효율적인 기계학습 자질 선별을 통한 한국어 운율구 경계 예측 모델의 성능 향상

(Performance Improvement of a Korean Prosodic Phrase Boundary Prediction Model using Efficient Feature Selection)

김 민 호 † 권 혁 철 ††
(Minho Kim) (Hyuk-Chul Kwon)

요 약 운율구 경계 예측은 대화체 음성합성을 실현하기 위한 주요한 자연언어처리 기술 중 하나이다. 본 논문은 자연스러운 한국어 운율구 경계 예측을 실현하고자 기존의 학습 자질을 대신할 새로운 학습 자질을 제안한다. 이 새로운 자질들은 기존의 학습 자질보다 실제 언어생활에서 운율구 경계 발생에 영향을 미치는 여러 요인을 더 잘 반영한다. 특히, 수작업으로 구축한 운율구 경계 예측 규칙을 이용하여 추출한 학습 자질은 높은 정확도 향상에 이바지한다. 본 논문에서 제안한 새로운 학습 자질을 바탕으로 CRFs(Conditional Random Fields)를 이용하여 운율구 경계 예측 모델을 만들었다. 그 결과 3단계 운율구 경계(강한 경계, 약한 경계, 운율구 내부 비경계) 예측에서 86.63%의 정확도를, 6단계 운율구 경계(상승조/하강조 강한 경계, 상승조/하강조/평탄조 약한 경계, 운율구 내부 비경계) 예측에서는 81.14%의 정확도를 보였다.

키워드 : 한국어 운율구 경계 예측, 규칙, 기계학습, Conditional Random Fields (CRFs)

Abstract Prediction of the prosodic phrase boundary is one of the most important natural language processing tasks. We propose, for the natural prediction of the Korean prosodic phrase boundary, a statistical approach incorporating efficient learning features.These new features reflect the factors that affect generation of the prosodic phrase boundary better than existing learning features. Notably, moreover, such learning features, extracted according to the hand-crafted prosodic phrase boundary prediction rule, impart higher accuracy. We developed a statistical model for Korean prosodic phrase boundaries based on the proposed new features.The results were 86.63% accuracy for three levels (major break, minor break, no break) and 81.14% accuracy for six levels (major break with falling tone/rising tone, minor break with falling tone/rising tone/middle tone, no break).

Key words : Korean Prosodic Boundary Prediction, Rule, Machine Learning, Conditional Random Fields (CRFs)

1. 서 론

화면의 내용과 자신이 입력한 키보드 정보나 마우스 좌표 등을 음성으로 알려 주어 컴퓨터를 사용할 수 있도록 도와주는 프로그램인 스크린 리더(screen reader)는 시각 장애(시력 0.3 이하 또는 시야가 10° 이하), 난독증, 언어장애, 학습장애 등을 가진 사람에게 유용한 프로그램이다. 스크린 리더의 사용자 만족도에 가장 큰 영향을 주는 부분은 자연스러운 음성합성(speech synthesis) 실현이다. 1990년대 초 음성합성 기술은 ARS 서비스처럼 사용자에게 들려줄 안내 멘트를 미리 녹음하였다가 재생하여 주는 간단한 기술이 주를 이루었다. 그러나 2000년대 중반에 들어 그 쓰임이 점차 늘어나면

· 이 논문은 2007년도 정부(교육과학기술부)의 재원으로 한국연구재단의 자원을 받아 수행된 연구임(2009-0083761)

† 학생회원 : 부산대학교 컴퓨터공학과
karma@pusan.ac.kr
†† 종신회원 : 부산대학교 정보컴퓨터공학부 교수
hckwon@pusan.ac.kr

논문접수 : 2010년 6월 4일
심사완료 : 2010년 9월 27일

정보과학회논문지: 소프트웨어 및 응용 제37권 제11호(2010.11)

서 스크린 리더나 E-Book처럼 임의의 문장에 대해서도 합성할 수 있는 기술이 시장에서 요구되고 있다. 이러한 무제한 합성 기술은 하나의 차분한 목소리로 합성음을 생성해내는 낭독체 음성합성뿐만 아니라 전달하고자 하는 메시지의 내용과 듣는 이의 감정에 따라 음색과 억양이 다르게 표현되는 대화체 음성합성을 목표로 한다. 대화체 음성합성을 위해서는 정확한 자연언어처리 기술이 뒷받침되어야 하는데, 가장 중요한 자연언어처리 기술 중 한 가지는 끊어 읽는 위치를 예측하는 운율구 경계 예측 기술이다.

운율구(韻律句)란, 실제 언어생활에서 화자가 긴 문장을 소리 내어 말할 때 생기는 발음 단위로써 문장 안에서 하나의 어절 또는 구절에 해당한다. 이러한 운율구는 화자가 가진 문법 지식이나 말의 길이 및 속도 등 여러 가지 요인에 의해 다양하게 형성되기 때문에 인간의 자연스러운 발화를 기계가 흉내 내기란 어렵다. 그러나 운율구 경계를 인간의 발화에 가깝도록 예측할 수 있다면 기계의 자연스러운 발화에 도움이 된다. 특히, 음성합성 시스템에서 정확한 운율구 경계 예측은 어조와 강세의 생성, 음소의 지속과 휴지 판단 등에 사용되어 자연스러운 음성합성에 큰 도움을 준다[1].

운율구 경계 예측에 대한 방법론은 크게 규칙 기반 접근법과 통계적 접근법으로 나뉜다. 전자는 운율구 경계 예측을 위한 언어정보를 규칙으로 만들어 이용하는 방법이다. 후자는 말뭉치로부터 이끌어낸 통계 수치를 활용하여 운율구 경계 예측에 적합한 통계 모형을 구축하여 이용하는 방법이다.

운율구 경계 예측 규칙은 전문가가 완전히 수작업으로 만들거나, 말뭉치를 이용하여 자동 혹은 반자동으로 만든다. 수작업에 의한 규칙은 특정 언어 자원에 대해 독립적이고 규칙의 정확도도 높지만, 규칙 구축에 많은 시간과 노력이 요구되므로 실제 언어생활에서 발생하는 많은 예외 현상을 처리하기 어렵다는 단점이 있다. 반면에 자동으로 만든 규칙은 규칙의 구축이 쉽고 적용 범위가 넓다는 장점이 있지만, 특정 말뭉치에 의존적이라는 문제점이 있다.

통계적 접근법은 대용량의 운율구 경계 분석 말뭉치를 이용하여 확장성이 좋고 적용 범위가 넓으며 전체적인 정확도가 비교적 높은 통계 모형을 구축할 수 있다는 장점이 있다. 하지만, 의미 있는 통계 정보를 추출하려면 일정 크기 이상의 신뢰할 수 있는 분석 말뭉치를 구축하여야 하기 때문에, 말뭉치 구축에 시간과 노력이 많이 요구된다. 특히, 운율구 경계 분석은 주석자의 주관적 판단에 많이 의존하므로 주석의 일관성 유지를 위해 더 많은 노력을 요한다.

본 논문은 통계적 접근법을 이용한 운율구 경계 예측에서 더 좋은 예측 결과를 도출할 수 있는 새로운 학습 자질을 제안한다. 2장에서 운율구 경계 예측의 국내외 여러 연구에 대해서 정리하고, 3장에서는 실제 언어생활에서의 운율구 경계 형성 특징을 반영한 학습 자질에 대해서 설명할 것이다. 4장에서는 3장의 여러 학습 자질에 대한 효율성을 실험적으로 증명하고, 마지막으로 5장에서 결론 및 앞으로의 연구 방향에 대해 논할 것이다.

2. 선행 연구

통계적 접근법을 이용한 운율구 경계 예측 시스템을 구현하고자 국내외 연구 기관에서는 운율구 경계 분석 말뭉치에서 다양한 학습 자질을 추출하여 기계학습 알고리즘에 적용하고 있다.

선행 연구에서 주로 사용한 학습 자질로는 (1) 어절의 품사 정보, (2) 어절의 길이, (3) 문장 내 특정 위치로부터 현재 어절까지의 거리 등이 있다.

Lee and Oh[2]에서 사용하는 품사 집합에 포함된 품사 개수는 23개, 김병창 외[3]에서는 32개로 세종 말뭉치에서 사용한 품사 집합보다 덜 세분된 품사 집합을 사용하였다. 정희정[4], 김현권[5]에서 밝힌 바와 같이, 명사와 부사는 통사적 기능에 따라 다시 하위범주화할 수 있다. 그러나 기존 연구에서는 품사 결합 정보를 자질로 사용할 때, 세분된 품사 집합을 사용하지 않아 문장의 구문적 특성을 파악하기는 어렵다. 문장의 통사적 구조가 운율 구조와 밀접한 관계를 맺었음을 볼 때 [1,6-8] 덜 세분된 품사 집합만으로는 운율구 경계 추정을 위해 의미 있는 학습 결과를 얻기 어렵다.

선행 연구에서 거리 자질은 문장의 처음 혹은 끝으로부터 현재 어절까지의 거리[9], 그것을 정규화한 수치[10,11], 현재 문장에 포함된 이전 구두점으로부터 현재 어절까지의 거리[9], 가장 가까운 선행 의존소/지배소로부터 현재 어절까지의 거리[2]와 같이 다양한 방법으로 추출되었다. 그러나 문장이 길수록 문장의 시작/끝에서 현재 어절까지의 거리보다는 현재 어절 이전에 발생한 운율구 경계로부터의 거리가 중요하다. 또한, 문장 기호의 용법이 매우 중의적이고 용법에 따라 발음과 휴지의 발생 여부가 달라져[12] 이러한 자질로부터 얻어진 정보가 의미가 없는 경우가 많다. 또, 지배소/의존소로부터의 거리는 구문 분석기의 성능에 크게 의존한다는 단점이 있다.

3. 운율구 경계 예측을 위한 학습 자질

운율구는 화자가 가진 여러 요인-발화 초점, 문장의 길이, 화자 개인의 신체 조건, 정서 상태, 발화 목적 및 스타일 등에 따라 그 경계의 종류와 위치가 결정된다. 본 장에서는 운율구 경계 형성에 영향을 미치는 여러

정보를 학습 자질로써 활용하는 방안에 대해 제시한다.

3.1 품사 정보

운율구 경계 형성에 영향을 미치는 정보 중 가장 중요한 정보는 현재 어절과 좌우 문맥에 나타난 어절의 품사 정보이다. 예를 들어, 현재 어절이 '은/는'과 같은 보조사로 끝날 경우 운율구 경계가 올 가능성이 크며, 다음 어절의 첫 형태소가 보조 용언으로 시작하면 현재 어절 다음에 붙여 읽을 가능성이 크다. 이러한 품사 정보를 학습 자질로 사용하려면 다음과 같은 사실이 먼저 고려되어야 기계 학습에 효과적으로 이용될 수 있다.

• 품사 세분화: 일부 명사는 다른 명사를 수식하는 관형적인 용법을 보이기도 한다. 예를 들어, '특별', '최종', '완전'과 같은 명사는 홀로 사용되지 않고 접두사처럼 다른 명사와 결합하여 복합명사를 이룬다. 따라서, 현재 어절이 접두명사로 끝나면 다음에 오는 명사와 붙여 읽을 가능성이 크다[13].

본 논문에서는 표 1과 같이 명사를 통사적, 의미적 성격에 따라 세분화하고 부사는 어떤 성분을 수식하느냐에 따라 세분화하였다. 또한, 연결어미도 '절간 연결만 가능한 연결어미'와 '어휘 간 연결도 가능한 연결어미'로 구별하였다.

표 1 품사 세분화

품사	세분화	
명사	명사성 명사	일반명사
		고유명사
	관형성 명사	접두명사[1]
	부사성 명사	시간명사
		정도명사
부사	명사 수식 부사	
	형용사 수식 부사	
	동사 수식 부사	
	문장 수식 부사	
연결어미	절 간 연결어미	
	통용 연결어미	

• 품사 정보 단위: 품사 정보를 학습 자질을 이용할 때 어절 전체의 형태소 결합을 하나의 값으로 볼지 아니면 어절을 이루는 각각의 형태소를 하나의 값으로 볼지를 결정하여야 한다. 그리고 만약, 어절을 이루는 각각의 형태소를 하나의 값으로 본다면 어절의 모든 형태소를 값으로 이용할지 아니면 첫 형태소나 끝 형태소와 같은 특정 형태소만을 값으로 이용할지도 고려되어야 한다.

운율구는 대부분 어절의 첫 형태소와 끝 형태소에 따라 그 경계의 종류와 위치가 결정된다. 따라서 본 논문에서는 어절의 첫 형태소와 끝 형태소의 품사 정보만을 이용하였다. 단, 첫 형태소와 끝 형태소가 아닌 형태소이면서 운율구 경계에 영향을 미치는 형태소를 위해 '지정사 포함 여부', '용언화접미사와 결합 여부' 등과 같은 별도의 학습 자질도 추가하였다.

• 품사 정보를 참조하는 어절의 수: 문장 내 의미정보 단위인 통사구와 문장 내 발음 단위인 운율구는 서로 밀접한 관계를 맺는다는 것은 이미 언어학 및 음성처리 분야의 여러 연구에서 보고되었다[1,6-8]. 예를 들어, '다른 통사구 사이의 경계', '구와 절 사이의 경계' 그리고 '절과 절 사이의 경계'는 강한 운율구 경계가 발생한다. 따라서, 현재 어절이 이전 어절과 통사구를 이루는지를 통계적으로 판단할 수 있다면 운율구 경계 예측에 많은 도움이 된다. 통사구 경계는 인접한 형태소의 품사 정보를 통해 간접적으로 판단할 수 있다. 이때 판단 근거가 되는 인접 형태소와의 거리는 실험을 통해 적절한 값을 찾아야 정확도를 높일 수 있다.

3.2 거리 정보

인간의 발음 단위는 비슷한 크기를 가진다. 이는 인간이 한 번의 호흡으로 말할 수 있는 물리적인 조건에 제약이 있기 때문이다[14-16]. 이러한 인간의 발화 특성을 반영한 것이 거리 정보를 이용한 자질이다.

선행 연구에서는 문장의 앞이나 혹은 뒤에 가까울수록 운율구 경계가 있을 가능성이 작으리라고 보았다. 이를 위해 문장의 시작에서부터 현재 어절까지의 거리, 문장 끝에서부터 현재 어절까지의 거리[9], 문장 내에서 현재 어절이 차지하는 위치를 정규화한 수치[10,11] 등을 학습 자질로 이용하였다. 그러나 2장에서 살펴본 바와 같이 문장의 시작/끝에서부터의 거리나 문장 기호로부터의 거리 정보만으로는 현재 어절 다음에 어떤 운율구 경계가 올지 정확하게 예측하기 어렵다.[2] 따라서, 문장 시작과 운율구 경계 모두 거리 정보의 기준점으로 정할 필요가 있다. 본 논문에서는 운율구 경계를 거리 정보의 기준점으로 정하고자 운율구 경계 예측 규칙을 활용하였다.

정영임[13]에서 자연스러운 한국어 운율구 경계를 예측하기 위해 세분화된 문장 성분 간 의존관계를 이용하여 통사구를 추출한 다음, 추출한 통사구의 유형에 따른 운율구 경계 예측 규칙을 수작업으로 구축하였다. 이렇게 구축된 규칙은 특정 언어 자원에 대해 독립적이고 규칙의 정확도도 높다. 특히, 다음의 상관관계를 바탕으로 한 강한 운율구 경계에 관한 규칙은 그 정확도가 90.75%이다.

1) 본 논문에서는 접두사처럼 다른 명사와 결합하여 복합명사를 이루는 명사를 '접두명사'라고 이름을 붙였다.

2) 이는 4.2절에서 실험을 통해 증명할 것이다.

- 다른 통사구 사이의 경계, 구와 절 사이의 경계, 그리고 절과 절 사이의 경계는 강한 운율구 경계가 발생한다.
- 수식언과 피수식언 사이에 다른 성분이 끼어들면 수식언 다음에 강한 운율구 경계가 발생한다.
- 독립언과 문장의 주절 사이에는 강한 운율구 경계가 발생한다.

3.3 연어 형성 정보

단어는 문장에서 무작위로 나타나는 것이 아니라, 다른 단어와 함께 나타나는 경향이 있다. 이렇게 단어들이 자주 함께 나타나는 현상을 연어(collocation)라고 한다. 여기서 "자주 함께" 나타난다는 것은 단순히 말뭉치상에서의 절대적 빈도로 많고 적음이 아니라, 그것이 예상보다 많은가 혹은 그렇지 않은가가 연어 관계 형성의 관건이 된다. 이러한 두 단어의 연어 형성 여부는 운율구 경계 예측을 위한 자질로 이용될 수 있다. 예를 들어, "~에 대하여", "~을/를 통하여", "~에 관하여" 등 연어 관계에 있는 두 단어는 붙여 읽을 가능성이 크다.

본 논문에서는 어절 사이를 기준으로 하여 이전 어절의 첫 형태소와 다음 어절의 끝 형태소 간 연어 형성 여부를 가설 검정(hypothesis testing)으로 판단하였다. 이는 3.1절에서 언급하였듯이 운율구는 대부분 어절의 첫 형태소와 끝 형태소에 따라 그 경계의 종류와 위치가 결정되기 때문이다.

독립성 검정을 이용하여 두 형태소의 연어 형성 여부를 판단하려면 다음과 같이 검정해야 할 가설을 설정하여야 한다.

- 가설 1. $P(w_2|w_1) = p = P(w_2|\neg w_1)$
- 가설 2. $P(w^2|w^1) = p_1 \neq p_2 = P(w^2|\neg w^1)$

가설 1은 형태소 w_2의 출현과 형태소 w_1의 출현이 상호독립적이라는 것을 뜻하고, 가설 2는 w_2의 출현이 w_1의 출현에 종속적이라는 것을 뜻한다. 이때, 가설 1이 가설 검정의 대상이 되는데 이 가설을 귀무가설이라고 한다. 반면 가설 2는 귀무가설이 기각되었을 때 받아들여지는 가설로써 대립가설이라고 한다. 독립성 검정에 의해 가설 1이 기각되면 가설 2가 받아들여지고 w_2의 출현이 어휘 w_1의 출현에 종속적이라는 것이 된다.

본 논문에서는 가설 검정을 위한 여러 방법 중에서 우도비(likelihood ratio)를 이용한 가설 검정 방법을 사용하였다. 우도비는 χ^2 통계치보다 좀 더 해석이 직관적이다. 즉, 하나의 가설이 다른 가설보다 얼마나 더 가능성이 있는지를 보여준다. 또한, 우도비를 이용한 가설 검정은 χ^2-test 보다 자료 부족 문제에 더 강하다.

우도비 λ는 전체 모수 공간상의 최대 우도 함숫값에 대한 가설에 따라 표현되는 일부 공간상의 최대 우도 함숫값의 비율로써 다음과 같이 표현된다.

$$\lambda = \frac{\max_{\omega \in \Omega_0} H(\omega; k)}{\max_{\omega \in \Omega} H(\omega; k)} \quad (1)$$

이때, ω는 모수 공간 Ω의 한 점(point)이고 Ω_0는 가설에 따라 표현되는 일부 공간이다. 우도비의 중요한 특징은 $-2\log\lambda$가 점근적으로 χ^2-분포를 따른다는 것이다. 특히, 이항분포에서 이 점근선은 매우 빠르게 접근하게 된다.

식 (1)은 이항분포 $b(k;n,x)$에서 식 (2)로 변환될 수 있다.

$$\log\lambda = \log\frac{L(H_1)}{L(H_2)} = \log\frac{b(c_{12},c_1,p)b(c_2-c_{12},N-c_1,p)}{b(c_{12},c_1,p_1)b(c_2-c_{12},N-c_1,p_2)} \quad (2)$$

실제로 가설 1과 가설 2에서

$$L(H_1) = b(c_{12};\, c_1,p)b(c_2-c_{12};\, N-c_1,p) \quad (3)$$

$$L(H_2) = b(c_{12};\, c_1,p_1)b(c_2-c_{12};\, N-c_1,p_2) \quad (4)$$

이므로 식 (2)는 식 (3), (4)에 의해 다음과 같이 변환된다.

$$\log\lambda = \log L(c_{12},c_1,p) + \log L(c_2-c_{12},N-c_1,p) - \log L(c_{12},c_1,p_1) - \log L(c_2-c_{12},N-c_1,p) \quad (5)$$

$$L(k,n,x) = x^k(1-x)^{n-k} \quad (6)$$

위에서 언급하였듯이 $-2\log\lambda$는 자유도가 1인 χ^2-분포를 따르고 가설 1의 기각 여부를 결정짓는 판단 근거로서 χ^2-분포를 사용할 수 있다. χ^2-분포에서 유의수준이 0.005일 때의 임계값(critical value)은 7.88이므로, 두 형태소의 $-2\log\lambda$가 7.88보다 크다면 '두 형태소 A와 B의 출현은 독립이다'라는 귀무가설은 기각되고 두 형태소는 연어 관계에 있다고 판단할 수 있다.

3.4 기타 정보

품사 정보, 거리 정보, 연어 형성 정보 외에도 어절의 길이가 길면 상대적으로 운율구 경계가 발생할 가능성이 크다는 것을 반영하고자 어절 길이를 학습 자질로 선별하였다. 또한, 앞에서 언급하였듯 인간이 한 번의 호흡에 말할 수 있는 물리적인 조건에 따라 일반적으로 운율구의 길이는 1~4 사이이다[18]. 따라서, 운율구 비경계가 연속해서 얼마나 나타났는지도 중요한 학습 자질이다.

4. 실험 및 결과

본 논문에서 제안하는 새로운 학습 자질을 바탕으로 CRFs를 이용하여 운율구 경계 예측 모형을 구축하였다. CRFs는 확률적 학습 방법의 하나로 은닉 마르코프 모델(HMM: Hidden Markov Model), 최대 엔트로피 모델(MEM: Maximum Entropy Model)과 마찬가지로 순서에 의미가 있는 데이터(sequence data)에 대해 좋은

성능을 보인다. 또한, 은닉 마르코프 모델과 최대 엔트로피 모델에서 발생하는 국소최대(local maximum) 문제와 label bias 문제를 극복할 수 있는 모델이다[17].

CRFs는 L1-CRFs와 L2-CRFs로 나뉘는데 L1-CRFs는 정규화 방법으로 라플라시안 사전 지식(laplacian prior)을 사용하며, L2-CRFs는 가우시안 사전 지식(gaussian prior)을 사용한다. 일반적으로 학습 데이터에 노이즈가 적으면 L2-CRFs가 더 좋은 성능을 보이지만, 학습 데이터에 노이즈가 많으면 L1-CRFs가 더 좋은 성능을 보인다.

4.1 운율구 경계 유형의 분류와 말뭉치 구성

선행 연구에서 사용하는 운율구 경계 유형은 보통 2단계(경계/비경계)와 3단계(강한 경계/약한 경계/비경계)이다. 세분화된 운율구 경계 유형을 사용할수록 시스템이 예제를 세분화된 유형 중 어느 클래스로 분류할 것인가에 대한 난도가 증가하고, 운율구 경계 태깅 단계에서 주석자 간 태깅 일치도와 정확도가 감소한다는 단점이 있다. 반면, 시스템이 바르게 예측하였을 경우, 유용하게 구분된 정보를 제공하고, 음성 합성 시스템에서 생성한 합성음의 명료성과 자연성을 향상할 수 있다.

운율구 경계에 발생하는 휴지는 어조와 결합하여 나타나므로, 본 논문에서는 표 2와 같이 정영임[18]에서 제안한 휴지와 어조를 결합하여 6단계로 세분화한 운율구 경계 유형을 모형화에 이용한다. 표 3은 실험을 위해 사용한 학습데이터와 평가데이터에 나타난 운율구 경계 유형의 분포를 보여준다.

운율구 경계 분석 말뭉치는 주석자의 개인차가 많이 작용하기 때문에 본 논문에서는 정영임[18]에서 구축한 다수 주석자 간 태깅 일치 신뢰도를 보장한 운율구 경계 분석 말뭉치를 학습데이터와 평가데이터로 이용하였다. 정영임[18]에서는 다수의 주석자가 KBS 뉴스(2005년 1월~2006년 6월)를 들으면서 해당 스크립트에 운율구 경계를 태깅하였다. 이때, 다수의 주석자가 태깅함으로써 생기는 주석자 간 운율구 경계 태깅의 불일치를 줄이고자 운율구 경계 주석 말뭉치 구축 과정을 세 단계-주석자 훈련 단계, 운율구 경계 태깅 단계, 말뭉치 신뢰도 추정 단계로 나누어 수행하여 말뭉치의 신뢰도를 높였다.

표 2 운율구 경계 유형의 분류

2단계	3단계	6단계
경계	강한 경계	상승조 강한 경계
		하강조 강한 경계
	약한 경계	상승조 약한 경계
		하강조 약한 경계
		평탄조 약한 경계
비경계	비경계	비경계

표 3 실험 데이터의 운율구 경계 유형 분포 (단위: 어절)

구 분	학습데이터	평가데이터
상승조 강한 경계	2,181(10.2%)	241(9.8%)
하강조 강한 경계	2,514(11.8%)	296(12.1%)
상승조 약한 경계	4,964(23.2%)	599(24.5%)
하강조 약한 경계	15(0.1%)	1(0.0%)
평탄조 약한 경계	55(0.3%)	5(0.2%)
비경계	11,653(54.5%)	1,305(53.3%)
전 체	21,382	2,447

4.2 학습 자질 평가 결과

실험을 위해 구축한 운율구 경계 예측 모형은 L2-CRFs를 이용하였으며, 하이퍼 파라미터 C는 1로 설정하였다. 하이퍼 파라미터의 값이 크면 클수록 학습 데이터에 과적합(overfitting)하는 경향이 있으므로 실험을 통해 적절한 값을 취하도록 한다.

학습 자질 평가 실험에서는 표 3의 학습데이터를 대상으로 10-fold cross validation을 사용하여 아래 수식에 의해 구해진 정확도의 평균값을 평가 측도로 활용한다.[3)]

$$\text{정확도}(accuracy) = \frac{tp+fn}{tp+tn+fp+fn}$$

tp: true positives, tn: true negatives

fp: false positives, fn: false negatives

표 4는 품사의 세분화가 운율구 경계 예측에 미치는 영향을 알아보기 위한 실험 결과이다. 표 4의 '기존 품사 집합' 열은 '21세기 세종 계획 말뭉치'에서 사용한 품사 집합을 이용한 결과로 품사 수는 45개이다. '새로운 품사 집합' 열은 '기존 품사 집합'을 표 1의 세분화를 통해 세분화한 품사 집합을 사용한 결과이다. 어절 사이를 기준으로 품사 정보를 참조하는 앞뒤 어절의 수를 바꾸어 가면 실험을 진행하였다. 단, 어절을 이루는 모든 형태소의 품사 정보를 이용하였다. 표 4의 실험 결과에서 알 수 있듯이 운율구 경계 예측을 위해 추가로 품사의 하위범주화를 한 결과 정확도가 더 높은 것을 알 수 있다. 이후의 실험에서는 추가로 품사의 하위범주화를 통한 새로운 품사 집합을 이용하여 진행하였다.

표 5는 품사 정보를 이용할 때, 어절을 이루는 모든 형태소의 품사 정보를 이용하였을 때와 첫 형태소와 끝 형태소의 품사 정보만을 이용하였을 때의 정확도를 비교한 것이다. 어절 사이를 기준으로 품사 정보를 참조하는 앞뒤 어절의 수를 바꾸어 가면서 실험을 진행하였다. 표 5의 실험 결과에서 알 수 있듯이 품사 정보를 이용할 때 어절을 이루는 전체 형태소의 품사 정보를 모두 이용하는 것보다는 어절의 첫 형태소와 끝 형태소의 품

3) 표 9의 실험에서는 운율구 경계 유형별 예측 성능을 비교하고자 정확도(precision)와 재현율(recall)을 이용하였다.

표 4 품사의 하위범주화에 따른 실험 결과

WindowSize	기존 품사 집합(A)	새로운 품사 집합(B)	B/A
앞, 뒤 1어절	71.23%	71.97%	1.010
앞, 뒤 2어절	70.88%	72.01%	1.016
앞, 뒤 3어절	70.57%	71.58%	1.014
앞, 뒤 4어절	70.64%	71.11%	1.007
앞, 뒤 5어절	70.22%	70.96%	1.011

표 5 품사 정보 단위에 따른 실험 결과

WindowSize	모든 형태소 (A)	첫, 끝 형태소 (B)	B/A
앞,뒤 1어절	71.97%	72.98%	1.014
앞,뒤 2어절	72.01%	72.94%	1.013
앞,뒤 3어절	71.58%	72.67%	1.015
앞,뒤 4어절	71.11%	72.12%	1.014
앞,뒤 5어절	70.96%	71.64%	1.010

사 정보만을 이용하는 것이 운율구 경계 예측을 위해 더 효율적이라는 것을 알 수 있다. 이후의 실험에서는 첫 형태소와 끝 형태소의 품사 정보만을 이용하여 실험을 진행하였다.

표 6은 품사 정보만을 이용하여 운율구 경계 예측 모델을 구성할 때, 최적의 모델을 찾고자 품사 정보를 참조하는 앞뒤 어절의 수를 달리하여 실험한 결과로써 첫 형태소와 끝 형태소의 품사 정보만을 사용하였다. 표 6에서 확인할 수 있듯이 이전 참조 어절 수가 늘어날수록 정확도가 증가하다가 이전 참조 어절 수가 3 이상일 때부터 정확도가 떨어진다. 이후 참조 어절 수 역시 너무 많은 어절을 참조하기보다는 이후 첫 어절의 품사 정보만을 이용하였을 때 정확도가 가장 높았다. 따라서, 품사 정보를 참조하는 최적의 앞뒤 어절 수는 어절 사이를 기준으로 앞 2어절과 뒤 1어절이다. 이후의 실험에서는 앞 2어절과 뒤 1어절까지의 품사 정보만을 이용하여 실험을 진행하였다.

표 7은 기존의 연구에서 사용하던 거리 정보와 본 논문에서 제안한 강한 운율구 경계를 기준으로 한 거리 정보를 비교 실험한 결과이다. 'DIST.start'는 문장의 시작에서부터 현재 어절까지의 거리를 이용한 결과이고 'DIST.end'는 문장 끝에서부터 현재 어절까지의 거리를 이용하였을 때의 결과이며 'DIS.nomal'은 문장 내에서 현재 어절이 차지하는 위치를 정규화한 수치이다. 그리고 'DIST.new'는 본 논문에서 제안한 강한 운율구 경계로부터 현재 어절까지의 거리이다. 표 7의 실험에서부터는 L1-CRFs와 L2-CRFs를 사용하였을 때를 비교하고자 두 개의 운율구 경계 예측 모형을 구형하였다. 표 7의 실험결과에서 알 수 있듯이 강한 운율구 경계로부터 현재 어절까지의 거리를 학습 자질로 이용하는 것이 기

표 6 앞뒤 참조 어절 수에 따른 실험 결과 (단위: %)

		이후 참조 어절 수				
		1	2	3	4	5
이전 참조 어절 수	1	72.98	72.98	72.63	72.16	72.01
	2	73.33	72.94	72.90	72.20	71.46
	3	73.25	72.74	72.67	72.20	71.89
	4	72.94	72.28	72.36	72.12	71.68
	5	72.90	72.20	72.11	72.01	71.64

표 7 거리 정보를 이용한 실험 결과

	DIST.start	DIST.end	Dist.normal	Dist.new
L1-CRFs	75.47%	75.64%	73.45%	80.12%
L2-CRFs	73.49%	73.49%	71.50%	78.11%

표 8 연어 형성 여부(A)와 연속된 운율구 비경계 수(B) 자질의 실험 결과

	사용 전	A 추가	B 추가	A, B 추가
L1-CRFs	80.12%	80.64%	80.83	81.14
L2-CRFs	78.11%	78.64%	78.34	78.68

존의 거리 정보를 이용한 것보다 더 좋은 결과를 보였다. 또한, L2-CRFs 보다 L1-CRFs를 이용하였을 때의 결과가 더 좋았다. 이는 약한 운율구 경계가 다른 경계보다 수의적인 경향이 가능하기 때문에 학습 데이터에 노이즈가 많을수록 더 좋은 성능을 보이는 L1-CRFs를 이용한 운율구 경계 예측 모형이 더 높은 정확도를 나타낸 것으로 판단된다.

표 8은 이전까지의 학습 자질에 더해 연어 형성 여부와 연속해서 나타난 운율구 비경계의 수를 학습자질로 사용한 결과이다. 그 밖에도 지정사 포함 여부', '용언화 접미사와 결합 여부' 등과 같은 어절의 중간에 있는 형태소의 품사 정보와 어절의 길이를 학습 자질로 추가로 이용하여 실험하였지만 대부분 0.1~0.3의 낮은 정확도 향상을 보였다. 이상의 실험 결과를 통해 본 논문에서 제안한 새로운 학습 자질들이 기존의 학습 자질보다 운율구 경계 예측에 대해 더 효율적이라는 것을 확인할 수 있었다.

표 9는 본 논문에서 제안한 학습 자질을 바탕으로 여러 기계학습 기법을 이용하여 운율구 경계 예측 모형을 구축한 결과이다. 이전의 실험들과는 달리 학습데이터를 통해 학습한 모형을 평가데이터를 이용하여 평가하였으며, 운율구 경계 유형별 예측 성능을 비교하고자 정확도(precision)와 재현율(recall)을 평가 척도로 활용하였다. 전체적으로 나이브 베이즈가 가장 낮은 성능을 보였으며, CRFs가 가장 높은 성능을 보였다. C4.5는 나이브 베이즈와 CRFs의 중간 정도 성능을 보였지만 비경계에 대해서는 가장 높은 재현율을 보였다.

표 9 규칙과 통계 기반의 복합적 접근법을 이용한 운율구 경계 예측 유형별 성능 평가

운율구 경계 유형	나이브 베이즈		C4.5		CRFs	
	정확도	재현율	정확도	재현율	정확도	재현율
상승조 강한 경계	42.6	43.2	53.7	32.2	64.5	57.3
하강조 강한 경계	75.3	61.8	71.1	70.2	74.0	78.7
상승조 약한 경계	57.6	54.2	61.4	58.3	70.1	63.3
하강조 약한 경계	0	0	0	0	0	0
평탄조 약한 경계	0	0	0	0	0	0
비경계	83.7	88.8	84.9	93.9	86.5	91.0

표 10 선행 연구와의 비교 실험 결과

운율구 경계 유형	본 논문의 예측 모델	김승원[11]의 예측 모델	정영임[13]의 예측 모델
강한 운율구 경계	92.68%	78.03%	90.75%
약한 운율구 경계	69.28%	55.40%	71.90%
운율구 내부 비경계	95.97%	86.51%	91.32%
전체	86.63%	77.32%	-

표 10은 본 논문에서 제안한 학습 자질로 구축한 한국어 운율구 경계 예측 CRFs 모형을 선행 연구에서 제안한 다른 두 가지 모형과 비교한 결과이다. 앞의 실험과 마찬가지로 학습데이터로 학습한 모형을 평가데이터로 평가하였다. 학습 기법의 차이에서 오는 특성을 배제하고 학습 자질의 효율성만을 분석하고자 김승원[11]에서 제안한 운율구 경계 예측 CRFs 모형과 비교를 하였으며, 규칙 기반 모형과 비교하여 어떠한 차이가 있는지 분석하고자 정영임[13]에서 제안한 규칙 기반 운율구 경계 예측 모형과 비교하였다.

김승원[11]에서는 어절을 이루는 전체 형태소의 품사 정보, 문장 내에서 현재 어절이 차지하는 위치를 정규화한 수치, 단어의 어휘를 그대로 사용한 자질을 제안하였다. 정영임[13]에서는 구문 정보와 운율 정보를 바탕으로 운율구 경계 예측 규칙을 구축하였다. 김승원[11]과 정영임[13]에서는 '강한 운율구 경계', '약한 운율구 경계', '운율구 내부 비경계'의 3가지 유형으로 나누어서 실험하였기에 여기서도 똑같이 3가지의 유형에 대해 실험을 진행하였다.

표 10의 결과에서 알 수 있듯이 본 논문에서 제안한 새로운 학습 자질로 구축한 CRFs 모형이 기존의 학습 자질로 구축한 CRFs 모형과 규칙 기반 모형보다 한국어 운율구 경계 예측에서 더 높은 정확도를 보였다. 그러나 표 9와 표 10의 결과에서 운율구 경계 유형별로 봤을 때, 약한 운율구 경계 예측은 다른 경계 유형과 비교하면 예측 성능이 더 낮았으며, 규칙 기반 모형의 약한 운율구 경계 예측 정확도와 비교하였을 때도 예측 성능이 더 낮았다. 이는 약한 운율구 경계가 다른 두 유형보다 화자에 따라 가변적으로 변할 가능성이 크기 때문에 의미 있는 통계 정보가 추출되지 않았기 때문이다. 가변적 운율구 경계 예측을 위해 화자의 발화 특징에 대한 좀 더 다양한 분석이 이루어져야 할 것이다.

5. 결론 및 향후 연구

본 논문에서는 통계적 접근법을 이용한 운율구 경계 예측에서 더 좋은 예측 결과를 도출할 수 있는 새로운 학습 자질을 제안하였다.

품사 정보를 효율적으로 활용하고자 품사의 하위범주화를 통해 선행 연구에서 사용한 품사 집합보다 좀 더 세분화한 품사 집합을 사용하였다. 또한, 어절의 품사 정보를 이용할 때 운율구 경계 예측에 더 많이 관여하는 첫 형태소와 끝 형태소의 품사 정보만을 이용하였다. 그 결과 선행 연구에서 사용한 품사 정보를 이용하였을 때보다 운율구 경계 예측의 정확도가 더 향상되었다.

거리 정보를 이용할 때는 문장의 시작이나 끝에서부터 현재 어절까지의 거리보다는 강한 운율구 경계로부터 현재 어절까지의 거리가 운율구 경계 예측에서 매우 중요한 역할을 한다는 사실 역시 실험을 통해 증명하였다.

앞으로는 규칙 기반 접근법과 통계적 접근법을 결합하여 서로 간의 결점을 보완하는 방법에 대해 연구를 진행할 예정이다.

참 고 문 헌

[1] Taylor, P., Black, A. W., "Assigning Phrase Breaks from Part-of-Speech Sequences," *In Proceedings of Eurospeech*, pp.995-998, 1997.

[2] Lee, S., Oh, Y., "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS systems," *Speech Communication*, vol.28, pp.283-300, 1999.

[3] Kim, B., Lee, G., G., "Implementation of Korean TTS System based on Natural Language Processing," *Malsori*, vol.46, pp.51-64, 2003. (in Korean)

[4] Jeong, H., *Study on Korean Nouns*, Hangug-munhwasa, 2002. (in Korean)

[5] Kim, H., "The Construction of Adverb Lexicon in Contemporary Korean - On Some Issues of the description and the Classification of Adverbs -," *Korean Journal of Linguistics*, vol.24, pp.109-144, 1999. (in Korean)

[6] Kwon, J., Kim, Y., Moon, Y., et al., "A Study on the Interface between Syntactic and Prosodic Structure with Special Reference to the Modes of Ambiguity Resolution," *Korean Journal of Linguistics*, vol.20, pp.103-109, 1997. (in Korean)
[7] Kim, S., *Rhythmic Units and Syntactic Structures in Korean: A Phonetic and Linguistic Study Aiming at Improving the Rhythmic Properties of Synthetic Speech*, Seoul National University, 2002. (in Korean)
[8] Lee, Chan-Do, "A Computation Study of Prosodic Structures of Korean for Speech Recognition and Synthesis: Predicting Phonological Boundaries," *The Transactions of the Korea Information Processing Society*, vol.4, no.1, pp.280-287, 1997. (in Korean)
[9] Hirschberg, J., Prieto, P., "Training International Phrasing Rules Automatically for English and Spanish Text-to-Speech," *Speech Communication*, vol.18, pp.281-290, 1996.
[10] Ostendorf, M., Veilleus, N., "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location," *Computational Linguistics*, vol.20, no.1, pp.27-54, 1994.
[11] Kim, S., Kim, B., Jeoung, M., Lee, G., "Using CRF (Contional Random Fields) to Predict," *Human & Cognitive Language Technology 2005*, pp.134-138, 2005. (in Korean)
[12] Yarowsky, D., "Homograph Disambiguation in Text-to-speech Synthesis," *Progress in Speech Synthesis*, pp.366-377, 1996.
[13] Jung, Y., Cho, S., Yoon, A., Kwon, H.-C., "Prediction of Prosodic Break Using Syntatic Relations and Prosodic Features," *Korean Journal of Cognitive Science*, vol.19, no.1, pp.89-105, 2008. (in Korean)
[14] Lee, S., Oh, Y.-H., "The Modeling of Prosodic Phrasing and Pause Duration using CART," *Korean Scientific and Cultural Studies Programme Workshop 1998*, vol.15, no.1, pp.81-86, 1998. (in Korean)
[15] Chun, Jin.-w., Kim, H. W., Kim, D. g., Lee, Y., "Prosodic-Boundaray Prediction for Korean Text-to-Speech System," *In Proceedings of Acoustical Society of Korea*, vol.22, no.1, pp.77-83, 2002. (in Korean)
[16] Ostendorf, M., Veilleus, N. "A hierarchical Stochastic Model for Automatic prediction of Prosodic Boundary Location," *Computational Linguistics*, vol.20, no.1, pp.27-54, 1994.
[17] J Lafferty, A McCallum, F Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Machine Learning-International Workshop then Conference*, 2001.
[18] Jung, I., *Reliable Prediction of Prosodic Breaks by Combining Rules and probabilities Obtained from Small-Scale Corpus*, Pusan National University, 2009.

김 민 호
2007년 부산대학교 정보컴퓨터공학부 학사. 2009년 부산대학교 컴퓨터공학과 석사. 2009년~현재 부산대학교 컴퓨터공학과 박사과정. 관심분야는 자연언어처리, 정보검색, 인공지능

권 혁 철
정보과학회논문지 : 소프트웨어 및 응용 제 37 권 제 1 호 참조