영조기계번역에서 어휘의 품사적모호성과 그 해소

박 명 철

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《나라의 과학기술을 세계적수준에 올려세우자면 발전된 과학기술을 받아들이는것과 함께 새로운 과학기술분야를 개척하고 그 성과를 인민경제에 적극 받아들여야 합니다.》(《**김정일**선집》 증보판 제11권 138~139페지)

나라의 과학기술을 발전시키고 세계선진과학기술을 제때에 받아들이는데서는 나라들사이의 언어장벽을 제거하는 문제가 중요한 과제로 나서며 여기에서 기계번역프로그람이 차지하는 위치와 중요성은 매우 크다.

이 글에서는 영조기계번역체계개발에서 중요한 위치를 차지하는 입구어의 품사적모호성과 그 해소를 위한 방법론적문제에 대하여 분석하려고 한다.

다른 언어들보다 형태변화가 제한되여있는 영어를 입구언어로 하는 영조기계번역에서는 어휘적준위에서 나타나는 모호성이 다른 언어들보다 비할바없이 복잡하며 이로부터 모호성해소의 과학적인 방법론을 세울것을 더 절박하게 요구한다.

일반언어학적인 견지에서 모호성의 류형을 크게 두가지 즉 어휘적모호성과 구조적모호성으로 구분할수 있다.

언어에 잠재하여있는 어휘적 및 구조적모호성들을 영조기계번역의 견지에서 합리적으로 해결하지 않는다면 기계번역의 처리에서 모호성해소의 과학성과 안정성을 보장할수 없다.

영조기계번역의 모호성에는 단어의 개념적의미가 나타내는 전통적인 모호성이 포함되지 않는다.

영조기계번역에서는 단어의 개념이 나타내는 모호성이 번역활동에서는 그 해소대상으로 되지 않는다는 관점으로부터 출발하여 어휘적모호성에서 그것을 배제하여 전통언어학의 모호성의 류형을 제한하였다.

이와 함께 기계번역에서는 그 특성상 어음론적특성에 따르는 동음이의어적단어부류에 존재하는 모호성도 배제한다. (실례; I/eye, no/know, one/won, sea/see)

영조기계번역에서의 모호성은 언어단위가 하나의 형식을 가지고 두가지이상의 어휘문법적기능을 수행하는 언어적특성으로 볼수 있다. 결국 모호성은 언어단위들의 내용과 형식사이에 일대일 대응이 성립하지 않는 자연언어의 특성과 엄밀한 기호적표기만을 인식하고 처리하는 콤퓨터의 특성을 다같이 고려하여야 하는 기계번역의 고유한 특성이라고 볼수 있다. 여기에서 말하는 어휘문법적기능에는 한 어휘적단위가 두개이상의 역어를 가지는 기능까지 포함된다. 결국 출구어의 다양성도 모호성의 한가지 표현방식으로 된다는것이다.

영조기계번역에서는 일반언어학적인 견지에서 본 어휘의 다의적모호성과 함께 품사적모호성을 더 포함시켜 두 부류로 나누고있다.

영어단어 for를 실례로 들어보자.

례①: He found it increasingly difficult to read, for his eyesight was beginning to fail.
(그는 책을 읽기가 점점 힘들었다. 왜냐하면 시력이 떨어지기 시작하였기때문이다.)

②: This is an English language course for foreign students.
(이것은 외국인학생들을 위한 영어과정안이다.)

우의 두 실례문장에서 전치사, 접속사의 두개 품사를 가지고있는 for를 어떻게 판정하고 그에 알맞는 대역을 설정하는가 하는것이 문제로 된다. for는 전치사로서의《~을 위한》의 의미와 접속사로서의《왜냐하면》이라는 의미를 가질수도 있다. 실례 ①에서는《왜냐하면》이라는 의미를 가진 접속사로 판정하고 번역해야 하며 실례 ②에서는 for가 앞의 단어 course를 수식하며《~을 위한》이라는 전치사의 의미를 가지고 번역되여야 한다. 이 문장들에서 볼수 있는바와 같이 영어에는 하나의 어휘적단위가 두개이상의 품사와 함께 그에 해당한 의미를 가지는것들이 있다.

영조기계번역에서 입구어인 영어어휘가 가지는 품사적모호성은 다품사어들에 의하여 발생하게 된다.

영조기계번역에서는 전통문법에서 품사전성이라고 명명한 단어조성수법에 의하여 생겨난 어휘들을 다품사라는 어휘부류에 소속시켜 분석하게 된다. 기계번역을 위하여 정한 어휘부류에는 다품사어라는 개념이, 언어단위에는 다품사어라는 대상이 더 들어가게 된다. 영조기계번역에서 다품사어란 결국 문맥속에서 둘이상의 어휘부류적특성 즉 둘이상의 품사자격을 가지고 실현될수 있는 언어단위이다.

현대영어에서 품사전성으로 다품사어를 만들어쓰는 기본원인은 어휘를 가지고 그와 비슷한 다른 현상과 상태를 대신하는 방법으로서 기성어휘를 가지고도 새로운 개념들을 많이 나타나게 하자는데 있다.

다품사어의 류형을《삼흥》전자대사전, 100만단어규모크기의 영조기계번역용기계사전 어휘들과 3 000만단어규모의 영어문장코퍼스를 분석한데 기초하여 분류하였다.

아래에 구체적인 다품사어류형을 제시한다.

№	류형	실례
1	동사/명사	view, control
2	동사/명사/부사	retail, plash
3	동사/명사/형용사	calm, fake
4	동사/부사	alight, plumb
5	동사/부사/형용사	smooth, upstage
6	동사/형용사	asperate, dizzy
7	명사/부사	whatsoever, yesterday
8	명사/전치사	vice, modulo
9	명사/접속사	the instant, the moment
10	명사/부사/전치사/형용사	opposite, past
11	명사/부사/형용사	northwest, backstage
12	명사/부사/형용사/동사	back, cool
13	명사/변화형	aiming, cooking
14	명사/변화형/형용사	touching, drunk
15	명사/한정어	semi, this, these
16	명사/한정어/부사	both
17	명사/형용사	novel, referent

표계속

№	류형	실례
18	부사/명사/접속사	any time
19	부사/전치사	over, natheless
20	부사/전치사/접속사	the year before
21	부사/전치사/형용사	nearer, aslant
22	부사/접속사	if ever, each time, hence
23	부사/특수어	as it were, honestly
24	부사/형용사	alike, crazy
25	변화형/동사/명사	underlay, saw
26	변화형/전치사	regarding
27	전치사/접속사	until, for, than, as
28	변화형/형용사	grown, bouncing
29	동사/변화형	forego
30	변화형/접속사	provided
31	전치사/동사/형용사/접속사/명사/부사	like

우의 표에서는 변화형, 한정어, 특수어와 같이 영어의 전통적인 품사체계에 속하지 않는 품사류형들이 있다. 이것들은 기계번역의 해석과정에 편리하게 기계번역체계용 기계사전에서 리용하는 품사류형들이다.

기계번역에서 품사류형의 분류는 기계번역체계의 기계사전작성으로부터 시작하여 형태부해석과 품사판정, 구문해석 등 번역을 위한 전반공정에 영향을 미치게 된다. 따라서 다품사어의 류형수는 해당 번역체계와 밀접한 련관을 가지고 해당 체계에 맞게 설정되여야 한다.

다품사어가 가지는 고유한 특성을 다음과 같이 갈라볼수 있다.

첫째로, 다품사어단어들의 대부분이 빈도수가 3 000단어에 속하는 단어들로서 문장에서의 쓰임빈도수가 매우 높다는것이다.

둘째로, 단순한 형태론적구조를 가진 영어단음절어들이 영어다품사어의 총수에서 큰 비중을 차지한다는것이다.

셋째로, 다품사어의 품사소속성은 고립상태에서가 아니라 문장속에서 나타난다는것이다.

례: Traffic slowed to a crawl as we approached the accident site.

(우리가 사고현장에 다가갔을 때 교통운행은 기여가듯이 느려졌다.)

실례에 쓰인 slow는 오직 문장속에서만 동사로 쓰이였다는것을 알수 있다.

현대영어에서 다품사적현상이 나타나게 된 원인은 어휘들의 품사형태적표식 즉 품사소속을 나타내는 형태가 뚜렷하지 않은것과 관련되며 현대영어의 주요단어조성수법으로서 합성법, 파생법 등과 함께 품사전성수법이 하나의 독자적인 단어조성수법으로 되고있는것과 관련된다.

기계번역의 해석과정에 해당 문맥에 맞는 하나의 품사를 결정하는것은 쉬운 일이 아니며 여러가지 규칙들과 언어자료기지들이 동원되게 된다.

종전의 기계사전에만 의거하던 전통적인 기계번역방식과는 달리 여러가지 언어자원

을 효과적으로 리용하는 현재의 기계번역체계에서 보다 높은 번역정확도를 달성하자면 품사코퍼스, 련어코퍼스, 구문나무코퍼스 등 방대한 량의 언어자료기지들을 구축하고 그것들을 효과적으로 리용하기 위한 방법론이 세워져야 한다.

기계번역흐름의 첫 공정인 형태부해석단계가 끝난 다음에 진행되는 품사판정단계에서 해석된 입구어어휘의 품사적모호성이 제기되며 품사결정과정에 해소된다.

품사결정은 문장의 매 단어(또는 형태부)들에 정확한 품사정보를 할당하는 처리이다.

품사결정은 자연언어에 존재하는 여러가지 모호성가운데서 품사적모호성을 해소하기 위한 처리로서 기계번역을 비롯한 많은 자연언어처리과제들에서 선행공정으로 된다. 품사결정체계는 품사적모호성으로 인한 구문해석단계에서의 무거운 부담을 덜어주며 정보검색체계에서의 색인어와 검색어추출, 언어정보획득, 철자검사, 자동사전구축 등에 리용될수 있다.

품사적모호성해소를 위한 품사결정방법들은 크게 규칙에 기초한 방법, 통계적방법, 변환규칙과 통계적처리를 결합한 방법으로 갈라볼수 있다.

규칙에 기초한 품사결정방법은 품사결정에 필요한 지식을 규칙의 형태로 표현하고 그것을 리용하여 품사결정을 진행하는 방법이다.

따라서 여기서는 지식을 표현하는 규칙의 형태와 규칙의 획득방법이 정확도향상과 체계개발에서 중요한 요소로 된다.

규칙에 기초한 품사결정체계들에서 품사결정과정은 크게 두 단계로 진행된다. 첫번째 단계는 사전을 리용하여 문장의 매 단어에 가능한 품사들의 목록을 할당하는 단계이다. 두번째 단계는 수동으로 작성된 애매성해소규칙들을 적용하여 문장의 매 단어에 대하여 가능한 품사목록을 하나의 품사로 축소하는 단계이다.

규칙에 기초한 방법의 부족점을 극복하기 위하여 통계적방법이 제기되였다. 통계적방법에서는 대량의 코퍼스(생코퍼스 혹은 품사코퍼스)를 분석하여 통계정보를 추출하고 얻어진 통계값들을 리용하여 품사적모호성해소를 확률적으로 진행한다.

통계적방법에는 리용되는 통계모형에 따라 HMM(숨은 마르꼬브모형)에 의한 방법, 신경망에 의한 방법, 품사 N–그람에 의한 방법 등이 있다.

통계적방법의 우점은 학습을 자동적으로 진행하며 례외적인 언어현상들을 효과적으로 처리할수 있다는것이다. 부족점은 통계자료기지를 구축하기 위해 많은 량의 품사코퍼스가 필요한것이다.

변환규칙과 통계적처리를 결합하여 품사결정을 수행하는 방법도 있다.

이 방법에서는 품사결정을 위한 언어지식을 변환규칙의 형태로 작성하고 리용하는데 변환규칙은 품사코퍼스로부터 통계적으로 그리고 자동적으로 추출된다.

여기서는 품사코퍼스를 준비하고 그에 대한 초기해석결과와의 차이를 줄여나가도록 변환규칙을 획득하고 적용하는 과정을 반복하게 된다.

우리는 영조기계번역체계의 성능을 높여나가는데서 제기되는 실천적문제들을 적극 해명해나감으로써 강성국가건설에 실제적으로 이바지할수 있는 과학기술적성과들을 이룩해나가야 할것이다.