형태소 어휘 문맥에 기반한 태깅 오류 정정

언어처리연구팀, 음성/언어정보연구센터, 한국전자통신연구원
김영길 · 양성일 · 홍문표 · 박상규

Tagging Error Correction Using Lexical Morpheme Context

Young Kil Kim, Sung Il Yang, Mun Pyo Hong, Sang Kyu Park
NLP Team, Speech/Language Technology Research Center, ETRI, Daejeon, Korea

요 약

본 논문에서는 형태소 분석 대상 어절의 좌우 어절내의 대표 형태소 어휘 문맥 정보에 기반한 형태소 오류 정정 방안을 제안한다. 현재까지 주변의 품사열 문맥 정보에만 의존하는 기존의 품사 태깅 모델과 달리 주변 어휘를 반영할 수 있는 좌우 어절 문맥을 이용해 형태소 태거의 성능을 향상시킬 수 있는 방법들이 제시되었다. 그러나 이러한 어절 문맥에 의한 지속적인 성능 향상을 위해서는 대량의 품사 태깅 문맥 정보를 필요로 한다. 따라서 본 논문에서는 이러한 자료 부족 문제를 해결하기 위하여 기존의 분석 대상 어절 좌우의 어절 단위의 어휘 문맥 정보가 아닌 좌우 어절내의 대표 형태소 단위의 형태소 어휘 문맥을 이용한 품사 태깅 오류 정정 방안을 제안한다. 실험을 통해, 형태소 어휘 단위의 문맥 정보의 적용성(Coverage)이 높고 기존의 품사 문맥 정보 기반의 형태소 분석기의 태깅 오류를 정정하여 그 정확성을 크게 향상시킬 수 있음을 보인다.

서 론

한국어 형태소 분석 결과들 중에서 정확한 분석 결과를 선택하는 형태소 태거는 특히 형태소 분석기의 높은 정확성을 요구하는 자동번역에 있어서 번역 시스템의 전체적인 성능을 좌우한다. 현재 널리 사용되고 있는 품사 문맥 정보를 이용한 형태소 품사 태거는 주위 품사 문맥 정보로서는 변별력이 없는 다양한 품사 중의성 문제로 인하여 형태소 분석기의 정확성을 크게 저하시킨다(임희동, 서영훈, 2000 ; 안영민, 서영훈, 2001 ; 임희석, 김진동, 임해창, 1997).

주어진 어절의 정확한 형태소 분석을 위해서는 문장에서의 주변 단어들의 어휘, 품사, 의미 및 문맥적인 공기 관계가 복합적으로 고려되어야 하지만 이를 반영하는 품사 태깅 모델을 결정하기가 쉽지 않다. 따라서 주변의 품사열 문맥 정보에만 의존하는 기존의 품사 태깅 모델과 달리 주변 어휘 규칙을 반영할 수 있는 어휘 문맥 기반 방식을 추가 적용하는 혼합형 태거들에 대한 연구가 활발히 진행되어 왔다(임희동, 서영훈, 2000 ; 안영민, 서영훈, 2001 ; 임희석, 김진동, 임해창, 1997 ; 이정규 등, 1997 ; 이상주 등, 1998). (1, 2)에서는 규칙 정보와 통계 정보의 상호 보완적 특성을 이용한 혼합형 방법을 제안한 바 있다. 규칙이 적용되는 중의성들에 대해서 높은 정확률로 태깅한 후 주위 조사와 어미의 통계 문맥 정보에 의해 규칙에 의해 해결되지 않는 중의성의 해소를 시도하였다. (3, 4)에서는 좌우 어절 문맥 규칙에 의한 품사 태깅 방안 및 규칙 추출 방안을 제안하였으며 (5)에서는 품사 태깅된 코퍼스에서 좌우 어절 규칙의 자동 획득 방안을 제안하였다. 좌우 어절에 의한 문맥 규칙에 의해 품사 태거의 성능이 향상될 수 있지만 자료 부족 문제로 인하여 지속적인 성능 향상을 위해서 비용과 시간이 많이 드는 대량의 품사 태깅 정보가 필요하다. 즉 품사 문맥만으로 변별력이 없는 실제 발생하는 언어현상에 대해 적용할 수 있는 어휘 문맥의 사용은 필수적이지만 이 어휘 문맥의 적용성(Coverage)을 높일 수 있는 방법 또한 고려되어야 한다.

따라서 본 논문에서는 이러한 자료 부족 문제를 해결하기 위하여 형태소 분석 대상 어절의 좌우 어절내의 대표 형태소 어휘 문맥 정보에 기반한 형태소 오류 정정 방안을 제

E-mail : kimyk@etri.re.kr
E-mail : siyang@etri.re.kr
E-mail : hmp63108@etri.re.kr
E-mail : parksk@etri.re.kr

안한다. 그리고 기존의 품사 문맥 정보 기반의 형태소 태거의 오류 정정을 통해 기존의 품사 문맥 기반의 형태소 분석기의 성능을 크게 향상시킬 수 있음을 보인다.

형태소 분석 오류

본 논문에서는 한중 자동번역 시스템 TELLUS-KC의 형태소 분석기 MAKUS를 이용하여 형태소 분석시 나타나는 오류 유형을 분석한다. MAKUS는 형태소 분석 이후 사용하게 될 상세한 번역 정보를 위해서 조사, 어미 및 보조용언들에 대한 문법적 기능별로 세분류한 137 품사세트에 대하여 형태소 접속 규칙에 기반하여 형태소를 분석하며 품사열 문맥 정보를 이용한 품사 태깅을 수행한다. 분석기의 정확률은 대상 문장에 따라 다를 수 있지만 평균 95~96% 정도의 정확률을 보인다. MBC 방송 뉴스 문장 500 테스트 문장들에 대한 형태소 분석을 수행하여 그 오류 유형을 살펴보면 크게 지식의 문제와 형태소 분석 엔진의 문제로 나눌 수 있는데, 전체 오류 중 지식 에러 및 부족으로 인한 경우가 20.6%를 차지하였다. '초중고교', '잡노동' 등의 경우에서처럼 접사 정보가 누락되어 분석이 실패하는 경우가 다소 발생하였고 그 외 누락되어 있던 단어들로는 '심대하다', '예단하다'와 같이 극히 드물게 사용되는 어휘들이었다. 형태소 간의 접속 가능 여부를 지정하는 접속규칙이 잘못되어 발생한 오류도 발생하였다. 분석 엔진의 문제로 인한 오류가 태깅문제, 고유명사 처리 문제 및 복합 형태소 처리 오류 등을 포함하여 72.3%, 기타 오류가 8.1%를 차지하였다. 엔진 오류 중 가장 큰 오류 원인은 태깅 에러로서 형태소 분석 전체 오류 중 57.1%를 기록하였다.

태깅 문제를 분석해 본 결과 의미적 컨텍스트가 고려되어야 해결할 수 있는 태깅 문제가 대부분이었다. 물론 '이라크군'의 경우처럼 접미사 '군'이 고유명사 뒤에 붙을 수는 있지만 인명 고유명사의 뒤에만 붙을 수 있다는 휴리스틱 규칙을 적용하면 해결할 수 있는 경우도 몇몇 있었다. 그러나 '(인물) 중의', '(3년) 만인' 등의 경우처럼 주위에 공기하는 어휘 또는 의미적인 고려 없이는 해결하기 어려운 문제들이 대부분이다. 다음은 테스트 문장들에서 대표적으로 나타나는 형태소 태깅 오류들의 유형을 보이고 있다. 각 품사별로 다양하게 태깅 오류가 발생하며 이러한 오류들은 주변 어휘 문맥 관계를 고려하지 않고서는 정확하게 태깅될 수 없는 현상들이다.

[용언 : 관형사]

오늘 주가지수는 어제보다 19포인트 오른 797.5를 기록했습니다.

19[숫자]+포인트[단위성의존명사] **오른[성상관형사]** 797[숫자]+.[기호]+5[숫자]+를[목적격조사]

[용언 : 체언]

문을 열자 양파 등 썩은 야채가 널려 있고.

문[보통명사]+를[목적격조사] **열자[보통명사]** 양파[보통명사]

[관형사 : 용언]

48개 정당은 막바지 유세에 온 힘을 기울였습니다.

유세[보통명사]+에[부사격조사] **오[너라불규칙동사]+ㄴ[관형사형전성어미]** 힘[보통명사]+를[목적격조사]

[부사 : 용언]

집 지붕이 채 무릎에도 미치지 않습니다.

지붕[보통명사]+가[주격조사] **채[규칙동사]+어[종속연결어미]** 무릎[보통명사]+에[부사격조사]

[용언+어미 : 체언+조사]

먼 데서 나는 소리를 사람보다 훨씬 더 잘 들을 수 있다고 합니다.

데[기타의존명사]+에서[부사격조사] **나[인칭대명사]+는[보조사]** 소리[보통명사]+를[목적격조사]

[체언 : 체언+조사]

까치부부는 화가 덜 풀린 듯 양쪽에서 협공을 계속합니다.

까치[보통명사]+부부[보통명사]+는[보조사] **화가[보통명사]** 덜[성상정도부사]

형태소 오류 정정 모델

1. 통계 확률 모델

분석 대상 어절의 좌우 어절 문맥은 품사 태깅된 말뭉치가 제한적이어서 데이터 희귀성 문제를 야기시킨다. 따라서 그 적용성을 높일 수 있는 형태소 어휘 단위의 문맥 정보를 고려할 수 있다. 형태소 어휘 문맥 정보의 윈도우 사이즈를 크게 할수록 정확성은 올라가지만 이 또한 데이터 부족 현상이 발생하므로 분석 어절의 좌우 어절을 그 대상 문맥으로 사용한다. 따라서 본 논문에서는 좌우 어절의 형태소 어휘와 분석 어절의 어휘 문맥을 반영하는 n-gram(n=4, 3, 2, 1) 형태소 어휘 문맥 정보를 사용한다.

[입력문] 유세에 온 힘을.

[형태소 태깅 정보]
유세[보통명사]+에[부사격조사]
온[성상관형사]
힘[보통명사]+를[목적격조사]
[형태소 어휘 문맥 정보]
uni-gram(온)→성상관형사.
bi-gram(에, 온), bi-gram(온, 힘)→성상관형사.
tri-gram(유세, 에, 온)→성상관형사.
tri-gram(에, 온, 힘)→성상관형사.
quad-gram(유세, 에, 온, 힘) →성상관형사.

위의 예에서 보듯이 좌우 어절 "유세에, 힘을"문맥을 보게 되면 "온"이 관형사임을 알 수 있지만 3-gram(에, 온, 힘) 또는 2-gram(온, 힘)에 의해서도 비교적 정확한 형태소 태깅을 위한 충분한 문맥 정보를 반영할 수 있다. 따라서 좌우 어절의 일부 형태소 어휘 문맥 정보에 의해 어휘 기반 형태소 품사 태깅 적용성(Coverage) 및 정확성을 높일 수 있다. Tir-gram과 4-gram에서 뒤 어절의 기능어 부분을 포함시키지 않은 이유는 형태소 태깅 오류를 분석해 본 결과 용언 또는 명사와 관련된 오류에 있어서 분석 어절의 형태소 결합에 영향을 미치는 문맥은 앞 어절의 형태소 헤드 어휘와 기능어 어휘 그리고 뒤 어절의 헤드 어휘임을 알 수 있었다. 한 문장이 다음과 같이 n개의 어절(word phrase)로 구성되고 각 어절은 형태소(morphological unit)들의 나열로 분석될 수 있다.

$S=w_1\ w_2\ \cdots\ w_{i-1}\ w_i\ w_{i+1}\ \cdots\ w_n$,where $w_{i-1}=m_{i-1,h}+m_{i-1,f}$, $w_{i+1}=m_{i+1,h}+m_{i+1,f}$, $m_{i-1,h}$: w_{i-1} 어절의 헤드 어휘. $m_{i-1,f}$: w_{i-1} 어절의 대표 기능어.

$m_{i-1,h}$는 i-1번째 어절 $w_{i-1}(=m_{i-1,1}+m_{i-1,2}+\cdots+m_{i-1,p})$가 p개의 형태소 열로 구성되어 있을 때 헤드 어휘를 나타낸다. 예를 들어 "선거유세에 온 힘을…."과 같은 문장에서 분석 어절 "온"앞 ㅇ절에서 복합명사 "선거유세"의 헤드어휘 $m_{i-1,h}$는 "유세"가 되며 대표 기능어 $m_{i-1,f}$는 부사격 조사 "에"가 된다. "10명 중의"에서 분석 어절 "중의"이전의 어절 "10명"에서 "명"이 $m_{i-1,h}$이며 $m_{i-1,f}$는 NULL값이 된다. 이 외에도 접미사 및 접두사 제거, 용언의 어간 어휘 사용 등을 통해 헤드 어휘의 적용성을 높인다.

그리고 어절의 대표 기능어는 복합 조사 및 복합 어미 등의 대표형을 취함으로써 그 적용성을 높인다. 예를 들어, "도시에서는 보기 힘든 장면"과 "도시에서 보기 힘든 장면"은 분석어절 "보기"의 앞 어절의 $m_{i-1,f}$이 "에서"로 대표격을 사용함으로써 그 적용성을 높일 수 있다. 이 때 분석 대상 어절 w_i의 형태소 태깅 결과는 k개의 후보가 가능하다고 가정한다.

$$w_i=c_1\,|\,c_2|\cdots|\,c_j\cdots|\,c_k$$

형태소 분석 후보를 결정할 수 있는 결정 요소들을 어절 빈도가 최대인 후보로 결정할 수 있는 어절 분석 확률, $P(c_j|w_i)$, 어휘 문맥 정보의 최대치 $MAX(P(c_j|m_{i-1,h}, m_{i-1,f}, w_i, m_{i+1,h}))$인 c_j를 분석 후보로 결정할 수 있는 어휘 문맥 정보로 크게 2가지로 볼 수 있다. 이를 가중치를 반영하는 어휘 문맥 태깅 함수 $T(m_{i-1,h}, m_{i-1,f}, w_i, m_{i+1,h})$를 확률 모델식으로 표현하면 다음과 같다.

$$T(mi-1,h,\ mi-1,f,\ wi,\ mi+1,h)=\boldsymbol{argmax}_{cj}[(P(c_j|w_i)+\varepsilon)\times(\alpha P(c_j|m_{i-1,h},\ m_{i-1,f},\ w_i,\ m_{i+1,h})+\beta_1\times P(c_j|m_{i-1,h},\ m_{i-1,f},\ w_i)+\beta_2\times P(c_j|m_{i-1,f},\ w_i,\ m_{i+1,h}))]+\gamma_1((P(c_j|m_{i-1,f},\ wi)+\gamma_2 P(c_j|w_i,\ m_{i+1,h})))],\ \text{where}\ \alpha>\beta_2>\beta_1>\gamma_2>\gamma_1$$

이 때 ε는 데이터 부족 현상을 보완하기 위한 분석 어절 확률값에 대한 Smoothing 인자이며 α, β_1, β_2, γ_1, γ_2는 문맥정보 가중치를 나타낸다. 문맥정보 가중치는 매칭되는 문맥의 길이와 및 형태소 대표어 또는 기능어간의 매칭 중요도 차이에 의해 그 값이 차이가 난다. 통계치의 γ_2가 γ_1보다 문맥 가중치가 높은 이유는 품사 태깅을 위한 분석 어절과의 문맥 범위는 같지만 γ_2는 문맥 형태소로 실질 형태소가 γ_1는 형식 형태소가 문맥 정보로 사용되었기 때문이다. 즉 매칭되는 문맥 어휘가 길고 실질 형태소인 대표 형태소가 포함될수록 문맥 가중치는 높아진다.

2. 데이터 자동 추출

형태소 품사 태그드 말뭉치에서 어절별 분석 빈도 및 앞뒤 어절에서의 형태소 헤드 어휘 및 대표 기능어 어휘 문맥 정보를 추출할 수 있다. 본 태그드 말뭉치는 STEP 2000 품사 태그드 코퍼스와 ETRI 품사 태그드 코퍼스를 한중 자동 번역 시스템 TELLUS-KC에서 사용하는 MAKUS 품사세트로 반자동적으로 변환한 말뭉치이며 그 크기는 1,980,000 어절 규모이다. 다음은 품사 태깅된 말뭉치에서의 "온"의 분석 일례를 보여준다.

개발[보통명사]+에[부사격조사]
온[성상관형사]
힘[보통명사]+를[목적격조사]

발전하[여불규칙동사]+어[종속연결어미]
오[너라불규칙동사]+ㄴ[관형사형전성어미]
네트워크[보통명사]+,[기호]

Table 1. 어절 분석 확률 및 어휘 문맥 정보의 일례

유형	문맥	분석결과	빈도
Uni-gram	온	오[너라불규칙동사]+ㄴ[관형사형전성어미]	59
			3
	온	온/성상관형사	88
Bi-gram	에_온	오[너라불규칙동사]+ㄴ[관형사형전성어미]	30
		온/성상관형사	10
	온_힘	온/성상관형사	11
	...		
Tri-gram	유세_에_온		0
	에_온_힘	온/성상관형사	2
	...		
Quad-gram	유세에_온_힘		0
	...		

[입력문] 선거 유세에 온 힘을 기울였습니다.
[MAKUS 형태소 분석 결과]
선거[보통명사] 유세[보통명사]+에[부사격조사]
오[너라불규칙동사]+ㄴ[관형사형전성어미] 힘[보통명사]+를[목적격조사] 기울이[규칙동사]+었[과거시제선어말어미]+다[평서형종결어미]+.[기호]

[형태소 어휘 문맥 기반 오류 정정]
$T(m_{2.1}, m_{2.2}, w_3, m_{4.1})$
$= argmax_{c_i}$ [(P(c_i | w_3)+0.01)*(100*(P(c_i | $m_{2.1}$, $m_{2.2}$, w_3, $m_{4.1}$) +10*P(c_i | $m_{2.1}$, $m_{2.2}$, w_3)+ 50*P(c_i | $m_{2.2}$, w_3, $m_{4.1}$) + 1*P(c_i | $m_{2.1}$, w_3) + 5*P(c_i | w_3, $m_{4.1}$)))], where α=100, β_2= 50, β_1= 10, γ_2= 5, γ_1= 1
c_1=오[너라불규칙동사]+ㄴ[관형사형전성어미]
c_2=온[성상관형사]

P(c_1)=P(온/(오[너라불규칙동사]+ㄴ[관형사형전성어미]))= (0.871+0.01)*(0+0+0+3/4+0)=0.661
{ P(c_1 | $m_{2.2}$, w_3, $m_{4.1}$)=P(c_1 | 에, 온, 힘) = 0/2 = 0.0
P(c_1| $m_{2.2}$, w_3) = P(c_1| 에, 온) = 30/40
P(c_1| w_3, $m_{4.1}$) = P(c_1| 온, 힘) = 0/11
P(c_1| w_3) = P((오[너라불규칙동사]+ㄴ[관형사형전성어미]) | 온) = 593/681 = 0.871 }

P(c_2) = P(온[성상관형사]) = (0.129+0.01)*(0+0+50*2/2+1*1/4+ 5*11/11)=7.680
{ P(c_1 | $m_{2.2}$, w_3, $m_{4.1}$)=P(c_1 | 에, 온, 힘) = 2/2 = 1
P(c_1| $m_{2.2}$, w_3) = P(c_1| 에, 온) = 10/40
P(c_1| w_3, $m_{4.1}$) = P(c_1| 온, 힘) = 11/11
P(c_1| w_3) = P((오[너라불규칙동사]+ㄴ[관형사형전성어미]) | 온) = 88/681 = 0.129 }

P(c_1) =7.680 > P(c_2)=0.653
Error Correction :: 오[너라불규칙동사]+ㄴ[관형사형전성어미] -> 온[성상관형사]

Fig. 1. 형태소 어휘 문맥 기반 오류 정정의 일례.

다음은 "온"에 관한 어절 분석 확률 및 문맥 정보를 추출한 일례이다. 태그드 말뭉치상에서 나타난 "온"어절에 관한 분석 결과는 온[성상관형사], 오[너라불규칙동사]+ㄴ[관형사형전성어미] 2가지로 형태소 분석된다. 다음 Table 1은 1절에서의 예문 "유세에 온 힘"에 대한 어휘 문맥 태깅을 위한 어절 분석 확률 및 어휘 문맥 정보를 나타낸다. 이 문맥 정보를 보면, Bi-gram 어휘 문맥 정보는 상당한 빈도로 나타나고 있음을 알 수 있다. 그러나 Tri-gram 이상에서의 문맥정보는 "에_온_힘"이 2번 나타나고 그 이외에는 어떤 어휘 문맥도 발생하지 않음을 알 수 있다. 즉 문맥의 범위가 넓은 3-gram 이상에서는 그 어휘 문맥 정보가 상당히 희소적으로 발생함을 알 수 있다. 그리고 3-gram 이상의 문맥 정보는 빈도가 적은 대신 상당히 정확한 태깅 정보를 제공해 준다.

3. 오류 정정 시뮬레이션

Fig. 1은 "온"에 관한 분석 과정을 나타낸다. 이때 문맥 가중치 α, β_1, β_2, γ_1, γ_2의 값은 태깅된 말뭉치의 적용 범위 등을 고려하여 실험적으로 그 값이 결정될 수 있겠지만, 본 실험에서는 n-gram간의 가중치 차이가 10배가 되도록, α=100, β_1=50, β_2=10, γ_1=5, γ_2=1 그리고 ε는 0.01로 설정하였다. 이 가중치의 차이가 클수록 최장 문맥 및 실질 형태소 문맥에 우선권을 주게 되는 방식이 된다. 이 가중치들은 학습 태깅 데이터의 적용성 및 확률 분포 등을 고려하여 실험적으로 결정될 것이다.

실험 및 평가

1. 주요 어절별 비교 태깅

태깅 오류가 빈번히 발생하는 대표적인 10어절에 대해 통계 학습에 사용된 태그드 말뭉치에 포함되어 있지 않는 각각 100개의 샘플들을 추출하여 대상 어절을 포함하는 문장에 대한 품사 문맥 기반 태깅(PBM : Pos Based Method)

Table 2. 주요 어절별 형태소 태깅

어절	분석 후보	PBM 정확률	PBM+MBM 정확률
온	동사+어미 관형사	83%(83/100)	86%(86/100)
사고	용언+어미 명사	93%(93/100)	95%(95/100)
채	용언+어미 의존명사 부사	91%(91/100)	98%(98/100)
화가	명사 명사+조사	13%(13/100)	85%(85/100)
나는	동사+어미 대명사+조사	64%(64/100)	80%(80/100)
전체 정확률		68.8%(344/500)	88.8%(444/500)

Table 3. 품사 문맥 기반 형태소 분석

총 어절수	6,672
형태소 분석 오류 어절수	288
형태소 분석 정확률	95.68%

Table 4. 형태소 어휘 문맥 적용률

Bi-gram	58.05%(3,873/6,672)
Tri-gram	21.49%(1,434/6,672)
Quad-gram	3.07%(205/6,672)

과 이에 형태소 어휘 문맥 기반 태깅에 의한 오류 정정 과정을 추가한 방법(PBM+MBM : Morpheme Based Method)의 정확도를 비교 분석하였다.

Table 1에서와 같이 품사 문맥 기반 태깅 방법의 정확률이 68.8%인 반면 어휘 형태소 문맥에 기반한 방법은 88.8%의 성능을 보이고 있다. 이는 대상 어절들이 태깅 오류가 빈번히 일어나는 어절임을 감안할 때 상당히 정확한 태깅 결과를 보여 주고 있다. 어휘 형태소 문맥에 의하여 태깅 오류 정정이 이루어지지 않는 경우는 주로 어휘 문맥 정보가 부족하기 때문이었으며 좌우 어절의 형태소 어휘 문맥의 범위를 벗어나는 경우도 있었다.

2. 형태소 어휘 문맥 태깅

본 논문에서 제안하는 형태소 어휘 문맥을 이용한 형태소 분석에 대한 정확률을 평가하기 위해 기존의 품사열 문맥 정보 기반 형태소 분석 방식과 이에 대해 형태소 어휘 문맥을 추가한 경우의 형태소 분석 정확률을 비교한다. 실험 문장은 문맥 학습에 사용하지 않은 총 6,672어절 규모의 500문장을 방송 뉴스 스크립트와 고등학교 국어 교과서에서 무작위로 추출하였다. 먼저 품사 문맥 기반 형태소 분석기에 의한 형태소 분석 결과를 평가하였다. Table 3에서와 같이 총 6,672어절 중 288어절에서 형태소 분석 오류가 발생하여 전체 형태소 분석 정확율은 95.68%를 기록하였다

그리고 품사 문맥 기반 태깅과 이에 형태소 어휘 문맥 기반 태깅에 의한 오류 정정 모듈을 추가한 방법에 대해 형태소 어휘 문맥 적용률 및 형태소 분석 정확률을 평가하였다.

Table 4에서의 어휘 문맥의 적용률을 보면 Quad-gram에서 3.07% 정도의 적용률 밖에 보이지 않는다. 따라서 Bi-gram과 Tri-gram의 형태소 어휘 문맥에 의존하여 형태소 오류 정정이 이루어지고 있음을 알 수 있다. Table 5에서는 n-Gram(n=2, 3, 4) 어휘 문맥 모두 사용하여 총 6,672어절 중 198어절에서 형태소 분석 오류가 발생하여 전체 형태소 분석 정확율은 97.03%를 기록하였다.

Table 5. 형태소 어휘 문맥 기반 형태소 분석

총 어절수	6,672
형태소 분석 오류 어절수	195
형태소 분석 정확률	97.08%
형태소 분석 정확률 향상	1.4%
형태소 에러 감소율(ERR)	32.29%(93/288)

101개의 형태소 분석 오류를 수정했으며 이 중 85개의 오류가 태깅 오류로 인한 수정이었다. 나머지 16개는 형태소 분석 오류를 수정하였다. 그리고 이 중 8개는 품사 문맥 기반 태거의 결과가 정확한 것이었다. 형태소 분석기의 전체 성능은 1.4%의 성능 향상을 이루었으며 에러 감소율(ERR)은 32.29%이었다.

결 론

형태소 분석 오류의 대부분을 차지하는 형태소 태깅 오류는 분석 어절의 주위 어휘 공기 문맥정보를 고려하지 못함으로 인하여 발생한다. 따라서 본 논문에서는 형태소 분석 시 주변 어휘의 공기 정보를 반영하여 형태소 분석의 정확성을 높이면서도 통계 데이터 부족 현상을 해소하기 위해 대표 형태소 단위의 문맥 정보를 이용한 태깅 방법을 제안하였다. 이는 어절 단위의 문맥 정보를 이용하는 방법에 비해 그 적용성이 높아 데이터 부족 현상을 해소할 수 있다는 장점이 있다. 또한 품사 문맥 정보에 의한 태깅에 비해 형태소 어휘 문맥에 의해 형태소 오류 정정을 수행함으로써 보다 정확한 형태소 분석이 가능함을 실험으로 밝혔다.

지속적인 형태소 분석 및 태깅의 정확성 향상을 위하여 다음과 같은 일이 계속 진행되어야 한다. 우선 태깅 모델 보완이 필요하다. bi-gram, tri-gram, quad-gram의 가중치를 실험을 통해 최적의 태깅 성능을 낼 수 있는 값으로 설정해 주어야 한다. 그리고 형태소 어휘 문맥 정보가 계속적으로 수정되어야 한다. 기존 품사 태그드된 말뭉치에 상당수 오류가 존재하며 각기 다른 품사 태그세트간의 변환으로 인한 오류 또한 다수 존재한다. 따라서 이에 대한 오류 수정 작업이 계속 이루어져야 한다. 또한 분석 속도를 고려하여 형태소 오류 정정의 대상이 되는 어절은 품사 문맥 태깅 방식의 분석기와 태깅 정확도를 비교하여 정확성이 높은 어절을 자동으로 기구축하여 사용해야 한다. 그리고 고빈도로 등장하는 태깅 오류의 가능성이 높은 어절들을 우선적으로 그 용례들을 추출하여 품사 태그드 코퍼스에 추가시킴으로써 지속적으로 형태소 어휘 문맥 데이터를 보강해 나가면 점진적인 형태소 분석 성능의 향상을 이룰 수 있을 것으로 기대한다.

REFERENCES

임희동, 서영훈(2000) : "어절간 문맥 정보를 이용한 혼합형 품사 태깅", 제12회 한글 및 한국어 정보처리 학술대회 논문집, *pp376-380*

안영민, 서영훈(2001) : "조사와 어미의 문법 기능을 활용한 품사 태깅 시스템", *pp97-100*

임희석, 김진동, 임해창(1997) : "언어지식과 통계 정보의 보완적 특성을 이용한 품사 태깅", 제9회 한글 및 한국어 정보처리 학술대회 논문집, *pp102-108*

이정규, 이상주, 임희석, 임해창(1997) : "규칙 기반 한국어 품사 태깅을 위한 어휘 규칙 획득의 수작업 최소화 방안", 제24회 정보 과학회 봄 학술발표 논문집, *pp479-482*

이상주, 류원호, 김진동, 임해창(1998) : "품사태깅을 위한 어휘규칙의 자동획득", 제10회 한글 및 한국어 정보처리 학술대회 논문집, *pp20-27*