한국어 연속음성 인식을 위한 형태론적 변형 처리

정경석° 박혁로
전남대학교 전산학과
{ksjung@dal.chonnam.ac.kr°, hyukro@chonnam.ac.kr}

Processing of Morphological Transformation for Korean Continuous Speech Recognition

Kyung Seok Jeong° Hyuk Ro Park
Dept. of Computer Science, Chonnam National University

요 약

한국어는 형태론적 변형 현상이 자주 일어나게 되어 최종적으로 음성인식의 성능에 좋지 않은 영향을 끼친다. 본 논문에서는 연속음성 인식의 성능 개선을 위해 형태론적 변형을 처리하는 방법을 제시하고 짧은 형태소를 결합하여 의사형태소를 추출하고자 한다. 이 방법은 음성인식의 성능 개선을 위하여 품사 세트와 사전을 다시 정의하고 텍스트 정규화를 수행한다. 그리고 불규칙 용언 처리의 규칙을 작성하고 나머지 형태론적 변형현상은 에러 패턴을 분석하여 빈출 어휘 중심 및 다단계로 규칙 처리하였다. 마지막으로, 단음절 형태소들을 결합함으로써, 최종적으로 원하는 의사형태소를 구할 수 있었다. 제안된 시스템은 오 인식률이 높은 단음절 형태소들을 결합하여 성능 향상이 기대됨은 물론, 형태론적 변형현상에서는, 9~10%의 높은 성능 향상을 가져올 수 있었다.

1. 서 론

한국어 음성 인식에 관련된 연구는 주로 음성신호의 패턴을 분석하여 그 패턴에 가장 근접한 글자를 찾아내는 음성신호 처리기술 기반의 고립단어 인식에 관한 연구에 치중되어 왔다. 그러나 화자독립 대용량 연속음성 인식이 대두 되면서 자연어 처리에 기반 한 학습 및 후처리 등의 중요성이 강조되고 있다.

일반적인 한국어 자연어 처리의 과정은 서구 언어와는 달리 그 구조적 특성상 형태소 분석이 선행되어야 한다. 한국어는 교착어로서 의미를 나타내는 실질 형태소에 조사와 어미 같은 어법적 관계를 나타내는 형식 형태소가 붙음으로써 문법 기능을 한다. 하지만, 문자 기반의 자연어 처리 방법을 그대로 음성인식 시스템에 적용하기란 쉽지 않다. 이는 각 낱말의 어미변화에 의해 문장의 성분이 결정되며, 첨용과 활용이 자유롭고, 불규칙 현상 및 음운 현상이 발달해 있어서 정확한 입력 값을 기대하기가 어렵다는 것이다. 또한, 일례로 형태소 단위의 음성인식 시스템에서는 다음과 같은 문제점을 볼 수 있다. 예를 들면, 'ㄴ', 'ㄹ', '이' 등과 같은 단음소와 대부분 의존명사, 접미사의 경우는 단음절로서 하나의 형태소이다. 이와 같은 형태소는 음성인식에서 매우 짧은 시간 동안에 발성되기 때문에 이를 인식하기에는 많은 어려움이 있다. 그리하여, 가급적으로 언어학적인 단위인 형태소를 유지하면서 음성인식에 그다지 무리가 가지 않은 범위 내에서 적절한 형태소들을 결합한 새로운 디코딩 단위인 의사 형태소(Pseudo-Morpheme)의 정의가 필요하게 된다.

한국어 대 어휘 연속음성 인식을 위해서는 형태소를 인식 단위로 하는 것이 일반적이다[1,2,3]. [1]의 연구는 형태소 자체를 인식단위로 선정하여 위와 같은 문제점을 발견할 수 있었다. 이런 이유로 형태론적 변형처리와 형태소를 인식 단위로 선정할 때 인접한 여러 형태소들을 결합하여 새로운 인식 단위를 생성하려는 시도가 진행되어 왔다[2][3]. 하지만 [3]은 한정된 불규칙 용언을 정의한 후, 같은 음소의 나열일 경우는, 빈도수가 높은 규칙만을 적용함으로써 좋지 않은 결과를 나타내고, 더욱 구체적인 규칙을 적용하지 못하여, 다양한 형태론적 변형을 고려하지 않고 있다. 예를 들어 '따라'를 형태소 분석하면 '따르+아'로 용언이 복원되는데, 단순히 불규칙 용언의 '르' 불규칙을 적용되어, '딸+라'로 되어 원하는 입력 값을 찾을 수 없게 된다. 이는 다단계 형태론적 변형을 이용해 다시금, 'ㄹ'의 중복임을 인지하고 'ㄹ'을 탈락시킴으로 얻고자 하는 의사형태소인 '따+라'를 얻을 수 있다.

본 논문은 [2][3]에서 제안한 의사형태소 사용을 채택하고 빈출 어휘 중심으로 다단계 형태론적 변형한다. 그리고 단음절을 줄여 오 인식률을 줄이고자 한다. 1장의 서론에 이어, 2장에서는 한국어의 형태론적 변형을 소개하고, 3장에서는 의사형태소 추출기를 논하며, 4장에서 실험 및 고찰, 5장에서는 결론을 기술한다.

2. 한국어의 형태론적 변형

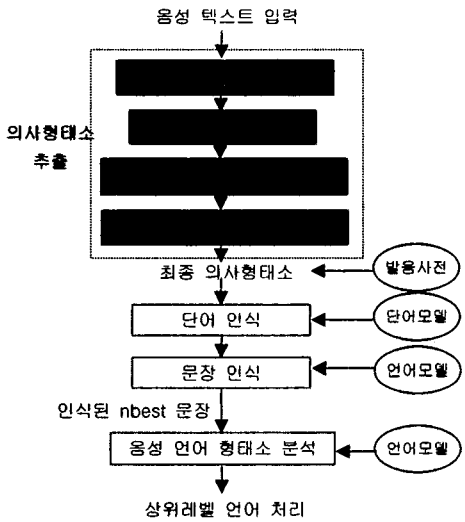
한국어에서 형태론적 변형은 형태소가 결합 시 어떤 형태소의 일부가 변형되는 언어 현상이다. 한국어에서는 형태소의 일부가 변형되는 경우뿐만 아니라 형태소가 탈락 또는 축약되는 경우도 형태론적 변형으로 간주한다. 영어와 같은 굴절어에 비해 교착어인 한국어의 어절 형성 규칙은 좀더 복잡한 양상을 띠고 있으며, 형태론적 변형이 일어나는 원인이 다양하다. 형태론적 변형의 대표적인 예로는 [표 1]와 같다[4].

[표 1] 한국어의 형태론적 변형 종류와 예

형태론적 변형 종류	예
불규칙 용언 활용	걸으니 -> 걷+으니(ㄷ 불규칙)
매개모음	같은 -> 같+ㄴ('으' 탈락)
준말과 존칭	에선 -> 에서+는(준말)

3. 의사 형태소 추출

연속음성 인식기는 실제 음성의 정답 텍스트 문서를 입력을 시작으로 의사형태소를 구성한 후, 각 모델을 통하여 인식 결과인 nbest 문장을 구한다. 그리고 후처리를 통해 최종적으로 인식된 문장을 얻어 낼 수 있다[3]. 본 논문의 의사형태소 추출기의 구조도는 [그림 1]과 같은 연속음성 인식기의 점선 모양과 같다.



[그림 1] 연속음성 인식기의 의사형태소 추출기 구성도

3.1 품사세트 및 사전 구성

품사의 수와 사전의 구성은 음성 인식 시 많은 영향을 끼친다. 음성에서 텍스트로의 변환만을 목적으로 할 경우, 10개 이내의 대분류 품사(명사, 조사, 동사, 어미 등)로도 충분하다. 하지만, 다음 단계의 음성이해를 염두에 두고 세분류에 의한 품사를 사용하고 일반형태소 결과를 사전 표제어로 사용하면 형태소 해석을 다시 할 필요가 없어지는 장점이 있다. 대부분 한국어 형태소 분석에 있어서는 40-50개 정도로 분류하는 것을 기준으로 하고 있다. 본 논문에서는 47개의 의사 형태소를 정의하였다[5].

사전 형태소의 길이가 길수록 발화의 길이가 길어서 인식 단위로는 좋지만, 사전의 크기가 커지고 이용범위도 떨어지게 된다. 형태소 분석기는 일부 복합명사 및 복합조사 등을 포함하는데 이들을 구성하는 각각의 형태소들은 빈번하게 사용되는 형태소들로, 각각의 작은 단위 명사로 줄였다. 특히, 복합명사의 경우 4음절 이상이면 사전에서 제거 하였다.

[표 2] 사전의 짧은 형태소를 구성하는 예

품사	형태소	짧은 형태소
복합명사	정보통신	정보+통신
조사	까지와는	까지+와+는
어미	었던	었더+ㄴ

3.2 텍스트 정규화

실제 입력되는 문장에는 특수기호 및 알파벳, 아라비안 숫자, 알파벳과 숫자가 혼용된 약어, 영문 단위명사 등이 포함 되어 있다. 게다가 아라비안 숫자의 경우 문맥에 따라 숫자를 읽는 방식이 달라지기 때문에 텍스트 정규화 내부에서 이 모든 현상을 수용하여 처리하는 것은 매우 어렵다. 그러므로 특정 시스템이라는 가정 아래에 몇 가지 특수 문자들은 임의로 제거하고 괄호 문자와 기호 단위, 아라비아 숫자, 알파벳 등을 처리 하였다. 예를 들어 'apple은 200₩이다.'라고 입력 한다면 '애플+은+이백+원+이다'로 한글화 된다.

3.3 형태론적 변형 처리

불규칙 용언의 활용의 종류는 총 15가지로 정의 된다[6].

ㄷ-불규칙, ㅂ-불규칙, ㅅ-불규칙, ㅎ-불규칙,
러-불규칙, 르-불규칙, 우-불규칙, ㄹ-탈락,
ㅓ-불규칙, 하여/해-불규칙, 외-불규칙, 으-불규칙,
서술격 조사 -이 축약, 거라-불규칙, 너라-불규칙

이에 따라 본 연구는 한국어 문법에 나타난 모든 불규칙 용언 활용에 관한 규칙을 정의하였고 그 외의 모든 형태론적 변형 규칙을 빈출 어휘 우선 순으로 작성하였다.

어떤 어절에서는 두 가지 이상의 형태론적 변형이 복합적으로 일어나기도 한다. 예를 들어 '아롱다워서'이라고 입력한다면 '아롱답+어서'으로 분석되어 'ㅂ'이 탈락되고 'ㅜ'라는 모음이 첨가된다('ㅂ' 불규칙 용언). 그 다음 'ㅜ'와 'ㅓ'의 모음들이 합쳐져서 다시 '아롱다+워서'라는 입력 열을 구할 수 있게 된다. 이는 두 가지 현상을 순서대로 처리하지 못할 경우 틀린 결과를 가져올 수 있다는 말이다. 그래서 불규칙 용언과 또 다른 형태론적 변형 규칙을 순서적으로 다단계로 적용해야 한다.

또한 기존의 한국어 형태소 품사태그 집합으로는 표기하기 힘든 경우들이 대화체 문장에서 자주 발생하게 되는데, 대표적으로 준말과 같은 경우가 그러하다. 하지만 이를 해결하기 위해 준말의 어휘와 태그를 사전에 수록하는 기준을 정하기가 어렵다. 그래서 본 논문은 대표 형태소만을 사전에 수록하고 변형 규칙에 의하여 처리할 것인지를 결정한다. 변형 규칙 또한, 예외적인 모든 변형 현상을 고려할 수 없으므로 에러패턴 중 코퍼스에서 빈출 어휘 중심으로 규칙화 하였다. 대표적인 변형규칙은 다음 [표 3]과 같이 정의 된다.

[표 3] 대표적인 변형 규칙 과 예

변형 규칙	예
매개모음 '으' 탈락	같은->같+ㄴ->같+은('으'추가)
하여/해 불규칙 특수 처리	했다->하+었+다->했+다('ㅐ'자소확인)
'하' 첨가	의도지만->의도+하+지만->의도+지만('하' 탈락)
다단계 변형	아름다워서->아름답+어서->아름다+우+어서('ㅂ'불규칙적용)->아름다워서(모음합성)

[표 3]의 2번째 같은 경우, 불규칙 '하였다'와 '했다'는 모두 '하+었+다'로만 분석되므로 [3]의 논문은 빈도수가 높은 규칙을 적용하지만, 본 논문은 입력된 텍스트의 특정 자소를 확인함으로써, '하+였+다'와 '했+다'를 구분지어 줄 수 있다. 마지막으로, 보편적인 변형규칙으로 정의할 수 없는 몇 가지 예외적인 어휘들도 또한, 빈출 어휘 중심으로 예외 규칙으로 추가 시켰다. 대표적인 예외적인 어휘는 [표 4]와 같다.

[표 4] 예외 어휘의 종류와 예

변형 현상	예
준말	뭔가 -> 무어+ㄴ가 -> 무언가(결합)
변형규칙오류	잡을 -> 잡+ㄹ -> 잡+울('ㅂ' 불규칙 적용)

3.4 결합형태소 구성

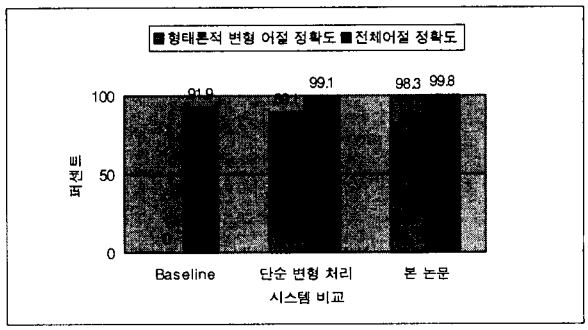
한국어 연속음성 인식에서는 짧은 형태소는 음향 모델에서 인식 시, 짧은 발화로 인해 오 인식될 확률이 매우 높아 최종 음성 인식에 좋지 않은 결과를 야기한다. 특히, 50%에 달하는 단음절 형태소들을 말할 수 있다. 인접한 단어들을 결합하는 방법으로는 규칙 기반 방법과 통계 기반 방법으로 나뉜다. 규칙 기반 방법은 각 형태소의 품사 정보를 이용하여 접속규칙을 생성하고 각 접속 규칙에 해당하는 형태소들을 결합하여 결합형태소를 생성하는 반면, 통계적 방법은 평가척도를 이용하여 연관성이 많은 형태소들을 결합해 간다[7].

이 논문에서는 [2]의 논문에서 정의한 규칙 기반 방법으로 하고 단 음절만 우선하는 제약 규칙을 적용한다. 이는 결합 형태소를 길이가 길어질수록 에러율은 줄어들지만, 학습 및 탐색 시 혼잡도를 증가시키는 원인이 되기 때문이다.

4. 실험 및 고찰

인식 실험에 사용된 음성 코퍼스는 방송뉴스로써 아나운서와 리포터 등이 발음한 문서로, OOV(Out of Vocabulary)를 포함하고 있다. 인식 대상의 문장 개수는 대략 15,000문장이고, 한 문장 당 평균 8~9어절로 구성되어 있다. 이 중, 형태론적 변형이 일어난 어절은 총 어절 중 8%로 한 문장 당 한 번은 형태론적 변형이 일어난다고 말 할 수 있다. CMU-SLM[8]을 이용한 결과, 의사형태소의 수는 2만 3천개정도이고 단음절의 형태소는 빈도가 30%정도로 감소되어 [7]의 시스템과 비슷한 성능을 나타낼 수 있었다.

또한, 형태론적 변형처리에 관한 어절 단위로 실험 평가를 수행하였다. 형태론적 변형처리를 하지 않은 시스템과 단순한 불규칙 용언 활용과 규칙만을 고려한 시스템 그리고, 이 논문에서 주장하는 에러패턴을 분석한 후, 빈출 어휘 중심으로 형태론적 변형의 다단계 적용한 결과는 아래 [그림 2]와 같다.



[그림 2] 어절 정확도 평가

본 논문에서 일어난 실패는 대부분이 OOV로 인한 형태소 분석 실패로 인한 오류였다. 예를 들면, '나말라에'라는 장소명의 입력이 들어 왔을 때 '나+아+말+라+에'로 잘 못 분석되어 음운 변이가 일어나 버리는 경우를 들 수 있다.

5. 결론 및 향후 연구과제

본 논문에서는 한국어 연속음성 인식을 위한 형태론적 변형을 효과적으로 처리하는 방법에 대해 소개하였다. 그리고 형태소를 결합함으로써 오 인식률이 높은 단 음절을 줄임으로써 음성 인식 시 좋은 성능을 기대할 수 있다. 향후 연구로는 통계적 척도를 이용한 형태소 결합을 통해 규칙 기반 방법을 보완하는 연구가 있다. 또한, 음성 인식 시 언어 모델에서 문법적 정보를 이용한 품사 단위의 트라이그램 등을 적용하여 후보 노드를 탈락시켜 학습이나 탐색의 비용을 줄일 수 있는 언어모델을 연구해 볼 계획이다.

참고 문헌

[1] Ha-Jin Yu, H. Kim, J.S. Choi, J.M. Hong, K.S. Park, J.S. Lee, H.Y. Lee, "Automatic recognition of Korean broadcast news speech," Proc. of ICSLP, 1998.

[2] Oh-Wook Kwon and Jun Park, "Korean large vocabulary continuous speech recognition with morpheme-based recognition units," Speech Communication, 2002. (in press)

[3] 이경님, 정민화, "의사 형태소 단위의 음성언어 형태소 해석.", 한글 및 한국어 정보처리 학회지, pp.396-404, 1998.

[4] 강승식, 한국어 형태소 분석과 정보 검색, 2002.

[5] 김재훈, 서정연, "자연언어 처리를 위한 한국어 품사 태그," 한국과학기술원, 인공지능센터, 1994.

[6] 남기심, 고영근, 표준 국어 문법론, 1998.

[7] 박영희, 정민화, "대어휘 연속음성 인식을 위한 결합형태소 자동생성," 한국음향학회지, 2002.

[8] The CMU Statistical Language Modeling (SLM) Toolkit, http://www.speech.cs.cmu.edu/SLM_info.html