음성 인식 후처리를 위한 띄어쓰기 오류의 교정

임동희, 강승식 장두성
국민대학교 컴퓨터학부 ㈜케이티 미래기술연구소
nlp@cs.kookmin.ac.kr, sskang@kookmin.ac.kr dschang@kt.co.kr

Word Spacing Error Correction for the Postprocessing of Speech Recognition

Dong-Hee Lim, Seung-Shik Kang Du-Seong Chang
Dept. of Computer Science, Kookmin University Advanced Technology Lab., KT

요 약

음성인식 결과는 띄어쓰기 오류가 포함되어 있으며 이는 인식 결과에 대한 이후의 정보처리를 어렵게 하는 요인이 된다. 본 논문은 음성 인식 결과의 띄어쓰기 오류를 수정하기 위하여 품사 정보를 이용한 어절 재결합 기법을 기본 알고리즘으로 사용하고 추가로 음절 바이그램 및 4-gram 정보를 이용하는 띄어쓰기 오류 교정 방법을 제안하였다. 또한, 음성 인식기의 출력으로 품사 정보가 부착된 경우와 미부착된 경우에 대한 비교 실험을 하였다. 품사 미부착된 경우에는 사전을 이용하여 품사 정보를 복원하였으며, N-gram 통계 정보를 적용했을 때 기본적인 어절 재결합 알고리즘만을 사용 경우보다 띄어쓰기 정확도가 향상되는 것을 확인하였다.

1. 서론

음성 인식 시스템은 자동 통역을 비롯하여 전자 제품의 작동 등 자연 언어 인터페이스의 그 활용 분야가 매우 다양하다. 하지만 음성 인식 시스템이 실용적으로 사용되기 위해서는 응용 시스템이 만족하는 수준의 정확도가 요구되고 있다.

음성 인식 시스템은 잡음 등 발화 환경에 따른 인식률 저하 문제와 더불어 음성 신호를 문자열로 변환하는 과정에서 다양한 유형의 오류들이 발생한다. 그 중에서도 음성인식 결과의 띄어쓰기 오류는 음성인식 결과에 대한 이후의 정보처리를 어렵게 하는 요인이 되며 띄어쓰기가 없거나 잘못된 경우 의미 전달이 잘 못 될 수 있다. 예로 한글 수사열에 대한 숫자 변환에 영향을 줄 수 있다. "삼 십"과 "삼십"은 각각 "3 10", "30"으로 서로 다른 처리 결과를 가진다.

자동 띄어쓰기 방법은 크게 통계 기반 방법, 규칙 기반 방법으로 구분할 수 있다[3,4]. 통계 기반 방법은 손쉽게 구현할 수 있지만 학습 말뭉치에 의존적인 결과가 나오며 자료 부족 문제도 발생한다. 규칙 기반 방법은 구현시에 어휘 지식이 필요하고 어휘 지식 구축 및 유지보수에는 많은 비용이 든다.

본 논문은 음성 인식 후처리를 위한 띄어쓰기 오류 교정을 위하여 규칙과 통계 기반 방법을 통합한 형태의 띄어쓰기 오류 교정 시스템을 구현하였다[2].

2. 음성 인식 결과

본 논문에서 사용한 음성 인식기의 출력은 인식 단위별로 분리되어 있고, 각각의 인식 단위에 품사 정보가 부착된 경우(1)와 그렇지 않은 경우(2) 두 가지를 제공한다. 인식 단위는 형태소 단위나 어절 단위로 고정된 형태가 아니며 음성 인식의 성능을 높이기 위해 인접한 형태소와 어절이 합쳐진 형태도 있고 형태소만으로 구성된 인식 단위도 있다. 그리고 품사 정보가 부착된 경우 형태소 단위로 다시 분할을 할 수 있는 정보(':')가 추가로 존재한다.

품사 정보를 음성 인식 결과와 함께 출력해 주기 위해서는 음성 인식기의 크기 및 시스템 자원 사용량이 그렇지 않은 경우보다 커지기 때문에 음성 인식기가 최종적으로 사용될 환경에 따라 품사 정보가 미부착된 결과를 출력 결과로 사용해야 할 경우도 있다. 음성 인식 결과의 구체적인 예는 다음과 같다.

품사정보 부착된 경우 - (1)
"나/대명사 잘/부사 할:수/동사+의존명사 있:다/동사+어미"
"백/수사 이:십:육/수사+수사+수사 쩜/수사 팔/수사"
"즐거운/형용사 한:주:되:세요/관형사+의존명사+동사+어미"
품사정보 미부착된 경우 - (2)
"나 잘 할수 있다"
"백 이십육 쩜 팔"
"즐거운 한주되세요"

3. 띄어쓰기 오류 교정 방법

품사 정보를 이용한 어절 재결합을 기본적으로 사용하고 음절 N-gram 정보를 이용해 오류를 수정하는 방법을 설명한다.

3.1. 품사 정보를 이용한 어절 재결합

음성인식 결과의 품사 정보를 사용하기 위해 우선 각각의 품사에 해당하는 형태소 단위로 인식단위를 분할한다. 구체적인 예는 다음과 같다.

조심:해:서/보통명사+조용사+어미 오/동사 시:오/선어말어미+어미
→ 조심/보통명사 해/조용사 서/어미 오/동사 시/선어말어미 오/어미

삼:월/수사+의존명사 이:십:구:일/수사+수사+수사+의존명사
→ 삼/수사 월/의존명사 이/수사 십/수사 구/수사 일/의존명사

형태소 단위로 분할된 결과를 가지고 다음과 같은 규칙에 따라 어절을 재결합한다.

품사가 조사, 어미, 선어말어미, 조용사, 접미사이면 앞 어절과 붙여쓰고 기타 품사이면 띄어쓴다. 단, 수사와 수사, 수사와 의존명사 사이는 붙여쓴다. 규칙에 따른 어절 재결합의 구체적인 예는 다음과 같다.

조심/보통명사 해/조용사 서/어미 오/동사 시/선어말어미 오/어미
→ 조심해서 오시오

삼/수사 월/의존명사 이/수사 십/수사 구/수사 일/의존명사
→ 삼월 이십구일

3.2. 품사 정보 복원

음성인식 결과에 품사 정보가 미부착된 경우 3.1.방법을 가지고 어절 재결합을 할 수 없게 된다. 따라서 후처리 단계에서 품사 정보 복원을 해주어야 한다. 복원을 위해 음성인식 단위에 대한 품사 사전을 사용하고 다음과 같이 구성된다.

음성인식어절 | 후보개수 | 후보1 | 후보2 | 후보3 | …

예) 가마 | 2 | 가마/보통명사 | 가:마/동사+어미
이월 | 2 | 이월/보통명사 | 이:월/수사+의존명사
해드립니다 | 1 | 해:드립:니다/조용사+보조동사+어미

품사 사전을 가지고 품사 정보를 복원을 할 경우 품사가 하나로 정해지지 않고 2개 이상으로 정해질 경우가 존재하기 때문에 3.1.에서 사용한 규칙만 적용한 어절 재결합 결과의 정확도가 낮다. 정확도를 높이기 위해 다음과 같은 품사 모호성 관련 어절 재결합 조건을 추가로 적용한다.

1) 2음절 명사+조용사 → 붙여쓰기
2) 부사+조용사 → 띄어쓰기
3) (2음절 이상 조사/어미)+조용사 → 띄어쓰기
4) 2음절 명사+"같"으로 시작하는 조사 → 붙여쓰기
5) 부사+"같"으로 시작하는 조사 → 띄어쓰기
6) "네"+접미사, "내"+접미사 → 띄어쓰기
7) "보고" → 띄어쓰기
8) 부사+조사 → 띄어쓰기
9) (2음절 이상 조사/어미)+조사 → 띄어쓰기
10) 부사+어미 → 띄어쓰기
11) 2음절 명사+"되"로 시작하는 어미 → 붙여쓰기
12) (2음절 이상 조사/어미)+어미 → 띄어쓰기

3.3. 음절 bigram 정보 적용

강승식[3]은 말뭉치에서 각 bigram 음절쌍 <X,Y>에 대해 공백의 출현 위치에 따라 좌공백 빈도, 우공백 빈도, 사이공백 빈도, 총 출현 횟수를 계산하여 임의의 두 음절 사이에 공백이 삽입될 확률을 계산하였다. 최종적으로 계산한 값이 실험에 의해 결정한 경험적 임계치를 넘으면 공백을 삽입하는 것으로 결정하였다.

$$P(X_i, X_{i-1}) = 0.25 \cdot P_R(X_{i-1}, X_i) + 0.5 \cdot P_M(X_i, X_{i-1}) + 0.25 \cdot P_L(X_{i+1}, X_{i-2})$$

오류를 수정하는 목적으로 위의 방법을 변경하여 적용하였다. 하나의 임계치만을 가지고 띄어쓰기를 결정하는 것이 아니라 확실하게 띄어 쓰는 임계치(t_{space}=0.6)와 붙여 쓰는 임계치($t_{nonspace}$=0.4)를 설정해 좌공백 확률, 사이공백 확률, 우공백 확률 세 개의 값 모두 띄어 쓰는 임계치 초과 또는 붙여 쓰는 임계치 미만일 경우에 대해서만 3.1.의 띄어쓰기 결과를 수정한다.

음절 X_i와 X_{i+1} 붙여 쓰는 조건 :

$$P_R(X_{i-1}, X_i) > t_{space} \ \& \ P_M(X_i, X_{i-1}) > t_{space} \ \& \ P_L(X_{i-1}, X_{i+2}) > t_{space}$$

음절 X_i와 X_{i+1} 띄어 쓰는 조건 :

$$P_R(X_{i-1}, X_i) < t_{nonspace} \ \& \ P_M(X_i, X_{i+1}) < t_{nonspace} \ \& \ P_L(X_{i+1}, X_{i+2}) < t_{nonspace}$$

3.4. 음절 4-gram 정보 적용

시스템이 결정을 잘못 내리는 오류들에 대해 에러 사전을 구축하여 성능을 향상시킬 수 있다. 기본적인 에러 사전 구축은 학습이 완료된 자동 띄어 쓰기 시스템에 학습 말뭉치를 입력으로 하여 출력된 결과에 대한 띄어쓰기 에러를 찾아낸다. 에러는 붙여 쓰는 곳을 띄어 쓴 삽입 에러, 띄어 쓰는 곳을 붙여 쓴 삭제 에러 두 가지가 있다. 각각의 에러에 대해 발생한 위치 앞뒤 2음절씩 총 4음절을 에러 종류 표시와 함께 에러 사전에 추가를 한다.

언어적으로 띄어쓰기가 두 가지 이상으로 허용되는 경우나 말뭉치의 띄어쓰기 오류로 인해 에러로 판단되어 에러 사전에 포함되는 것을 피해야 한다. 이를 위해서 에러로 판단되는 음절열을 가지고 말뭉치에서 띄어 쓴 경우와 붙여 쓴 경우의 빈도를 조사하였다. 그 결과를 바탕으로 빈도를 비교하여 높은 쪽으로 결정되도록 에러 사전을 구축하였다.

에러 사전에 있는 음절열에 대해서 삽입 에러로 표시된 경우는 무조건 붙여 쓰고, 삭제 에러로 표시된 경우는 무조건 띄어 쓴다.

4. 실험 및 결과 분석

1,051,481개의 음성 인식 단위로 분할된 문장에 대해 학습(1,040,966 문장)과 테스트(10,515 문장) 말뭉치로 나누어 사용하였다. 품사가 부착된 경우와 미부착 된 경우에 대한 성능 비교를 위해 말뭉치는 품사 정보가 부착된 말뭉치에서 품사를 제거하여 미부착 말뭉치를 만들어 실험을 하였다. 학습 말뭉치와 테스트 말뭉치 각각에 대해 띄어쓰기 정답 말뭉치가 존재한다.

품사 미부착 말뭉치에 대해서는 음성인식 단위에 대한 품사 사전을 이용하여 품사 정보를 복원한 경우와 일반적인 형태소 분석기의 품사 사전을 이용한 경우로 구분하여 실험을 하였다.

"품사 정보를 이용한 어절 재결합", "음절 bigram 정보 적용", "음절 4-gram 정보 적용"을 방법(1), 방법(2), 방법(3)으로 하여 단계별로 적용하였을 경우 어절 단위 recall, precision, 그리고 F-score 값의 변화를 확인하였다. F-score는 다음과 같이 계산된다.

$$F\text{-}score = \frac{2 \times precision \times recall}{precision + recall}$$

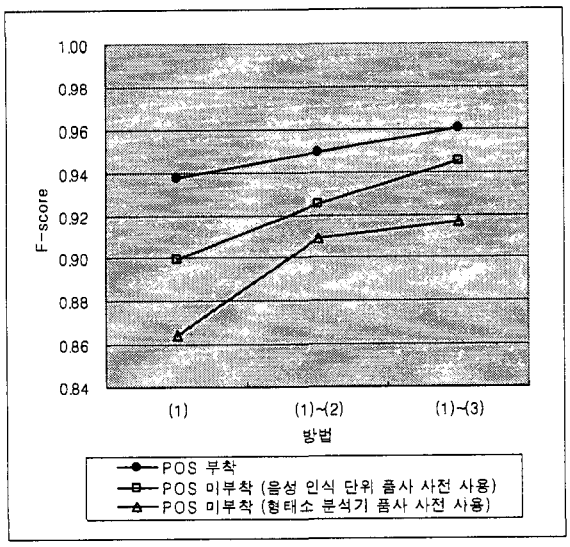

그림 1. 품사 부착 방법에 따른 성능 비교

결과를 살펴보면 방법(1)만을 적용할 때보다 방법(2), 방법(3)을 단계별로 적용할 때 성능이 좋아지는 것을 확인할 수 있고 품사 정보가 부착된 테스트 말뭉치에 대한 띄어쓰기 결과가 전체적으로 좋은 결과를 보인다. 그리고 품사 정보가 미부착된 테스트 말뭉치에 대한 실험에서 일반적인 형태소 분석 사전을 이용한 결과보다 음성인식 단위에 대한 품사 사전을 사용하였을 때의 성능이 상대적으로 높았다.

표 1. 품사 부착 말뭉치에 대한 결과

방법	recall	precision	F-score
(1)	0.9964	0.8859	0.9379
(1) ~ (2)	0.9869	0.9140	0.9491
(1) ~ (3)	0.9766	0.9448	0.9605

표 2. 품사 미부착 말뭉치에 대한 결과
(음성인식 단위에 대한 품사 사전 사용)

방법	recall	precision	F-score
(1)	0.9561	0.8492	0.8995
(1) ~ (2)	0.9598	0.8931	0.9253
(1) ~ (3)	0.9604	0.9311	0.9455

표 3. 품사 미부착 말뭉치에 대한 결과
(일반적인 형태소 분석기의 품사 사전 사용)

방법	recall	precision	F-score
(1)	0.9342	0.8044	0.8644
(1) ~ (2)	0.9438	0.8768	0.9091
(1) ~ (3)	0.9342	0.9003	0.9169

5. 결론

음성 인식 결과에 대한 띄어쓰기 오류를 교정하기 위하여 어절 재결합 규칙과 음절 바이그램 및 4-gram 기법을 적용하는 방법으로 시스템을 구현하여 그 성능을 확인하였다. 기본적인 어절 재결합 규칙 방법만을 사용한 결과보다 음절 N-gram을 적용한 결과의 성능이 높았다. 품사 정보가 부착된 경우에 어절 단위 F-score가 0.023의 성능 향상, 미부착 테스트 말뭉치의 경우는 0.046과 0.053의 성능 향상을 확인할 수 있었다. 향후 성능을 높일 수 있는 기법들을 추가하여 현재의 규칙들을 보완할 필요가 있다.

참고 문헌

[1] 최재혁 외 4인, "연속 음성 인식 후처리를 위한 음절 복원 rule-based 시스템과 형태소 분석 기법의 적용", 전자공학회 논문지, 제36권, C편, 제3호, pp.47-57, 1999.

[2] 박미성 외 6인, "형태소 분석 기법을 이용한 음성 인식 후처리", 전자공학회 논문지, 제36권, C편, 제4호, pp.65-77, 1999.

[3] 강승식, "음절 bigram을 이용한 띄어쓰기 오류의 자동 교정", 음성과학회논문지, 제8권, 제2호, pp.83-90, 2001.

[4] 이도길 외 3인 "한글 문장의 자동 띄어쓰기를 위한 두 가지 통계적 모델", 정보과학회논문지, 소프트웨어 및 응용 제30권, 제4호, pp. 358-370, 2003.

[5] Nakagawa and Tetsuji, Chinese and Japanese Word Segmentation Using Word-Level and Character-Level Information, In Proceedings of COLING, pp.466-472, 2004.