KTAG99: 새로운 환경에 쉽게 적응하는 한국어 품사 태깅 시스템[1]

김재훈*, 선충녕**, 홍상욱**, 이성욱**, 서정연**, 조정미***

*한국해양대학교, 컴퓨터공학과

**서강대학교, 컴퓨터학과

***삼성종합기술원 휴먼인터페이스 Lab.

jhoon@hanara.kmaritime.ac.kr

KTAG99: Highly-Adaptable Koran POS tagging System to New Environments[1]

Jae-Hoon Kim*, Choong Nyoung Sun**, Sang Wook Hong**, Songwook Lee**, Jungyun Seo**, and Jeong Mi Cho**

*Department of Computer Engineering, Korea Maritime University

**Department of Computer Science, Sogang University

**Human & Computer Interaction Lab., Samsung Advanced Institute of Technology

요 약

한국어 정보처리를 위한 언어정보는 응용 분야에 따라 큰 차이를 보인다. 특히 말뭉치를 이용한 연구에서는 언어정보가 달라질 때마다 시스템을 새로 구성해야 하는 어려움이 있다. 본 논문에서는 이와 같은 어려움을 다소 완화시키기 위해 새로운 환경에 잘 적응할 수 있는 한국어 품사 태깅 시스템에 관해서 논한다. 본 논문에서는 이 시스템을 KTAG99라고 칭한다. KTAG99는 크게 실행부와 학습부로 구성되었다. 한국어 품사 태깅을 위한 실행부는 고유명사 추정기, 한국어 형태소 분석기, 통계기반 품사 태거, 품사 태깅 오류교정기로 구성되었으며, 실행부에서 필요한 언어정보를 추출하는 학습부는 고유명사 추정규칙 추출기, 형태소 배열규칙 추출기, 사전 추출기, 확률정보 추정기, 품사 태깅 오류수정 규칙 추정기로 구성되었다. KTAG99에서 필요한 언어정보의 대부분은 학습 말뭉치로부터 추출되거나 추정되기 때문에 아주 짧은 시간 내에 새로운 환경에 적응할 수 있다.

1. 서론

일반적으로 한국어 품사 태깅 시스템은 크게 형태소 분석기와 품사 태거로 구성되어 있다. 형태소 분석기는 주어진 어절에 대해서 가능한 형태소 해석 결과를 출력하는 시스템이다. 즉 형태소 분석의 중의성을 생성하는 시스템이다. 반면에 품사 태거는 형태소 분석기에서 생성된 형태소 분석의 중의성을 해소하는 시스템이다. 이와 같이 두 시스템의 역할이 분명히 구별되어 있으나, 두 시스템은 아주 밀접한 관계를 가지고 있다. 예를 들면, 형태소 분석의 중의성이 작으면 작을수록 품사 태거의 정확률은 높아진다. 또한 형태소 분석기의 성능(과분석이나 오분석 정도, 미등로어 추정 능력 등)은 품사 태거의 성능에 커다란 영향을 준다. 따라서 한국어 품사 태거의 성능을 정확하게 평가하는 일은 대단히 어려운 일 중에 하나이다. 또한 한국어 품사 태깅에 관한 연구는 이미 여러 연구 기관에서 이루어졌으

[1] 본 연구는 한국과학재단의 핵심전문연구(과제번호: 981-0922-115-2)와 한국전자통신연구원 교환전송기술연구소의 연구용역비(과제명: 고유명사 추정을 이용한 형태소 태거의 성능 개선) 지원으로 수행되었다.

나[1][2][3], 이들 연구를 객관적으로 비교할 수 있는 여건(학습 말뭉치의 크기, 품사 태그 집합, 어휘사전의 규모 등)이 조성되지 않았다.

이와 같은 어려운 여건에도 불구하고 한국전자통신연구원(ETRI)의 주최로 금번에 개최된 "제1회 형태소 분석기 및 품사 태거 평가 대회(MATEC99)"[4]는 서로의 시스템을 객관적으로 비교할 수 있는 좋은 계기가 되었다고 생각한다. 왜냐 하면, 같은 학습 말뭉치와 같은 품사 태그 집합 등을 이용해서 형태소 분석기 및 품사 태거에 필요한 정보를 학습함으로써 어느 정도 유사한 환경 하에서 서로의 시스템을 비교할 수 있는 기회를 제공해주었기 때문이다[2].

본 연구팀은 이 대회에 명사 추출기 뿐 아니라 형태소 분석기와 품사 태거 분야에 참가하게 되었으며, 본 논문에서는 이들 각 시스템의 핵심 기술과 각 시스템의 구조에 대해서 간략히 기술한다. 본 연구팀에서 출품한 각 시스템은 독립적으로 구별된 시스템이 아니라 하나의 시스템에서 선택인자(option)에 따라, 명사 추출기, 형태소 분석기, 품사 태거의 기능을 선택할 수 있다. 따라서 본 논문에서는 출품된 각 시스템을 개별적인 구조에 대해서는 설명하지 않고 전체 시스템 구조에 대해서 설명함으로써 각 시스템의 구조를 파악할 수 있도록 한다.

특히 명사 추출기는 품사 태거를 이용하였다. 즉, 품사 태거의 출력에서 보통명사(ncn)[5]에 해당하는 명사만 출력하도록 선택인자를 추가함으로써 구현되었기 때문에 추가적인 설명은 하지 않을 것이다.

본 논문은 2장에서 KTAG99의 구조에 대해서 기술한다. 3장에서 6장까지는 전처리기로서의 고유명사 추정기, 형태소 분석기, 품사 태거후처리기로서의 오류교정기에 대해서 차례로 기술한다. 7장에서 본 논문에서 기술된 각 시스템의 특징에 대해서 기술하고, 마지막으로 8장에서 결론 및 MATEC99에 출품 소감을 간략히 기술한다.

[2] 대회의 의도와는 조금 다르게 각 시스템의 연구환경이 너무 달라서 객관적으로 서로의 시스템을 비교할 수 있는 연구환경을 조성하기가 어려웠다고 생각된다.

2. KTAG99: 한국어 품사 태깅 시스템

본 연구팀이 MATEC99에 출품한 한국어 품사 태깅 시스템을 KTAG99라고 칭한다. 그림 1에서 표현된 바와 같이 KTAG99는 크게 학습부와 실행부로 나눈다.

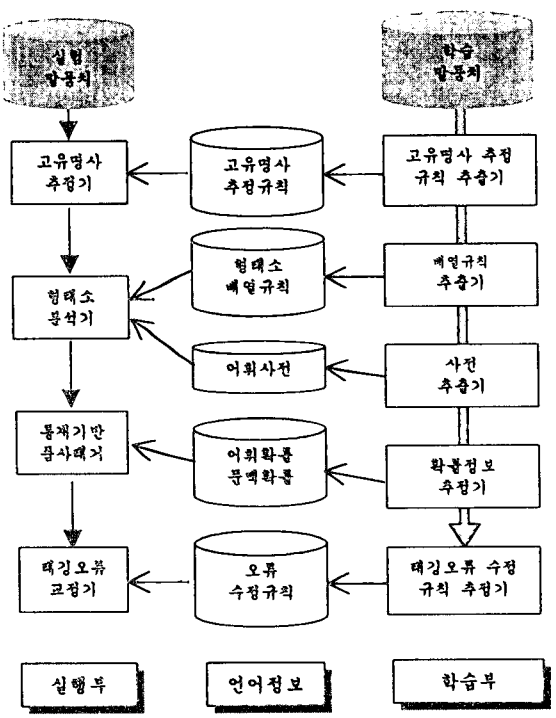

그림 1. KTAG99의 구조

학습부는 실행부에서 필요한 언어정보를 추출한다. 이것이 KTAG99의 가장 큰 장점 중 하나이다. 학습부는 고유명사 추정규칙 추출기, 형태소 배열규칙 추출기, 어휘사전 추출기, 확률정

보 추정기, 품사 태깅 오류 수정규칙 추정기로 구성되었다. 고유명사 추정규칙 추출기는 미등록어에 효과적으로 대처하기 위해서 고유명사를 추정하기 위한 규칙을 추출한다. 형태소 배열규칙 추출기는 학습말뭉치로부터 형태소 배열규칙을 추출한다. 어휘사전 추출기도 학습 말뭉치로부터 형태소 분석에 필요한 어휘 목록을 추출한다. 확률정보 추정기는 품사 태깅을 위한 각종 확률 정보를 학습 말뭉치로부터 추정한다. 품사 태깅 오류 수정규칙 추정기는 통계기반 품사 태거의 오류를 수정하기 위한 오류수정 규칙을 추정한다.

고유명사 추정규칙 추출기를 제외한 모든 학습 시스템은 사람의 손이 개입되지 않고 완전 자동으로 이루어진다. 고유명사 추정규칙 추출기는 고유명사를 추정하기 위한 규칙이 대부분 어휘 문맥을 사용해야 하고 심지어 의미 정보를 요구하기 때문에 완전 자동으로 구축되는 것은 대단히 어려운 일이므로 현재 반자동으로 수행된다. 고유명사 추정규칙 추출기를 이용해서 기본적인 틀을 추출한다. 그리고 나서 기본적인 틀을 전문가에 의해서 더욱 일반화된 규칙으로 정련한다.

실행부는 학습부에서 추출된 언어정보를 이용해서 입력된 문장에 대해서 품사 태깅을 수행한다. KTAG99의 실행부는 고유명사 추정기, 형태소 분석기, 통계기반 품사 태거, 오류교정기로 구성되었다. 고유명사 추정기는 한국어 처리에 고질적인 문제 중 하나인 미등록어 처리한다. 고유명사 추정기는 모든 미등록어를 처리하기 위한 시스템은 아니다. 미등록어 중에서 빈도가 가장 높은 고유명사를 추정하여 형태소 분석기의 과분석과 오분석을 완화한다. 형태소 분석기는 각 어절에 형태소를 분석하며, 형태소 분석기에서 일반적인 미등록어 추정 기능을 담당한다. 통계기반 품사 태거는 형태소 분석기에서 생성된 형태소 분석의 중의성을 해소한다. 마지막으로 품사 태깅 오류 교정기는 여러 형태의 문맥 정보를 효과적으로 사용할 수 없는 통계기반 품사 태거의 단점을 다소 보완하기 위해서 통계기반 품사 태거의 오류를 교정한다.

KTAG99에서 사용되는 언어정보는 그림 1에 대부분 표현되었다. 그러나 불규칙 용언 사전과 고유명사 추정을 위한 단서단어 사전(clue word dictionary) 그리고 품사 태깅 오류 수정규칙을 추정할 때, 필요한 동작성 명사 사전은 표현되지 않았다. 이들 언어정보는 학습 말뭉치로부터 추출되는 것이 아니며, MATEC99을 위해서 한국전자통신연구원으로부터 제공된 학습 말뭉치로는 추정할 수 없는 정보들이 대부분이다.

3. 고유명사 추정기

한국어 형태소 분석에서 가장 고질적인 문제 중 하나는 미등록어 추정이다. 이 미등록어의 대부분은 고유명사이다. 따라서 고유명사를 형태소 분석에서 쉽게 추정할 수 있도록 고유명사 추정기를 사용한다. 본 논문에서 제안된 고유명사 추정기는 매우 초보적인 수준이며, 4장에서 설명될 형태소 분석기에서 형태소 분리 기능의 일부를 담당하게 된다. 고유명사 추정규칙은 정규표현(regular expression)으로 표현되며, 고유명사 추정기는 유한 상태 변환기(finite-state transducer)로 구현되었다. 즉, 현재 고유명사 추정을 정규문법의 표현능력을 벗어나지 못하고 있다. 유한 상태 변환기의 실질적인 구현은 flex[6]를 이용한다. flex는 어휘분석을 목적으로 개발되었으며 공개된 프로그램이다. 아래에는 flex의 정규표현으로 기술된 고유명사 추정규칙의 일부를 보인 것이다.

```
{LEFT_SP}{NAME}/(씨|님|양|군)({RIGHT_SP}|{JOSA})
        printNQ_LEFT_SP(yytext);
}
```

를 찾는 문제로 품사 태깅 문제를 해결한다. 가중치 망은 한국어 형태소해석에서 발생되는 제반 문제를 자연스럽게 표현할 수 있으며, 고차 문맥정보 뿐 아니라 어휘 문맥정보도 쉽게 사용할 수 있다[8]. 가중치 망 모델은 은닉 마르코프 모델[9], 퍼지망 모델[10] 등의 모델에서 추출된 매개변수를 가중치로 이용할 수 있다. 따라서 한국어 품사태깅을 위해 고차 문맥정보를 이용하기 위해서 기존에 제안된 여러 가지의 모델을 수정없이 그대로 한국어에 적용할 수 있다. MATEC99에 출품된 KTAG99에서는 가중치로 확률정보를 사용하였다.

6. 품사 태깅 오류 교정기

통계기반 품사 태거는 다양한 문맥정보를 자연스럽게 적용할 수 없다. 그러나, Brill에 의해서 제안된 변형 규칙(transformational rule)[11]은 다양한 문맥정보를 효과적으로 적용할 수 있는 방법 중 하나이다. 본 논문에서는 Brill에 의해서 제안된 변형 규칙을 한국어에 적합하도록 수정하였다. 변형 규칙을 한국어에 적용한 예는 [2]와 [3]이 있었다. 본 논문에서는 [2]의 결과를 최대한 이용하였고, MATEC99에 참가하기 위해서 몇 종류의 규칙 틀을 더 추가하였다. Brill은 변형 규칙을 자동으로 추출하는 방법으로 오류를 토대로 변형 규칙을 학습한다[11].

7. KTAG99의 특성

가. 새로운 환경에 잘 적응한다.

KTAG99는 품사 태그가 변화하거나 새로운 실험 환경에 아주 쉽게 적응할 수 있다. 일단 새로운 학습 말뭉치만 준비되면, 쉽게 새로운 환경에 적응된다. 즉, 품사 태그 집합를 변경하고, 학습하면 바로 새로운 환경에서 수행되는 시스템을 얻을 수 있다.

나. 범용이다.

KTAG99는 특수한 목적을 가지고 개발된 것은 아니고, 여러 응용 분야에 두루 사용되고 있다. 실질적으로 음성인식, 음성합성, 기계번역 등에 사용하고 있다.

다. 학습 말뭉치에서 추출된 사전만 사용한다.

KTAG99는 학습 말뭉치에서 추출된 표제어 22,353개를 기본 사전으로 사용하고 있으며, 이 외에 불규칙 용언 사전으로 표제어 728개가 추가되었다.

라. 형태소 배열규칙이 불충분하다.

KTAG99는 품사 쌍을 형태소 배열 규칙으로 사용한다. 현재 품사 태그의 수는 27개이다. MATEC99에서 제공된 학습 말뭉치는 언어학적으로 모순이 되는 형태소 배열규칙이 포함되어 있다. 예를 들면, (명사 + 동사) 등이 그것이다. KTAG99에서는 이를 위해서 어휘 형태소 배열규칙[8]을 사용하고 있다. 불충분한 형태소 배열규칙 때문에 많은 형태소 해석결과를 출력하고, 형태소 분석 속도도 그다지 빠르지 못하다. 품사 태그 수가 부족하여 충분한 제약조건으로 사용할 수 없었다.

마. 미등록어 추정 기능의 보완이 필요하다.

미등록추정 기능을 보완하기 위해서 고유명사 추정 기능을 추가하였으나 많은 미등록어가 발생되고 있다. 더구나 미등록어 추정 알고리즘은 기능어의 부분 분석 결과를 이용하고 있기 때문에 내용어의 부분 분석 결과는 이용하고 있지 않다. 그러나 미등록어는 기능어 뿐 아니라 내용어의 부분 해석 결과를 이용할 수 있도록 개선되어야 한다.

바. 과분석 문제가 심각하다.

형태소 분석기의 배열규칙 정보가 불충분한 관계로 형태소 분석기에서는 매우 많은 형태소 분석 결과를 출력한다.

사. Trigram 확률 정보를 이용한다.

대부분의 한국어 품사 태깅 시스템은 bigram의 확률정보만 이용하고 있다. 그러나 trigram 정보는 아주 쉽게 정확률을 높일 수 있는 좋은 문맥정보이다.

아. 후처리 모델(Brill's model)의 학습 속도가 대단히 늦다.

Brill의 오류를 기반으로 한 변형 규칙 학습 방법은 학습 시간이 매우 오래 걸린다.

7. 결론 및 토의

본 논문은 MATEC99에 출품한 명사 추출기, 형태소 분석기, 품사 태거의 핵심 기술 및 시스템의 구조에 대해서 기술하였다. 분 연구팀에 의해서 출품된 시스템은 KTAG99라고 칭하며 이 시스템의 가장 큰 장점은 새로운 환경에 잘 적응할 수 있다는 것이다.

최근 몇 년 동안 한국어 품사 태깅에 관한 연구는 이미 여러 연구 기관에서 이루어 졌으며 [1][2][3], 한국어 품사 태깅에 대한 연구가 어느 정도는 성숙되었다고 생각된다. 그러나, 서로 다른 연구 환경(학습 말뭉치의 크기, 품사 태그, 사전 등)으로 각각의 시스템을 객관적인 비교할 수 있는 기회는 거의 없었으나 금번에 개최된 MATEC99는 각각의 시스템을 객관적이고 공정하게 비교할 수 있는 좋은 기회를 제공해 주었다고 생각된다.

MATEC99에서 제공된 학습 말뭉치는 공개된 여러 학습 말뭉치[12]들 보다는 좋은 질을 가졌다고 판단된다. 그러나, 품사 태그 집합은 형태소 분석이나 품사 태깅을 위해서 충분한 정보를 가지지 못한 것으로 판단된다. 이와 같은 이유로 MATEC99에 참석한 많은 팀들이 품사 태깅 시스템을 학습시키는 데에 여러 가지 어려움이 있었던 것으로 생각된다.

"제2회 형태소 분석기 및 품사 태거 대회"에서는 좀더 객관적이고 공정한 평가 방법을 개발하여 각 시스템을 평가했으면 한다. 즉, 모든 연구환경을 일치시킨 상황 하에서 매우 공정한 평가 방법을 이용해서 각 시스템을 평가되었으면 한다.

참고 문헌

[1] 김재훈, "가중치 망 모델을 이용한 한국어 품사 태깅," 한국정보과학회논문지, 제25권, 제6호, pp. 951-959, 1998년.

[2] 임희석, 김진동, 임해창, "어절 태그 변형 규칙을 이용한 한국어 품사 태거," 한국정보과학회논문지, 제24권, 제6호, pp. 673-684, 1997년.

[3] 신상현, 이근배, 이종혁, "통계와 규칙에 기반한 2단계 한국어 품사 태깅 시스템," 한국정보과학회논문지, 제24권, 제2호, pp. 160-169, 1997년.

[4] http://aladin.etri.re.kr/~nlu/STANDARD/matec99.html

[5] ETRI, 자연어 정보처리 기술 표준화 - 1차년도 초기 표준안, 1998년.

[6] http://www.gnu.com/

[7] 김재훈, 서정연, 김길창, 실용적인 한국어 형태소 해석, 한국과학기술원, 전산학과, CS-TR-95-98, 1995년.

[8] 김재훈, 오류-보정 기법을 이용한 어휘 모

```
{LEFT_SP}{NAME}/({SPA}*{JOB})
      printNQ_LEFT_SP(yytext);
}
{LEFT_SP}{SURNAME}/({SPA}*{JOB}) {
      if (strstr(yytext, "전") || strstr(yytext, "부"))
          REJECT;
      printNQ_LEFT_SP(yytext);
}
{LEFT_SP}{LOC}/("지역")?({RIGHT_SP}|{JOSA}) {
      printNQ_LEFT_SP(yytext);
}
{LEFT_SP}{HAN}+("위원회")/({RIGHT_SP}|{JOSA}) {
      printNQ_LEFT_SP(yytext);
}
"(주)"?{COMPANY}/({SPA}*{JOB}) {
      printNQ(0, yytext);
}
{GROUP}/({SPA}*{JOB}) {
      printNQ(0, yytext);
}
```

여기서 {LEFT_SP}와 {RIGHT_SP}는 단어의 경계를 나타낼 수 있는 여러 가지 기호들을 표현한다. {NAME}은 한국에서 사용되고 사람 이름을 표현하며, 현재에는 성을 포함한 세 자 혹은 네 자로 구성될 수 있는 경우에만 표현한다. {JOSA}는 가능한 모든 조사들을 표현하고 한다. {SURNAME}는 성을 표현하고 있다. 이를 인식하는 규칙에서 현재 오류를 조금 범하고 있다. 현재에는 '전 대통령'과 같은 문구에서 오류를 범하기 때문에 이와 같은 패턴을 인식할 수 없도록 REJECT 기능을 추가하였다. 그 밖에 앞에서 언급한 고유명사 추정 규칙에 기술된 정의는 독자들이 쉽게 추측할 수 있을 것으로 생각된다.

4. 형태소 분석기

본 논문에서의 형태소 분석기의 기본 알고리즘은 CYK 알고리즘에 기반하고 있다[7]. 형태소 분석기는 아래와 같은 절차들로 구성되었으며, 입력의 기본 단위는 어절[3]이다.

토큰 분리 : 영문자, 숫자 기타 기호 등 한글과 그 이외의 문자들로 분리한다.

불규칙 처리 : 불규칙 현상과 각종 음운 현상을 처리한다. 불규칙 처리를 위한 규칙은 절차적으로 기술된 것이 아니고, 하나의 지식 베이스로 구성되어 있다. 그래서 새로운 지식이 발견되면 지식 베이스에 새로운 내용을 추가함으로써 쉽게 확장할 수 있다.

형태소 분리 : 형태소 분리는 사전과 불규칙 처리 규칙에 바탕을 두고 있다. 분리된 형태소는 이차원 배열에 표현된다.

형태소 배열규칙 검사 : 형태소 배열 규칙을 이용해서 형태소 배열 구조에 적합한 형태소 분석을 출력한다. 형태소 배열 규칙은 품사의 쌍으로 표현되었다. MATEC99의 학습 말뭉치에 부착된 품사 태그[5]의 수는 27개로 사실상 형태소 배열 규칙을 사용하기에는 매우 부적합하다. 또한 이 학습 말뭉치에는 언어학적으로 문제가 될 수 있는 형태소 배열 구조를 가지고 있다. 예를 들면 (명사+동사) 등과 같은 것이다. 엄밀히 말해서 이와 같은 구조는 띄어쓰기 오류이거나 복합동사에 해당하는 것들이다.

미등록어 추정 : 형태소 분석의 부분 분석 결과를 이용해서 미등록어를 추정한다. 어절의 오른쪽의 부분 분석 결과에는 대부분이 기능어에 대한 해석이다. 형태소 분석기에는 이와 같은 기능어의 부분 분석 결과를 이용해서 정확하게 분석되지 않은 내용어의 품사를 추정한다.

5. 통계기반 품사 태거

본 논문에서 사용되는 품사 태거는 가중치 망 모델[1]을 이용한다. 가중치 망 모델은 형태소해석 결과를 격자구조로 표현하고, 이 격자구조에 적절한 가중치를 적재하여, 가중치 망을 형성하고, 이 가중치 망으로부터 가장 적절한 경로

[3] 본 논문에서의 어절은 공백을 분리자로 간주하였다.

호성 해소, 한국과학기술원, 전산학과, 박사 학위 논문, 1996년.

[9] 김재훈,. 임철수, 서정연, "은닉 마르코프 모델을 이용한 효율적인 한국어 품사 태깅," 정보과학회논문지, 제22권 1호, pp. 2118-2125, 1995년.

[10] 김재훈, 조정미, 김창현, 서정연, 김길창, 퍼지망을 이용한 한국어 품사 태깅," 제5회 한글 및 한국어 정보처리 학술대회 발표 논문집, pp. 539-603, 1993년.

[11] Brill, E. "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging," Computational Linguistics, vol. 21, no. 4, pp. 543-565, 1995.

[12] KAIST, 대한민국 국어정보베이스 II, 1998년