형태소분석기 및 품사 태거 평가대회(MATEC99) 개요

이재성, 박재득, 차건회, 박세영
ETRI 컴퓨터소프트웨어연구소 지식정보연구부
{jasonl, jdpark, chakh}@etri.re.kr, sypark@computer.etri.re.kr

Morphological Analyzer and Tagger Evaluation Contest (MATEC 99) Overview

Jae Sung Lee, Jay Duke Park, Keon Hoe Cha, Se Young Park
Knowledge Information Department,
Computer and Software Technology Laboratories, ETRI

요약

한국어 정보처리에서 기본 모듈로 많이 사용되는 형태소분석기, 태거 및 명사추출기에 대한 객관적인 평가를 위해서는 실제 사용되는 언어에 대한 평가기준과 방대한 양의 평가자료 구축이 필수적이다. 전자통신연구원(ETRI)에서는 표준적인 평가기준과 평가자료(말뭉치)를 구축하여 "제 1 회 형태소분석기 및 품사 태거 평가대회"을 개최하였으며, 이 대회는 학습기간을 포함하여 1999 년 6 월 7 일부터 10 월 1 일까지 진행되었다. 평가에는 총 15 개팀이 참가하였고, 명사추출, 태거, 형태소분석기의 각 분야에 대해 약 25 만 4 천어절의 학습 말뭉치를 제공한 후, 시험말뭉치 약 3 만 3 천어절에 대해 평가가 이루어졌다. 이 글에서는 이 대회의 취지, 진행과정, 평가방식, 평가결과 등에 대해 소개한다.

1. 서론

지식정보화 시대를 맞이하여 정확하고 효율적인 한국어 정보처리의 필요성이 높아지고 있으며, 특히 인터넷의 보급과 더불어 방대한 양의 문서 처리가 필요하게 되어 그 중요성이 더욱 부각되고 있다. 그러나, 한국어 정보처리의 기본 기술인 한글언어처리에 대한 객관적인 평가가 부족하여, 현재의 기술이 어느 정도 수준에 와 있고, 어떤 방향으로 연구가 되어야 하는지 자세하게 파악하기 힘든 실정이다. 방대한 양의 언어처리에 대한 객관적인 평가를 하기 위해서는 대량의 말뭉치를 구축하여, 각각의 언어현상이 나타나고 실제로 사용되고 있는 문장에 대한 평가기준이 필요하다. 이런 객관적인 평가와 평가기준은 기술의 진보방향을 명확히 제시해 주며, 그에 따라 기술의 발전을 가속화시켜 주고, 개발 역량을 집중할 수 있도록 해 줄 것이다.

외국의 경우에도 이미 정보검색, 정보추출, 요약, 의미구분 등과 같은 언어관련 기술의 평가 컨소시엄인 TREC(TREC web), MUC(MUC web), SUMMAC(SUMMAC web), SENSEVAL(SENSEVAL web) 같은 것이 구성되어 대량의 언어정보처리에 대한 객관적인 평가 기준을 제공해 주고 있으며, 이를 통해 개발자들의 정보교환과 기술발전을 유도하고 있다. 현재 한국어에 대한 자연언어 처리의 여러 기술 중에서 한국어 형태소 분석기 및 품사 태거 기술은 오랫동안 개발되어왔고, 어느 정도 객관적 평가를 할 수 있을 정도로 성숙한 단계에 이르렀다고 생각된다.

평가를 위해서는 표준화된 태그의 사용이나 출력 양식이 필요한데, 이를 통해 참가팀들은 자연스럽게 표준화된 입출력형식을 구현하게 된다. 따라서, 표준화된 자연언어 처리 프로그램들은 요소화가 가능하게 되고, 다른 큰 시스템의 한 구성 요소로 손쉽게 대체되어 추가될 수 있으므로 개발의 효과를 증폭시킬 수도 있다.

이러한 배경과 취지에서, 우선 한국어 정보처리의 공통 기본 모듈로 정보검색 및 기계번역 분야 등에서 광범위하게 사용되고 있는 명사추출, 태거, 형태소 분석 기술을 평가하기 위해, "형태소 분석기 및 태거의 성능평가 대회 (Morphological Analyzer and Tagger Evaluation Contest: MATEC)"를 개최하였다. 그 동안 전자통신연구원에서는 "자연언어 처리기술 표준화" 과제를 통해 약 29 만어절의 품사부착 말뭉치를 구축하였고, 여기에서 사용된 기준과 구축된 말뭉치를 이용하여 대회의 평가기준으로 사용하였다. 물론 말뭉치 구축 기준이나 평가기준은 학계나 업계의 의견을 수렴하여 결정된 내용이므로, 비교적 객관적인 평가 기준으로 볼 수 있다(전자통신연구원 1999a, 1999b, 1999c).

다음절에서는 각 분야별 연구동향을 간단히 살펴보고, 이어서 대회진행, 사용된 품사부착 말뭉치에 대한 설명, 평가 결과 및 맺음말 순으로 설명한다.

2. 각 분야별 연구동향

2.1 형태소분석기

한국어 형태소 분석에 관한 연구로는 최장일치법, head-tail 분리 기억방법(최형석 이주근, 1984), tabular parsing 과 접속정보를 이용하는 방법(김성용, 1987; 이은철 1992), 음절정보에 의한 방법(강승식, 1991, 1992, 1993a, 1993b), 2 단계 모델에 기반한 방법(이성진, 1992), 다층 형태론에 기반한 방법(강승식 1994) 등이 있고, 규칙이나 활용형을 사전에 넣어 처리하는 방식(조영환 1993; 김재한 1994) 등이 있었으며, 기계학습을 통한 방식(장병탁 1990)이 시도되기도 했다. 또한 형태소해석에서 나타나는 중의성 해결을 위한 방법(임희석 1993a, 1993b; 김충원 1994), 접속정보를 확장 보완한 방법(이은철 1992; 김병희 1993)이나 의미적으로 한 단어가 되는 어휘들에 대한 처리를 하는 방법(김민정 1991, 허윤영 1994) 등이 연구되었고, 또한 미등록어에 대한 처리를 보강하여 보다 강건한 형태소해석기를 구축하기 위한 연구(강승식 1993a; 임희석 1993a; 김재훈 1995) 등도 이루어 지고 있다. 범용 또는 철자검색용으로 만들어진 형태소 해석기 외에 특별한 용도를 위해 만들어진 것들도 있는데, 문자인식 오류 검출을 위한 연구(김윤호 1992), 구문분석에 용이한 자질구조를 생성하기 위한 연구(송연정 1994), 연속음성 인식을 위한 연구(김병창 1997) 등이 있다.

2.2 명사추출

명사추출은 대개 정보검색을 위한 색인어 추출을 위해 사용되며, 형태소 해석 기법을 이용하여 명사(또는 색인어 후보)를 추출해야만 보다 정확하게 추출할 수 있다(강승식 1995). 그러나, 보다 빠른 속도를 고려하거나 특수 목적에 따라 효과적일 수 있는 색인어 추출을 위해 어미 및 조사 사전만을 이용하는 방법(이영주 1989)이나, 형태소해석 과정 중 불필요한 어미분석 단계 등을 생략하여 빠른 속도로 처리하는 방법(최재혁 1993a, 1993b) 등도 연구되었다.

2.3 태거

한국어 태거에 관한 연구로는 HMM(Hidden Markov Model)을 이용하여 어절단위로 태깅을 한 연구 (이운재 1993), 형태소 단위의 태깅을 위해 HMM 을 적용한 연구 (임철수 1994)등의 초창기 연구와 한국어의 어절과 띄어쓰기 특성 등을 고려하여 설계한 연구(신중호 1994; 김진동 1997, 1998), 퍼지망을 이용한 연구(김재훈 1993), 변형 규칙을 이용한 연구(임희석 1996, 1997), 가중치 망을 이용한 연구(김재훈 1998), 최대 엔트로피 모델을 이용한 연구(강인호 1998), 통계와 규칙을 혼합하여 사용한 연구(신상현 1997; 임희석 1998) 등이 있으며, 어휘 규칙의 추출과 이를 중심으로 태깅을 하기 위한 연구(이정규 1997; 이상주 1998, 1999) 등이 있다.

3. 대회진행

3.1 평가종목

평가종목은 명사추출, 태거, 형태소분석기의 3 분야로 나누었다. 명사추출 분야는 문장에서 명사를 얼마

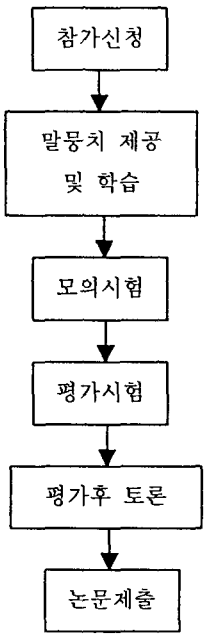

그림 1. 대회진행순서

나 정확하게 추출해 내는가를 평가한다. 이 종목은 주로 정보검색분야에서 하부 모듈로 사용될 경우에 그 성능이 어떤지를 알 수 있도록 하기 위한 것이다. 태거는 문장에서부터 각각의 형태소를 분리해 내고, 그 다음에 태그를 얼마나 올바르게 부착하는가를 평가하는 종목이다. 형태소분석기 종목은 어절에 대해 출력한 분석가능한 결과들을 태거입장과 참가팀이 공통적으로 인정하는 정답의 입장에서 평가하기 위한 것이다.

3.2 진행 방식

전체적인 진행순서는 그림 1과 같다. 주최측인 한국전자통신연구원에서는 참가신청을 받고, 참가신청자들에게 학습에 필요한 말뭉치 약 26만어절과 품사부착 말뭉치의 작성지침을 제공했다. 품사부착 말뭉치는 주최측에서 지난해 연구과제로 수행한 결과 구축된 것으로, 정보처리 기술의 표준화를 목적으로 만들어 졌으며, 여러 분야에서의 의견들이 수렴된 결과이다(전자통신연구원 1999a, 1999b, 1999c). 이번 대회를 통해 실제 표준화된 자료를 자연언어 처리 시스템에서 사용해 보고 문제점을 수정하도록 했다. 품사부착 말뭉치는 각 시스템의 정답을 제시해 주는 자료로 사용된다. 즉, 태거의 경우, 학습데이터로 직접 사용되며, 출력해야 할 답을 제시해주고, 명사추출기에게는 명사추출 유형을 제시해 주며, 형태소 분석기에게는 형태소 분석된 결과 중 가장 올바른 결과의 형태를 보여준다. 학습기간 중 수정되는 말뭉치는 계속 새로운 버전으로 참가팀에게 제공되었다.

학습의 마지막 단계에서 실제 시험을 대비하여 최종점검을 하기 위한 모의시험 말뭉치를 제공했다. 모의시험 말뭉치는 시험에 사용될 형태와 똑같은 입력파일과 정답파일로 구성된다. 참가자는 이 파일을 이용하여 각 시스템을 시험해 보고, 미리 문제점을 수정할 수 있도록 했다. 모의시험 말뭉치는 매우 작은 규모로서 약 1,000어절 정도이지만, 시험용 말뭉치의 문장과 비슷한 유형들을 포함시켜 시스템 조정을 효과적으로 할 수 있도록 했다.

시험은 공개되지 않았던 약 3만 3천어절의 시험용 말뭉치를 온라인으로 제공하고, 각 참가자들은 이를 각자 받아 스스로 시스템에서 수행해 본 후, 그 결과를 수정없이 주최측에 제공함으로서 이루어졌다. 보내준 결과는 평가기준에 따라 주최측에서 평가하여 그 결과를 참가자들에게 통보하였다. 이 평가 결과를 토대로 각 참가팀들이 함께 모여 문제점 등을 토론하고 정리하여 논문으로 제출하였다.

3.3 평가방식

평가시험은 시험기간에 정해진 웹사이트에서 시험용 말뭉치를 각자 다운로드받아서 수행하고, 그 실행결과물을 대회운영팀에게 기한내에 제출해야 하며, 모든 평가는 평가시험 결과로 제출한 제출물을 근거로 이루어졌다. 각 종목별 평가 방식을 간단히 소개하면 다음과 같다.

1) 명사추출

명사의 추출 여부는 일정한 범위내에 나타나는 명사의 재현률과 정확률로 평가했다. 즉, 시험용 코퍼스를 일정한 크기로 나누고, 그 크기내에서 찾아내야 할 명사를 나열하고 그 명사들을 어떤 비율로 찾아내는가를 측정했다. 이번 평가에서는 빈도수를 무시했다.

$$재현률 = \frac{올바르게\_찾아낸\_명사수}{전체\_명사수}$$

$$정확률 = \frac{올바르게\_찾아낸\_명사수}{찾아낸\_전체\_명사수}$$

2) 태거

태거는 두가지 방법으로 평가를 했다. 첫째는 "어절별 정확률" 로 어절별로 정답과 비교하여 일치하면 맞는 것으로 하고 그렇지 않으면 틀린 것으로 했다. 평가기준은 간단하게 다음과 같이 표시된다.

$$정확률 = \frac{올바르게\_태깅된\_어절수}{전체\_어절수}$$

둘째는 "품사별 정확률" 로 전체 코퍼스에서 나타나야 할 품사에 대해 올바르게 각 품사 개수가 나왔는가를 측정하는 것이다. 평가기준은 다음과 같이 두가지로 측정된다.

$$(품사별)재현률 = \frac{(품사별)올바르게\_찾아낸\_개수}{(품사별)전체\_개수}$$

$$(품사별)정확률 = \frac{(품사별)올바르게\_태깅된\_개수}{(품사별)태깅된\_전체\_개수}$$

3) 형태소 분석기

형태소 분석기는 대개 어절단위로 분석하여 가능한 모든 분석을 출력한다. 그러나 가능한 모든 분석을 모든 어절에 대해 정답으로 만들어 내기도 어려울 뿐만 아니라, 실제 문맥에서는 거의 불필요한 분석까지 포함하고 있어서 평가하기에 어려움이 있다. 이런 문제를 피해서 간단하게 평가하기 위해 다음과 같이 2가지로 평가했다.

첫째는 "K-best 식"으로 형태소 분석기의 분석결과를 임의의 개수 K만큼 출력하여 그 결과중에 정답이 포함되어 있으면 맞는 것으로 한다. 이는 태깅정답제시율(일종의 재현률)로 다음과 같은 수식으로 표시된다.

$$태깅정답제시률 = \frac{정답과\_일치되는\_분석이\_포함된\_어절수}{전체\_어절수}$$

둘째는 "과반수 지지식" 으로 문맥에 관계없이 하나의 어절만을 보고, 그 어절에 대한 분석결과를 판단하는 평가이다. 즉, 한 어절에 대해 참가자의 과반수 이상이 정답으로 인정하는 분석결과가 포함된 비율을 계산하는 것이다. 이 평가의 경우는 앞에서와 다르게 말뭉치에 나타나는 모든 어절(instance)을 평가하지 않고, 모든 어절을 미리 분석하여 같은 종류(type)의 어절을 하나의 어절로 만들어 이 어절들에 대해서만 평가를 했다. 평가기준을 수식으로 표현하면 다음과 같다.

$$재현률 = \frac{과반수\_지지의\_정답을\_포함한\_분석결과수}{과반수\_지지를\_받은\_모든\_분석수}$$

$$정확률 = \frac{과반수\_지지의\_정답을\_포함한\_분석결과수}{생성된\_전체\_분석수}$$

3개의 전분야에 걸쳐 명사에 대한 평가는 단순한 일치로 판단하지 않고, 복합어에 대해 고려하도록 했다. 예를 들어 "사과나무" 의 경우, 어떤 경우는 하나의 복합어로 보기도 하고, 또 어떤 경우는 단순명사 2개를 붙여 쓴 것으로 취급하기도 한다. 따라서 관점에 따라 정답이 달라질 수 있다. 물론, 경우에 따라, 복합어의 판단 기준이 명확할 수도 있지만, 그렇지 못한 경우도 많으므로 이를 고려하여 좀더 완화된 기준을 가지고 평가하였다. 즉, 명사가 연속하여 나올 경우, 이를 복합어로 처리하거나 단순명사의 연속으로 처리하거나 모두 맞는 것으로 평가하도록 했다.

평가의 결과가 재현률과 정확률 두가지로 나올 경우, 두개의 시스템을 직접 비교하는데 어려움이 있다. 이를 해결하기 위해 F측정(F-measure: Manning 1999)을 사용했다. 즉 재현률 R과 정확률 P에 대한 가중치 α를 0.5로 두어, 두가지 모두를 같은 비율로 중요하게 취급하여 평가했다. 이를 수식으로 표현하면 다음과 같다.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{2PR}{(R+P)}$$

표 1. 각 말뭉치의 분야별 어절수

	학습	모의시험	시험
소설류	158,049(62%)	286(29%)	9,012(27%)
비소설류	96,365(38%)	316(32%)	12,194(36%)
뉴스	0(0%)	379(39%)	12,649(37%)
합계	254,414(100%)	981(100%)	33,855(100%)

이외에도 평가를 위해 시스템별 특성(처리 속도, 예외 처리 등)은 각 참가팀들이 보고하도록 했다. 하지만, 시험결과에 대한 비공개 원칙 때문에 각 시스템의 특성별 비교는 직접적으로 하지 않았다.

4. 품사부착 말뭉치(태그 코퍼스)

4.1 말뭉치 구성

품사 부착 말뭉치는 KAIST 말뭉치를 새로 표준화된 태그세트와 태깅지침에 따라 수정한 것으로 그 구성은 표 1와 같다. 학습 말뭉치는 약 25만 4천어절로 소설류가 62%, 비소설류가 38%를 차지했다. 말뭉치 구축시 다양한 원문정보의 확보가 어려워 분야가 소설류로 많이 치우쳐 있다. 앞으로 보다 균형잡힌 말뭉치를 구축하여 이런 문제를 개선할 여지가 있다. 그렇지만, 이 학습 말뭉치는 평가대회에 참가한 모든 팀에게 동등하게 제공되어 평가됨으로 최소한 상대적인 평가에는 별 무리가 없을 것으로 생각된다.

실제 시험은 약 3만3천어절로 학습말뭉치 크기의 약 13%에 해당되는 비교적 많은 양으로 시행했다. 또한 시험말뭉치는 학습 말뭉치와는 다르게 뉴스(방송뉴스 및 신문기사) 분야를 포함시켰다. 이는 새로운 분야에 대한 적응력을 시험해 보기 위한 것이다. 또, 실제 시험을 대비하여 모의시험 말뭉치를 제공했는데 그 구성은 실제 시험 말뭉치에서 일부를 추출한 것으로 실제 시험 말뭉치의 구성과 유사하게 했다.

4.2 태깅(품사부착) 기준

태깅 기준은 표준화 과제를 통해 이루어진 결과를 사용했다. 즉, 태깅의 성격상 기준의 적용이 여러가지 관점에 따라 다르게 이루어 질 수 있으므로, 기존의 한국어 사전을 참조하고, 불일치하는 경우에 대해, 기준 사전을 정하여 결정하거나 토론을 통해 결정했다. 또한 언어학적인 관점과 전산학적인 관점이 차이가 있을 경우, 전산학적인 처리 효율성을 원칙적으로 따르기로 했다. 태깅 기준에 관한 자세한 기준과 문제점 등은 표준화자료(전자통신연구원 1999a, 1999b, 1999c)를 참조하기 바란다.

4.3 학습말뭉치 수정 및 오류분석

대회에서 사용된 학습말뭉치는 자연언어 처리기술 표준화 과제의 1차년도 결과로 나온 것이며, 많은 부분이 표준화회의를 통해 결정된 것이다. 그러나, 다양한 시스템과 다양한 용도에 대해 구체적으로 문제없이 사용될 수 있는지에 대한 검증은 없었다. 따라서, 대회를 통해 참가팀이 직접 사용해 보고 문제점이나 의견을 보내오면, 이를 반영하여 학습말뭉치를 수정하되, 대회의 원활한 진행을 위해 가능하면 수정의 정도를 최소화했다.

수정된 것이 반드시 오류이고, 수정 안한 것이 올바르게 태깅된 것으로 단정할 수는 없지만, 수정된 것이 올바른 태깅의 결과라고 가정하면, 대략적으로 그 태깅 오류의 원인을 파악할 수 있다. 오류의 종류를 크게 7가지로 분류하였다: "붙띄" 는 원문에서 붙여 써야 할 것을 띄어 쓴 오류, "띄붙" 은 원문에서 띄어 써야 할 것을 붙여 쓴 오류, "원문" 은 원문정보를 잘못 입력한 오류, "문법" 은 말뭉치 구축 형식과 다르게 만들어진 오류, "분리" 는 형태소 단위의 분리를 잘못한 경우, "복원" 은 형태소 복원을 잘못한 경우, "태그" 는 형태소에 대해 잘못 태그를 붙인 어절수를 각각 나타낸다. 말뭉치 관리의 어려움으로 수정된 내용을 모두 기록하지 못했으며, 수정된 내용의 분석은 간단한 프로그램을 통해서 처리했다. 즉, 구버전의 원문과 신버전의 원문을 비교해 보아 다르면, 일단 "원문" 오류로 판정하고, 또 다시 원문 오류중에서 이전이나 다음 어절과 붙이거나 띄어쓴 경우로 구분하여 "붙띄", "띄붙" 오류가 수정된 것을 계산했다. 따라서 "원문" 오류의 갯수에는 "붙띄", "띄붙" 오류의 갯수를 포함하고 있다. 또, 간단한 문법확인 프로그램으로 형식이 틀린 것을 확인하여 "문법" 오류 수를 계산하였다. "분리" 오류는 구버전과 신버전에서 분석된 각각의 형태소들을 연결한 어휘들은 같으나 분리된 형태소 갯수만

표 2. 버전별 수정된 내용(수정 어절수)

버전	붙띄	띄붙	원문	문법	분리	복원	태그	합계
0.2	83	52	141	13	1086	1291	1900	4505
0.3	3	1	15	0	360	170	254	799
0.4	0	0	1	0	134	92	29	256

다를 경우, "복원" 은 분석된 형태소들의 내용이 다를 경우, "태그" 는 분석된 태그들중 하나라도 틀린 경우를 계산했다.

학습말뭉치는 4 차례에 걸쳐 수정된 버전을 참가팀에게 배포했는데, 각 버전이 배포될 때마다 수정된 내용을 대략 분류하여 나타낸 것이 표 2 이다. 0.1 버전에서 0.2 버전으로 수정되어 제공될 때는 비교적 많은 수의 "붙띄" , "띄붙" 및 "원문" 오류가 수정되었는데, 이것은 태깅 원칙이 처음에는 원문에 있는 오류를 그대로 말뭉치에 반영하기로 했다가, 나중에는 띄어쓰기 오류를 수정하여 입력하기로 원칙이 바뀌었기 때문이다. 또 사소한 "문법" 오류도 일부 있었다. 그러나 대부분의 오류는 "분리" , "복원" , "태그" 로 실제 태깅에 관련된 것이었다. 구체적인 태깅의 오류는 분석하지 못했지만, 대략 그 원인을 살펴보면, 가) 잘못된 태깅을 한 경우, 나) 같은 유형에 대해 일관성이 없이 태깅한 경우, 다) 태깅 기준의 변화로 바뀐 경우(예: "그러다보니" 에서 "다" 를 "다가" 로 복원하여 처리하던 것을 복원하지 않고 바로 연결어미 "다" 를 인정) 등으로 나눌 수 있다. 초기 버전에서 가) 및 나)형의 오류가 많았으나, 나중에는 상대적으로 분리와 복원에 관련된 다)형의 수정이 많았다. 나)형의 오류는 보는 관점에 따라서 다르게 판단할 수 있으므로, 최종 말뭉치에서도 이런 유형의 오류가 포함되어 있을 수 있으므로 이에 대한 분석을 하여, 보다 객관적으로 인정받는 말뭉치를 만들 필요가 있다.

5. 참가팀 및 평가결과

대회에 참가한 팀은 총 18 개 팀이었으나, 최종 평가에 참가한 팀은 15 개 팀이다. 표 3 은 참가팀의 참가종목을 나타낸 것으로, 명사추출에 14 개팀, 태거에 8 개팀, 형태소분석기에 7 개팀이 참가했다.

14 개팀이 참가한 명사추출의 경우, 전체 평균은

표 3. 참가팀 및 참가종목 (접수순)

팀이름	참가종목			소속
HUMAN	N	x	x	충북대
SKOPE	N	T	M	포항공대
KLE	N	T	M	포항공대
미리내	N	x	x	부산대
MORANY	N	T	x	연세대
SeeKore	N	T	M	서강대/한국해양대
TU-HI	N	T	x	동경대/(주)히다찌
청실	x	T	x	KAIST
홍실	N	x	M	KAIST
LGKMA	N	T	M	LG 종합기술원
KUNLP	N	T	M	고려대
CNUCS	N	x	x	충남대
Cypher	N	x	M	전북대
Cadenza	N	x	x	한국마이크로소프트
CBUCE	N	x	x	전북대

N-명사추출, T-품사태깅,
M-형태소분석기비교, x -참가안함

재현률 0.85, 정확률 0.80, F=0.83 이다. 최고는 F=0.92 이며, F 값에 따라 순위별로 표시한 그래프가 그림 2 이다. (평가결과는 개별성적에 대해 비공개를 원칙으로 했으므로 각 팀을 익명으로 나타낸다.)

태거의 평가에는 원래 8 개팀이 참여하였으나, 1 개팀은 평가형식에 맞지 않아 7 개팀만이 평가의 대상이 되었다. 태거의 어절별 평가는 평균 정확률 0.39 를 보였으며, 각 팀의 정확률을 순위별 그래프로 나타낸 것이 그림 3 이다. 또, 품사별 평가 결과를 F 값의 순위그래프로 나타낸 것이 그림 4 이다. 품사별 평가의 순위는 어절별 평가와 같았다. 전체 F 값 평균을 품사 태그별로 나타낸 것이 표 4 이다. 품사별로 평가 결과를 보면 어휘적으로 혼동의 여지가 없는 기호나, 속격조사(예: '의'), 선어말어미(예: '었', '겠', '시' 등) 등을 정확하게 판단해 내어, 기호(F=0.99), 속격조사(F=0.98), 선어말어미(F=0.98) 순으로 높은 정확도를 보였다. 이와는 반대로 문맥에 따라서 판단해야 하는 어휘가 많이 포함된 감탄사(F=0.73)나 관형사(F=0.78) 등이 정확도가 낮게 나왔다. 예를 들어 감탄사 '아니' 는 부정을 나타내는 동사 '아니'로 혼동되기도 하고, '그래'는 동사 '그러다+어'로 분석되기도 한다. 또, 관형사 '이' 나 '그' 등은 대명사와 많이 혼동되는 예이다.

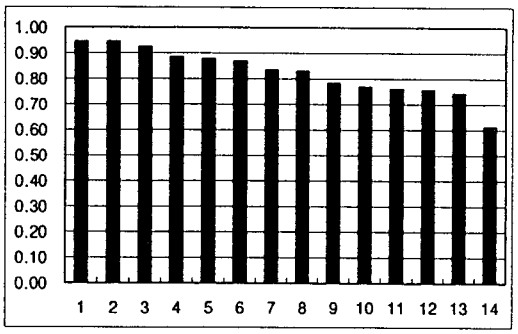

그림 2. 명사추출의 평가 결과

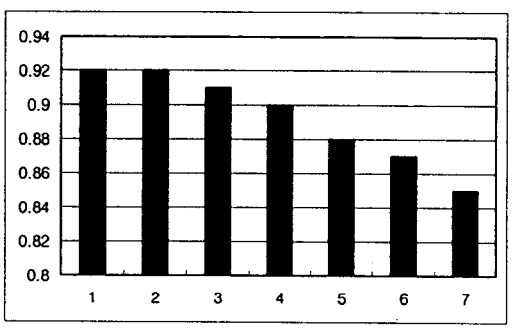

그림 3. 태거의 어절별 평가 결과

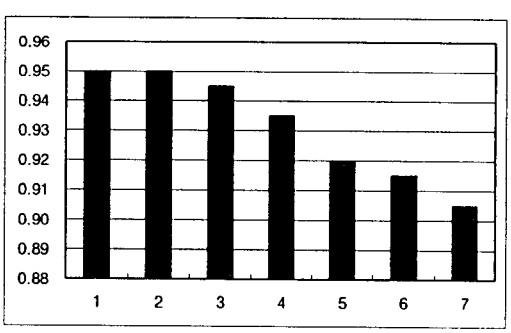

그림 4. 태거의 형태소별 평가 결과

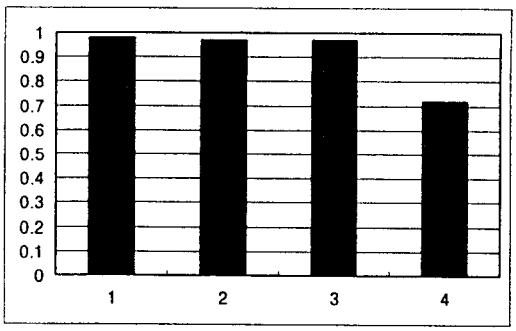

그림 5. 형태소 분석기 K-best 평가 결과

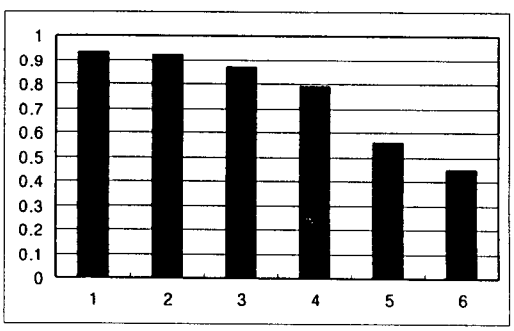

그림 6. 형태소분석기 과반수지지식 재현률 결과

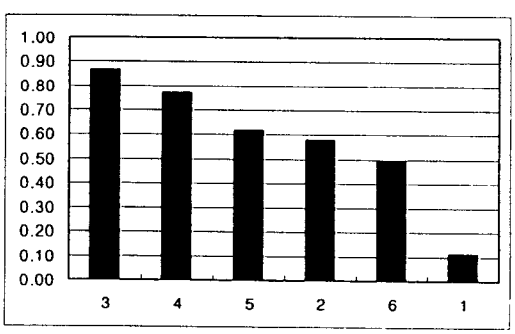

그림 7. 형태소분석기 과반수지지식 F 값 결과

형태소분석기의 평가는 모두 7개팀이 참가했으나, 출력 형식이 잘못된 팀을 제외하여, K-best 방식은 4개팀, 과반수 지지식은 6개팀만을 평가에 포함시켰다. K-best 의 평균 재현률은 0.91로 비교적 높았으며, 평균 어절당 후보수는 8.18개였다. 각 팀의 재현률 그래프는 그림 5이다.

형태소분석기 평가의 과반수 지지식은 재현률 및 정확률 2가지로 평가하였다. 이를 재현률 순위로만 나타낸 것이 그림 6이고, 이때의 평균 재현률은 0.75이다. 또 앞에서 정한 F값으로 계산하여 순위를 나타낸 것이 그림 7이며, 이 경우 각팀의 평균값들은 재현률 0.75, 정확률 0.56, F값 0.57이다. X축의 숫자는 재현률 순위의 그림 7에서 나타낸 팀번호로 재현률순위가 1위인 팀이 최하위인 6위로 나타났다. 그 이유는 그 팀이 가장 많은 분석후보를 생성했기 때문이다. 즉, F값 계산시 정확률은 후보생성을 많이 할수록 불리하

표 4. 각 품사별 F값

태그	태그이름	재현률	정확률	F값
s	기호	0.99	1.00	0.99
jm	속격조사	0.98	0.99	0.98
ep	선어말어미	0.97	0.99	0.98
jc	격조사	0.96	0.97	0.96
f	외국어	0.94	0.97	0.96
jx	보조사	0.94	0.95	0.95
etm	관형형어미	0.95	0.94	0.95
xsv	동사파생접미사	0.93	0.95	0.94
nc	자립명사	0.93	0.93	0.93
ef	종결어미	0.91	0.94	0.93
maj	접속부사	0.88	0.98	0.93
xsn	명사파생접미사	0.93	0.92	0.92
ec	연결어미	0.93	0.92	0.92
nn	수사	0.91	0.91	0.91
mag	일반부사	0.93	0.90	0.91
etn	명사형어미	0.95	0.86	0.90
pa	형용사	0.93	0.88	0.90
co	지정사	0.91	0.89	0.90
pv	동사	0.91	0.88	0.90
np	대명사	0.87	0.91	0.89
xsm	형용사파생접미사	0.84	0.92	0.88
nb	의존명사	0.85	0.90	0.88
px	보조용언	0.86	0.89	0.88
xp	접두사	0.86	0.86	0.86
jj	접속조사	0.80	0.78	0.79
mm	관형사	0.79	0.77	0.78
ii	감탄사	0.73	0.73	0.73

게 되어 결국은 재현률이 1위이더라도 F값 순위에서는 최하위를 나타냈다. 대개 형태소분석기의 역할은 하나 또는 두개의 어절 내에서 가능한 모든 분석결과를 출력하는 것이고, 그 결과 중에 적합한 결과는 태거나 상위 수준의 다른 프로그램이 선택한다. 이런 관점에서 볼 때, 형태소 분석기의 역할은 올바른 분석결과를 포함하는 것이 중요하고, 평가도 정확률보다는 재현률을 중심으로 하는 것이 올바른 평가로 생각된다.

말뭉치의 분야별 성능차이를 비교해 보기 위해 정리한 것이 표 5이다. 각 종목별 성능은 형태소분석기의 과반수 지지식 중에서 재현률로 측정한 것만 제외하고는 비소설, 뉴스, 소설 순으로 우수한 결과를 내었다. 소설분야가 학습 말뭉치에 포함되지 않은 뉴스분야보다 더 못 분석해 낸 것은 소설분야 자체가 구어체가 많고, 생략 등이 많아, 분석이 어려웠기 때문으로 생각된다.

표 5. 각 말뭉치 분야별 평가 결과

평가종목	소설	비소설	뉴스	측정기준
명사 추출	0.79	0.85	0.83	F
태거(어절별)	0.87	0.91	0.90	정확률
태거(형태소별)	0.92	0.95	0.93	F
형태소(K-Best)	0.89	0.92	0.92	정답제시율
형태소(과반수지지)	0.63	0.64	0.64	F
형태소(과반수지지)	0.72	0.76	0.78	재현률

6. 맺음말

한국어 정보처리에 대한 연구가 국내외적으로 많이 진행되고 있음에도 불구하고, 공개적인 성능평가나 표준화에 노력은 많지 않았다. 이번 대회는 15개의 많은 팀이 참여하였고, 여러 분야에서 사용되는 자연언어처리 프로그램을 공통의 평가기준으로 측정하는 자리가 되었으며, 각 팀들의 개발 상황과 상대적인 기술수준을 파악할 수 있었다.

또한 표준화된 품사부착 말뭉치를 제공하여 이를 학습용 말뭉치로 사용하였고, 그 결과를 표준적인 형식이나 태그세트를 사용하여 출력하였다. 따라서, 시스템이 하나의 표준형식에 맞추어 작동될 수 있음을 보여주었다.

그러나, 비교적 짧은 기간동안 대회를 운영했으며, 이로 인해 학습용 말뭉치의 태깅이나 정답 기준에 대한 대회 참여자들의 의견 수렴에 다소 부족한 점이 있었다. 즉, MATEC99의 품사부착 기준과 차이가 많은 시스템은 이미 구현된 것을 고치는 시간이 많아, 전처리나 후처리 등으로 결과를 출력한 경우가 있었다. 또 명사추출의 기준을 한 말뭉치가 주로 태깅의 목적으로 만들어져서, 정보검색용으로는 부적합한 정답을 포함한 것도 있었다. 이러한 문제점을 해결하기 위해서는 실제 시스템 구현에 필요한 보다 객관적이고 효과적인 말뭉치를 구축하고 정답 기준 설정을 하여 평가할 필요가 있다. 또한 각 팀의 평가결과를 익명으로 제공하기로 결정함에 따라 각 팀의 구현방식과 성능과의 관계를 분석하여 발표하지 못한 점이 아쉽다.

이러한 문제점을 해결하고, 구문분석이나, 의미해석, 정보검색 등 다른 정보처리관련 분야를 포함하여 다음 대회를 계속적으로 개최할 경우, 관련분야의 정보교환 및 기술개발 향상에 크게 도움이 될 것으로 생각한다.

감사의 글

1. 본 논문은 정보통신부 "자연언어처리기술 표준화"과제의 일환으로 수행한 결과입니다.

2. MATEC99를 성공적으로 마칠 수 있도록 도와준 표준화 위원님들과 대회 참가팀들께 감사드립니다. 특히, 평가를 해주신 고려대학교의 임해창교수님과 김진동씨, 그리고 품사부착 말뭉치 구축을 해주신 (주)SnL의 이현아씨와 충남대의 임선숙씨에게 감사의 말씀을 드립니다.

참고문헌

강승식, 김영택, 1991, "한국어 형태소 분석기에서의 선어말어미의 분석 모형," 한국정보과학회 논문지, 1991. 9, Vol. 18, No. 5.

강승식, 김영택, 1992, "한국어 형태소 분석기에서의 불규칙 용언의 분석 모형," 한국정보과학회 논문지, 1992. 3, Vol. 19, No. 2.

강승식, 1993a, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석," 서울대학교 컴퓨터공학과 박사학위 논문.

강승식, 1993b, "음절 특성을 이용한 한국어 불규칙 활용 어절의 형태소 분석 방법," 제 5회 한국어 정보처리 학술발표 논문집.

강승식, 1994, "다층 형태론과 한국어 형태소 분석 모델," 제 6회 한국어 정보처리 학술발표 논문집.

강승식, 권혁일, 김동렬, 1995, "한국어 자동 색인을 위한 형태소 분석 기능," 한국정보과학회 봄 학술발표논문집, Vol. 22, No. 1.

강인호, 김재훈, 김길창, 1998, "최대 엔트로피 모델을 이용한 한국어 품사 태깅," 제 10회 한글 및 한국어 정보처리 학술발표 논문집.

김민정, 권혁철, 1991, "한국어 형태소 분석에서의 수사처리," 제 3회 한국어 정보처리 학술발표 논문집.

김병희, 임권묵, 송만석, 1993, "형태소 접속 특성과 인접 말마디 정보를 이용한 형태소 분석기," 제 5회 한국어 정보처리 학술발표 논문집.

김윤호, 이종국, 김항준, 이상조, 1992, "형태소 분석을 이용한 문자인식 에러의 검출," 제 4회 한국어 정보처리 학술발표 논문집.

김재한, 안미정, 옥철영, 1993, "활용 형태소에 기반한 한국어 형태소 분석기," 한국정보과학회 가을 학술발표논문집, 1993, Vol. 20, No. 2.

김재한, 옥철영, 1994, "어절 사전을 이용한 한국어 형태소 분석," 한국정보과학회 봄 학술발표논문집, Vol. 21, No. 1.

김재훈, 조정미, 김창현, 서정연, 김길창, 1993, "퍼지 망을 이용한 한국어 품사 태깅," 제 5회 한글 및 한국어 정보처리 학술발표 논문집.

김재훈, 서정연, 김길창, 1995, "실용적인 한국어 형태소 해석," 한국과학기술원 전산학과, 기술문서(CS-TR-95-98).

김재훈, 1998, "가중치 망을 이용한 한국어 품사 태깅," 한국정보과학회 논문지, Vol. 25, No. 6.

김진동, 임희석, 임해창, 1997, "Twoply HMM: 한국어의 특성을 고려한 형태소 단위의 품사 태깅 모델," 한국정보과학회 논문지, Vol. 24, No. 12.

김진동, 이상주, 임해창, 1998, "어절 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델," 제 10회 한글 및 한국어 정보처리 학술발표 논문집.

박영환, 김경서, 송만석, 1991, "말뭉치에 기반한 형태소 분석기 및 철자 검사기의 구현," 한국정보과학회 가을 학술발표논문집, 1991, Vol. 18, No. 2.

송연정, 배우정, 이기오, 이용석, 1994, "형태소 분석기의 자질구조 생성에 관한 연구," 한국정보과학회 봄 학술발표논문집, Vol. 21, No. 1.

신상현, 이근배, 이종혁, 1997, "통계와 규칙에 기반한 2단계 한국어 품사 태깅 시스템," 한국정보과학회 논문지, Vol. 24, No. 2.

신중호, 한영석, 박영찬, 최기선, 1994, "어절구조를 반영한 은닉 마르코프 모델을 이용한 한국어 품사 태깅," 제 6회 한글 및 한국어 정보처리 학술발표 논문집.

이상주, 류원호, 김진동, 임해창, 1998, "품사 태깅을 위한 어휘 규칙의 자동획득," 제 10회 한글 및 한국어 정보처리 학술발표 논문집.

이상주, 류원호, 김진동, 임해창, 1999, "품사 태깅을 위한 어휘문맥 의존규칙의 말뭉치기반 중의성주도 학습," 한국정보과학회 논문지, Vol. 26, No. 1.

이성진, 김덕봉, 서정연, 최기선, 김길창, 1992, "Two-level 모델을 이용한 한국어 용언의 형태소 해석," 한국정보과학회 가을 학술발표논문집, 1992, Vol. 19, No. 2.

이영주, 1989, "자동색인을 위한 한국어 형태소 분석 알고리듬," 한글 및 한국어 정보처리 학술발표 논문집.

이운재, 1993, "한국어 문서 태깅 시스템의 설계 및 구현," 한국과학기술원 전산학과 석사학위논문.

이은철, 이종혁, 1992, "계층적 기호접속 정보를 이용한 한국어 형태소 분석기의 구현," 제 4 회 한국어 정보처리 학술발표 논문집.

이정규, 이상주, 임희석, 임해창, 1997, "규칙 기반 한국어 품사 태깅을 위한 어휘 규칙 획득의 수작업 최소화 방안," 한국정보과학회 봄 학술발표논문집 Vol. 24, No. 1.

임철수, 1994, "HMM을 이용한 한국어 품사태깅 시스템 구현," 한국과학기술원 전산학과 석사학위논문.

임희석, 1993a, "어절의 중의성 유형 분류에 근거한 한국어 형태소 분석기," 고려대학교 전산학과 석사학위논문.

임희석, 이호, 임해창, 1993b, "형태소 분석 단계에서 발생하는 어절의 중의성 분석 방안," 한국정보과학회 봄 학술발표논문집 Vol. 20, No. 1.

임희석, 김진동, 임해창, 1996, "변형 규칙 기반 한국어 품사 태거의 개선," 제 8 회 한글 및 한국어 정보처리 학술발표 논문집.

임희석, 김진동, 임해창, 1997, "어절 태그 변형 규칙을 이용한 한국어 품사 태거," 한국정보과학회 논문지, Vol. 24, No. 6.

임희석, 김진동, 임해창, 1998, "통계 정보와 언어 지식의 보완적 특성을 고려한 혼합형 품사 태깅," 한국정보과학회 논문지, Vol. 25, No. 11.

장동수, 서영훈, 1993, "음절에 기반한 한국어 형태소 분석기," 제 5 회 한국어 정보처리 학술발표 논문집.

장병탁, 김영택, 1990, "다중언어 형태소 분석 및 합성을 위한 언어규칙의 기계학습," 한국정보과학회 논문지, 1990. 7, Vol. 17, No. 4.

전자통신연구원, 1999a, "품사부착 말뭉치 구축을 위한 품사 태그세트 지침서", 전자통신연구원 컴퓨터소프트웨어 기술연구소.

전자통신연구원, 1999b, "전자사전 표제어 선정 지침서", 전자통신연구원 컴퓨터소프트웨어 기술연구소.

전자통신연구원, 1999c, "품사태그 부착 말뭉치 구축 지침서", 전자통신연구원 컴퓨터소프트웨어 기술연구소.

조영환, 차희준, 김길창, 1993, "확장 사전 환경에서의 한국어 형태소 해석과 생성," 제 5 회 한국어 정보처리 학술발표 논문집.

최재혁, 이상조, 1993a, "양방향 최장일치법을 이용한 형태소 분석기," 한국정보과학회 봄 학술발표논문집, Vol. 20, No. 1.

최재혁, 이상조, 1993b, "양방향 최장일치법에 의한 한국어 형태소 분석기에서의 사전 검색 횟수 감소 방안," 한국정보과학회 논문지, 1993. 10, Vol. 20, No. 10.

허윤영, 권혁철, 1994, "'의미적 한단어' 유형 분석 및 형태소 분석 기법," 제 6 회 한국어 정보처리 학술발표 논문집.

C. D. Manning, H. Schutze, 1999, "Foundations of Statistical Natural Language Processing," p. 269, The MIT Press, Cambridge, Massachusetts, London, England.

MUC web home page, http://www.muc.saic.com/

Senseval web home page, http://www.itri.bton.ac.uk/events/senseval/

SUMMAC web home page, http://www-nlpir.nist.gov/related_projects/tipster_summac/

TREC web home page, http://trec.nist.gov/