영조기계번역에서의 형태론적해석과정에 대한 분석

박 명 철

경애하는 **김정은**동지께서는 다음과 같이 말씀하시였다.

《나라의 과학기술을 발전시키자면 과학기술정보사업체계를 정연하게 세워야 합니다. 과학연구사업에서 독창성과 창조성을 발휘하는것도 중요하지만 다른 나라들에서 이룩한 과학기술성과들을 우리 실정에 맞게 받아들이기 위한 사업도 잘하여야 합니다.》

나라의 과학기술을 하루빨리 세계적수준에 올려세우자면 발전된 과학기술을 우리의 실정에 맞게 받아들이기 위한 사업을 방법론있게 진행해나가야 한다.

이 글에서는 영조기계번역체계의 첫 공정으로서의 형태론적해석에서 제기되는 문제들에 대하여 분석하려고 한다.

콤퓨터언어학의 견지에서 볼 때 형태론에는 굴절과 파생, 합성의 세가지 주요령역들이 포함된다.

형태론적인 해석의 첫번째 대상으로서 굴절을 들수 있다.

굴절은 일반적으로 말줄기의 일부의 변화 또는 덧붙이의 바뀜 등에 의하여 단어의 어휘문법적의미를 나타내는것과 같은 문법적현상이다.

굴절은 말줄기의 형태론적변화, 개별적인 어미들의 변화의 뜻으로 쓰이며 굴절어미의 총칭적인 뜻으로도 쓰인다. 그리고 굴절은 어미, 어간, 어두에서 표현되면서 어미의 굴절, 어간의 굴절, 어두의 굴절로 쓰인다.

실례로 -ing로 끝나는 단어를 동사의 현재분사형으로 인식하는 영어의 간단한 규칙을 들수 있는데 이 규칙에서는 동사의 뿌리를 얻기 위하여 해당 단어의 어미를 떼여버리게 되여있다. 이것은 foaming, trying, testing과 같은 단어들에서는 잘 들어맞지만 have, hop, tie와 같은 단어는 -ing가 붙을 때 어미의 e가 없어지는것으로 하여 -ing를 떼여버리면 각각 *hav, *hopp, *ty만 남게 되므로 잘 맞지 않는다. 그리고 -ing가 붙은 단어라고 하여도 bring이나 fling과 같은 단어들처럼 현재분사형으로 해석되지 말아야 하는 단어들이 있는 경우도 있다.

우의 실례에서 보는바와 같이 단순한 어미제거규칙이라고 할지라도 보다 정교한 언어학적인 제한정보들이 반영되여야 한다. 때때로 형태론적인 해석만으로도 문법적범주들과 구조적기능을 쉽게 식별할수 있는데 실례로 영어에서 뒤붙이 -ize는 보통 동사를 가리킨다는 데로부터 이러한 뒤붙이들이 붙은 단어를 동사로 해석할수 있다. 이러한 해석이 콤퓨터상으로는 복잡하지 않다고 하더라도 실지 언어처리에서는 자주 제기되는 문제이며 따라서 일정한 주의를 돌릴것을 요구한다.

형태론적인 해석의 두번째 대상으로서 파생을 들수 있다.

파생은 주로 단어의 기본형태에 덧붙이를 붙여 한 단어로부터 다른 단어를 만들어내는 단어조성수법의 하나이다.

많은 언어들이 풍부한 파생형태체계를 가지고있으며 파생체계에서의 규칙성들은 사전의 크기를 줄이는데 리용될수 있다. 영어에서는 un-이나 non-과 같은 부정앞붙이들과 형용사들에 뒤붙이 -ly가 붙어 부사로 된다.

그러나 파생현상에서 silly와 같이 -ly로 끝나는 모든 단어들이 다 부사로 되는것은 아닌것처럼 례외적인 경우들이 적지 않게 존재하는것으로 하여 심중하게 연구되여야 한다. 영

어뒤붙이 -er는 dancer나 walker와 같이 흔히 동사가 나타내는 행동을 수행하는 사람을 의미하는 명사를 만들기는 하지만 그외의 의미들을 더 가지는 경우들도 있다. 실례로 computer는 동사 compute로부터 나온 전자설비를 나타내며 revolver는 동사 revolve로부터 이러한 동작을 하는 기계장치를 가지고있는 무기를 나타내기도 한다.

형태론적인 해석의 세번째 대상으로서 합성을 들수 있다.

합성은 두개이상의 단어들이 서로 결합하여 새로운 단일한 의미를 가진 단어를 만들어내는 단어조성수법의 하나로서 모든 품사들에서 다 나타난다. 실례로 명사에서는 car park, 형용사에서는 heartbreaking, 동사에서는 babysit, 전치사에서는 into 등을 들수 있다.

영어어휘들은 흔히 steam engine이나 steam hammer와 같이 명사들의 단순한 병렬배렬에 의하여 만들어지며 매개 구성요소로 되는 명사들은 사전에서 찾아볼수 있다. 시간이 지나면서 병렬배렬된 일부 명사들은 융합되여 steam-ship과 같이 단일한 명사를 이룰수도 있으며 사전에 전체적인 형태로 오르게 되군 한다.

그러나 조선어에서는《증기선》,《증기보이라》와 같이 병렬배렬보다는 융합현상이 보편적이며 이로부터《증기타빈자동조절체계》와 같이 새로운 합성어들이 쉽게 만들어진다.

새로운 합성어들은 기계번역체계에서의 형태론적인 해석에서 난관을 조성한다. 이러한 합성어들을 미등록어로 취급하는것은 비현실적이다. 왜냐하면 그의 의미와 정확한 대역은 그 구성요소로 되는 단어들로부터 파생될수 있기때문이다. 문제는 가능한 선택적인 토막이 여러개로 되는데 있다. 그 복잡성을 영어실례들을 가지고 설명한다면 coincide가 coin+cide로, cooperate가 cooper+ate로, extradition이 ex+tradition 또는 extra+dition으로, mandate가 man+date로 되는것과 같이 형태론적인 해석이 정확치 못하게 되는것을 들수 있다. 구체화된 형태론적인 정보가 반영된 해석프로그람에 의하여 해석이 진행되는 경우도 있지만 실지로 해소하기 힘든 모호성이 있는 경우도 있다.

기계번역에서 형태론적해석은 콤퓨터를 리용하여 자연언어본문에서 형태단어의 문법적정보를 탐색해내는 과정이다.

영어는 자모글자로서 단어가 일반적으로는 공백으로 구별되기때문에 단어의 구별이 명확하다. 그러나 공백을 경계로 단어단위를 구별하는 경우 일부 오유가 발생할수도 있는데 영어의 쓰기관습에 따르면 일반적으로 반점《,》, 반두점《;》, 옹근두점《:》, 감탄기호《!》, 물음표《?》, 옹근점《.》, 생략기호《…》 등은 모두 직접 단어뒤에 오며 단어사이에 공백이 없다. 그리고 단어가 생략되는 경우(실례로《prof.》,《corp.》,《doc.》,《Mr.》,《inc.》,《can't》,《let's》 등)도 있다. 그러므로 단어뒤의 표점부호와 단어의 생략을 어떻게 구별하는가 하는것은 곧 영어의 형태부분석에서 첫번째로 처리하여야 할 문제이다.

영어의 줄임형들은 she's – she is/she has와 같이 모호성을 띤다. 그러나 이러한 몇개의 모호성을 제외하고 영어의 줄임형들을 정확히 갈라내는것은 웃반점이 있는것으로 하여 단순하다고 볼수 있다. 품사판정이나 형태부해석, 기타 다른 과제들을 수행하기 위하여서는 이러한 줄임형들을 떼여내야 한다. 즉 영어문장에 들어있는 특수한 단어에 대한 합리적인 구분을 진행하여 매개의 단어단위를 생성하는 과정으로 되게 하여야 한다.

이밖에도 영어는 굴절이나 파생과 같은 형태변화를 하는것으로 하여 이러한 처리를 거친 다음에야 해당 단어가 원래 가지고있던 사전정보를 읽어들일수 있다.

영어의 형태부에 관한 정확한 리해와 처리는 기계번역을 위한 문법분석에서 필요한 기

본처리과정으로서 그의 처리결과는 전체 영어문법분석에 기초정보를 제공하며 각종 자연언어처리체계의 기초부분으로 된다.

영어본문에 쓰인 다음과 같은 문장들을 실례들어 보자.

례:① Mr. Kim is a good teacher. (김선생은 좋은 선생이다.)

② I'll see you home after the concert. (공연후에 집에서 당신을 만나겠다.)

우의 실례 ①에서 단어《Mr.》의《.》은 단어생략형식의 한 부분으로서 그것을 문장끝점으로 인식하여 단어와 분리시킬수 없다. 그러나《teacher》의《.》은 문장부호로서 반드시 단어와 갈라서 처리하여야 한다. 이 문장의 정확한 구분은《Mr./Kim/is/a/good/ teacher/.》로서 문장부호를 포함하여 총 7개 단위가 있다.

실례 ②에서 단어《I'll》은 원래 두개 단어이며 뒤단어가 생략형식을 취함으로써 가상적인《하나의 단어》를 구성하였다. 만일 공백을 단위로 하여 문장을 구분한다면 8개의 단위가 나오게 된다. 그러나 정확히 가르면《I/will/see/you/home/after/the/concert/.》로서 9개 단위이다.

영어에는 우에서 서술한 정황과 대응하는 단어수량이 제한되여있으므로 아래와 같이 처리를 진행할수 있다.

우선《prof.》,《Mr.》,《inc.》,《Co.》,《Jan.》등의 영어단어들을 함께 모아놓고 하나의 표화일로 보존한다. 단어구분처리를 진행할 때 현재 처리하는 단어가 표화일에 있는가를 조사하고있다면 한개 단어로 결정한다. 그뒤의《.》은 문장부호가 아니므로 가를수 없고 응당 한개 단어로 만든다.

또한 Let's와 let's를 let+us로 가른다.

또한 I'm을 I+am으로 가른다.

또한 {it, that, this, there, what, where}'s에 대하여 우선 's를 원형 is로 환원하고 다시 {it, that, this, there, what, where}+is로 가른다.

또한 {he, she}'s에 대하여서는《's》가 is일수 있고 has일수 있다. 보통 {he, she}+is로 하고 구분할 때 규칙을 써서 한 단계의 구분을 더 진행한다.

또한 WORD've, WORD'll, WORD're 에 대하여서는 단어뒤의《've》,《'll》,《're》에 따라 WORD+have, WORD+will, WORD+are로 구분한다.

또한 WORD'd는 한개 단어뒤에《'd》가 붙었다는것을 표시한다. 여기서《'd》는 would일수도 있고 had일수도 있다. 이때 그뒤의 단어가 동사원형인가 아닌가 등의 형태특징에 따라 규칙을 써서 WORD+would 또는 WORD+had로 구분한다.

또한 {is, was, are, were, has, have, had}n't에 대하여 먼저 n't를 원형 not로 하고 다시 {is, was, are, were, has, have, had}+not로 구분한다.

또한 can't는 can+not로 구분한다.

또한 won't는 will+not로 구분한다.

영어를 입구어로 하는 기계번역에서의 형태부해석처리는 영어의 형태론적변화에 의거하여 아래와 같이 진행할수 있을것이다.

우선 규칙적인 변화를 가지는 단어에 대하여 영어단어의 형태변화규범에 따르는 단어의 원형처리를 진행할수 있다.

《ed》로 끝나는 동사의 과거시칭에서《ed》를 없애면 원형으로 된다.

《ing》형동사 현재분사는 어미굴절을 떼고 대응한 형태정보를 단어마디에 기록한다.

《ly》가 어미로 붙어 어미굴절된 단어는 품사를 부사로 기록한다.

《er/est》어미의 단어는 갈라서 어미굴절을 떼고 대응한 비교급 또는 최상급을 기록한다.

《s/ses/xes/ches/shes/oes/ies/ves》어미의 단어는 어미굴절을 떼고 그때의 품사는 명사로 기록한다. 복수정보를 단어마디에 써넣는다. 이때 《ies/ves》어미의 단어는 굴절시 대응한 변화를 하여야 한다.

명사소유격 《WORD's》 또는 《WORDs》형식에 대하여 〈's〉 또는 〈s〉의 굴절을 진행한 후 형태정보를 현재 단어마디에 기록한다.

또한 동사, 명사, 형용사, 부사의 불규칙변화형식에 대하여 불규칙변화단어표를 만들고 형태굴절처리할 때 먼저 그에 대한 조사를 진행한다.

또한 년대, 퍼센트, 순서수사의 수자에 대해 처리할 때에는 대응한 형태부정보를 단어마디에 써넣는다. 아래에 실례들어 설명한다.

《2000s》는 년대를 표시한다. 《s》를 제거하고 원형을 《2000》으로 하고 현재품사를 시간명사로 확정한다.

《98.7%》는 퍼센트로서 굴절처리후 단어마디에 《percentage》+WORD(98.7%)라고 쓰고 현재단어의 품사를 수사로 써넣는다.

《$100》은 화폐를 표시하므로 굴절처리후 단어마디에 《WORD(100)》+《딸라》라고 기록하고 현재단어의 품사를 명사로 한다.

《85th》는 순서수사이므로 《th》를 굴절처리후 단어마디에 《제》+WORD(85)라고 기록하고 현재품사를 수사(순서수사)로 확정한다.

또한 련결부호 《-》를 가지는 합성단어에 대하여 각이한 형태부구성견본에 따라 처리를 진행한다.

one-fourth와 같이 수량수사와 순서수사가 합성된 분수사에 대하여 굴절후 단어마디에 《수사(4)》+《분의》+《수사(1)》이라고 기입하며 현재품사를 수사로 기록한다.

《명사+명사》, 《형용사+명사》, 《동사+명사》, 《명사+명동사》, 《동사+부사》 등은 합성명사를 구성한다.

《형용사+명사+ed》, 《형용사+현재분사》, 《부사+현재분사》, 《명사+과거분사》, 《명사+형용사》 등은 합성형용사를 구성한다.

《명사+동사》, 《형용사+동사》, 《부사+동사》 등은 합성동사를 구성한다.

우와 같은 굴절처리를 진행한 후 만일 사전에서 찾을수 없는 단어이면 새 단어(미등록어)처리를 한다. 통계적방법을 사용하여 품사추측을 할수 있고 가능한 품사를 확정할수 있다. 그러나 대역을 비롯한 기타 사전적세부정보는 확정할수 없다.

이처럼 기계번역에서의 형태론적해석은 입구문에 대한 처리의 첫 공정으로서 다음단계의 품사판정과 구문해석, 나아가서 대역선택과 합성에 이르는 전 공정의 성과여부를 좌우하는 중요한 공정으로 된다.

우리는 앞으로 영조기계번역체계의 성능을 높여나가는데서 제기되는 실천적문제들에 대한 연구를 심화시켜 강성국가건설에 실질적으로 이바지할수 있는 과학기술적성과들을 이룩해나가야 할것이다.