ETRI 방송 뉴스 자막 처리 시스템을 위한 미등록어 검출기의 개발

윤승*, 정의정*, 박준*, 이영직*

*한국전자통신연구원 음성언어팀

Unknown Word Extractor Development for ETRI Broadcast News Caption System

Yun Seung*, Eui_Jung, Jung*, Jun Park*, Youngjik Lee*

*Spoken Language Processing Team, Electronics and Telecommunications Research Institute

{yunseung, euijoung, junpark, ylee}@etri.re.kr

요 약

본 논문에서는 ETRI 방송 뉴스 자막 처리 시스템의 성능 향상을 도모하기 위해 개발된 미등록어 검출기에 대해 기술한다.

음성 인식 성능 하락에 큰 영향을 미치는 요인들 중 하나로 꼽히는 미등록어 문제를 해결하기 위해 ETRI 방송 뉴스 자막 처리 시스템에서는 오프라인으로 동작하는 미등록어 검출기를 채택하였다.

이 미등록어 검출기는 방송 뉴스 자막 처리 시스템 가동 전에 미리 인터넷을 통해 최신 신문 기사와 방송 뉴스를 수집해와 이를 토대로 두 단계에 걸쳐 미등록어를 사전에 추출하여 인식 어휘 사전에 포함시킴으로써 미등록어로 인한 방송 뉴스의 인식 성능 저하 문제를 해결하도록 하였다.

1. 서론

기존 미등록어 관련 연구는 형태소 분석기, 태거, 철자 및 맞춤법 교정기, 정보 검색 시스템, 기계 번역 시스템 등 주로 자연어 처리 시스템 위주로 진행되어 왔다. 그러나 이러한 자연어 처리 시스템에서 연구 대상으로 삼는 미등록어와 음성 인식에서 문제가 되는 미등록어는 그 성격 및 연구 관점이 다르므로 기존의 연구를 그대로 음성 인식에 적용하기에는 무리가 따른다. 따라서 음성 인식에서의 미등록어 문제 해결을 위해서는 기존의 자연어 처리 관련 연구에서와는 다른 접근 방법이 필요하다고 하겠다.

음성 인식에서의 미등록어 문제도 어떤 것을 인식 대상으로 삼는가와 인식 결과를 어디에 이용할 것인가에 따라 여러 가지로 나누어 살펴볼 수 있다. 본 논문에서는 연구 대상을 방송 뉴스 자막 처리 시스템에 한정해서 미등록어 문제의 해결 방안에 대해 살펴보기로 한다.

언제나 새로운 것을 대상으로 삼는 방송 뉴스의 특성으로 인해 방송 뉴스에는 인명, 지명, 회사·단체명 등의 새로운 어휘가 많이 나타나게 된다. 또 이러한 부류의 어휘들을 제외하고도, 다양한 분야를 다루는 방송 뉴스의 특성 때문에 인식 어휘 사전에 포함되어 있지 않은 빈도가 낮은 어휘들도 방송 뉴스에는 상당수 등장하게 된다. 따라서 미등록어 문제는 방송 뉴스 인식에도 심각한 영향을 미친다고 할 수 있다.

본 논문에서는 이러한 문제를 해결하기 위해 5가지 신문사와 2가지 방송사의 홈페이지에서 최신 기사 및 뉴스를 수집해 이들을 대상으로 미등록어를 검출해 인식 어휘 사전에 추가함으로써 미등록어로 인한 문제를

해결하고자 한다.

2. 신문 기사 및 방송 뉴스 수집

ETRI 방송 뉴스 자막 처리 시스템에서 인식 대상으로 삼는 뉴스는 KBS 9시 뉴스와 SBS 8시 뉴스이다. 이들의 경우 일부 속보성 뉴스를 제외하고는 대부분 그날 하루 동안 있었던 여러 가지 일들 중에서 비중이 높다고 생각되는 뉴스들을 방송하게 된다. 신문사 홈페이지의 메인페이지 역시 그날의 여러 기사들 중 가장 중요하다고 생각되는 기사들이 위치하게 되므로 그날 방송될 뉴스들과 신문사 홈페이지의 메인 페이지에 올라와 있는 기사들의 내용은 높은 유사성을 가진다고 할 수 있다. 따라서 방송 뉴스 자막 처리 시스템 동작 전에 신문사 홈페이지에서 최신 중요 기사들을 수집해 와 이들을 대상으로 미등록어를 검출해내면 방송 뉴스에 나타날 가능성이 있는 미등록어의 대부분을 처리할 수 있다.

2.1. 최신 중요 신문 기사 및 방송 뉴스의 수집

본 논문에서 제안하고 있는 미등록어 검출기는 이러한 처리를 위해 인터넷을 통해 조선일보, 동아일보, 한겨레신문, 중앙일보, 한국일보 등 5가지 신문사 홈페이지에서 메인 페이지에 존재하는 최신 중요 기사들을 수집해 오게 된다. 5가지 신문사를 대상으로 하는 것은 속보성 기사의 경우 연합뉴스를 받아 그대로 전재하거나 약간의 수정을 거쳐 홈페이지에 올리게 되므로 각 신문사의 내용이 거의 비슷하지만, 각 신문사가 중요하다고 판단해 메인페이지에 올리는 기사들이 서로 다를 수 있다는 점을 반영해 주고 또 각 신문사에서 자체 취재한 기사들 역시 미등록어 검출 대상에 포함시켜 주기 위해서이다. 또 신문 기사 외에 KBS, SBS 홈페이지의 뉴스 부분 메인 페이지에 올라와 있는 뉴스들 역시 수집 대상에 포함시켜줌으로써 각 방송사의 자체 취재 기사나 신문 기사와 방송 뉴스의 언어 특성 차이에 따라 다르게 나타나는 미등록어들도 반영할 수 있도록 하였다.

2.2 HTML 파싱 및 텍스트 변환

미등록어 검출기는 수집된 신문 기사와 방송 뉴스를 대상으로 HTML 파싱과 텍스트 변환 작업을 수행한다. 먼저 HTML 파싱이 이루어지게 되는데 이는 가져온 기사와 뉴스들이 HTML 문서이므로 문서 내의 HTML 태그를 제거해주고 필요 없는 정보들을 제거한 후 제목과 본문만을 정확하게 추출해내는 작업이 필요하기 때문이다.

HTML 파싱을 완료한 후에는 텍스트 프로세싱 과정을 거치게 된다. 여기에서는 기자 이름, 통신사 명칭 등의 필요 없는 정보를 제거하고 외래어, 숫자의 한글 변환 및 각종 기호 처리를 통해 수집된 신문 기사와 방송 뉴스를 실제의 음성 인식 결과와 유사한 형태로 변환하게 된다.

3. 2단계 미등록어 검출

2장에서는 주로 텍스트를 수집하고 처리하는 과정에 대해서 논의했다. 3장에서는 정제된 텍스트를 대상으로 두 단계에 걸쳐 미등록어를 검출해내는 방법에 대해 설명한다.

3.1. 태깅된 텍스트와 어휘 사전을 이용하는 방법

ETRI 방송 뉴스 자막 처리 시스템은 인식 단위로 의사형태소를 사용하고 있으므로 인식 어휘 사전도 의사형태소 단위로 작성되어 있다. 따라서 HTML 파싱과 텍스트 변환 과정을 거쳐 정제된 텍스트를 대상으로 형태소 분석을 실시해 그 결과와 인식 어휘 사전을 비교해 보면 인식 어휘 사전에 등록되어 있지 않은 미등록어들을 검출해낼 수 있다.

검출된 미등록어 집합에서 동일한 어휘들을 삭제한 후에 이 어휘들을 미등록어 사전에 추가함으로써 이후 방송 뉴스 자막 처리 시스템에 반영될 수 있도록 한다.

3.2. 미등록어로 태깅된 형태소를 추가하는 방법

3.1.의 논의에서 제외된 것이 형태소 분석기에서도 미등록어로 분석된 어휘에 관한 것이다. 형태소 분석기에서 미등록어로 분석된 것은 실제 미등록어일 수도 있으나 오타, 맞춤법 오류 또는 형태소 분석기 내부 문제 등으로 인한 오류일 가능성도 함께 존재한다. 이러한 경우 무조건 인식 어휘 사전에 추가할 경우에는 필요 없는 어휘들로 인해 사전 크기가 늘어나 전체 시

스템에 부담을 줄 수 있고 또 무조건 제외할 경우에는 인식 어휘 사전에 포함되어야 할 어휘들마저도 제외되는 문제가 발생한다.

이를 해결하기 위해서 본 논문에서는 전체 텍스트에서 2회 이상 미등록어로 분석된 어휘만을 인식 어휘 사전에 추가하는 방법을 제안한다. 현재 7군데에서 가져온 텍스트를 대상으로 작업을 수행하므로 중복되는 내용이 존재하는 경우가 많아 실제 미등록어라면 2회 이상 나타나는 경우가 대부분이고 또 핵심적인 미등록어의 경우 동일 기사나 뉴스 내에서도 반복적으로 쓰이는 경우가 많다. 만약 실제로 미등록어가 아닌데 미등록어로 잘못 분석된 경우라면 같은 상황이 두 번 발생하는 경우가 흔치 않으므로 이 경우에는 미등록어로 처리되지 않을 것이다.

텍스트에서 2회 이상 나타나 실제 미등록어로 분류된 어휘들은 최장 조사 분리 방법에 의해 조사를 잘라낸 후 미등록어 사전에 추가해 이후 인식 어휘 사전에 반영될 수 있도록 한다. 이는 방송 뉴스의 경우 나타나는 미등록어 대부분이 명사이기 때문에 가능한 방법이다. 물론 최장 조사 분리 방법을 채택할 경우 아래 예에서와 같이 미등록어가 잘못 잘라내어지는 경우가 있을 수 있다.

● 최장 조사 분리에 의해 오분석 되는 예

벨기에는 → 벨기에 + 는(O) 벨기에는 → 벨기 + 에는(X)

위의 예는 '벨기에는'이 '벨기에+는'으로 분석되어야 하나 최장 조사 분리 방법을 적용함으로써 '벨기'+'에는'으로 분석되는 경우이다. 그러나 이러한 오류의 경우 다른 시스템에서는 문제가 될 수 있으나 방송 뉴스 자막 처리 시스템의 경우 결과에 영향을 미치지 않는다. 두 경우 모두 최종 인식 결과는 '벨기에는'이 되기 때문이다.

3.3. 방송 뉴스 자막 처리 시스템에의 반영

3.1.과 3.2.의 과정을 거쳐 미등록어 사전에 추가된 어휘들은 발음 변환 과정을 거쳐 인식 어휘 사전에 추가됨으로써 최종적으로 방송 뉴스 자막 처리 시스템에 반영된다.

미등록어 사전에 존재하는 어휘들을 인식 어휘 사전에 추가할 때에 사람의 판단에 따라 선별적으로 추가할 수도 있고 사람의 개입 없이 일괄적으로 추가할 수도 있다. 사람의 개입 없이 추가할 경우에는, 형태소 분석기에서 미등록어로 추정된 어휘의 경우 단일 태그가 붙은 채로 시스템에 반영된다.

4. 실험 및 분석

미등록어 검출기를 이용했을 경우 실제로 방송 뉴스 자막 처리 시스템의 인식 성능 향상이 얼마나 이루어지는 지를 알아보기 위해 실험을 진행하였다. 이틀 분의 방송 뉴스로 실험을 위한 테스트 셋을 구성한 후 이를 대상으로 각각 미등록어 처리 전과 미등록어 처리 후로 나누어 방송 뉴스의 인식률을 측정한 후 그 결과를 비교해 보았다.(표1)

엄밀한 실험을 위해서는 실험 과정에서의 미등록어 검출이 2장에서 언급한 7군데의 홈페이지에서 가져온 기사 및 뉴스를 대상으로 이루어져야하나 당일의 방송 뉴스를 대상으로 하는 인식 실험 환경 구축에 어려움이 있어 음성 파일이 존재하는 방송 뉴스의 전사문을 대상으로 미등록어 검출을 실시하였다. 실험이 실제 환경과는 다르게 이루어졌으나 이 경우에도 미등록어 검출기의 평균 재현율이 94.7%에 이르는 점을 감안할 때 그 오차는 그리 크지 않을 것으로 예상된다.

표1. 미등록어 처리 전과 후의 인식 결과 비교(%)

	미등록어처리전	미등록어처리후	ERR
1999-12-17	81.0	81.6	3.16
2000-01-17	76.2	77.1	3.78

〈표1〉을 살펴보면 평균 0.75%의 인식률 향상이 있었음을 알 수 있다. 기대보다는 작은 수치이지만 평균 3.47%의 ERR 향상은 충분히 유의미한 수치로 볼 수 있을 것이다.

실험에서 미등록어 검출기의 역할은 그다지 두드러지게 나타나지 않았지만 미등록어 검출기가 미등록어가 많이 존재하는 뉴스가 언제 방송될지 모르는 상황에서 그에 대한 대비 수단으로서의 역할을 주기능으로 한다는 점을 염두에 둔다면 안정적인 인식 성능 확보를 위한 수단으로서의 미등록어 검출기의 존재 가치는

충분해 보인다.

5. 결론 및 향후 과제

본 논문에서는 방송 뉴스 자막 처리 시스템을 위한 미등록어 검출기에 대해 살펴보았다. 방송될 뉴스의 내용과 유사한 웹문서에서 두 단계에 걸쳐 미등록어를 검출해 방송 뉴스 자막 처리 시스템에 적용하는 본 미등록어 검출기는 인식률 향상에 효과적으로 작용할 수 있음을 알게 되었다. 그러나 제안된 미등록어 검출기는 아직 개선의 여지가 남아 있다.

첫째, 현재의 시스템은 미등록어를 검출하는 두 가지 방법 모두 태거에 의존하고 있기 때문에 태거 자체의 오분석으로 인한 오류는 검출해낼 방법이 없다. ETRI 음성언어팀에서 보유하고 있는 태거가 방송 뉴스에 최적화되어 있어 오류율이 낮기는 하지만 그럼에도 불구하고 이 점은 현재 시스템의 한계로 남는다.

둘째, 현재의 시스템에서는 검출된 미등록어가 언어 모델에 반영되지 못하고 있다는 점을 들 수 있다. 현재 이와 관련해 검출된 미등록어를 여러 클래스로 분류하여 각 클래스마다 임의의 대표 어휘의 언어 모델 값을 복사해와 적용하는 방안이 연구되고 있다.

향후 이러한 문제점들을 해결한 미등록어 검출기가 개발되어야 할 것이다.

감사의 글

이 연구는 정보통신부 출연 "청각 및 시각 장애인을 위한 디지털 방송기술 개발"과제의 일부로 수행되었습니다.

참고문헌

[1]David Yarowsky(1995), "Unsupervised Word Sense Disambiguation Rivaling Superivised Methods", Proceeding on 33rd Annual Meeting of the Association for Computational Lingustics, pp 189-196.

[2]Tetsuya Nasukawa(1996), "Full-text processing: improving a practical NLP system based on surface information within the context", Proceedings of 16-th International Conference on Computational Linguistics, pp.824-820.

[3] 박준, 김승회, 이영직, 양재우(2000), "방송뉴스 자막처리 시스템 개발", 제 17회 음성통신 및 신호처리 학술 대회 17권 1호.

[4] 박봉래, 황영숙, 임해창(1998), "용례 분석에 기반한 미등록어의 인식", 정보과학회논문지(B) 제25권 제2호.

[5] 박봉래(2000), "전문분석에 기반한 한국어 미등록어의 인식", 박사학위논문, 고려대학교.

[6] 양장모, 김민정, 권혁철(1996), "언어 정보를 이용한 한국어 미등록어 추정", 한국정보과학회 봄 학술발표논문집 Vol. 23, No. 1, pp.957-960.

[7] 차정원, 이원일, 이근배, 이종혁(1997), "형태소 패턴 사전을 이용한 일반화된 미등록어 처리", 정보과학회 인공지능연구회 춘계학술대회 논문집, pp.37-42.