한국어 형태소 분석 시스템에 대한 평가 방법 및 적용 사례 분석

김진동*, 임해창*, 박재득**, 이재성**
*고려대학교 컴퓨터학과 자연어처리연구실
{jin, rim}@nlp.korea.ac.kr
**한국전자통신연구원 지식정보연구부
{jdpark, jasonl}@etri.re.kr

Evaluation Method for Korean Morphological Analysis System and it's Application to MATEC99

Jin-Dong Kim*, Hae-Chang Rim*, Jay Duke Park**, Jae Sung Lee**
*NLP Lab., Dep. of Computer Sci.&Eng., Korea University
**Dept. of Knowledge Information, ETRI

언어계통상 교착어에 속하는 한국어는 형태소 분석 결과가 복잡하게 주어지기 때문에 형태소 분석 시스템에 대한 효과적인 평가가 쉽지 않다. 본 논문에서는 한국어 형태소 분석 시스템에 대한 평가 방법을 제시한다. 또한 이를 MATEC99에 적용한 사례를 분석하여 이에 대한 타당성을 입증하고 보완점을 기술한다.

1. 서론

과학 기술 분야에서 개별 연구 결과에 대해 객관적으로 평가할 수 있는 평가 방법의 존재는 연구 의욕을 고취시키며 연구의 방향을 제시해 주는 역할을 한다. 자연 언어 처리 분야의 경우 몇몇 연구 기관에 의해 인간의 언어 지식이 첨가된 다양한 언어 자원(resource)들이 구축되어 있어서 자연 언어 처리 시스템 평가의 기준 역할을 한다. 이러한 언어 자원의 대표적인 예로 Brown Corpus[1], Pen Teebank[2] 등을 들 수 있다. 또한 세계적으로 권위와 객관성을 인정받는 평가 대회들이 개최되어 자연 언어 처리 연구의 의욕을 고취시키고 효과적인 평가 항목, 평가 기준 등을 보급하고 있다. 이러한 평가 대회의 대표적인 예로 SENSEVAL[3], MUC[4], TREC[5] 등을 들 수 있다.

그러나 이러한 평가 자원, 평가 대회 등은 대부분 영어권 언어를 대상으로 한 것이기 때문에 언어적 특성상 영어와 많은 차이점을 가지는 한국어를 대상으로 한 시스템 평가에는 거의 도움을 주지 못한다. 현재 한국어를 대상으로 하는 언어 자원들이 몇몇 존재하나 객관적인 평가 자원으로 공인받기에는 그 양과 질에서 크게 미흡한 실정이다. 또한 이러한 평가 자원, 평가 항목 등을 개발하고 보급하는 역할을 담당해야 할 대규모 평가 대회가 개최된 적이 없기 때문에 그동안 한국어 처리 분야에서는 개별 연구 기관 별로 소규모 언어 자원에 대해 자체 고안된 평가 항목으로 시스템을 평가해 온 실정이다.

본 논문에서는 한국어 형태소 분석 시스템에 대한 평가 대회인 MATEC99에서 사용하기 위해 고안된 한국어 형태소 분석 시스템에 대한 평가 항목들에 대해 소개하고 이를 MATEC99에 실제로 적용해 본 경험을 바탕으로 보완점과 개선점을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서 한국어 형태소 분석 단계에 대해 간략하게 기술하고 있으며 이 절을 통해 본 논문에 걸쳐 사용된 용어를 소개하고 있다. 3절에서 한국어 형태소 분석 시스템의 평가를 위해 고안된 평가 항목에 대해 소개하며 4절에서는 이를 실제로 MATEC99에 적용한 사례를 기술한다. 5절에서는 간략한 요약 후에 개선점을 기술한다.

2. 한국어 형태소 분석 단계

자연 언어 처리(natural language processing)의 작업을 독립적인 몇 개의 단계로 나눌 때, 가장 하부에 위치하는 것이 **형태소 분석(morphological analysis)** 단계다. 형태소 분석 단계에서는 분석의 대상이 되는 **단어(word)**를 형태론적인 관점에서 분석하여 그 단어(word)의 **통사적 부류(syntactic class)** 또는 **품사(part-of-speech)**를 결정하는 작업을 한다.

그러나 자연 언어에서의 단어는 실제 발현되는 언어의 다양한 **문맥(context)** 가운데서 각기 다른 통사적 부류로 나타나는 경우가 많다. 그렇기 때문에 어떠한 단어 하나만을 독립적으로 놓고 보았을 때 그 단어에 대한 형태소 분석 결과는 일반적으로 유일하게 결정되지 못하고 몇 개의 가능한 후보가 주어지게 된다. 예를 들어 영어 단어 *note*의 품사는 *NOUN*도 될 수 있고 *VERB*도 될 수 있다. 하나의 단어가 하나 이상의 품사로 사용될 수 있는 것은 자연 언어가 가지는 특징이며 자연 언어의 단어가 가지는 이와 같은 성질을 단어의 **품사 중의성(part-of speech ambiguity)**이라고 한다. 또한 어떤 단어에 대한 형태론적인 분석을 통해서 그 단어가 가질 수 있는 중의

$$\text{형태소 분석 재현률} = \frac{\text{정답과 일치하는 응답 형태소 구조의 개수}}{\text{정답 형태소 구조의 개수}} \quad [\text{식 1}]$$

$$\text{형태소 분석 정확률} = \frac{\text{정답과 일치하는 응답 형태소 구조의 개수}}{\text{응답 형태소 구조의 개수}} \quad [\text{식 2}]$$

$$\text{태깅 정답 제시율} = \frac{\text{형태소 분석 결과 중에 정답과 일치하는 형태소 구조가 있는 어절의 개수}}{\text{어절의 개수}} \quad [\text{식 3}]$$

$$\text{평균 후보 수} = \frac{\text{모든 어절에 대한 형태소 분석 결과에 포함된 형태소 구조의 개수}}{\text{어절의 개수}} \quad [\text{식 4}]$$

적인 모든 품사를 찾아내는 도구를 **형태소 분석기(morphological analyzer)**라고 한다.

한편, 품사 중의성을 갖는 단어라고 할지라도 그 단어가 실제로 사용되었을 때의 문맥을 고려하면 그 단어의 품사를 하나로 결정할 수 있다. 예를 들어 "*Please note the phenomena.*"라는 문장에서 사용된 단어 *note*의 통사적 부류는 *VERB*임을 알 수 있다. 이렇게 주위의 문맥을 고려해서 품사 중의성을 가지는 단어의 품사를 결정하는 작업을 **품사 태깅(part-of-speech tagging)**이라고 하며 품사 태깅을 위한 도구를 통상 **품사 태거(part-of-speech tagger)**라고 한다.

한국어의 경우 형태론적 분석의 대상은 한국어 띄어쓰기의 단위인 **어절(eojeol)**이 된다. **교착어(agglutinative language)**인 한국어는 **형태소(morpheme)**의 종류가 많고 이들의 결합 형태도 다양하기 때문에 한국어 어절에 대한 형태론적 분석 결과가 매우 복잡하게 주어진다. 예를 들어 한국어 어절 *나는*을 형태론적으로 분석하였을 때 이의 결과를 **형태소 구조(morphological parse)**와 통사적 부류로 나타내면 [표 1]과 같다.

[표 1] 어절 '*나는*'에 대한 형태소 분석 결과

어절	형태소 구조	통사적 부류
나는	나[대명사]+는[보조사]	주격대명사
	나[동사]+는[관형형어미]	동사관형형
	날[동사]+는[관형형어미]	동사관형형

한국어는 어절의 통사적 기능이 매우 다양하기 때문에 이를 효과적으로 분류하기가 쉽지 않다. 이러한 이유로 한국어에서는 형태소 분석의 결과를 통사적 부류보다는 형태소 구조로 표현하는 경우가 많으며 품사 태깅의 결과 또한 형태소 구조로 표현된다.

형태소 분석 단계에서 요구되는 또 하나의 중요한 도구로서 **명사 추출기(noun extractor)**를 들 수 있다. 명사 추출기는 자연 언어로 기술된 문서에서 명사를 인식하기 위해 사용되며 주로 정보 검색 분야에서 많이 요구된다.

3. 한국어 형태소 분석 시스템 평가 방법[1]

한국어 형태소 분석 시스템을 평가하기 위해서는 먼저 실제로 평가가 이루어질 **평가 공간**이 규정되어야 하고 평가 공간에 대해 이상적인 시스템이 내어줄 결과가 미리 작성된 **평가 정답**이 마련되어야 한다. 예를 들어 품사 태거의 경우 일정한 원칙에 의해 수집된 원시 말뭉치가 평가 공간의 역할을 할 것이고 이에 대해 품사 정보를 첨가한 품사 태깅된 말뭉치가 평가 정답의 역할을 할 것이다. 평가에 참여하는 각 시스템들은 평가 공간에 대하여 나름대로의 **평가 응답**을 작성하여 제출한다. 실제의 평가는 준비된 **평가 항목**에 따라 평가 응답과 평가 정답을 비교함으로써 이루어진다.

3.1. 형태소 분석기 평가

이상적인 형태소 분석기는 어떠한 어절이 주어지더라도 그 어절에 대해 가능한 모든 형태소 구조를 생성해 주어야 하며 이 때 그 어절에 대해 가능하지 않은 형태소 구조를 생성해서는 안된다. 이러한 형태소 분석기를 가정할 때 실재의 형태소 분석기들에 대한 평가는 각 형태소 분석기가 이상적인 형태소 분석기에 비해 어느 정도의 다른 결과를 내는지 즉 어느 정도의 오류를 범하는지를 측정함으로써 이루어질 수 있다.

형태소 분석기들이 범하는 오류는 두 종류로 나누어질 수 있다. 하나는 형태소 구조 **미생성 오류**로서 이상적인 형태소 분석기라면 생성해 줄 형태소 구조를 생성해 내지 못하는 오류이다. 또 하나는 형태소 구조 **과생성 오류**로서 이상적인 형태소 분석기라면 생성하지 않을 형태소 구조를 생성해 내는 오류이다. 결과적으로 형태소 구조의 미생성 오류와 과생성 오류를 최소화하는 형태소 분석기일수록 성능이 좋은 형태소 분석기라고 할 수 있다.

이러한 사항에 기초하여 형태소 분석기의 성능에 대한 척도로서 **재현율**과 **정확률**을 사용할 수 있다. 재현율은 [식 1]과 같이 계산되며 형태소 구조 미생성 오류를 얼마나 최소화하였는지를 나타낸다. 또한, 정확률은 [식 2]와 같이 계산되며 형태소 구조 과생성 오류를 얼마나 최소화하였는지를 나타낸다.

형태소 분석 재현율과 정확률의 평가는 한국어 어절 목록을 평가 공간으로 하여 이루어진다. 또한, 이 어절 목록에 대한 기분석 사전이 평가 정답의 역할을 한다. 기분석 사전은 어절 목록에 포함된 각 어절에 대해서 이상적인 형태소 분석기가 내어줄 것이라고 생각되는 결과를 미리 등록해 놓은 것을 말한다.

한 편, 형태소 분석기의 대표적인 응용 시스템으로 품사 태거를 들 수 있기 때문에 품사 태거의 입장에서 형태소 분석기를 평가하는 것도 가치있는 평가라고 할 수 있다. 품사 태거의

1) 본 논문에서는 정량적인 평가만을 다룬다.
2) 삭제(소)→삽입(론)→삽입(적)

$$\text{어절 단위 태깅 정확률(어절 단위 태깅 재현률)} = \frac{\text{정확히 태깅된 어절의 개수}}{\text{어절의 개수}} \quad [\text{식 5}]$$

$$\text{형태소 단위 태깅 재현률} = \frac{\text{정답과 일치하는 응답 형태소의 개수}}{\text{정답 형태소의 개수}} \quad [\text{식 6}]$$

$$\text{형태소 단위 태깅 정확률} = \frac{\text{정답과 일치하는 응답 형태소의 개수}}{\text{응답 형태소의 개수}} \quad [\text{식 7}]$$

$$\text{명사 태깅 재현률} = \frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{정답 명사의 개수}} \quad [\text{식 8}]$$

$$\text{명사 태깅 정확률} = \frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{응답 명사의 개수}} \quad [\text{식 9}]$$

$$\text{명사 추출 재현률} = \frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{정답 명사의 개수}} \quad [\text{식 10}]$$

$$\text{명사 추출 정확률} = \frac{\text{정답과 일치하는 응답 명사의 개수}}{\text{응답 명사의 개수}} \quad [\text{식 11}]$$

입장에서 보았을 때 이상적인 형태소 분석기는 각 어절에 대해서 그 어절이 포함된 문맥에 맞는 형태소 구조를 반드시 생성해야 하며 될 수 있으면 적은 형태소 구조를 생성해야 한다. 품사 태거의 입장에서 본 형태소 분석기의 성능에 대한 척도로서 ***태깅 정답 제시율***과 어절당 ***평균 후보 수***를 사용할 수 있다. 태깅 정답 제시율은 품사 태거의 정확률에 직접적으로 영향을 미치며 [식 3]으로 계산된다. 또한 평균 후보 수는 품사 태거의 탐색 공간의 크기와 직접적으로 관련이 있으며 [식 4]로 계산된다.

태거의 입장에서 형태소 분석기를 평가하는 태깅 정답 제시율과 평균 후보 수 평가는 한국어에서 사용되는 실제 문장을 모아놓은 원시 말뭉치를 태깅 공간으로 하여 이루어진다. 또한 이 원시 말뭉치에 형태소 구조 정보를 첨가한 품사 태깅된 말뭉치가 평가 정답의 역할을 한다.

형태소 분석기의 재현율과 정확률 평가 항목은 형태론적으로 발생 가능한 모든 형태소 구조를 동등하게 취급하기 때문에 언어학적인 연구를 위한 응용으로서의 시스템 평가 항목으로 적합한 반면, 태깅 정답 제시율과 평균 후보 수 평가 항목은 실제로 발생 빈도가 높은 형태소 구조에 많은 가중치를 주게 되므로 실제로 한국어를 처리하는 응용으로서의 시스템 평가 항목으로 적합하다.

3.2. 품사 태거 평가

품사 태거의 평가 공간으로는 한국어 문장들을 모아 놓은 원시 말뭉치가 사용되며 평가 정답으로는 품사 태깅된 말뭉치가 사용된다. 이 때, 품사 태거의 평가는 원시 말뭉치의 각 어절에 대해 품사 태거가 결정한 형태소 구조를 정답과 비교함으로써 이루어지는데, 비교의 단위는 어절이 될 수도 있고 형태소가 될 수도 있다.

어절 단위 평가의 경우 품사 태깅의 **재현율**과 **정확률**은 [식 5]와 같은 공식에 의해 똑같이 계산되기 때문에 일반적으로 정확률만 측정된다.

형태소 단위 평가는 해당 어절의 응답 형태소 구조와 정답 형태소 구조를 형태소 단위로 비교함으로써 이루어진다. 따라서, 응답 형태소 구조와 정답 형태소 구조가 일치하면 응답 형태소 구조에 포함된 형태소들이 모두 맞는 것으로 계산된다. 응답 형태소 구조와 정답 형태소 구조가 일치하지 않는 경우는 정답 형태소 구조에 비해 응답 형태소 구조에서 미생성된 형태소와 과생성된 형태소를 찾아내어 이를 이용해 **재현율**과 **정확률**을 계산한다. 이를 위해 편집 거리(edit distance)를 구하는 동적 알고리즘(dynamic algorithm)[6]을 변형하여 활용할 수 있다.

편집 거리는 원래 두 개의 문자열간 거리를 구하기 위해 자주 사용된다. 문자열1과 문자열2간의 편집 거리는 문자열1을 문자열2로 바꾸기 위해 필요한 문자 단위 편집의 최소 비용으로 정의된다. 예를 들어 문자열 "형태소분석"을 문자열 "형태론적분석"으로 바꾸기 위해서는 한 문자를 삭제하고 두 문자를 삽입하는 편집 작업이 필요하다.[2] 이 때 두 문자열간의 거리를 $1\times\alpha+2\times\beta$로 나타낼 수 있다.[3]

편집의 단위를 문자 대신 형태소로 하면 비슷한 방식으로 응답 형태소 구조와 정답 형태소 구조간의 편집 거리를 구할 수 있다. 예를 들어 어떤 문맥에서 어절 '*나는*'에 대한 응답 형태소 구조가 '*나[동사]+는[관형형어미]*'인데, 정답은 '*날[동사]+는[관형형어미]*'라면 응답 형태소 구조를 정답 형태소 구조로 교정하는데 필요한 비용은 삭제 한 번, 삽입 한 번이므로[4] 응답에 과생성된 형태소가 한 개 미생성된 형태소가 한 개 있는 것으로 판단할 수 있다. 또한 이 응답에서 맞는 형태소는 한 개

2) 삭제(소)→삽입(론)→삽입(적)

3) 원래 편집 거리를 구할 때는 편집 명령을 '삽입', '삭제', '치환'의 세 종류로 구분하여 비용을 계산하나, 여기에서는 편의상 '한 번의 치환'을 '한번의 삭제 + 한번의 삽입'으로 대체하여 계산하였다.

4) 삭제(*나[동사]*)→삽입(*날[동사]*)

이다.

품사 태거의 형태소 단위 평가는 **품사별**로 이루어질 수도 있다. 예를 들어 명사의 태깅 **재현율**과 **정확률**은 각각 [식 6]과 [식 7]로 구해지며, 다른 품사에 대해서도 같은 방식으로 재현율과 정확률을 구할 수 있다.

태거의 품사별 성능 평가는 해당 태거의 특성을 파악하는 좋은 자료가 된다. 예를 들어 어떤 태거가 다른 성능은 떨어지지만 명사 태깅 재현율과 정확률이 높다면 정보 검색을 위한 도구로 유용하다는 판단을 내릴 수 있다.

3.3. 명사 추출기 평가

명사 추출기는 정보 검색을 위한 목적으로 많이 사용되므로 일반적인 정보 검색 환경을 고려하면 일정 크기의 문서 집합을 평가 공간으로 사용하는 것이 바람직하다. 이러한 경우 평가 공간의 각 문서에 대해 미리 작정된 정확한 명사 목록이 평가 정답으로서의 역할을 한다.

명사 추출기에 대한 평가는 평가 공간의 각 문서에 대해 명사 추출기가 응답한 명사 목록을 정답과 비교함으로써 이루어진다. 일반적으로 명사 추출기에 대한 평가 항목으로 **재현율**과 **정확률**이 사용된다. 명사 추출기의 재현율은 [식 8]과 같이 사용되며 명사 미생성 오류를 얼마나 최소화하였는지를 나타낸다. 또한 정확률은 [식 9]와 같이 사용되며 명사 과생성 오류를 얼마나 최소화하였는지를 나타낸다.

4. MATEC99에서의 적용 사례

MATEC99에서는 형태소 분석기, 태거, 명사 추출기의 세 부문에 걸쳐 참가 시스템들의 성능을 평가하였다. 평가 수행을 위해 ETRI에서 구축된 30만 어절 규모의 품사 태깅된 말뭉치가 활용되었다. 이 말뭉치 중 3만 어절 정도의 말뭉치가 평가용으로 사용되었고 나머지는 학습을 위해 참가팀들에게 배포되었다. 결국 평가용으로 활용 가능한 자원이 3만 어절의 품사 태깅된 말뭉치뿐이었으므로 실제 평가에 필요한 여러 종류의 평가 공간, 평가 정답들은 이 말뭉치를 기반으로 생성되었다.

형태소 분석기의 평가를 위해 두 종류의 평가 공간과 평가 정답이 필요하다. 하나는 형태소 분석 정확률과 재현율을 평가하기 위한 어절 목록과 이에 대한 기분석 사전이며, 또 하나는 형태소 분석기의 태깅 정답 제시율과 평균 후보 수를 평가하기 위한 원시 말뭉치와 이에 대한 품사 태깅된 말뭉치이다.

MATEC99에서는 먼저 형태소 분석 정확률과 재현율을 평가하기 위하여 ETRI 3만 원시 말뭉치에 포함된 어절들의 목록을 추출하여 이를 평가 공간으로 사용하였다. 문제는 평가 공간에 대한 기분석 사전을 준비하는 일이었는데, 품사 태깅된 말뭉치만을 통해서는 기분석 사전을 추출할 수 없고 이를 수작업으로 보완하는 데도 상당한 비용이 들기 때문에 부득이하게 완전한 기분석 사전 대신에 과반수지지 분석 사전을 구축하였다.

과반수지지 분석 사전은 평가 공간인 어절 목록의 각 어절에 대해서 평가에 참여한 형태소 분석기들이 내어준 결과를 취합하여 이들 중에서 과반수 이상의 분석기들이 생성한 형태소 구조만을 모아서 이를 해당 어절에 대한 과반수지지 분석으로 등록한 것이다[5]. 과반수지지 분석 사전은 평가에 참여한 팀들 중 과반수 이상이 지지한 형태소 구조는 맞을 가능성이 많고 반대로 과반수 이상의 지지를 얻지 못한 형태소 구조는 틀릴 가능성이 많다는 가정에 근거한 것이다. 그러나 이 가정이 객관적인 지지를 얻기에는 설득력이 충분치 못하기 때문에 과반수지지 분석 사전에 기초한 형태소 분석 재현율과 정확률 평가는 정확한 시스템 성능 평가라기 보다는 참가 시스템들의 상대적인 비교 자료로서의 의미를 지닌다.

형태소 분석기의 태깅 정답 제시율과 평균 후보 수 평가를 위해서는 평가용 ETRI 원시 말뭉치가 그대로 평가 공간으로 사용되었으며, 이에 따라 평가용 ETRI 품사 태깅된 말뭉치가 평가 정답으로 사용되었다.

형태소 분석기 평가를 위해 서로 다른 두 가지의 평가 공간이 사용되었으므로 평가시에는 형태소 분석기 평가에 참여한 각 팀들에게 두 가지의 평가 공간을 제시하여 이에 대한 응답을 얻어야 한다. 그러나, MATEC99에서는 태깅 정답 제시율과 평균 후보 수 평가를 위한 평가 공간, 즉 원시 말뭉치만을 참가팀들에게 제시하여 이에 대한 응답만을 얻었다. 그리고 이를 정렬하고 유일화[6]한 결과를 형태소 분석 재현율과 정확률 평가를 위한 평가 응답으로 대신하였다. 이렇게 한 이유는 어차피 형태소 분석기는 문맥을 고려하지 않고 어절별로 형태소 분석 결과를 낼 것이므로 이를 유일화하면 유일 어절 목록에 대한 형태소 분석 결과와 같을 것이기 때문이다.

품사 태거 평가를 위한 평가 공간으로는 ETRI 평가용 원시 말뭉치를 사용하였고, 평가 정답으로는 ETRI 평가용 품사 태깅된 말뭉치를 사용하였다.

명사 추출기의 평가를 위해서는 ETRI 평가용 원시 말뭉치에서 어절 수 100~120인 문서 309개를 추출하여 이를 평가 공간으로 사용하였으며, 이에 대응하는 품사 태깅된 말뭉치에서 명사로 태깅된 단어들의 목록이 평가 정답으로 사용되었다.

한국어 형태소 분석 단계에서는 어절을 형태소 단위로 분리하여 분석하는 것이 일반적이다. 그러나 형태소의 경계에 대한 해석 차이 때문에 형태소 분석의 결과가 다르게 나올 수 있다. 대표적인 예가 복합명사에 대한 분석인데, 예를 들어 어절 '*형태소분석*'에 대한 응답 형태소 구조가 '*형태소[명사]+분석[명사]*'인데 정답은 '*형태소분석[명사]*'일 수가 있다.[7] 이런 경우, 응답과 정답이 완전 일치(exact matching)하지 않는다고 해서 응답이 틀렸다고 할 수는 없다. 이런 문제는 형태소 분석기의 평가시 또는 품사 태거의 평가시에 발생할 수 있으며, 특히 복합명사와 관련해서는 명사 추출기의 평가시에도 발생할 수 있다. MATEC99에서는 단일 명사간 경계에 대한 해석 차이 때문에 발생하는 불일치 문제에 한해 각 평가 부문 모두에서 맞는 것으로 인정하였다.[8]

5. 평가 결과

MATEC99에 참여한 팀은 형태소 분석기 평가 부문 8팀, 품사 태거 평가 부문 8팀, 명사 추출기 평가 부문 14팀이었으나 제출한 응답의 포맷 문제로 형태소 분석기 평가 부문 전 항목

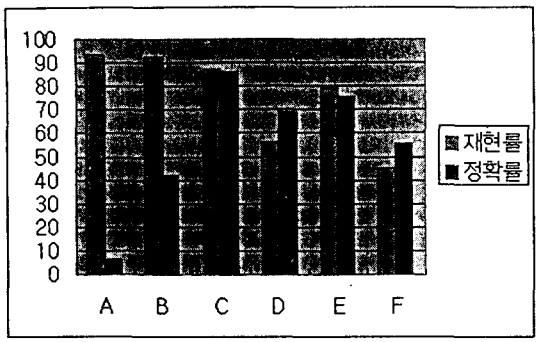
---

5) MATEC99의 형태소 분석기 평가에 참가한 팀들의 응답에 기초한 과반수지지 분석 사전은 평균 후보 수가 1.85개였고 형태소 분석이 전혀 없는 어절이 446개였다. 실제 평가시에는 이 사전을 보완하기 위하여 ETRI99 태깅된 말뭉치의 해당 어절에 태깅된 형태소 구조를 추가하였다.

6) UNIX 시스템에서 sort & uniq 명령어를 통해 이를 수행하였다.

7) 이와 비슷한 문제가 *[명사]+[접미사]*, *[어미]+[어미]*, *[조사]+[조사]*, *[동사]+[어미]+[동사]* 등의 경우에도 발생할 수 있다. 그러나 MATEC99에서는 가장 큰 비중을 차지하는 *[명사]+[명사]*의 경우만을 고려하였다.

8) 응답과 정답간 비교 전에 *[명사]+[명사]*의 형태로 분리된 분석이 있는지 먼저 검사하였다. 이러한 형태가 발견되면 두 개의 명사를 하나의 명사로 합친 후 응답과 정답간 비교를 수행하였다.



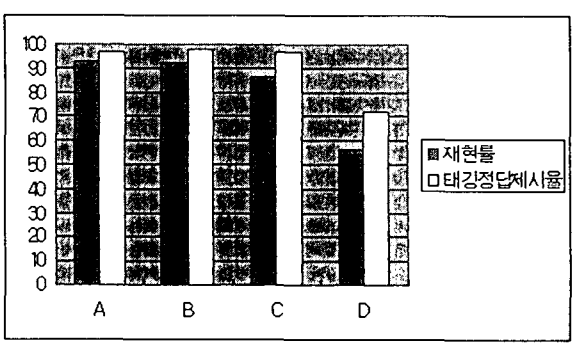
[그림 1] 형태소 분석기의 재현율, 정확률 평가 결과

에서 1팀이, 태깅 정답 제시율과 평균 후보 수 평가 항목에서 2팀이 실격 처리 됐으며, 품사 태깅 평가 부문에서도 1팀이 실격 처리되었다.

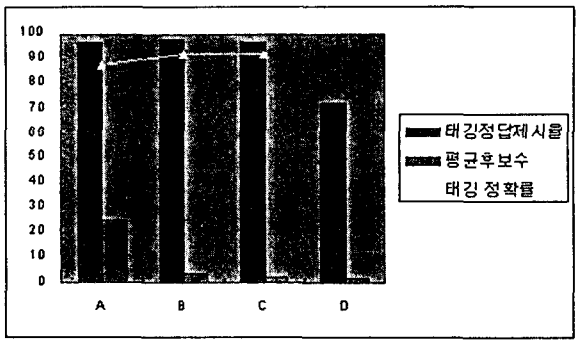
[그림 1]에 형태소 분석기 평가에 참가한 팀들의 재현율과 정확률이 나타나 있다. A팀과 B팀은 가장 높은 재현율을 보이지만 정확도에 있어서는 낮은 수치를 보이고 있다. A팀의 경우 어절 당 평균 24.79개에 달하는 분석의 개수가 정확도를 현저히 떨어뜨리는 원인이 되었으며, B팀 역시 여타 다른 팀들의 2배정도 되는 어절 당 평균 4.13개의 형태소 구조를 형태소 분석 결과로 내주고 있었다. C팀의 경우 재현율과 정확률이 비교적 고루 높은 수치를 보이고 있다. 그러나, 재현율과 정확률 평가가 정확한 기분석 사전을 토대로 이루어진 것이 아니라 과반수지지 분석 사전을 토대로 이루어진 것이기 때문에 이러한 평가 수치를 통해 시스템의 성능을 정확히 파악하기에는 무리가 있다. 그보다는 참가팀들 중에 C팀이 가장 보편적인 성능을 지녔기 때문에 평가 수치가 가장 높게 나왔다고 보는 것이 더 타당하다.

[그림 2]에서는 형태소 분석기의 재현율과 태깅 정답 제시율을 비교해서 보여주고 있다. C팀의 경우 재현율이 A팀과 B팀에 비해 떨어짐에도 불구하고 태깅 정답 제시율에서는 비슷한 성능을 보여주고 있다. 이와 같은 결과는 태깅 정답 제시율이 발생 빈도가 높은 형태소 구조에 크게 영향을 받기 때문에 발생한 것으로 풀이된다. 즉, C팀이 A팀과 B팀에 비해 형태론적으로 분석 가능하지만 발생 빈도는 낮은 형태소 구조를 생성하지 않는 경향이 있다고 볼 수 있다.

[그림 3]에서는 형태소 분석기의 태깅 정답 제시율과 평균



[그림 2] 형태소 분석기의 재현율과 태깅 정답 제시율 비교



[그림 3] 형태소 분석기의 태깅 정답 제시율, 평균 후보수 평가 결과

후보수 평가 결과를 보여주고 있으며, 실제 태깅 성능과의 상관관계 비교를 위해 어절 단위 태깅 정확률이 함께 보여진다.[9] A팀, B팀, C팀은 평균 후보 수의 차이와는 별 상관없이 높은 태깅 정답 제시율을 보이고 있다. 그러나 태깅 정확률에 있어서 A팀이 상대적으로 낮은 성능을 보이는데 그 이유는 A팀의 형태소 분석 평균 후보 수가 대단히 많기 때문에 품사 태거의 탐색 공간이 크게 된 것이 원인으로 작용한 것으로 풀이된다.

[그림 4]에서는 품사 태깅의 형태소 단위 재현율을, [그림 5]에서는 정확률을 보여준다. [그림 4]와 [그림 5]에 나타난 특징을 살펴보면 한국어의 경우 감탄사, 관형사, 접미사의 태깅 재현율이 그리 높지 않음을 알 수 있다. 또한 F팀의 경우 접두사에 대한 재현율과 정확률이 모두 0으로 나타났는데 이것으로 P팀의 품사 집합이 접두사를 포함하지 않는다는 것을 알 수 있다. C팀의 경우는 접속조사의 재현율이 아주 낮고, 격조사의 정확률도 상대적으로 낮다. 일반적으로 접속조사는 격조사의 일종인 부사격 조사와 혼동되는 경향이 많은 것으로 미루어 C팀은 접속조사를 부사격 조사로 잘못 태깅하는 경우가 많음을 알 수 있다. 품사 태거의 평가에 있어서 어절 단위 정확도 평가가 기본적으로 사용될 수 있으나 품사 태거의 특성을 파악하거나 품사 집합의 불일치 등의 문제에 유연하게 대처하기 위해서는 형태소 단위 평가가 더 효과적임을 알 수 있다.

6. 결론

한국어 형태소 분석 단계의 경우 분석 결과가 복잡하기 때문에 단순한 비교로는 효과적인 평가가 되기 힘들다. 본 논문에서는 한국어 형태소 분석 시스템에 대한 평가 방법을 제시하였다. 이 평가 방법은 MATEC99에서 비교적 성공적으로 수행되었으나 평가 과정에서 드러난 문제점도 무시할 수 없다.

먼저 형태소 분리에 대한 견해 차이에서 발생하는 형태소 구조간 불일치 문제가 매우 심각하기 때문에 이에 대한 해결 방안이 좀 더 보완되어야 한다. 또한 형태소 분석기의 재현율과 정확률 평가를 위한 평가 정답으로 기분석 사전이 아닌 과반수 지지 분석 사전이 사용된 것은 형태소 분석기 평가의 신뢰성을 떨어뜨리는 원인이 되었으므로 형태소 분석기의 정확한 분석을 위해서는 기분석 사전의 구축 작업이 이루어져야 한다. 과반수 지지 분석 사전이 이를 위한 기초 자원이 될 수 있을 것으로 기대된다.

9) D팀은 태깅 평가 부문에 참가하지 않았기 때문에 태깅 정확률 평가 수치가 없다.

참고 문헌

[1] Francis, W. N. and Kucera, H., *Frequency Analysis of English Usage: Lexicon and Grammar*, Houghton Mifflin, Boston, 1982

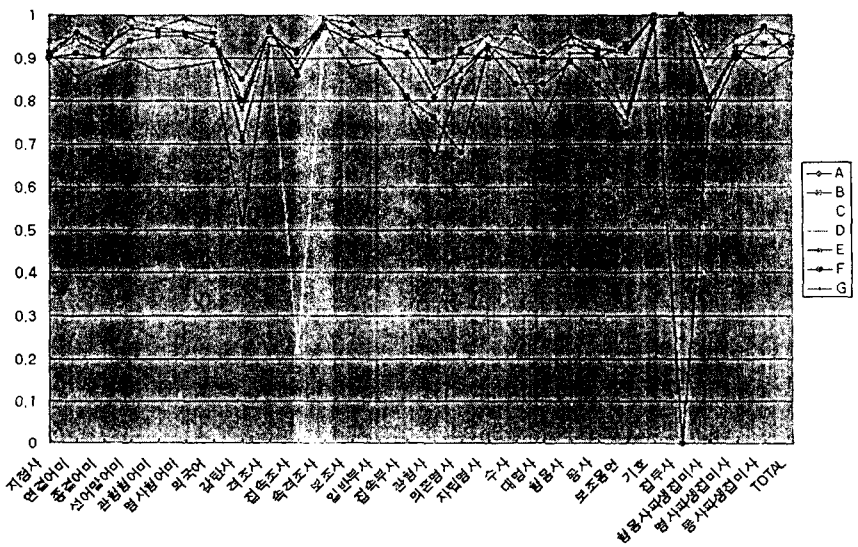
[2] Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A., "Building a large annotated corpus of English: the Penn Treebank," *Computational Linguistics* 19(1993), 313-330

[3] Senseval home page, http://www.itri.bton.ac.uk/events/senseval/

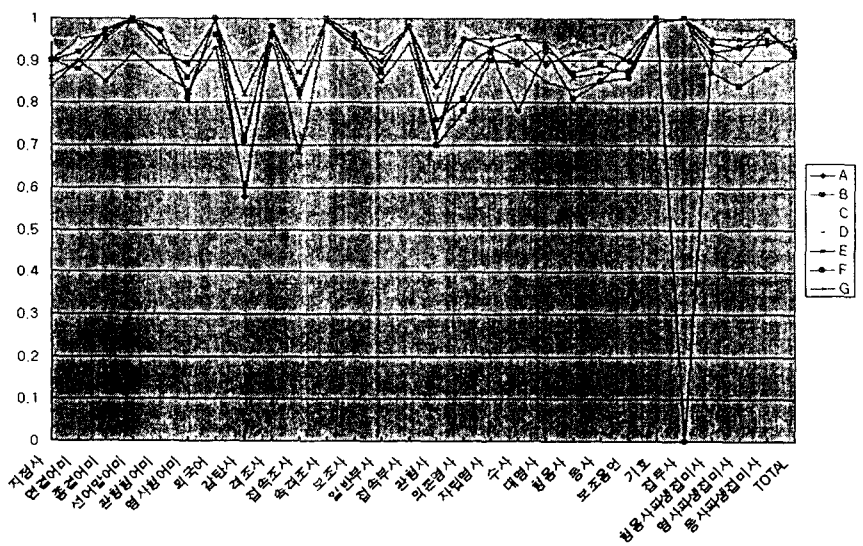
[4] TREC home page, http://trec.nist.gov/

[5] SAIC Information Extraction, http://www.muc.saic.com/

[6] Wagner, R. and Fisher, M, "The string to string correction problem," *JACM* 21(1974), 168-173



[그림 4] 품사 태거의 형태소 단위 재현율



[그림 5] 품사 태거의 형태소 단위 정확률