0
SRF7	0	0	0	0	0	0	0	2.9	0

<표 4>에서 '15', '13'는 그 자질과 자질값에 따라 선택되는 SRF가 단일하지만 '9(1)'과 '11(1)'의 자질값은 이음표의 문자

2) 본 논문에 사용된 영어 약자는 아래와 같이 사용되었다. 'SRF(Symbol Reading Formulae)', 'LAC(Left-Associated collocation)', 'RAC(Right-Associated Collocation')

3) 자질값 중 default는 명세화되지 않은 값을 모두 포함한다.

화가 SRF1, SRF3, SRF6, SRF7로 복수로 선택된다. 이러한 경우, 단일 자질만으로 이음표 문자화를 중의성을 해결할 수 없으며 자질의 가중치에 따라 다음 단계에서 다른 자질을 누적 적용해야 한다.

3.2 결정 트리의 적용

앞에서 분석한 자질은 이음표 문자화를 결정하는데 상이한 가중치를 가진다. 문자화를 결정하는 자질의 가중치는 각 문자화를 구별하는 정보값(Information Gain)으로 나타낼 수 있다. 따라서 각 자질의 정보값을 얻고 자질의 적용 순서를 위해 C4.5 알고리즘을 이용하였다.

$$info(S) = -\sum_{j=1}^{k} \frac{freq(C_j, S)}{|S|} \times \log_2\left(\frac{freq(C_j, S)}{|S|}\right)$$

$$info_X(T) = \sum_{i=0}^{n} \frac{|T_i|}{|T|} \times info(T_i)$$

$$gain(X) = info(T) - info_X(T)$$

S: 이음표 예제 집합, X: 자질, T: training 집합

C_j: 예제가 속하는 클래스

예제 집합 S의 엔트로피 info(S)에서 training을 통해 얻어진 각 속성값의 엔트로피의 합을 빼면 X자질이 가지는 정보값을 구할 수 있다. 이 정보값이 높을수록 그 자질은 클래스를 보다 잘 구별할 수 있는 자질이며 C4.5 알고리즘은 정보값이 최대가 되는 자질을 선택하여 다음 단계로 나가게 된다[6, 7].

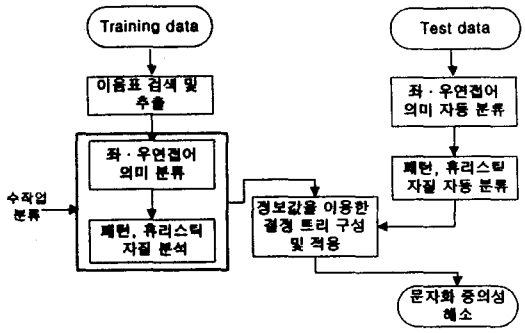
<표 4>에서 단일 문자화가 되지 않았던 자질 '9(1)', '11(1)'은 결정 트리에서 출력값에 따라 T를 {T_1, T_2, ...T_n}로 세부 분류하고 각각의 노드를 구별할 수 있는 다른 자질을 분석해서 각 {T_1, T_2, ...T_n}가 {C_1, C_2, ..C_n}의 단일 클래스로 분류될 수 있도록 트리를 구성한다. <표 5>는 다른 자질을 누적 적용하여 이음표 문자화가 결정되는 것을 보여준다.

<표 5 자질의 누적 적용>

이음표 문자화	자질과 자질값							
	9(1)			11(1)				
	2(0)	10(2)	15(0)	10(2)	8(12)	8(19)	8(20)	8(14)
SRF1	0.1	2.9	0	1.2	0	0	0	0
SRF2	0	0	0	0	0	0	0	0
SRF3	0	0	0	0	7.7	0	0	0
SRF4	0	0	0.3	0	0	0	0	0
SRF5	0	0	0	0	0	0	0	0
SRF6	0	0	0	0	0	8.4	13.4	0
SRF7	0	0	0	0	0	0	0	2.9

3.3 시스템 구현

<그림 1>은 본 시스템의 학습 및 학습한 모델을 실제 데이터 적용하는 단계를 간략히 보여준다. training 데이터에서 수작업으로 분석된 이음표 문자화 자질은 각각의 정보값과 적용 순서를 가지고 결정 트리를 구성한다. 실제 테스트 데이터에 나타나는 이음표의 패턴, 휴리스틱 및 좌우연접어 자질은 구성된 결정 트리의 적용을 통해 이음표 문자화가 중의성이 없이 결정된다.



<그림 1 시스템 구성도>

4. 실험 및 평가

실험은 분석한 데이터와 동일한 신문 기사에서 숫자와 결합한 이음표 12,692개 어절을 임의추출하여 10-fold cross-validation을 측정하였고, 다시 테스트 데이터 1000건을 임의추출하여 실험하였다. 10-fold cross-validation 결과는 96.23%의 정확도를 보였으며, 테스트 데이터 정확도는 97.7%를 보였다.

5. 결론 및 향후 연구

이상 본 연구에서는 신문 텍스트에서 이음표('-', '~')의 문자화를 결정하는 이음표의 좌우문맥, 이음표와 숫자의 결합 패턴 및 휴리스틱 자질을 자동 추출하고 C4.5 알고리즘을 이용해 각 자질의 정보값을 얻어 결정 트리를 구성하였다. 추출한 자질과 트리를 적용하여 모델을 학습한 결과 96.2~97.7%의 정확도를 보였다. 이는 규칙 기반 시스템 정확도인 94.5~96.1%보다 0.1~3.2% 높은 결과이며 다양한 문맥에 사용되는 이음표의 중의성을 해소하는데 학습을 통한 방법이 비용과 시간을 절약하며 규칙에 기반한 방법보다 정확도 역시 더 높을 수 있음을 보여준다.

본 연구 결과를 바탕으로 신문 텍스트에서 이음표 외 기호로 중의성을 가지는 '.', ':' 및 '/' 기호 및 아라비안 숫자의 문자화 중의성 해소를 위한 학습에 대한 연구가 향후 지속되어야 할 것이다.

참고 문헌

[1] 이정철(2003), "음성합성(Text-to-Speech) 기술", 10월 24일 부산대학교 항공관(제8공학관) 음성합성 기술 세미나 발표 자료.

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

[3] 윤애선, 권혁철(2003), "한국어 텍스트에 사용된 이음표의 자동 전사", 한국언어정보학회, 『언어와 정보』, pp.23~40.

[4] 정영임, 정휘웅, 윤애선, 권혁철(2003), "한국어 음성 합성을 위한 '이음표'의 문자 전사", 『제30회 한국정보과학회 춘계 학술발표회 논문집(B)』, pp.558~560.

[5] 이희승, 안병희(2001) 『새로 고친 한글 맞춤법 강의』, 신구문화사, 서울.

[6] J. Ross Quinlan(1993), 『C4.5 : programs for machine learning』, Morgan Kaufmann Publishers, San Mateo, Calif.

[7] Tom M. Mitchell(1997), 『Machine Learning』, McGraw-Hill