

# 예외 단어 선별 작업을 이용한 자동 발음열 생성 시스템의 성능향상에 관한 연구

광운대학교 컴퓨터공학과,<sup>1</sup> 광운대학교 음성정보처리기술연구센터<sup>2</sup>

안주은<sup>1</sup> · 김선희<sup>2</sup> · 김순협<sup>1</sup>

본 연구는 텍스트 분석 기반 자동 발음열 생성 시스템<sup>1)</sup>의 성능을 향상시킬 수 있는 방법으로 예외 단어 선별 작업을 제안한다.

제안하는 자동 발음열 생성 시스템은 형태음운 규칙 적용, 예외 규칙 적용, 일반음운 규칙 적용의 순서로 이루어진다. 이 때 예외 규칙은 예외사전을 검색하여 해당 어휘가 있을 때에만 적용하게 되는데, 예외 단어는 일정한 예외 환경에서만 나타난다. 본 시스템의 예외 사전에 있는 예외 단어는 5만여 어휘 연세사전<sup>2)</sup>에서 예외 환경에 해당하는 28,486개의 어휘 중, 총 2,931개의 예외 단어를 선별하였다.<sup>1)</sup> 본 시스템의 성능 평가는 ETRI 2,385, PBS 1,637 어절을 사용하였는데 실험 결과 각각 99.8%, 99.9%의 정확도를 보였고, 0.1%의 오류는 예외 규칙 중에 하나인 어휘적 경음화 현상에서 나타났다.<sup>1)</sup> 따라서 예외 사전을 보다 정확하게 만드는 방법인, 예외 단어 선별 작업을 통해 본 시스템의 성능을 향상시킬 수 있다.

본 연구에서 제안한 성능 평가 실험의 진행 순서는, (1) 입력 텍스트에서 예외 환경 추출, (2) 예외 단어 선별 작업, (3) 기존 예외 사전에 선별된 예외 단어 추가, (4) 형태음운 규칙 적용, (5) 예외 규칙 적용, (6) 일반음운 규칙 적용이다. 본 논문에서 사용한 실험 텍스트로는 ETRI 신문사설 19만여 어휘인데, 그 가운데 기존 예외 환경에 해당하는 2,931개의 어휘를 제외한 54,931개의 어휘를 먼저 뽑은 후, 각각의 예외환경 당 발생한 어휘 수 및 선별된 예외 단어 수를 추출하였다. 예외 단어로 선별된 78개의 어휘를 예외 사전에 추가시킨 후, 발음열을 생성한 결과 100%의 정확도를 보였다.

본 연구에서 제안하는 자동 발음열 생성 시스템에서 발생한 적은 오류의 주된 원인은 어휘적 경음화이다. 이와 같은 오류는 기존에 검사했던 예외환경을 제외한 나머지 예외환경에서 예외 단어를 선별하여 예외 사전에 추가한 후 발음열을 생성함으로써 줄일 수 있다. 또한, 특정한 예외 환경에서 예외 단어가 주로 발생한다. 이 점을 이용하여, 이후에 많은 양의 텍스트에 대한 예외 단어 선별 작업을 할 때에 보다 효율적으로 예외 사전을 보강시킬 수 있게 되었다.

## REFERENCES

- 1) 김선희 · 안주은 · 김순협. "텍스트 분석 기반 한국어 자동 발음열 생성 시스템", 음성신호처리학회 학술대회 논문집:2003.
- 2) 연세대학교 언어정보교육원. 연세한국어사전, 두산동아:1998.