영한 기계번역에서의 복합어구 인식

장 두 성, 김 덕 봉, 최 기 선
한국과학기술원 전산학과

Complex Phrase Recognition in English-to-Korean Machine Translation : MATES/EK

Du-seong Chang, Doek-bong Kim, Key-Sun Choi
Dept. of Computer Science, KAIST

요 약

복합어는 여러개의 단어가 하나의 의미를 나타내는 단어를 말한다. 이 논문에서는 번역시 구성단어들의 의미의 합이 아닌 다른 또 하나의 의미를 나타내는 단어를 대상으로 한다. 이러한 복합어는 구문해석단계에서 많은 애매성의 원인이 되며, 유형에 따라 숙어 처럼 새로운 의미로 항상 같이 쓰이는 복합어와 복합어의 형성이 복잡하여 규칙으로서 단어를 이해할 필요가 있는 단어로 구분할 수 있다.

첫번째 유형은 단어의 형성이 단순하여 하나의 사전 엔트리로 등록될 수 있다. 이때 이들 복합어가 가지는 개별 어휘 규칙을 같이 사전에 등록하여 사전을 효과적 이용할 수 있다. 두번째 유형은 규칙에 의한 처리를 하여야 한다. 이러한 복합어에 대한 인식을 구문분석이전에 행함으로서 적은 노력으로 복합어로 인한 전체 문장의 애매성을 감소시키고, 문장내 단어의 수를 감소시킴으로서 전체 번역시스템의 효율을 증대하며, 복합어의 처리는 번역문을 자연스럽게 생성하는 데 큰 효과를 나타낸다.

I. 서론

복합어는 *'state-of-the-art'*나 *'5 hundred and five'*처럼 둘이상의 단어가 하나의 의미를 나타내어 문장내에서 단일어의 특성을 가지는 단어를 말한다. 과학문헌이나 기술서적에는 많은

복합어가 쓰인다. 이러한 복합어가 많이 쓰이게 되는 이유는 어떠한 개념을 나타내는 데 효과적으로 복합어가 쓰일 수 있기 때문이다. 이들 복합어가 개개단어의 의미의 합이 아닌 다른 새로운 의미를 나타낼 때 기계번역시스템에서 자연스러운 번역문을 얻기 위해서는 이러한 복합어의 의미에 대한 분석이 이루어져야만 한다. 그러나 이러한 복합어에 대한 분석을 구문분석에서 행할 때 다양한 형태의 복합어는 구문분석단계에서 많은 애매성의 원인이 된다.

특수한 영역에 대한 기계번역시스템의 경우는 사용가능한 복합어의 유형과 번역시 대응될 수 있는 단어들을 유지하여 구문분석단계에서의 애매성을 줄일 수 있다[1]. 그러나, 이러한 방법은 과학문헌이나 기술서적과 같은 넓은 번역영역을 대상으로 할 때는 문장속에서 쓰일 수 있는 복합어를 모두 사전(해석사전과 번역사전)에 유지하는 것은 불가능하다. 그러므로 이러한 복합어의 분석은 어떠한 형태로든지 이루어져야만 한다. 이 논문에서는 복합어를 처리함에 있어서 구문분석이전의 단계에서 복합어의 유형에 따라 사전과 규칙을 이용하여 복합어를 처리함으로서 분석단계에서의 애매성을 줄이고 사전을 효율적으로 사용하는 방법을 논의한다.

2장에서는 이 논문에서 다루고 있는 복합어의 유형에 대하여 논하고 3장에서는 사전에 의존한 복합어의 분석방법과 여기에 사용되는 복합어규칙과 규칙적용에 사용되는 우선 순위에 대하여 논한다. 4장에서는 규칙에 기반한 복합어의 분석방법에 대하여 논한다.

II. 복합어의 유형

우리는 위에서 복합어를 둘이상의 단어가 하나의 의미를 나타내어 문장내에서 단일어의 특성을 가지는 단어로 정의 했다. 실제 문장내에서 쓰이는 복합어는 그 정의대로라면 *'computer science'*나 *'user interface'*처럼 의미가 단순히 개개단어의 의미의 합으로서 나타내어질 수 있는 복합어와 하나의 새로운 의미를 나타내는 복합어로 구분할 수 있다. 여러개의 단어가 새로운 의미를 나타내는 복합어는 다시 그 쓰임에 따라 다음의 2가지로 세분류할 수 있다.

1. *'5 hundred and five'*나 *'Georgia Institute of Technology'*처럼 복합어의 형성이 복잡하여 복합어의 형성규칙으로서 그 의미를 이해할 필요가 있는 단어열. 예를 들어 수사열은 여러개의 수사와 단어가 모여 하나의 수사의 의미를 나타내며 이 단어열들이 수사열을 이루는 방법은 하나의 규칙으로 표현될 수 있다. 또한 고유명사의 품사를 지니게 되는 복합어는 그 복합어를 이루는 단어들에서 그 단어열이 고유명사를 나타낸다는 단서를 찾을 수 있다. 이러한 복합어들을 사전에 일일히 등록한다는 것은 불필요한 일이다.

2. *'state-of-the-art'*나 *'give up'*처럼 이들이 문장내에서 쓰일 때 새로운 의미로 항상 갑

이 쓰이는 복합어. 이들은 명사구의 형태일 수도 있고 동사구나 형용사구의 형태일 수도 있다. 이들은 복합어를 이루는 방법이 단순하여 항상 문장내에서 같은 의미로 쓰이므로 하나의 단일어처럼 쓰인다. 그러므로 이러한 단어들은 하나의 사전 엔트리로 등록될 수 있다.

III. 사전에 의존한 복합어의 인식

여러개의 단어가 모여 하나의 새로운 의미를 나타낼 때 이 복합어의 새로운 의미를 개개의 단어의 의미에서 추출하기 위해서는 단어들간의 관계를 분석하여야만 한다. 이러한 분석은 개개 단어의 구문정보와 의미정보, 또는 그 이상의 정보를 이용한 분석이 있어야만 한다. 이러한 분석을 행하는 데는 시간과 비용이 많이 들 뿐더러 분석에 의해 추출된 의미또한 정확한 의미인지가 확실치 않다. 즉, 동사 *'give'*와 전치사 *'up'*의 의미로 부터 복합어 *'give up'*의 의미를 추출한다는 것은 어려움이 따를 뿐더러 설사 성공한다하더라도 그 추출된 의미의 신뢰성은 떨어질 것이다. 그러므로 이러한 개개단어의 의미로 부터 그의 의미를 추출하기 어려운 복합어에 대해서는 번역의 효율을 위해 그 의미를 사전에 등록하여만 한다.

그러나 동사 *'give'*와 전치사 *'up'*이 문장내에서 같이 쓰이다고 해서 이들이 항상 복합어 *'give up'*을 이룬다고는 할 수 없다. 그러므로 이러한 복합어를 사전에 등록하기 위해서는 복합어의 구성단어와 그 형태의 제약을 동시에 등록할 수 있는 숙어 사전과 같은 기능을 하는 것[1]이 있어야 할 것이다.

복합어들은 문장내에서 일정한 형태로 나타난다. 즉 특정한 품사가 복합어를 구성하는 단어들 사이에 오거나 단어의 일부가 생략이 가능하기도 한다. 이러한 복합어가 나타나는 형태를 복합어와 같이 등록하여 일련의 단어열을 단어 각각의 의미로 해석할 것인지, 아니면 하나의 새로운 의미를 나타내는 복합어로 해석할 것인지의 여부를 결정한다. 다음은 사전에 복합어와 같이 등록하는 규칙들이다.

- **Space rules** : leave off에서와 같이 구성 어휘 사이에 다른 어휘가 삽입될 수 없는 복합어를 나타내는 규칙들.
- **Dash-n-Dash rules** : ought -n- to에서와 같이 구성 어휘들 사이에 부정 어휘(not,

[1]여기에서 숙어사전이라 명하지 않은 이유는 여기에서 다루는 복합어가 비록 숙어처럼 여러개의 단어가 하나의 새로운 의미를 나타내는 것이긴 하지만 *'state-of-the-art'*에서처럼 숙어의 범주에 속하지 않는 것이 있고, 또한 *'not only ~but also'*에서와 같은 구문지식을 필요로 하는 숙어는 포함하고 있지 않기 때문이다.

scarcely, hardly 등)만이 삽입될 수 있는 복합어를 나타내는 규칙들.

- **Dash-POS-Dash rules** : turn -PRON- on이나 -det- added performance에서와 같이 구성 어휘들 사이에 지정한 품사만이 삽입될 수 있는 복합어를 나타내는 규칙들.
- **Dash-Number-Dash rules** : such -3- as에서와 같이 구성 어휘 사이에 정해진 수이하의 다른 어휘가 삽입될 수 있는 복합어를 나타내는 규칙들.
- **Optional rules** : 0be -n- associated with에서와 같이 구성 어휘 사이에 생략가능한 단어가 있는 복합어를 나타내는 규칙들.

복합어를 규칙으로 기술하므로 같은 단어열에 하나 이상의 복합어 규칙이 적용될 수있고, 하나의 단어열이 두개의 복합어 규칙에 포함될 수도 있다. 이러한 경우 복합어 규칙의 적용에 우선 순위를 주어야 할 필요가 있다. 이러한 우선 순위의 적용은 다음의 휴우리스틱 규칙에 의한다.

1. **최장 일치 우선** : 한 단어열에 여러 복합어 규칙이 적용될 때 적용되는 단어의 경계가 가장 긴 복합어 규칙을 적용한다. 예를 들어 *"... in contrast with ..."*의 단어열에 in contrast보다는 in contrast with를 적용한다.
2. **최다 일치 우선** : 하나의 단어열이 두개이상의 복합어 규칙에 포함될 때 적용되는 단어의 수가 가장 많은 복합어 규칙을 적용한다. 예를 들어 *"... in contrast to date ..."*의 단어열에 to date보다는 in contrast to를 적용한다.
3. **근접 우선** : 하나의 단어열에 두개 이상의 복합어 규칙이 적용될 때 적용되는 단어의 수가 같다면 문장 내에서 더 좁은 범위에 적용되는 복합어 규칙을 우선 적용한다.

몇개의 단어가 뭉치어 개개의 단어의 의미의 합과는 다른 하나의 의미를 표현할 때 우리는 이러한 단어열을 보통 숙어라 한다. 이런 숙어를 인식하기 위해서는 단어 사전과는 다른 숙어 사전을 두어야 한다. 그러나 모든 숙어를 이러한 사전에 의존해서 완벽하게 해석할 수는 없다. 그러나 또한 모든 숙어를 구문해석이나 의미해석에 의존한다는 것은 과도한 양의 문법이 필요하고 파싱과정에 많은 부담을 주게 된다. 그러므로 구문지식이 필요치 않는 숙어(복합어)는 사전에 등록하여 사전에 기반한 처리를 파싱에 앞서 미리함으로서 파싱시 문장내 단어의 수를 줄이고, 개개의 단어를 해석할 때 생기는 많은 애매성을 줄여 파싱의 효율을 높일 수 있다(그림 1 참조). 또한 복합어를 그들의 형태에 따른 복합어 규칙과 같이 단일어 사전에 등록함으로서 개별단어를 사전에서 찾을 때 복합어의 가능성을 검사하게 됨으로서 사전을 효과적으로 이용한다.

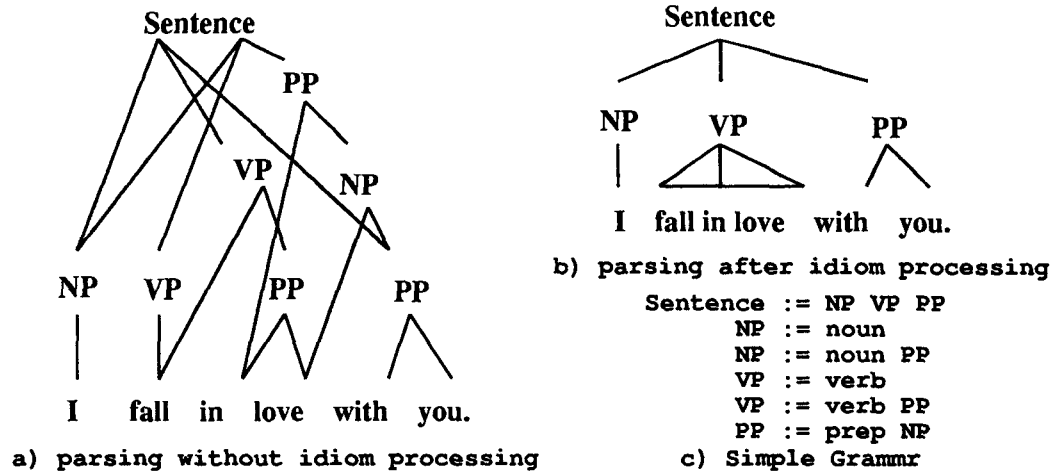

그림 1: 복합어 처리의 효과 : 파싱 결과

IV. 규칙에 의존한 복합어의 인식

우리는 앞장에서 개개의 단어로 부터 새로운 의미를 추측하기 힘든 복합어는 사전에 그 복합어의 제약규칙과 함께 등록하여 번역의 효율을 추구하여야 한다고 주장했다. 그러나 새로운 의미를 나타내는 복합어를 모두 사전에 등록한다는 것은 매우 불필요한 일이다. 이것은 새로운 의미를 생성하는 복합어라 할지라도 그 복합어를 구성하는 방법이 어떠한 규칙에 의하고 있으면 그 규칙에 의해(단어들간의 의미 또는 구문적인 관계[2]가 아니라 복합어를 구성하는 단어 형태에 의해) 새로운 복합어의 의미를 추측할 수 있기 때문이다. 이러한 복합어를 인식하기 위해서는 이런 복합어를 만드는 규칙을 구하고 이들 규칙을 효율적으로 복합어의 의미를 찾는 데 적용하여야 한다.

1 유한상태 오토마타의 이용

단어들의 구문 정보나 의미정보가 아닌 단어의 결합 순서만을 의존하여도 복합어의 의미를 정확하게 알 수 있다. 이러한 예로는 수사열을 들 수 있다. 수사열은 여러개의 수사와 특정한 단어들이 어울려 하나의 수사의 의미를 나타내는 복합어를 말한다. 이러한 복합어는 그 구성단어가 한정되어 있으므로 이들 단어들을 구성성분으로 하는 정규 문법을 작성할 수 있다(그림 2 참조). 실제 이러한 정규문법을 사용하면 정수, 분수, 소수, 기수, 서수의 모든 수사열에 대한 의미를 구문분석이나 의미분석없이 단순한 형태 규칙만으로도 알 수 있다. 수사열의 의미는 이러한 정규문법에 대응하는 유한상태 오토마타(Finite State Automata)를 구성하여 효율적으로 알아낼 수 있다. 이렇게 알아낸 의미는 더이상의 분석 단계를 거치지 않고 번역문의 생성에

바로 사용될 수 있다.

```
NUMBER    = ONE | TEN | HUNDRED | THOUSAND | MILLION | BILLION | TRILLION
ONE       = one | two | ... | nine
TEN       = ten | eleven | ... | nineteen
TWENTY    = ({twenty | thirty | ... | ninety} [-] ONE) | TEN
HUNDRED   = {a | (0-9)* | ONE} HUNDRED [and] TWENTY
HUNDRED2  = ONE hundred [and] TWENTY
THOUSAND  = {a | (0-9)* | HUNDRED} thousand [,] HUNDRED2
THOUSAND2 = HUNDRED thousand [,] HUNDRED2
MILLION   = {a | (0-9)* | HUNDRED} million [,] THOUSAND2
MILLION2  = HUNDRED million [,] THOUSAND2
BILLION   = {a | (0-9)* | HUNDRED} billion [,] MILLION2
BILLION2  = HUNDRED billion [,] MILLION2
TRILLION  = {a | (0-9)* | HUNDRED} trillion [,] BILLION2
```

그림 2: 영어에서 정수를 나타내는 정규표현

2 복합어의미의 단서이용

복합어의 의미는 그를 구성하는 단어의 형태적인 특성에 의해 그 의미를 짐작할 수도 있다. 이러한 예로 복합어가 고유명사의 품사를 지니게 되는 고유명사열이 있다. 고유명사열을 이루는 단어들은 대부분 대문자를 포함한 단어(Capitalized Word)이며, 이러한 단어들의 나열은 고유명사열이라는 단서가 된다. 다음과 같은 예에서 추출된 고유명사열은 더 이상의 분석없이도 번역문의 생성에 바로 이용될 수 있다.[2]

"*As*[3] **Apple's Alan Kay**[4] and others have pointed out..."

복합어의 의미를 판단하는 데 단어의 형태적인 특성을 이용하는 또다른 예는 복합어를 이루는 단어중 어느 한 단어(sub-word)가 다른 주단어(head-word)의 접사역할을 하는 복합어이다. *'high-performance'*나 *'mid-night'*에서 처럼 접사의 역할을 하는 단어는 주단어의 사전정보에 약간의 의미를 더하거나 품사의 변형을 가져온다. 그러므로 이러한 단어들을 단서로 이용하여 복합어의 의미를 파악할 수 있다.

[2]모든 고유명사열이 번역문의 생성에 그대로 이용될 수 있는 것은 아니다. 예를 들면, "*The* **Georgia Institute of Technology's UIDE** is a research system." 에서 처음의 고유명사열은 생성에 그대로 사용되지 않을 수도 있다.

[3]이텔릭체 단어는 고유명사처리가 되지 않는 예이다.

[4]볼드체는 고유 명사처리가 되는 예이다

3 복합어 형성의 패턴 이용

이 밖에도 년도와 같은 특정한 의미를 나타내는 단어열은 어떠한 복합어의 패턴이 어떠한 의미를 나타내는 지를 많은 양의 자료에 의해 얻을 수 있으므로 이러한 패턴을 이용하여 복합어의 의미를 알 수 있다. 연도를 나타내는 복합어에는 다음과 같은 예가 있다.

- *"... in 1980"*나 *"... is nineteen-ninety-one"*에서처럼 단순히 연도 그자체를 의미하는 형태
- *'the 1980s'*나 *'the nineties'*에서 처럼 연대를 의미하는 단어열
- 이 밖에 *'3 decades'*와 같은 몇년의 기간을 의미하는 형태가 있다.

과학 기술 문헌에는 *'kg'*, *'kilogram'*, *'cm'*, *'centimeter'*, *'dB'*, *'decibel'*, *'dollars'*, *'MIPS'*등과 같은 단위를 나타내는 단어가 비교적 많이 나온다. 이러한 단어는 거의 모두가 앞에 정도를 나타내는 단어(예를 들면 *'2 cm'*) 혹은 구(예를 들면 *'150 to 600 Mbps'*)를 동반하여 하나의 복합어를 이룬다. 이러한 단어들은 복합어를 인식하는 것만으로도 그 의미를 알 수 있다.

V. 평가및 결론

우리는 위에서 기계번역 시스템에서 구문분석 이전에 유형에 따라 사전이나 규칙을 이용하여 복합어를 인식하는 것을 제안하였다. 이러한 방법은 구문분석이나 의미분석없이 단순한 복합어의 형태만을 이용하여 적은 노력으로 복합어를 인식하므로 다양한 복합어를 구문단계나 그이후

		형태소 분석결과			복합어 처리결과		
Sample text	No. of Sentence	Words /Sents	Cat. /Word	Total Words	Words /Sents	Cat. /Word	Total Words
I	259	7.5	1.53	1940	7.1	1.33	1840
II	909	6.7	1.40	6087	6.6	1.30	6009
III	539	15.6	1.48	8389	14.4	1.31	7749
IV	315	25	1.50	7898	23	1.31	7235

표 1: 복합어 처리의 효과 : 문장내 단어수의 감소

의 단계에서 해석할 때 생기는 많은 애매성을 줄일 수 있다. 구문분석 단계이전에 복합어를 인식함으로서 여러개의 단어로 구성된 복합어가 하나의 의미를 가지는 단일어처럼 분석에 참여하게 된다. 그림 1에서 복합어의 인식으로 문장내 단어의 수가 적어짐을 알 수 있다. 문장내 단어의 수는 문장의 애매성에 비례하고 이는 전체 번역시스템의 속도에 비례하게 되므로[3] 결론적으로 구문분석이전의 복합어의 인식은 번역시스템의 효율을 크게 증대시키고, 이러한 복합어의 처리는 번역문을 자연스럽게 생성하는데 큰 효과를 나타낸다.

참고 문헌

[1] Pierre Isabelle *"Another look at Nominal Compounds"* COLING, pp. 509-516, 1984

[2] Judith N. Levi *"The Syntax and Semantics of Complex Nominals"* pp. 75-222, Academic press, 1978

[3] Carl G. de Marken *"Parsing the LOB Corpus"* 28th Annual Meeting ot the ACL, pp. 243-251, 1990