표준안에 따른 품사 부착 말뭉치 구축

이현아, 이원일*, 임선숙**, 허은경, 이재성*, 차건회*, 박재득*
㈜에스엔엘, *ETRI 지식정보연구부, **충남대 국어국문학과
대전시 유성구 어은동 1, 우:305-333
{halee,redann}@snl.co.kr, {wonilee, jasonl, chakh, jdpark}@etri.re.kr

Part-of-speech Tagged Corpus Construction for ETRI Standardization

Hyun-A Lee, Wonil Lee*, Sun-Suk Lim**, Eun-Kyung Her, Jae Sung Lee*, Keon-Hoe Cha*, Jay Duke Park*
Speech and Language Inc., *ETRI Knowledge Information Department, **ChungNam National University
1 Eoeun-dong, Yusong-gu, Taejon, 305-333

요약

본 논문에서는 한국전자통신 연구원 지식정보 연구부에서 제안하는 자연어 정보처리 기술 표준안을 적용하여 품사 부착 말뭉치를 구축하는 과정에서 논란의 여지가 있었던 대표적인 사항들에 대해 기술한다. 아울러 ETRI 표준안이 도출된 원칙과 취지 등을 품사 부착 말뭉치 구축과 관련하여 설명하고, 현재의 ETRI 표준안이 앞으로 어떤 식으로 개선되어야 할 지에 대해 제안한다.

1 서론

본 논문에서는 한국전자통신연구원에서 제안하는 자연어 정보처리 기술 표준안에 따른 품사 부착 말뭉치인 'ETRI 품사 태그 부착 말뭉치 Ver. 2.0' (이하 ETRI Corpus)을 구축하는 과정에서 논란의 여지가 있었던 대표적인 사항들에 관해 기술한다.

ETRI Corpus 구축에 관한 기술에 앞서 ETRI 표준안의 도출 원칙과 그 취지에 관해 먼저 기술하고자 한다.

ETRI 표준안은 크게 3개의 원칙 하에 도출되었다. 첫째, 기존의 연구 결과를 적극 수용하고 최대한 호환이 가능한 표준안을 도출한다. 이를 위해 주요 5개 기관을 선정하고 각 기관의 연구 결과를 수집하여 서로 비교, 분석하는 과정을 수행했다. 둘째, 실제 처리 시스템에 적용하기에 앞서 가장 중립적인 표준안으로 도출한다. 따라서 현재의 ETRI 표준안은 응용시스템(형태소 해석기 등)에 바로 적용하기에는 적합하지 않은 부분도 분명히 포함하고 있다. 이런 부분은 ETRI 표준안에 대한 검토 결과를 모아서 응용 시스템에 적합하게 수정, 보완되어야 할 것이다. 마지막으로 의미적 판단을 요하는 부분은 되도록 배제하자는 원칙을 적용하였다. 이에 따라, 대명사의 하위분류, 고유명사 분류 등 의미적으로 달라지는 품사 카테고리를 가능한 한 배제하였다.

이러한 세 원칙 하에 ETRI 표준안의 도출 과정은 크게 세 단계로 이루어진다. 1단계에서는 기존의 연구 결과를 수집하여 서로 비교하여 공통점과 차이점을 분석해 내었고, 특히 차이가 나는 부분에 대한 타협점을 찾는 데에 주력하였다. 2단계에서는 1단계의 분석 결과를 바탕으로 표준화 위원회에서 여러 차례 회의를 통해 초기 표준안을 도출하였다. 3단계에서는 초기 표준안을 정비 및 보완하기 위하여 표준안에 따른 품사 부착 말뭉치를 구축하였다. 3단계의 결과물이 바로 ETRI Corpus이고, ETRI Corpus는 ETRI 초기 표준안을 응용 시스템에 적합하도록 정비, 보완하기 위한 시범 구축에 그 목적이 있다.

본론에서는 먼저 ETRI Corpus의 일반적 사항에 대해 기술하고, ETRI Corpus의 구축 과정에서 오류로 나타나기 쉬운 현상이나 논란의 여지가 있는 사항들에 대해 기술한 후, 앞으로 ETRI 표준안의 개선 방향을 제안한다.

2 ETRI Corpus의 개요

ETRI Corpus는 약 30만 어절로 97년에 구축되었던 SERI/KAIST corpus 중 약 26만 어절을 표준안에 따라 재구축하였고 신규로 약 3만 어절을 추가 구축하였다.

ETRI Corpus는 소설, 비소설(설명문, 논설문, 전문 분야 문헌 등), 신문, 방송 뉴스 스크립트 등으로 이루어져 있다[표2-1].

[표2-1] ETRI Corpus의 분야비

분야	비율	분야	비율
소설	46.6%	신문	2.7%
비소설	50.0%	방송 뉴스 스크립트	0.7%

ETRI Corpus의 파일 형식은 텍스트로 되어 있고, 파일 이름의 형식은 소설의 경우 'NOVEL#.txt', 비소설은 'EXPL#.txt', 신문의 경우 'NEWSP#.txt', 방송 뉴스 스크립트는 'NEWS#.txt'로 정했다.

전반적인 외형은 SERI/KAIST corpus의 형식을 대부분 수용하였다. 한 줄에 한 어절씩 두는 vertical 방식을 수용했고 한 어절은 '어절 번호, 원어절, 태깅 결과'의 세 필드로 이루어진다. 각 파일의 헤드 정보도 SERI/KAIST corpus의 형식을 대부분 수용하였다.

구축 방식은 원시 말뭉치를 기존의 자동 태거로 초벌 태깅한 후, 작업자가 표준안에 따라 수작업 형태로 정련하는 방식을 취하였다.

3 ETRI Corpus 구축과 표준안에 대한 고찰

본 장에서는 ETRI Corpus의 구축 과정 중 ETRI 표준안을 적용하면서 논란의 여지가 있는 사항에 대해 부분적으로 기술한다. ETRI Corpus는 ETRI 표준안을 1차 검증, 보완하기 위해 시범 구축된 것으로 구축 과정에서 초기의 ETRI 표준안을 부분적으로 수정, 보완하는 작업을 병행하였으나, 응용시스템에 직접 적용하기에 부적합한 부분은 그대로 남아 있다. 앞서 기술한 대로 가장 중립적인 표준안에 따라 ETRI Corpus가 구축되면서 논란의 여지가 되었던 대표적인 사항들에 대해 기술함으로써 ETRI 표준안이 앞으로 어떤 방향으로 수정, 보완되어야 하는지에 대한 방향을 제시하고자 한다. 본 논문에서 기술하는 내용은 전체 중 일부 대표적인 내용만을 예로 든 것이다.

3.1 형식 오류

ETRI Corpus 구축의 작업 방식이 수작업이었기 때문에 형식 오류를 포함하였다. 형식 오류는 태깅 결과를 표현하는 형식을 벗어난 오류와 태깅 결과와 원어절이 일치하지 않는 오류 등이 나타났다. ETRI Corpus에서는 태깅 결과를 나타낼 때 형태소와 품사는 '/'로 구분하고, 형태소와 형태소는 '+'로 연결한다. 이러한 약정 기호가 빠졌거나 품사를 표현하는 심볼이 잘못된 경우 등이 오류로 나타났다.

이러한 형식 오류는 자동 검사 프로그램 등을 활용하여 수정하였다.

3.2 복합명사의 태깅

본 절에서는 ETRI 표준안에서 태깅 시 복합명사와 관련하여 고려한 사항을 기술한다.

첫째, 기본적으로 복합명사는 복합명사를 이루는 단위명사로 분할하여 태깅하였다. '계급투쟁, 계급의식, 신문기사, 공중전화, 공중변소' 등이 이에 해당된다.

그러나 복합명사 중에는 분할하지 않고 하나의 명사로 봐야 하는 예외의 경우가 있다. 이런 예외에 속하는 경우는 크게 세 가지로 들 수 있는데, 고유명사 역할을 하는 어휘가 여기에 해당되고 완전히 하나의 단어로 굳어진 경우가 또한 해당되며, 분할한 어휘가 독립적으로 쓰이지 않는 경우가 해당된다.

고유명사 역할을 하는 어휘나 완전히 하나의 단어로 굳어진 어휘는 분할을 하면 전체의 의미를 상실하여 정보를 잃게 되므로 반드시 하나의 단위로 처리해야 한다.

고유명사 역할을 하는 어휘에는 회사명, 지역명, 제품명 등이 해당되는데, '한국통신, 삼성물산, 남아프리카, 뉴욕시, 동아시아, 해태제과, 구룡반도' 등을 예로 들 수 있다. 그러나 여기에서도 다시 예외의 경우가 생기는데, 이런 어휘들이 보통명사의 역할도 같이 하는 경우이다. 예를 들어 '한국통신'이 회사 이름이 아닌 '한국의 통신'이라는 뜻이 되면 '한국/nc+통신/nc'로 분할하는 것이 더 적합하다. 결국 이 과정에서 ETRI 표준안의 원칙 중 의미적인 기준을 가능한 배제하겠다는 부분이 어긋나게 된다. 즉, ETRI 표준안에서 의미적 판단을 요한다고 해서 '고유명사' 카테고리를 배제했는데, 실제로 복합명사를 고유명사 역할을 하는 경우와 그렇지 않은 경우로 일일이 구별하면서 태깅을 했기 때문에 '고유명사'를 제외한 의미를 어느 정도 상실했다고 할 수 있다. 그러나 '고유명사'와 관련한 이슈는 복합명사뿐만 아니라 단일명사 중에서도 어디까지를 '고유명사'로 볼 것인지에 대한 명확한 기준이 없기 때문에 '고유명사'를 제외시킨 의미를 모두 상실한 것은 아니다.

완전히 하나의 단어로 의미가 굳어진 어휘 역시 하나의 명사로 봐야 한다. '고등학교, 중학교, 초등학교, 입밖, 뜻밖, 눈앞, 코앞, 백색가루, 엉망진창, 시계바늘' 등은 전체를 하나의 명사로 태깅하였다. 이런 어휘들은 단위명사 각각이 독립적으로 쓰일 수 있지만 분할을 했을 때와 전체의 의미가 서로 다르기 때문에 하나의 명사로 태깅해야 한다. '백색가루'는 '마약'을

뜻하는 다른 표현으로 쓰였기 때문에 하나의 명사로 보았고, '입밖, 뜻밖, 엉망진창'은 '입밖에 내지 마라, 뜻밖의 행운, 엉망진창이다.'처럼 관용적인 표현에만 쓰이는 표현이므로 하나의 명사로 보았다. '시계바늘' 같은 어휘도 '시계'와 '바늘'과는 전혀 관계없는 객체를 가리키는 말이므로 하나의 명사로 보았다.

'동구밖' 같은 어휘는 '동구'가 국어사전에 표제어로 등록되어 있기는 하지만 실제 '동구'가 독립적으로 쓰이는 예는 거의 없고 항상 '동구밖'으로만 쓰이므로 하나의 명사로 보았다. '여비서, 여직원' 등은 '여'를 접두사 역할을 하는 것으로 보아 하나의 명사로 태깅하였다.

이 외에 복합명사로 보기는 어렵지만 사자성어도 분할을 하면 전체의미를 상실한다고 보아 하나의 명사로 태깅하였다. '양자택일, 대서특필, 용맹정진, 양자역학' 등이 해당된다.

본 절의 처음에 제시하였던 복합명사의 태깅의 기본 원칙 아래 지금까지 여러 가지의 예외 원칙을 기술하였지만 말뭉치에 나타나는 복합명사의 태깅은 매번 혼란을 가져올 만큼 뚜렷한 원칙과 명확한 기준을 정하기 곤란한 부분이었다. 따라서 복합명사의 태깅은 의미적인 기준과 operational한 기준을 모두 적용하는 것이 적합하다고 생각된다. 의미적 판단이 작업자마다 주관적일 수 있지만 예외가 없는 기준이고, operational한 기준은 객관적이지만 예외가 있기 때문에 두 가지 기준을 모두 고려해야 할 것이다.

3.3 '명사+용언'의 태깅

ETRI 표준안이 가장 중립적인 입장에서 도출된 것을 대표적으로 반영하는 부분이 '명사+용언'의 유형이다. '맛있다, 맛없다, 끝나다, 끝내다, 사기당하다, 도둑맞다' 등이 해당되는데, 흔히 하나의 용언이라고 생각되는 어휘 중 명사와 용언 사이에 조사가 생략되고 동시에 공백이 없어지면서 결합한 어휘가 상당수 포함되어 있다.

이런 유형에 대해 ETRI 표준안은 '명사+용언'로 분리하여 태깅하는 것을 기준으로 정했다. 가장 중립적인 입장을 유지하기 위함이고, 따라서 ETRI 표준안 중 응용시스템에 적합하도록 보완되어야 할 부분이기도 하다. '맛있다'를 하나의 용언으로 보는 것은 어색하지 않지만 '맛없다'를 하나의 용언으로 보는 것은 어색하고, '사기당하다, 폭행당하다'와 같은 어휘를 모두 하나의 용언으로 보는 것이 바람직한가에 대한 판단이 그리 쉽지 않기 때문이다. 또한 각자가 공통된 기준 없이 어떤 어휘는 하나의 용언으로 보고 어떤 어휘는 그렇지 않게 보고 있는 부분이었기 때문에 ETRI 표준안에서는 이를 모두 '명사+용언'으로 분리하여 태깅하는 것으로 지침을 정했다. 물론 이후 단계에서 반드시 이런 유형에 대한 공통 기준을 도출하고 좀더 응용시스템에 적합하도록 보완되어야 한다는 전제가 있다.

여기에도 예외의 경우가 있다. 생략된 조사를 넣어서 간단한 예문을 만들어 봤을 때 의미가 상실하는 경우는 분리하지 않고 하나의 용언으로 보아야 한다. '손쉽다, 색다르다' 등이 해당된다. 이것 역시 operational한 기준과 의미적 판단이 모두 필요한 경우이다.

3.4 파생어의 태깅

ETRI 표준안에서는 용언과 결합하는 접사는 전혀 인정하지 않고 명사와 결합하는 접사도 아주 제한적으로 인정한다. 따라서 파생용언은 항상 전체를 하나의 용언으로 태깅하고, 파생명사는 인정하는 접사만을 분리하고, 인정하지 않는 접사와 결합한 경우에는 전체를 하나의 명사로 태깅한다.

접사를 원칙적으로 인정하지 않는 것은 모든 접사가 모든 명사와 전부 결합하지 않고, 그 결합 관계가 규칙적이지 않기 때문이다. 따라서 ETRI 표준안에서는 기존 연구 결과에서 공통적으로 인정하는 것을 기본적으로 수용하고 접사가 결합된 형태 전체를 하나의 명사로 보기 어려운 경우 접사로 인정하였다. 순수 우리말 접사인 '어치, 짜리, 씩, 께나' 등이 결합된 어휘를 명사로 보기에 어려우므로 접사로 인정하였다.

가장 중립적인 표준안이 되려면 모든 접사를 인정하는 것이 맞겠지만, 기존 연구 결과들이 대부분 접사를 제한적으로만 인정하고 있고, 국어학/언어학 분야에서의 접사에 대한 연구 결과도 명확하게 나와 있지 않기 때문에 기존 연구 결과를 적극 수용하는 원칙을 우선 적용하였다.

3.5 조사와 어미의 태깅

ETRI Corpus에서 가장 응용시스템에 적용하기 어려운 부분이 조사와 어미에 대한 기준일 것이다. 기존의 연구결과에서 대부분 한 가지로 보았던 부분을 ETRI Corpus에서는 여러 경우로 나누었고, 그 기준이 명확하지 않는 것이 사실이다. 그러나 분명한 것은 자동화하기 어렵다고 해서 모든 것을 하나로만 통일시키는 것이 해결책이 되어서는 안될 것이다.

대표적인 예로 '라는/이라는'의 태깅이다. '황진이라는 기생은 당대의 명물이었다'에서 '라는'은 '라고 하는'으로 주제화의 역할을 하는 것으로 보아야 한다. '그녀가 황진이라는 사실은 변함이 없다'에서 '라는'은 지정사 '이'가 생략된 형태로 문장 내에서 '그녀가'의 술어 역할을 한다. 즉, '라는/이라는'은 크게 '라고 하는'으로 보아야 하는 경우와 술어 역할을 하는 것으로 보아야

하는 경우의 두 가지가 있다. 여기에 '라는'의 준말인 '란'의 경우에는 '라는 것은'의 준말 형태인 보조사 '란'과도 그 쓰임새를 구별해야 한다. 즉, '학교란(학교라는 것은) 공부를 하는 곳이다.'에서는 '란'이 보조사이고 '학교란 곳은 공부를 하는 곳이다.'에서 '란'은 '라는'의 준말로 쓰인 경우이다. 이런 각각의 경우를 구별해서 태깅하는 것이 결코 쉽지 않지만, 문장 내에서 술어 역할을 하는 경우와 주제화의 역할을 하는 경우는 분명 구별해야 하는 사항이다.

4 결론

지금까지 ETRI Corpus 구축과 ETRI 표준안이 도출된 기본 원칙과 취지에 관해 기술하였고, ETRI 표준안에 따라 시범 구축된 ETRI Corpus와 관련하여 논란의 여지가 많았던 대표적인 사항 몇 가지에 대해 기술하였다.

현재의 ETRI 표준안은 바로 응용 프로그램에 적용하기에 부적합한 부분이 포함되어 있고 이런 부분은 관련 연구자들이 각자의 시스템에 ETRI 표준안을 적용해 본 후 그 결과로부터 개선점을 찾아내어 ETRI 표준안을 개선시켜야 한다. 즉, 현재의 ETRI 표준안이 표준화 위원들의 그간 연구 경험과 기존 연구 결과의 비교/분석을 통해 1차 도출되었다면 2차 단계에서는 실제 여러 시스템에 적용한 후, 그 결과로부터 개선된 표준안이 도출되어야 한다.

앞으로 ETRI 표준안은 관련 연구자들이 각자의 방법과 각자가 정한 기준들을 모아서 조정하고 통합하는 역할을 해나가야 한다고 감히 제안한다. ETRI 표준안은 또 하나의 방법론이 아니라 기존의 연구 결과들이 모여서 조정되고 통합된 결과물로 보아야 한다. 따라서 서로가 중복되는 연구를 가능한 한 피하고, 서로의 연구 결과를 재활용, 공유할 수 있는 매개체로서의 역할을 해야 할 것이다.

5 참고 문헌

[1] "자연어 정보처리 기술 표준화" 1차년도 표준안 지침서, 1998, ETRI 지식정보연구부