MATEC99 의 Test Suites 작성을 위한 표준안 도출 과정 상의 문제점 및 개선방안

박재득, 이재성, 차건회, 박세영, 이현아
{jdpark,jasonl,chakh,sypark}@etri.re.kr, halee@snl.co.kr

Problems and Direction of Improvement in the Standardization Process for the Test Suites of MATEC99

Jay Duke Park, Jae Sung Lee, Keon-Hoe Cha, Se Young Park, and Hyun A Lee
Knowledge Information Department,
Electronics and Telecommunications Research Institute & SnL Co.

요약

이 논문에서는 MATEC99 에서의 형태소 분석기 및 품사 태거의 평가를 위한 Test Suites, 즉, 평가용 시험문제와 문제에 대한 정답을 정하기 위하여 마련한 여러가지 표준안의 도출과정상에서 발견된 문제점의 유형을 기술하고자 한다. 그리고, 각각의 문제점 유형별로 현실적 제약으로 인한 어려움을 살펴보고, 한편으로는 일부 문제에 대해서는 시간적 제약 때문에 현재의 표준안에는 반영되지 않았지만 향후 점진적인 개선을 위한 방안을 제안하고자 한다. 대부분의 문제에 완전한 해결책이 발견되지 않고 있어 관련 전문가들의 조언을 이끌어 내기 위하여 문제를 공론화하고자함에도 목적이 있다.

1. 서론

1999 년 국내 처음으로 15 개팀이 참가한 가운데 MATEC99 를 개최하여 국내의 형태소 분석기술 및 태거 그리고 자동색인을 위한 명사추출기의 성능을 비교해 보았다. 그러나, 모두 알고 있는 것처럼 대회 개최전에도 예상되었던 몇가지 문제점들이 아직은 해결되지 않은 채로 있다. 그렇다고 모든 문제가 다 풀릴 때까지 기다릴 수는 없을 것이다. 오히려, 불완전하지만 이러한 대회를 통하여 발전적이고 현실적인 개선방안을 더 빨리 찾아낼 수 있을 것으로 생각된다.

이 대회 개최를 통해서 명료하게 드러난 문제점 및 개선 대책을 표준안과 연관하여 논하고 이를 바탕으로 후의 평가대회에서 더욱 공정하고 합리적인 평가기준으로써, 한편으로는 훨씬 좋은 품사 부착 말뭉치 구축의 지침으로 활용할 수 있는 계기로 삼고자 하는 것이다.

평가대회는 표준안 지침서, 학습용 품사 태그 부착 말뭉치, Test Suites, 채점 방법 등을 마련하여 이루어졌다. 이 중에서 가장 개선의 여지가 많은 것이 Test Suites 의 올바른 작성에 관한 것이라고 본다.

Test suites 도 품사 태그 부착 말뭉치의 일종이며, 학습용 품사 태그 부착 말뭉치와 대별된다. 둘 다 표준안에서 정한 지침대로 품사 태그를 원시 말뭉치에 부착한 것이다. 학습용 말뭉치는 평가 대회에 참가하는 팀들에게 평가 전에 표준안의 숙지 또는 개선된 표준안 제기, 시스템 튜닝 등의 연습에 활용하도록 제공되는 것이다. 학생들의 시험에 비유하자면, 학습용 품사 태그 부착 말뭉치는 모의 고사 문제 및 정답의 짝이고, Test Suites 는 본고사 문제 및 정답의 짝으로 볼 수 있다.

이 Test Suites의 바람직한 구축에 관련된 문제는 세부적으로 살펴보면 다음과 같은 표준안에 관련된 문제와 결부되어 있다.

1) 표준화를 위한 명확한 전제 조건 및 기본 원칙과 지침의 수립
2) 표준안 도출 과정상의 공정하고 합리적인 절차나 방법의 명확한 수립
   - 적절한 품사 태그 세트 설정 문제
   - 전자사전의 표제어 선정의 기준 설정 문제
   - 말뭉치에 품사 태그를 부착하는 태깅방식 설정 문제
   - 공정한 출제 범위 및 난이도 선정의 문제
   - 표준안 이의제기 처리 및 결과 평가 방법 문제

다른 문제들보다도 이러한 문제는 표준안 도출 및 표준안을 적용하여 Test Suites와 학습용 말뭉치를 구축하는 과정에 자주 대두되는 문제로써, 그동안 불명확하고 암묵적으로 통용되던 것을 명확히 언급하고 개선해야 되는 것들이다. 이들 각각에 대한 기존의 방안들을 검토해보고 부분적 개선방안을 차례대로 제시하고자 한다.

2. 표준화를 위한 명확한 전제 조건 및 기본 원칙과 지침의 수립

2.1 품사태그세트 용도설정: 전산처리적 응용

MATEC99와 관련하여 마련한 품사 부착 말뭉치는 언어학적인 다양한 자료 조사 및 다양한 목적으로 태깅을 한 것이 아니다. 이러한 다양한 목적을 위한 태깅은 비용과 시간이 많이 들 것이나 분명히 필요한 것이며, 이러한 용도의 말뭉치는 문화부에서 추진하는 세종계획에서 구축하고 있다. MATEC99의 품사 부착 말뭉치는 형태소 분석 이후에 통계적 자료를 이용한 애매성 해소 및 구문 분석으로의 입력 또는 명사, 동사 등의 추출을 위한 자동 품사 태깅에 응용하는 것을 목표로 한다.

2.1.1 1차적 용도: 자동 품사 태거(tagger)의 학습용 말뭉치 구축

이 평가와 관련된 표준안에서 제시하는 품사 태그 세트의 가장 주된 용도는 자동 품사 태거를 위한 학습자료로 활용되는 품사 부착 말뭉치 작성용이다. 이러한 용도는 의미론적인 어휘분류나 구문론적 또는 형태론적인 품사분류만 연구하는 사람들의 분류체계와 관점이 다르다는 것을 전제하는 것이다.

문장분석 과정에서 형태소 분석이 끝나고 구문 분석을 시작하기 전에 구문분석에서의 애매성 폭발을 줄이기 위하여 품사 태거를 활용, 형태소 분석의 출력인 어절별 분석결과들을 순위화(ranking)하여 상위 몇 개의 분석결과들만 구문분석의 입력으로 넘겨주기 위한 용도를 자연언어처리 분야의 연구자들은 주지하고 있을 것이다.

따라서, 태깅용 또는 태그 부착 말뭉치 작성용 품사 태그 세트는 어절 내에서의 다양하고 복잡한 형태소들의 결합 제약을 반영하는 형태론적 분류보다 덜 정밀하며, 문장 내에 멀리 떨어져 있는 어절간 또는 구절간의 순서, 공기 제약 관계 등의 현상을 반영하는 구문론적 분류보다도 덜 정밀하다. 그래서, 태깅용 품사 태그 세트는 형태소 분석과 구문 분석 사이의 중간 단계의 처리를 위한 '형태-통사분석론'적인 분류를 따른다고 말할 수 있다.

그리고, 정보검색에서의 명사 추출의 정확도를 높이기 위해서 자동 품사 태거를 사용할 수 있다는 것을 주지하고 있을 것이며, 품사 태거는 형태소 분석기와 밀접한 관련을 가지고 있고 대부분 형태소 분석의 결과를 활용할 수 있다. 따라서, 형태소 분석기의 출력에서의 품사 분류 체계의 형태도, 어절 내의 형태소간의 세밀한 결합 제약관계를 표현하는 세분된 품사체계보다 더 단순한 분류가 품사 태거에서의 활용을 위해서는 더욱 적절할 것이다.

2.2 태그 세트 활용을 위한 부수적 고려 사항

2.2.1 태그세트 차이에 대한 고려: 태그 세트와 규칙과의 관계 설정

태그 세트에 어휘의 모든 정보가 온전히 담겨 있다고

생각하지는 않을 것이다. 대부분 태그 세트에서 다루지 않은 거시적인 또는 세부적인 언어현상들은 각 시스템의 (통계적) 문법 규칙이나 전자사전의 자질-값에서 반영한다는 것을 가정한다. 품사 태그 세트는 어휘들을 전산처리적으로 유용한 기준에 의하여 범주화(categorization)한 결과의 일종으로 볼 수 있다. 일반적으로 범주(category)의 설정은 각 범주의 구성요소들이 갖는 공통적 성질을 표현하는 규칙설정의 수반을 전제하고 있다.

규칙과 같은 일반화된 경향이나 성질(예를 들면, '조사'라는 범주는 '명사'라는 범주의 오른 쪽에 한 어절 내에서 인접하여 나타난다')의 표현을 도출하지 않을 바에는 범주화의 필요가 전혀 없다고 볼 수 있기 때문이다. 따라서, 품사 태그 세트의 설정은 일반적으로 같은 품사태그가 할당된 어휘들의 공통적 속성을 나타내는 접속 및 공기 제약 관계의 경향을 대략적으로 명시하는 문법규칙을 설정하는 것을 수반한다. 다만, 말뭉치 기반의 규칙의 도출은 기존의 규범적이고 선험적이며 직관적인 방법과 달리, 관찰적이고 경험적이며 귀납적인 방법으로 이루어진다는 것이다. 그리고, 이 규칙들의 확실성의 정도를 확률통계적 값으로 표현하거나 계산한다.

문법규칙은 지나친 일반화에 의해 많은 예외를 유발하여 정확도를 떨어뜨리게 하는 위험도 있겠지만, 품사 태그에 속하는 어휘부류들 간의 접속 및 공기제약의 일반적 특징을 포착하여, 시스템이 이전에 관찰하지 못했지만 규칙에 의해 예견할 수 있는 새로운 표현을 받아들일 수 있는 포용력을 갖게 해준다. 범주화의 부작용인 예외현상을 좀 더 자세히 포착하여 관찰 데이터와의 일치도를 높이기 위하여 보통 하위범주화를 시도한다. 즉, 명사를 하위범주화하여 '의존명사', '자립명사' 등으로 분류하는 것과 같은 것을 말한다.

그리고, 하위범주화의 표현은 여러가지로 가능하다. 품사 세분류 또는 자질-값 등으로 표현한다. 예를 들면, '마리'와 같은 하위 범주인 '의존명사'를 명사('n')의 하위분류로 'nb'와 같은 품사태그로 뭉뚱그려 표현할 수 있고, 자질-값의 형태로 표현하려면, '마리'라는 사전 표제어의 정보로써 품사는 '명사('n'), 품사세분류 자질-값으로 '의존('b')과 같은 형태로 나누어서 정보를 표현할 수 있으며 각각은 장단점을 가지고 있다.

그러나, 규칙에 부합하는 언어현상보다 예외현상이 너무 많아져서, 부합도를 높일 목적으로 규칙성이 없는 예외도 포섭하려고 규칙을 세분하다 보면, 엄청나게 복잡해진 규칙의 기술(description)에 드는 메모리의 추가 부담으로 인해 규칙 설정의 주된 의의인 표현의 경제성이 규칙화하지 않은 것보다 훨씬 상실되어 버린다. 그러므로, 이러한 경우에는 억지로 하위범주화하여 어휘범주들간의 관계로 일반화 시키므로써 표현상의 효율성과 관찰 데이타와의 부합도를 떨어뜨리지 말고, 어휘들간의 접속관계 그 자체가 축퇴된 규칙(degenerated rule)이 되도록 해야 한다.

이러한 경우의 예로써, 연어(collocation) 및 숙어적(idiomatic) 표현에 대한 처리에서 이들의 구성 요소 어휘를 범주화하지 않고 전체 표현을 한 단위로 등재하는 것과 같은 것이 있다. 그리고, 한국어 품사 '조사'는 여러가지 복합 패턴을 보여주는데, 초기에 복합조사의 패턴을 규칙화 해보려는 시도가 점점 많은 예외 현상의 발생으로 인하여, 규칙의 생산성이 떨어지게 되어 규칙화를 포기하고 복합조사 하나하나를 퇴화된 규칙으로서 그대로 사전에 등재하게 된 것을 예로 들 수 있다. 복합어에 대한 처리 문제도 이와 비슷한 변천을 보여 주고 있는 듯 하다.

이것은 나무는 보되 숲은 보지 못한다든지, 높이는 날되 자세히 보지 못하는 우를 범하지 않고 적절한 균형을 유지해야 한다는 것을 시사한다. 그래서, 관찰된 언어현상을 제대로 시스템의 처리 능력에 반영하기 위해서는, 말뭉치의 관찰 데이터의 분포를 잘 관찰하여 그 분포의 규칙성의 성질에 따라 적절하고 분별력 있게 표현해야 한다는 것이다.

2.2.2 태그 세트 차이에 대한 고려: 태그 세트와 사전과의 관계

예외가 많은 규칙에 대한 예외현상을 변별해주는 역할도 해주는 것이 사전의 각 어휘에 대한 자질-값이다. 자질-값도 어휘들에 대한 하위범주화 정보를 담고 있다. 이 자질-값을 갖는 모든 어휘들이 이 자질-값으로 이름지어지는 하위범주 태그의 집합을 이룬다고 할 수 있다(이현아,1996). 이러한 하위범주는 품사태그의 세분류로 표현될 수 있다. 즉, 태그세트의 세분류로 편성시키는 것과 사전의 자질-값으로 편성시키는 것에는 효과면에서 큰 차이가 있다. 일단, 태그 세트에 편성시킨 것은 이러한 하위범주가 가지는 접속 및 공기 제약적 정보를 태거의 학습에 활용하겠다는 적극적 의지가 담긴 것이라는 점이다.

예를 들면, 어떤 태그 세트에는 '명사'라는 대분류만 있고 자립명사와 의존명사의 하위범주를 태그 세트에 편성하지 않은 채, 사전의 자질-값 등으로만 편성하여 등재한 경우에는 명사의 자립성 여부에 따른 좌우 접속 및 공기 제약적 관계의 차이를 품사 부착 말뭉치의 통계적 품사천이 정보에 반영하지 않겠다는 의지가 강한 것으로 볼 수 있다.

어떤 시스템에서는 품사 태그를 5개 정도의 대분류만 설정하고 하위의 세분된 범주 정보는 사전의 각 어휘에 자질-값으로 등재하여 정보를 분담할 수도 있고, 표준안에서처럼 총 40여개의 품사 태그로 표현할 수도 있다. 표현은 다르지만 이들이 나타내는 어휘범주들과 그들의 속성은 같은 것을 가리킬 수 있으므로, 이들 간에는 매핑이 이루어질 수 있다고 본다. 따라서, 각 시스템이 채택하고 있는 태그 세트를 표준안과 비교할 때는 이러한 점을 잘 고려하여야 한다고 본다.

2.2.3 태그 세트 설정 및 태그 할당 작업시의 지침

ㄱ) **형태-통사적 분포상의 성질에 의한 분류 기준 유지**

용도가 형태-통사적 성질에 따른 품사 태깅이므로 태그 세트의 설정기준에 어휘의 의미적인 성질에 우선하여 품사를 분류하려는 경향을 경계해야 한다. 즉, 어휘의 접속 및 공기 제약적 분포의 성질이 우선되어 반영되어야지, 어떤 어휘가 어떤 의미적 부류에 속하기 때문에 어떤 품사에 속해야 한다는 성질이 우선되어 반영되어서는 안된다는 것이다.

일반적으로 형태-통사적으로 비슷한 성질을 갖는 범주의 어휘들은 의미적으로도 비슷한 성질을 갖는다는 개연성이 있지만, 항상 그렇지는 않기 때문에 품사 분류에 혼동이 올 경우에는 이러한 의미적 범주화를 개입시켜서는 안된다는 것이다.

ㄴ) **태그 할당의 일관성의 유지**

이것은 표준안을 적용하는 태깅과정에서(특히 수동 태깅에서) 자주 발생할 수 있는 오류를 주의해야 한다는 것이다.

예) "끌어오던"이란 어절을 어떤 문장에서는 "끌어오/pv+던/etm"으로, 어떤 문장에서는 "끌/pv+어/ec+오/px+던/etm"로 태깅

그러나 다의성을 갖는 어휘가 문맥에 따라 다른 통사적 용법이나 의미를 갖는다면, 다른 태깅을 허용해야 한다.

예 1) '산의 높이(높이/nc)를 측량한다'

예 2) '해가 높이(높이/mag) 떠올랐다'

일관성을 유지해야 하는 이유를 굳이 말하자면, 태깅이 틀렸더라도 일관성있게 틀리면 나중에 올바른 태깅으로 일괄 수정할 수 있는 장점이 있기 때문일 것이다.

일관성 오류의 회피 및 탐지, 복구를 위한 지원 도구가 일부 마련되어 있지만 이에 대한 자세한 언급은 생략하기로 한다.

ㄷ) **태깅의 기본 단위 규정**

현재 태깅의 기본 단위는 형태소로 보고 있다. 그리고, 현재의 기준안은 어절을 분석의 기본 단위로 삼고 있다. 의미적 최소 단위, 즉 형태

소를 분리해내는 것이 형태소 분석기 및 이와 관련된 태거의 주요 기능 중의 하나이다.

그런데, 다어절 한단위가 되는 형태소(복합어, 고유명사, 속담, 숙어 등)가 있다. 따라서 여러 어절에 걸쳐 하나의 의미 단위를 이루는 것에 대한 분석은 현재 고려하고 있지 않다.

그러나, 사전에 다어절 한단위 표제어 등록이 가능하고, 분석알고리즘이 필요에 따라 이 사전을 활용하여 여러 어절을 한 단위로 인식할 수 있는 시스템이 많이 개발되고 있다. 따라서, 말뭉치 태깅에서도 이러한 다어절 한 단위를 하나의 단위로 태깅할 수 있도록 지침을 변경할 필요가 있다고 본다.

ㄹ) **태깅 표현방식, 용인 가능한 문장/표현의 범위 규정**

태깅 표현에 있어서, 현재는, 어절에 대해서 하나의 정답, 즉 하나의 태깅 결과만 명시하게 되어 있다. 그러나, 이번 평가 대회에 참가한 각 팀의 표제어 선정 상의 기준에 있어서의 취향 차이로 인하여, 하나의 어절에 대해서 여러 방식의 태깅을 인정할 수 있는 메커니즘의 필요성이 대두되었다.

표준안에서 우선 대상으로 삼는 표현은 현재 한국어 표준어법에 맞는 것으로 하고 있다. 그러나, 표준어법을 약간 벗어난 범위로 확대하고자 할 때, 한국어에서 용인가능한 문장(즉, 정문)의 범위를 정하는 것은 여러 전문가들의 견해차에 의하여 쉽지 않다고 본다. 용인가능한 문장의 판정에 대해서는 국어학 전문가의 도움을 빌거나 응용목적별 참가자들의 다수결로 판정할 수 밖에 없을 것이다.

ㅁ) **적합하고 충분한 자료 바탕 언어현상별 특수성의 세밀한 반영**

말뭉치는 종류별로 많으면 많을수록 좋고, 많이 수집하여 관련 언어현상을 많이 보는 것이 좋지만, 현실적으로 수집과 관찰에 허용되는 노력과 시간에는 제약이 주어지고 언어는 시간에 따라 변화하기 때문에 가용자원-제약적(resource-limited) 환경을 고려한 점진적인 축적 방법을 채택할 수 밖에 없다. 즉, 현재까지 관찰된 현상만으로 일단 정리를 하여 잠정적 결론을 도출하여 언어지식을 구축하되, 나중에 반례(counter-example)가 생겼을 때 적절히 튜닝할 수 있는 적극적인 준비 태세를 갖추는 것을 의미한다.

따라서, 언어현상의 성질을 묘사할 때도 "절대, 항상, 전혀" 등과 같은 그러한 표현은 말뭉치 언어처리의 관점에서는 적절하지 않은 표현이다. "말뭉치에서 발견하지 못했고, 예문을 생각해내기도 쉽지 않다라는 표현을 써야 한다". 즉, 나중에 어떤 반례가 발견되거나 용법이 새로 생길 수 있는 가능성의 여지가 항상 남아 있기 때문이다.

말뭉치 구축에 소요되는 제약으로 인한 자료의 빈곤(data sparseness)문제를 무시할 수 없기 때문에 자료로부터 수집된 말뭉치에만 의존하여 언어현상을 분석할 수는 없다. 기존에 국어학이나 언어학에서 주로 활용하였듯이, 기억해내거나 새로 만들어낸 문장(negative example, positive example)도 말뭉치 후보에 포함시켜야 한다.

관찰된 다양한 언어현상으로부터 어휘들의 유사성 뿐 아니라, 세밀한 차이점도 포착하여 품사 태그 설정에 반영할 수 있어야 한다.

관찰 데이타와 이에 대한 현상을 기술(description)하는 언어지식(즉, 품사 태그 세트 및 수반되는 규칙, 사전 표제어, 자질-값 등)과의 부합도(fitness)를 높이기 위하여 어휘 범주를 필요한 만큼 적절히 세분하는 것이 필요

하다(세분된 어휘 범주의 표현은 앞에서 언급하였듯이, 자질-값 또는 품사태그 세트 등으로 표현되고 이들 범주의 성질은 규칙으로 표현된다).

또 한가지 중요한 것은 관찰된 데이터와의 부합도 뿐 아니라, 관찰된 데이터로부터 일반화된 규칙을 도출하여, 아직 관찰되지는 않았지만 잠재적으로 용인가능한 예문도 규칙에 의하여 커버될 수 있는 여지를 마련해야 한다는 것이다. 즉, 시스템이 관찰하거나 학습하지 않은 미지의 현상에 대해서도 어느 정도 대처 능력을 갖도록 해야 한다는 것이다.

범주화에 있어서 계층적 범주화를 도입하여 각 계층의 범위와 중요성의 순위를 반영하기 위하여, 범주의 표시도 계층적인 표현을 쓴다. 예를 들면, 명사는 'n', 의존명사는 'nb', 자립명사는 'nc'로 표현한다.

ㅂ) 전산처리의 효율성 고려

언어학적 또는 국어학적 입장에서도 의견이 일치되지 않거나, 전산처리적 관점에서 많은 비효율이 예상되는 경우에는 전산처리적 효율성 측면을 우선적으로 고려해야 한다. 이것은 태거나 태깅 뿐 아니라 이와 불가분의 관계에 있는 형태소 분석기의 품사분류나 전자사전, 어절 구성 규칙(접속 규칙)등의 결과를 반영해야 되기 때문에 태그 세트 그 자체만의 문제가 아니다(이현아, 1996).

예를 들면, 접사 및 복합명사의 처리 문제는 생산성과 데이터와의 부합도 문제, 즉 ㅁ)과 ㅂ)의 상충되는 조건을 적절히 만족시켜야 하는 부담을 초래한다. 접사를 분리하여 규칙으로 정하다 보면 너무 많은 예외가 발생되므로, 이들을 커버하려고 세부규칙을 자꾸 만들다 보면, 규칙이 너무 복잡해져서 규칙화하지 않은 것보다 비효율적일 수 있게 되기 때문이다. 이런 경우에는 접사나 복합명사를 분리하지 않고 한 단위로써 표제어로 등록하는 것이 더 효율적이다. 즉, 이것은 지나친 범주의 세분을 동반한 규칙화와 표현 복잡도와의 상충되는 조건의 균형을 맞추어야 한다는 것을 의미한다.

태그 세트를 정한다는 자체가 범주화를 전제하는 것이고, 범주화 자체가 갖는 고유의 성질을 알아야 하며, 범주의 용도를 생각해야 한다. 앞에서 언급하였다시피, 범주화의 용도는 규칙을 만들고자 함에 있다. 그렇지만, 규칙성이 없는 현상으로부터 억지로 범주화를 하는 과정에서 비효율성과 관찰 데이터와의 비일치성의 문제가 발생한다.

전산처리적 효율성은 주로 소요 메모리와 수행속도로 평가된다. 그리고, 부수적으로 사전이나 규칙의 확장/변경의 편리성, 가독성, 재활용성 등을 들 수 있지만 이런 요소들은 정량화하기가 힘들다. 그런데, 이들은 자주 상충적인 관계에 있을 수도 있고, 이들의 상대적 중요도는 응용 목적과 사용 환경 및 취향의 차이에 따라 달라질 수 있다. 즉, 플랫폼(platform)이 어디냐에 따라서 메모리나 수행속도의 중요도의 비중이 달라진다.

최근에 CPU나 메모리의 속도나 용량은 대폭 늘어나면서도 가격은 하락하므로써 이들은 별로 중요하지 않은 것처럼 보이지만 환경에 따라서 중요해질 수도 있다. 예를 들면, PDA와 같은 비교적 제한된 용량의 메모리를 쓰는 환경에 탑재하는 경우에는 메모리를 가능한 적게 쓰는 방안을 고려해야될 것이기 때문이다. 그리고, 수행속도에 민감한 셋톱박스 같은 실시간 응용의 경우에는 속도를 우선하기 위해 고성능 CPU나 메모리를 늘리는 방안을 채택할 수도 있다.

효율성에 관련된 세부요소로 다음과 같은 것을 포함시킬 수 있다.

a) 소요 메모리 양(memory storage): M
- •전자사전 표제어 수*M(품사태그+자질-값)
- •규칙수*M(품사태그와 관련된 규칙(R))

b) 수행 속도(computation speed)
- •사전표제어 액세스 평균시간*평균 표제어 수(어절당)
- •사전 정보 해독 시간(엔코딩된 표현을 쓰는 경우)
- •해당규칙 평균탐색시간*평균관련규칙수
- •규칙 적용/해석(Interpretation) 시간

c) 추가/변경/튜닝/재사용 등에 드는 평균시간
- •사전표제어 및 정보, 규칙 평균 작성 및 컴파일 시간
- •오류의 발견 및 이의 진단 및 교정에 소요되는 평균시간(가독성(readability)의 측면도 포함됨)
- •타 응용을 위한 형태의 변경 및 재사용에 소요되는 평균시간

c)와 같은 요소들은 정확하게 측정하기는 곤란하지만, 이러한 측면도 태그 세트와 이와 관련된 전자사전, 규칙의 표현 및 분석기와 시스템 전체의 성능, 효율성, 진취성과도 관련 있다. 따라서, 이러한 요소들에 대한 비중 배분을 위한 세팅(setting)은 각 연구자들의 응용 목적마다 다를 수 있다. 이러한 부분의 고려에 따른 취향과 용도의 차이로 인한 태그세트나 표제어 선정방법이 달라질 수 있는 경우에는 모두를 인정할 수 있는 방안을 찾아야 한다고 본다.

그러나 표준안을 제안하거나 평가에 참여하는 팀은 각자의 세팅이 어떠한지를 언급하고, 그 세팅하에서 태그 세트나 관련된 분석/태깅방식이 부합하는지를 합리적으로 설명하여야 주장에 설득력이 있을 것이다.

ㅅ)다른 레벨의 언어처리와의 상호부합성(coherence)의 유지

형태론적으로 어절단위만을 분석과 분류의 대상으로 한다면(즉, 형태소 분석기에서의 접속 체크 대상의 범위를 어절 내로만 한정한다면), 의미적 한단위로 볼 수 있는 고유명사(예, '참을 수 없는 존재의 가벼움'), 다어절 복합명사(고유명사) 등과 숙어에 대한 처리는 구문/의미분석기에서 형태소 분석기가 출력한 어절 단위 분석 결과를 바탕으로 한단위의 의미로 분석해내어야 한다.

이러한 한단위들은 대부분 사전에 한단위로 등록되어 있어야 하기 때문에 형태소 분석/태깅 단계에서 하나의 단위로 인식하는 것을 표준안에 반영하도록 해야 된다고 본다.

현재의 표준안에서 보는 태깅 방침에서 전제하는 형태소분석기는 얼밀히 말하면 어절단위 분석기이다. 그런데, 형태소라는 것이 '의미적 불가분의 최소단위'이고, 다어절 한단위 의미의 어휘도 하나의 형태소이므로 형태소 분석/태깅이 되어야 할 필요는 없다고 본다.

그리고, 형태소 분석/태깅 이후에 구문/의미분석 등의 처리를 고려한다면 구문/의미분석 단계에서의 처리 방침과 손발이 맞는 처리를 하도록 하기 위해, 구문/의미분석 단계에서 더 효율적이고 정확하게 처리할 수 있는 것은 구문/의미분석 단계의 몫으로 넘겨야 한다.

주의해야할 것은 의미적으로 전성된 다어절 복합 어절의 의미를 제대로 파악 또는 정의할 수 있어야 표제어 선정이 명확해진다는 것이다. 그런데, 이 의미를 파악하고 정의하는 것은 자의적인 요소가 많이 포함되므로 원칙이 있을 수 없다고 인식하고 있다.

그러나, 이러한 부분에도 자의적인 요소를 줄이고, 객관적인 방법으로 선호도의 기준을 마련할 수 있는 여지가 있다고 본다.

그리고, 신중히 검토해보아야 할 중요한 사항 중의 나머지 하나는 응용 목적이 어디에 있는냐, 즉 기계번역, 정보검색, 퇴고, 분류, 요약, 질의-응답 등 중의 어떤 것이냐에 따라서 태그세트나 태깅방식 등이 크게 달라질 수 밖에 없는가 하는 점이다. 이와 관련된 문제를 부분적으로 정보검색을 위한 색인용 명사 추출기가 통상의 형태소분석기나 품사태거와 많이 달라져야만 하는가에 대한 논의로 국한시켜서 4.3.3 의 명사추출기 평가에서 다루기로 한다.

3. 표준안 도출 과정상의 공정하고 합리적인 절차의 명확한 수립

Test Suites를 구축하기 위해서는 여러 참가팀들에 대해 중립적이면서 합리적인 이유와 기준에 입각한 표준적 품사 태그 세트의 설정이 중요하다. 그런데, 다량의 말뭉치를 바탕으로 표준안을 도출하는 과정에서 다음과 같은 중요한 문제점이 자주 봉착되어서, 이에 대한 체계적인 접근 방법을 갖추는 것이 필요하다고 생각된다. Test Suites 및 학습용 말뭉치 구축에 관한 표준안에 관련된 문제는 4가지로 들 수 있다.

1) **적절한 태그 세트의 결정 문제**
   임의의 품사 태그의 새로운 추가 설정 및 기존 태그의 삭제 여부의 결정 문제
2) **품사태그의 명확한 분포적 성질 또는 정의의 설정 문제**
   이것은 어떤 임의의 어휘에 대하여 어떤 품사 태그를 부착하는 것이 가장 적합한 것인가를 판정하게 하는 품사 태그별로 정의된 할당 기준을 명료하게 설정하는 문제이다.
3) **표제어의 선정방안에 대한 대안들의 평가 방법 설정의 문제**
   어절에서 사전에 등재되는 기본 단위들을 분할해내는 작업을 말한다.
4) **이형태, 준말, 방언, 원문 철자 및 맞춤법 오류 처리에 대한 문제**
   - 원형 복원과 원문 코퍼스의 원어절 오류에 대한 입장과 처리 방법에 대한 문제이다.

이들 문제는 말뭉치 태깅 과정에서 자주 대두되는 문제로써, 평가에도 많은 영향을 미치는 요소인데도 아직 합리적이고 누구나 납득할 만한 객관적인 방법이 제시되지 않고 있으며, 현재는 다수가 동의하거나 이견이 분분할 때는 평가기관 자체의 취향을 택하고 있을 뿐이다. 그러나 이러한 문제가 단시간에 완벽하게 해결될 것으로 기대하는 것은 무리일 것이다. 다만, 여기서는 이들에 관련된 문제를 실례를 통해서 제기하고, 이들에 대한 해결방안을 부분적으로 제시하고자 한다.

ㄱ) 임의의 어휘범주(즉, 품사) 태그의 새로운 설정 여부의 결정 문제

여기서 '품사'라는 용어 대신에 어휘범주라는 포괄적인 용어를 사용한 것은, 국어학적인 관점에서 분류한 기준을 참고로 하여 전산처리적 관점도 반영하여 어휘를 분류하겠다는 의지를 담은 것이다. 즉, 자연언어 처리에서는 처리의 효율성 측면에서 형태소 또는 단어가 아닌 복합단위 등에 대해서도 범주를 부여하기도 하기 때문에, 품사보다 광의의 뜻을 갖는 어휘범주라는 용어를 사용하고자 한다.

예1) "튼튼", "깨끗"과 같은 어휘는 "몸도 튼튼(깨끗) 마음도 튼튼(깨끗)", "튼튼도(은) 한다", "튼튼(깨끗)하다" 등이 가능한데, 일반 형용사와는 다른 성질을 갖는다고 볼 수 있다. 이러한 경우에, '튼튼', '깨끗' 등으로만 표제어를 선정할 것인지., 또는 '튼튼하', '깨끗하'등의 표제어도 인정할 것인지, 전자의 경우에 새로운 범주를 설정해주는 것이 좋은지 아니면, 기존 범주 중에 가장 가까운 쪽(예를 들면, "용언 겸용 명사")에 편성시키고 자질-값 또는 하위범주

정보를 주는 것이 좋은 지를 어디에 근거하여 결정할 것인가 하는 문제가 있다.

예2)표준안 품사태그에 '수사'는 있는데, '수관형사'는 없다. 자립성이 없어보이는 "한두", "서너" 등의 어휘는 수사로 분류되는 다른 어휘들(예, "일", "다섯", "스물")과 유사한 성질도 갖지만, 형태론적으로 중요한 다른 성질도 갖는다. "한두", "서너" 등은 "한두 개", "서너 명" 등과 같이 단위성 (의존)명사를 수반하는 경향이 강하지만 일반 명사를 수반하는 경우도 있다. 이러한 경우에 수관형사 품사태그는 전혀 필요 없는가 하는 문제가 있다.

예3) 소위 '단위성 의존명사'라는 어휘범주도 표준안 품사 태그에 포함되지 않고 있다. '마리, 개, 필, 대, .." 등은 '한, 두, 네, 열,.."등의 어휘 등이 좌측에 인접하거나, 오른 쪽에 '당'이라는 어휘가 인접하여 '개당, 두당, 마리당..." 등으로 표현되어야 하는 의존성이 매우 강하다. 즉, 이 어휘들은 좌측이든 우측이든 위와 같은 부류의 다른 어휘들에 의해 인접하지 않고 독립적으로 표현하는 경우를 찾기가 매우 힘들다. 그리고, "'대'는 단위명사이다."와 같은 예문에서 따옴표가 접속된 것도 이것의 의존적 특성의 하나라고 볼 수 있다. 이러한 경우에 '단위성 의존명사'라는 어휘범주의 품사태그는 필요 없는가 하는 문제이다.

ㄴ) 어휘 범주의 할당기준 설정의 문제

어떤 어휘 범주이든지 간에 그 범주의 어휘들 특성을 다른 범주의 어휘들 특성과 대별하여 잘 설정해두어야, 어떤 임의의 어휘에 대해서도 애매하지 않게 올바른 어휘범주를 할당할 수 있다. 따라서, 어휘범주 할당기준의 설정은 품사부착 말뭉치와 이와 밀접한 관련을 갖는 전자사전의 품질을 좌우하는 중요한 문제이다.

예1) "묵다"라는 어휘에 관련된 "묵다, 해묵다, 케케묵다"를 일반 시중 사전에서는 모두 동사로 등록하고 있다. 하지만, 형용사 쪽이 더 가깝다는 의견이 있을 수 있다. 그렇다면, 형태론적으로 동사, 형용사라는 어휘범주에 대한 할당기준은 무엇인가? 교과서에 나오는 동사, 형용사에 대한 정의는 이를 부정하는 반례를 쉽게 만들어 낼 수 있을 것이기 때문에 할당 기준으로서는 불완전하다고 볼 수 있다. 그러면, 도대체 이 할당 기준이라는 것은 어떻게 도출되어야 하는 것인가?

예2)

(a)연구원이라는 직업은 훌륭한 직업이다.

(b)자동차라는 문명의 이기는 인류에게는 필요악 중의 하나이다.

(c)'바람과 함께 사라지다'라는 작품은 마가렛 미첼이 지은 소설이다.

(d)지금 당장 돈을 내라는 말입니까?

(a), (b), (c)의 경우에는 '격조사'로 보고, (d)의 경우에는 '어미'로 흔히 분류한다. 그런데, 다음의 예문에서 '라는'은 '어미'로 분류하는 것이 애매한 경우이다.

(e) 이 마크는 취사는 금지라는 표시이다.

이 경우 "라는"이 조사인지 어미인지 구분이 안되는데, 조사와 어미에 대하여 통상적으로 우리가 알고 있는 할당 기준에 정확하게 부합하는 지가 확실하지 않다.

ㄷ)표제어의 선정(즉, 어절에서 사전에 등재되는 기본 단위들을 분리해내는 작업)방안에 대한 대안들의 평가방법 설정의 문제

- 말뭉치 내의 어절들에 대해 어떤 단위들로 분리하여 어휘범주 태깅을 할 것인가 다양한 대안들이 제시되고 있으나, 이들 각각의 대안들에 대한 우월성을 평가할 공통적으로 적용가능한 기준이나 방법론이 체계적으로

만들어지지 않아 일관성과 객관적 타당성에 문제가 있다.

예) "먹음직하다"에 대해서 "먹+음+직+하+다", "먹+음직+하+다", "먹+음+직하+다"

위와 같이 태깅하는 대안들이 제시되고 있는데, 이것은 "음직"이나 "직하", "음직하"를 하나의 사전표제어로 등록할 것인가, 말 것인가를 결정하는 중요한 사안이기 때문에 신중하게 결정해야 하는 문제이지만 아직 뚜렷하게 어떤 대안이 우월하다고 결정할 기준이나 방법론이 마련되어 있지 않다. 실제로 시중의 사전에도, 이들에 대한 표제어 등록 상태가 가지각색이어서 혼동을 초래한다.

ㄹ) 이형태, 준말, 방언, 원문 철자 및 맞춤법 오류 처리에 대한 문제

이형태는 표층 형태 그대로를 모두 인정하고 있다. 준말은 대부분 복원을 원칙으로 하고, 많은 예외 조항을 두고 있다. 방언도 대부분 원문 그대로 태깅하고, 철자 및 맞춤법 오류는 교정 후에 태깅한다. 그런데, 이러한 기본 원칙과 예외를 두는 것에 대해 더 근본적인 이유가 명시되지 않고 있어서 배경 원리를 이해 못한 태깅 작업자가 잦은 실수를 유발할 수 있다. 그리고, 이러한 배경 원리의 합리성 관점에서 이 기준안의 기본 방침의 재고가 필요하다고 본다.

이상과 같은 문제에 대해 차례대로 접근방법을 생각해보기로 한다. 일반적으로 태깅을 하기 이전에 우선 어절의 분할(segmenting)을 어떻게 하는 것이 바람직한가를 결정하는 ㄷ)의 표제어 선정기준을 결정하는 것이 우선적으로 수행되어야 할 작업이다. 이러한 결정을 바탕으로 그 이후에 각 분할단위(segment)에 대한 품사태그 할당을 어떻게 할 것인가를 결정하는 ㄱ)과 ㄴ)의 문제를 해결해야 한다.

ㄷ)에 대해서는 의미의 최소단위 또는 한단위를 분리하는 것과 효율적인 처리를 추구하는 관점에서 접근하고자 한다.

일단, 한 어절이나 어휘의 구성문자열을 두개 이상의 단위로 분할하여 태깅할 수 있으려면, 적어도 다음 두가지의 조건을 만족 한다.

1) 보통 간단히, '분리되었을 때의 의미의 합이 전체 하나로서의 의미가 같다, 즉 전성되지 않았다'라고 말한다. 엄밀하게 말하면, '분할된 어휘들의 의미적 관계를 의미분석단계에서 함수적으로 또는 일정한 알고리즘에 의하여 복원할 수 있어서, 원래의 분할되지 않았을 때의 의미와 같게 할 수 있다'라고 할 수 있다.
2) 분리된 어휘들 각각이 같은 의미를 가지고 독립적으로 사용될 수 있거나, 같은 의미를 가지고 다른 어휘들과 접속가능하다.

그렇지만, 위와 같은 조건을 만족하더라도 다음과 같은 경우에는 분리하지 않고 하나의 단위로 두는 것이 더 효율적이거나 안전하다고 생각한다.

1) 결합관계의 규칙성과 생산성이 적은 경우, 즉, 분리할 수도 있지만, 분리하여 결합규칙을 제공하는 것이 분리하지 않고 결합된 형태로 처리하는 것보다 비효율적일 경우(각 시스템의 세팅 하에서)
2) 분리된 어휘들의 의미 관계를 하나로 결합된 의미와 동치 관계로 복원하는 알고리즘이 복잡하거나 불확실한 경우

ㄱ), ㄴ)에 관련된 문제는 현상의 범주화와 관련된 문제로 보고 접근 방향을 제시하고자 한다.

ㄴ)과 같은 문제에 대한 접근 방법은 다음과 같이 정할 수 있다고 본다.

형태소 분석기나 태깅에 유효하고 적합한 할당기준은 범주의 외연적 기준(extensional criteria)이다. 외연적 기준은 "어휘의 접속관계에 의한 분포적 특성"을

따르는 것이다. 이러한 접속관계는 왼쪽과 오른쪽의 접속상의 분포를 살펴보아야 한다. 이들을 편의상 어휘범주의 L-Dist와 R-Dist라고 하자. L-Dist나 R-Dist는 그 어휘범주와의 공기정보(Mutual Information, Manning, 1999) 값이 어떤 임계치를 넘는 그 어휘범주 좌우에 인접한 어휘들만 포함된 것이다. 이들에 대한 예와 사용법이 예를 통하여 나중에 설명된다.

보통 어휘의 의미적 특성을 나타내는 내포적 기준(intensional criteria)은 참고로만 활용해야 한다. 한 범주의 내포적 기준은 그 범주의 원소 어휘들에 공통적이면서 다른 범주의 성질들과 두드러지게 차이가 나는 속성들을 기술한 것이며, 외연적 기준에서 묘사하는 속성은 제외한 것이다.
예를 들어, '동사'라는 범주의 외연적 기준과 내포적 기준의 일부를 다음과 같이 예시할 수 있다.

a) 외연적 기준
L-Dist = {매우,훨씬,점점,.... },
R-Dist = {ㄴ,ㄹ,는,고,며,지,고,게,었,더...}
b) 내포적 기준
사물의 동작이나 행위를 나타내는 어휘 집합

형용사에 대해서도 말뭉치를 바탕으로 위와 같은 것을 작성할 수 있을 것이다. 이러한 경우에 그 어휘가 보여주는 현상과 가장 흡사하게 매치되는 외연적 기준을 갖는 품사 태그를 할당하는 것일 것이다. 그래서, 여기서 중요한 것은 어떤 품사의 외연적 기준을 다른 품사의 기준과 명확하게 구분하여 애매하지 않게 표현하느냐 하는 것이다. 임의의 어휘를 동사 또는 형용사 둘 중에 어떤 것을 할당해야될 지는 그 어휘의 L-Dist와 R-Dist가 어떤 범주의 L-Dist와 R-Dist와 차이가 가장 적은 지를 판정하여 결정해야 할 것이다.

첫번째 문제인 ㄱ)에 대해서는 여러가지 접근 방법이 있을 수 있다. 첫째는 주어진 코퍼스로부터 통계적 기법(Schuze97, Manning99)을 적용하여 처음부터 완전히 새로운 태그 세트를 도출하는 방법이 있다. 그러나, 이러한 방법이 대규모의 데이터에 대해서 신뢰성 있는 결과를 도출할 수 있을지가 의문이고, 무엇보다도 이미 어느 정도 인정된 태그 세트를 이용할 필요도 있다는 점에서 현재로서는 바람직한 방법이 아닌 것 같다.

두번째 방법은 현재 정의된 태그 세트를 기반으로 이것을 예외 현상이나 오류의 발견을 단서로 점차 개선해나가자는 것이다. 즉 , 이것은 새로 발견된 특이현상 또는 예외 현상을 포섭하기 위해 언어정보를 수정하는 방법으로 주어진 세팅(setting) 하에서, 앞에서 언급한 관찰 데이타와의 부합도(fitness to observed data)와 효율성의 척도를 가지고, 여러 가지 대안 중에 어떤 것을 선택하는 것이 바람직한 지를 결정하는 문제이다. 여기서 대안은 다음과 같은 두가지가 있을 수 있다.

1) 기존의 가장 유사한 범주에 편성시킨다
2) 완전히 새로운 하나의 품사태그를 태그 세트에 추가한다.

이 각각의 대안에 따르면서 부합도와 효율성의 평가척도를 적절히 만족시키려면, 부가적인 작업이 필요하다. 그 각각에 대응하는 작업은 다음과 같다.

1) 차이점에 해당하는 성질은 자질-값이나 하위범주 품사태그로 설정한다. 새로운 자질값이나 하위범주 품사태그를 같은 범주로 발견된 모든 어휘의 사전정보로 등록해야 한다. 하위범주간 또는 자질-값 간의 접속 및 공기제약 규칙을 정의 또는 확률 통계적으로 도출한다.
2) 새로 설정된 품사태그를 같은 범주로 판별된 모든 어휘의 전자사전 정보로 등록해야 한다. 이 새로운 품사태그와 다른 품사 태그와의 새로운 접속 및 공기관계 규칙을 정의하거나 또는 확률 통계적으로 도출한다.

이러한 각각의 대안에 대한 우월성 평가는 2장에서 정의한 주어진 언어현상 데이터에 대한 부합도와 효율성의 척도에 기반하여 평가해야 한다. 그런데, 이 우월성은 세팅을 어떻게 정하였느냐에 따라 가변적이며

나중에 발견될 지도 모를 다량의 반례의 현상들에 의해서 뒤집어질 수도 있으므로 잠정적이라는 점이다. 그러나, 이러한 가변적인 과정을 겪으면서 언어지식은 점차 안정화될 것이라고 추정한다.

여기서 가변적이고 잠정적인 것은 예를 들면, 관찰 데이타와 세팅일 뿐이고, 대안의 결정 방법과 절차는 명확하게 정의할 수 있다. 이러한 방법이나 절차가 절대적인 것은 아니지만, 표준화에서 추가해야 할 활동은 여러 가지 제시된 태그 세트들을 다수결에 의해서 결정하기 이전에 이와 같은 대안 선택의 프로세스를 투명하게 정하는 것이라고 본다.

흔히 말하는 지정사 '이', 즉 '사람+이+다'라고 할 때의 '이'와 표준안 품사 태그인 '형용사 파생접미사'로 분류한 '하'의 성질을 비교하기 위해 형태론적 분포를 살펴볼 때,'이'는 '하'와 유사성을 많이 보이고 있다.

그러나, 유사성도 있지만 차이점도 분명히 있다.

1) '이'
   L-Dist = {사람, 사상.... 'ㄴ', ... 것,바,... 공부, 장만, ... 가난, 청결,.....}
   R-Dist = {ㄴ, ㄹ,... 고, 며,... 시,겠,었,더,...라는, 다, 다는....}
2) '하'
   L-Dist = {공부, 장만,..... 가난,청결,간단,총명...}
   R-Dist = {ㄴ, ㄹ,.... 고,며,....시, 였,더,겠,... 라는, 다,다는,... }

'이 두개를 하나의 범주로 설정하여 유사성을 반영하고, 이 범주의 각기 다른 하위범주나 각기 다른 자질-값으로 차이점을 반영하자는 대안'과 기존의 '지정사', '형용사 파생 접미사' 등의 범주를 고수하자는 대안 간의 옳고 그름이나 우월성을 어떻게 어디에 근거하여 판단하는 것이 좋은가 하는 문제가 제기될 수 있다. 앞에서 예를 든 '튼튼, 깨끗, 잠잠..'등에 대한 품사 태그 설정 문제도 같은 방식으로 따져볼 수 있을 것이다.

국어문법에서 또는 사전에서 그렇게 되어 있기 때문에 따라야 한다는 주장만 한다면, 우리가 굳이 말뭉치를 이용해서 용례를 찾고 품사를 태그를 다시 재조정하는 작업의 의미를 부정하는 것이 될 것이다. 기존의 판단에도 나름대로 이유가 있을 것인데, 앞에서 예를 든 것과 같은 분석적 방법으로 재조명 또는 재확인해 보자는 것일 뿐이다.

그리고, 위에서 언급한 차이점을 반영하는 방식에서 다음과 같은 선택의 문제도 고려해볼 필요가 있을 것이다.

1) 하위범주 태그로 표현하여 말뭉치 태깅에 하위범주화를 반영시켜 통계적 정보추출에 직접적으로 영향을 미치게 하는 것이 좋은가?
2) 아니면, 사전의 자질-값으로만 표현하여 말뭉치에는 태깅이 되지 않게 하여 통계적 정보 추출에 직접 영향을 주지 않는 것이 좋은가?

그리고, 이러한 선택은 어디에 근거해서 어떤 방법으로 이루어져야 하는지도 단순하게 생각할 문제는 아닌 것으로 보이므로, 이 문제에 대해서도 부분적인 접근 방법을 제안하고자 한다.

한 어절에 대한 여러가지 형태소 분석의 결과의 애매성 해결, 즉 태깅에 도움이 되는 정도의 레벨까지만 품사 태깅용 품사 태그로 설정하는 것이 좋다고 본다. 이러한 생각을 비교적 극단적인 두가지 예를 통하여 설명하고자 한다.

형태소 분석에서 인접 형태소 간의 음운적 결합 제약을 걸러내기 위하여 세분한 '유종성 명사', '양성고음 용언어간'과 같은 범주는 이미 형태소 분석 단계에서 충분히 활용하여 형태소를 분리해내는 것으로 효용가치를 다했으며, 그 이후에 품사간 애매성을 해소하는 태깅의 단계에서는 전혀 도움을 주지 못하므로, 이러한 세분된 범주는 당연히 품사 태그로 설정할 이유가 없을 것이다. 반면에, '한두', '서너'와 같은 '수관형사' 같은 하위 범주는 형태소 분석 단계에서는 세분되어서 어절

내의 접속 체크 등에 쓰이는 효용성은 없다. 그러나, 품사 태깅을 위해서 오른 쪽에 '명,개' 등의 '단위성 의존명사'가 올 확률이 높다는 공기 제약 관계를 활용하는 데에 도움이 되므로 품사태그로 인정이 되는 것이 좋다고 본다.

르)에 대한 수정된 개선방안과 그에 대한 이유를 다음과 같이 제시한다.

이형태는 표층 형태와 기준 형태를 같이 표시한다. 이유를 말하자면, 이형태는 단지 음운적인 결합의 제약으로 다른 대안적인 표현 형태를 제시할 수 없어서 불가피하게 다른 형태로 표현되는 차이가 있을 뿐, 구문, 의미 등의 기능상 전혀 차이를 발견할 수 없는 경우에 사용되기 때문이다.

그렇다면, 사전 표제어로는 이형태가 모두 등재되지만, 형태소 레벨의 정보는 차이가 있어도 구문/의미 레벨 이상의 정보는 차이가 없으므로, 구문/의미 레벨의 사전 정보를 공유하는 것이 언어정보의 일관성과 메모리 공간의 절약을 위해서도 정규형(canonical form)으로 변환하는 것이 더 좋다고 보기 때문이다. 그러나 입력어절 그대로도 표시해주는 것이 원어절 복원 및 원어절과의 대응 관계의 자료를 추출하기 위해서도 좋다고 본다.

예)'사람+이', '철수+가'에서 '이','가' 대신에 '사람+가', '철수+이'와 같이 대치할 수 없고 음운 결합적인 제약에서만 불가피하게 발생한 대안적인 어휘로 볼 수 있다.

→ 사람/nc+이{가/jx}/jx

준말 및 축약은 본디말과 원형을 같이 태깅하는 방안이 필요하다고 본다. 준말은 본디말에 대해서 선택적 대안으로 사용할 수 있는 표현으로서 이반적으로 구어적 표현이나 간결체 등의 문체에 사용하므로, 본디말과는 달리 필자의 태도(attitude)가 다르게 담긴 것으로 볼 수 있으므로, 본디말로 복원을 하더라도 준말로부터 복원되었음을 표기하여 원래부터 본디말로 쓴 표현과 차별을 둘 필요가 있다고 본다. 그리고, 준말과 본디말과의 대응관계 등을 이용하여 자동적으로 사전을 구축하는데에도 활용할 수 있게 준말과 본디말로 복원된 형태를 같이 태깅하도록 한다. 이렇게 하여 놓고, 평가시에는 각 팀이 이 중에 하나의 방식으로 하여도 맞다고 채점해 주는 데에도 사용할 수 있다.

예) 회복케 회복+케(하/xsv + 게/ec)/xsvc

방언도 준말에 대한 처리와 비슷한 방법으로 처리하는 방안이 좋지 않을까 한다. 원시 말뭉치의 철자 오류나 맞춤법 오류등에 대해서도 현재 오류를 교정한 후에 태깅을 하는 방식으로 하는데, 퇴고시스템 등에의 응용을 위해서는 준말처리와 같은 방식으로 처리를 하는 것도 고려해 볼 수 있는 방안이라고 생각한다.

4. 평가의 공정성과 형평성에 관련된 문제

4.1 대회 참가자들의 의견 제시의 중요성

대회 참가팀에게는 대회 개최 이전에 표준안 지침서와 이 지침서의 내용이 반영된 태그부착 말뭉치를 배포해준다. 이렇게 배포하는 지침서나 태그부착 말뭉치는 절대 준수해야만 하는 요지부동의 규정이 아니라 평가를 위한 초안으로서, 참가팀 누구나 각 시스템의 특장점이 잘 부각되고 정당하게 인정받을 수 있도록 설득력 있는 이의나 개선된 룰을 제안할 수 있는 기회를 제공하여 공정하고 합리적인 평가체제로 다듬어 나가기 위한 논의를 유발시키고자 하는 것이다.

평가 대회 개최 이후에 제기된 또 다른 문제는 평가방식의 개선이다. 평가 개최 이전에 평가주최측은 평가에 참여하기 위해서만 필요한 부가적 노력을 최대한 줄여주는 노력을 해야 했다. 그런데, 요구사항이나 불만사항을 사전에 많이 제기하였으면 완전하지는 않더라도 어느 정도는 경감시킬 수 있는 방안이 강구될 수 있었을 것이다. 즉, 평가방식 뿐 아니라 참가팀의 적극적 의견 개진에 있어서도 개선의 여지가 있었다고 본다.

평가대회에 참가하는 팀들은 평가 전에 태깅 지침 표준안(ETRI,1999)을 충분히 검토하여 숙지를 하고, 이의가 있으면 대안을 제시하는 절차를 충분히 거쳐 평가가 자기 팀에 불리하지 않게 미리 노력을 해야 한다고 본다. 그러나, 태깅 지침서의 분량이 비교적 많은 편이어서 대부분 읽어보지 않는 것 같다. 그래서 대부분 일단 평가가 행해지고 난 후에 오류에 대한 불만과 지적이 많이 발생한다.

Test Suites나 학습용 말뭉치의 분포에 있어서, 자기 팀 시스템의 특징을 잘 보여주는 문장이나 언어현상이 빠져있는 것에 불만이 있을 수 있다. 이러한 경우에는 자신의 시스템의 특징을 두드러지게 보여줄 수 있는 말뭉치나 예문을 표준안에 맞추어 태깅한 결과를 평가주최측에 제시하여 Test Suites에 포함될 수 있는 기회를 놓치지 않아야 한다. 이 새로운 예문이나 말뭉치에 표준안과 다른 태그나 표제어 선정방식이 있으면, 이에 대한 정당화를 표준안 대안의 형식으로 아울러 제시하여야 한다.

그리고, 평가주최측에서 배포한 학습용 말뭉치를 자신의 시스템의 결과와 잘 비교하는 것이 필요하고, 비교하여 차이가 나는 것을 자세히 분석하는 것이 필요하다. 그리고, 여기서 자신의 시스템의 오류는 개선의 자료로 삼고 표준안이나 학습용 말뭉치에 오류가 있는 경우에는 개선안이나 수정 건의안, 또는 오류 리포트를 제출하는 것이 바람직하다.

표준안에 불만이 있을 때는 표준안에 대한 의견을 다음 항목에 대하여 논하고 제출하는 것이 바람직하다고 본다.

ㄱ) 문제제기의 대상 표준안 명시
ㄴ) 개선 또는 교정된 표준안 제시
ㄷ) 근거 자료 및 각 시스템의 주요 세팅의 제시
예문과 예문의 표준안에 따른 태깅 결과 제시
(분석 결과의 태깅 및 세그멘팅은 표준안을 따르되, 표준안을 따를 수 없으면 그 이유를 설명하고 새로운 기준안으로 제시해야 함)

4.2. 표준안 대안들의 이견 조정 및 선택과정

표준안에 대해서 이견이 있어서 다양한 대안이 제시되는 경우에는 이들에 대한 적절한 판단과 조치가 필요하다고 본다. 그리고, 이러한 대안들의 성격은 다음과 같이 분류될 수 있다.

1) 합리적이고 설득력 있는 이유가 제공 되는 것
   - 자신이 설정한 세팅 하에서 부합도와 효율성이 있음을 입증하는 자료가 제시되는 경우이며
   - 이 경우에 옳고 그름이나 우월을 가릴 수 있는 경우가 있고, 그렇지 않고 취향의 차이를 인정하지 않을 수 없는 경우가 있다.
2) 언어적 직관, 표준어법, 사전 등에서 보통 채택하고 방식을 따른 것
   - 합리적이고 설득력 있는 이유가 제공되지 않고 단지 어떤 관습이나 직관에 따라 제시되는 경우에는, 대부분의 사람들이 동의하지 않으면 채택되기 힘든 것이다.

대안들의 내용이 의견상 분분하고, 단시간에 우열을 가리기 위한 검증을 하기 위해서 관련 용례를 많이 확보해야 되며, 많은 시간의 실험을 해보아야 되는 경우에는 여러 가지 대안을 잠정적으로 모두 수용할 수 있도록 할 수는 있겠지만 이것은 쉽지 않을 것이다. 왜냐하면, 태그부착 말뭉치를 구축할 때, 한 어절에 대하여 여러 가지 방식으로 태깅을 한 결과를 만들어야 하므로 시간과 노력이 많이 소요되기 때문이다. 그래서, 이것이 불가능하면 다수결로 하나를 결정할 수 밖에 없다. 채택되지 않은 대안들을 다음번 표준안 개선에 반영시키고자 하면, 우월성을 뒷바침하는 자료를 그동안 마련해야 할 것이다.

합리적인 이유가 있고 대안의 우열을 가리기 힘든 경우에는 취향에 따라서 다른 면을 평가에서는 인정할 수 있도록 해야할 것이다.

4.3 문제출제 범위 및 평가방식 상에서의 개선점

4.3.1 태거 평가

Test Suite에는 한두개의 문제도 아니고 몇 만개 이상의 문제와 정답을 대부분 수작업으로 작성하기 때문에 오류가 많이 발생할 수 있다. 평가주최측에서는 오류의 최소화를 위해 최대한 노력해야겠지만 오류를 완전히 없앨 수는 없다.

그래서, 평가결과 즉, Test Suite에 기준하여 각 팀의 답안지를 채점한 것을 각 팀은 돌려 받아서 이러한 Test Suite에 있는 오류를 검출해내고 자신의 관점에서 자신의 답안지의 채점을 다시 할 수 있는 시간을 충분히 주고 자체평가 결과를 Report할 수 있도록 하는 것도 필요하다는 참가팀들의 주장은 타당하다고 본다.

그런데, 태그부착 말뭉치를 검토하고 태깅에 오류가 있다고 이의제기를 하는 것의 유형을 자세히 따져보면 다음과 같이 세분할 수 있다.

a) 태깅 지침 표준안에 어긋나거나 태깅에 일관성이 없어서 말뭉치를 수정해야 되는 것이다.
b) 해당 표준안의 불합리성을 예증할 수 있는 예문(counter-example)을 들 수 있어서 표준안을 수정해야될 오류이다.
c) 취향이나 세팅에 따라서 달라질 수 있는 사안이기 때문에, 표준안과는 다르지만 자신들이 견지하고 있는 방식도 충분히 가능하기 때문에 자신들의 방식을 표준안에 추가해야 될 사항이다.
d) 앞에서 언급했던 대안의 객관적 선택 기준인 효율성이나 부합도 등의 측면에서 우월성을 입증할 수는 없지만, 표준 어법이나 시중의 사전 또는 여러 사람들이 채택하는 방식과는 다르기 때문에 표준안을 수정해야 한다.

태그부착 말뭉치 구축 과정에 여러 사람이 분업을 하고 실수를 유발할 수 있으므로 말뭉치에 오류가 있을 수 있으므로 a)와 같은 경우도 많이 발생하지만, 이 중의 대부분의 경우는 태깅 지침서를 충분히 읽어보지 않고서 판단하기 때문에, 위 네가지 경우에 어디에 해당하는지를 잘 모르는 경우가 많다.

이러한 종류의 오류에 대한 파악을 하고 표준안의 개선방안을 적극적으로 제시하여야 협조와 경쟁에 의한 바람직한 표준안으로 발전할 수 있을 것이다.

태깅에 있어서 정답이 하나가 아니고 여러 가지가 가능하거나 이 중의 어떤 것이 더 우월하다고 판단하기 힘든 경우에는 이 모든 경우를 정답으로 채점할 수 있는 방안도 적극적으로 고려해보아야 한다.

4.3.2 형태소 분석기 평가

형태소 분석기의 절대적 평가는 Test Suite를 구축하기가 품사 태그 부착 말뭉치 구축보다 몇곱절 어렵기 때문에 현실적으로 매우 힘들다. 그래서, 현재로서는 상대적 비교 정도로 그치고 있는데 이에 적극적인 평가방법이 요구되고 있다. 유럽의 MorphOlympics(Linguist List,1994)에서처럼 Glass Box Test와 같은 방법의 적용도 검토해보아야 한다. 이번 평가에서는 정확성(관찰데이타와의 부합도) 부분만 평가에 반영되었고, 효율성 측면은 평가에 반영 시키지 못한 면도 개선해야 할 것이다.

4.3.3 명사추출기 평가

명사추출기는 태그세트에 별로 민감하지 않아서 표준안과의 마찰은 크게 없다고 본다. 다만, 미등록어의 처리 문제와 관련하여 복합명사나, 고유명사, 숫자표현(예, '1.21 사태')이나 외국어(예, 'Y2K 문제')가 들어 있는 표현들의 처리는 이들을 한단위로 처리할 수 있도록 하는 태깅방식으로 표준안을 변경해야할 것이다. 이를 뒷바침하기 위하여, Test Suites에만 나타나는 복합명사, 고유명사, 숫자표현, 외국어 등의 리스트를 미리 만들어 이 중의 일부만 남기고 평가 전에 미리 참가팀들에 배포하는 방안도 생각할 수 있다.

형태소 분석이나 태깅이 제대로 되지 않아도 명사부분만 잘 분석하면 된다는 견해가 있는데 이것은 재고의 여지가 있다고 본다. 복합명사의 분해 같은 것은 형태

소 분석의 기능이 아니고 정보검색만을 위한 추가 기능이라고 본다. 원형복원 등은 하지 않아도 좋지만, 비 체언부의 분석도 아울러 잘 이루어져야 한다고 본다.

예를 들면, '가지 않았다'와 같은 예문에서 명사 추출만 신경쓴다면 '가/동사+지/어미'의 분석에 의하여 동사를 걸러낼 수 있는 여지가 없으므로 색인어의 과생성을 초래할 것이다. 그리고, 명사 이외의 부분에 대해서도 분석을 제대로 하여야, 즉 그 어절이 어떤 구성성분으로 구성되었는지 형태소 분석이 되어야 확신을 가지고 그 어절의 색인어로의 등록을 기각할 수 있다. 분석이 제대로 안되면, 용언부가 미등록어 명사로 색인될 가능성이 높아지기 때문이다.

4.3.4 기타

입력 오류에 대한 대처 능력, 즉 *exception handling*의 능력 또는 강건성(robustness)과 효율성 부분에 대한 평가 방안을 좀 더 체계적으로 마련할 필요가 있다. 그리고, 말뭉치의 영역도 소설이나 에세이 중심에서 신문기사, 방송스크립트, 학술지 등 더 실용적인 응용 영역으로 확대하는 것도 필요한 개선방향이 될 것이다.

5. 결론

이 논문에서는 MATEC99에서의 형태소 분석기 및 품사태거의 평가를 위한 Test Suites를 만들기 위하여 마련한 여러가지 표준안의 도출과정 상에서 발견된 문제점의 유형을 기술하고 해결방안을 모색하였다.

일부 문제에 대해서는 체계적인 해결 방안을 제시하려고 노력하였고, 일부 문제에 대해서는 현재로서 누구나 만족할 만한 해결방안을 제시하기가 힘들어서 부분적인 개선 방안을 제시하였다. 이러한 문제에 대해서는 여러 연구자들이 많은 관심을 가지고 다양한 어프로치에 의한 연구를 통하여 개선된 표준안이 도출하면 이에 입각한 Test Suites의 구축으로 더욱 공정하고 합리적인 평가대회를 개최할 수 있을 것으로 기대한다.

한가지 중요한 것은 표준안을 제안하고 수정요구할 수 있는 권리는 누구에게나 열려 있다. 다만, 기존 표준안에 명백한 잘못이 있고, 새로운 대안과 그것에 대한 설득력 있는 이유를 제시할 수 있으면 누구나 표준안 수정을 요구할 수 있는 것이다. 그리고, 대안은 없더라도 현재 표준안에 문제가 있을 때에는 유감만 표현할 것이 아니라 수정의 방향이라도 제시될 수 있도록, 반례라든지 모순되는 이유를 설명하여 주는 것이 필요할 것이다.

이 표준안은 정보통신 표준협회의 검토를 거쳐 대한민국 표준안으로 제정될 수 있기 때문에 여러 전문가들의 심도있는 검토의견이 제시되기를 바라마지 않는다. 이 표준안은 일단 제정되더라도 여러 연구자들의 건설적인 개선안의 제시로 계속 수정 보완될 것이다. 이 표준안에 따라 품사 태거 및 형태소 분석기의 평가가 이루어질 것이고, 또한 태그 부착 말뭉치가 대량으로 구축되어 보급될 것이므로 장차 이 말뭉치를 사용하게 될 잠재 수요자인 전문가들께서는 가능한 구축되기 이전에 좋은 대안을 제시해주어야 할 것이다. 이 논문에서 다루어지지 않은 더 중요한 문제가 있거나 다루어진 문제에 대한 좋은 개선 방안이 있다면, 표준화 홈페이지(http://aladin.etri.re.kr/~nlu/STANDARD)를 통하여 의견을 제시할 수 있다.

이러한 작은 노력들이 쌓여서 좋은 풍부한 연구 기초자료로 또 우리 시대의 문화적 자산으로 평가될 수 있는 것을 다 함께 구축하여 공유할 수 있을 것이다.

<감사의 글>

이 논문은 정보통신부가 지원하는 "자연어 정보처리 기술 표준화 연구('98~'99)"과제의 결과물 중의 하나이다. 과제수행을 지원한 정보통신부에 감사를 표한다.

참고문헌

[1] 이현아, 박재득, 장명길, 박수준, 박동인, 1996, "구문적 언어지식 획득과정의 문제점 분석 및 지원도구 설계', 한글 및 한국어 정보처리학술 대회 논문집.

[2] 마리 노엘르 가리-프리외(이재영 역), 1992, 문법과 언어학의 만남-문장연구 통사론, 한불문화출판

사.

[3] [3] Linguist List, 1994, FYI:Morphological Analyzer for German, http://listserv.linguistlist.org/ issues/5/5-463.html, LINGUIST List 5.463.

[4] K. S. Jones and J. Galliers, 1996, Evaluating Natural Language Processing Systems:An Analysis and Review", Lecture Notes in Artifical Intelligence, Vol. 1083, Springer, pp.228.

[5] E. Charniak, 1993, Statistical Language Learning, The MIT Press, pp. 21-38.

[6] H. Schuze,1997, Ambiguity Resolution in Natural Language Processing: Computational and Cognitive Models, CSLI Publications, pp. 27-63.

[7] C.D. Manning and H. Schuze,1999, Foundations of Statistical Natural Language Processing, The MIT Press.

[8] ETRI, 1999, 품사 부착 말뭉치 구축 지침서, http://aladin.etri.re.kr/~nlu/STANDARD.

[9] ETRI, 1999, 전자사전 표제어 선정 지침서, http://aladin.etri.re.kr/~nlu/STANDARD.

[10] ETRI, 1999, 품사 부착 말뭉치 구축을 위한 품사 태그 세트 지침서, http://aladin.etri.re.kr/~nlu/STANDARD.