통계정보와 WordNet을 이용한 복합명사 분석

류민홍[0] , 나동열 , *장명길
연세대학교 전산학과, *한국전자통신연구원 언어공학연구부
{mhlyu, dyra}@magics.yonsei.ac.kr, *mgjang@etri.re.kr

Nominal Compound Analysis Using Statistical Information and WordNet

Min-Hong Lyu[0] , Dong-Yul Ra , *Myung-Gil Jang
Computer Science Dept., Yonsei University
*Language Engineering Dept., ETRI

요 약

복합명사의 한 구조는 구성 명사간의 수식관계의 집합이라고 본다. 한 복합명사에 대하여 가능한 여러 구조 중에서 올바른 구조를 알아 내는 것이 본 논문의 목표이다. 이를 위하여 우리는 최근에 유행하는 통계 기반 분석 기법을 이용한다. 먼저 우리의 복합 명사 분석 문제에 알맞은 통계 모델을 개발하였다. 이 모델을 이용하면 분석하려는 복합명사의 가능한 분석 구조마다 확률값을 얻게 된다. 그 다음 가능한 구조들 중에서 가장 확률값이 큰 구조를 복합명사의 구조로 선택한다. 통계 기반 기법에서 항상 문제가 되는 것이 데이터 부족문제이다. 우리는 이를 해결하기 위해 개념적 계층구조의 하나인 워드넷(WordNet)을 이용한다.

1. 서론

영어나 한국어에서 명사구는 매우 다양한 구조를 가진다. 그 중에서 특히 여러 개의 명사가 연달아 나타나서 생긴 명사구를 복합 명사라 부른다. 예를 들면 명사구 "a table cover cloth" 는 (관사 'a' 뒤에) 3 개의 연속된 명사 열로 구성되어 있다. 따라서 이 명사구는 복합명사 (nominal compound)이다. 복합명사를 구성하는 명사 사이에는 여러 가지 관계가 존재할 수 있는데, 앞의 예의 경우, "a cloth with which a table is covered"와 같은 의미를 전달한다고 볼 수 있다. 즉, 2개의 술어-인자(predicate-argument) 관계를 내포하고 있다. :

"a cloth is the instrument of action *covering*"
"a table is the patient of action *covering*".

결국 복합 명사 "a table cover cloth"를 이해하려면 위와 같은 내포된 의미를 파악하여야 한다. 복합명사의 분석이란 이러한 복합명사의 내포된 의미를 파악하는 작업이다[2]. 복합명사를 분석하여 그 내포된 의미를 완전히 파악하기 위해서는 각 명사간의 관계의 종류(격 또는 역할) 까지도 파악하여야 한다. 그러나 이와 같은 깊은 의미까지 파악하는 것은 매우 어렵다. 본 논문에서는 이와 같은 깊은 의미까지의 분석(deep analysis) 대신 서로 관계가 되는 명사쌍들을 파악하는 정도의 낮은 수준의 분석(shallow analysis)에 만족하고자 한다. 명사쌍을 이루는 명사 중에서 하나는 수식하는 명사, 하나는 수식을 받는 명사로 구분하여 위의 예에서는 table이 cover를 수식하며, cover가 cloth를 수식한다고 분석하는 것이다 [1] . 최근의 복합명사에 관한 연구는 대체로 이러한 정도의 분석을 목표로 하고 있다[3][4][6]. 정보 검색 시스템에서 이와 같은 복합명사의 분석 정보를 이용하는 경우 보다 정확한 검색이 가능하게 된다.

본 연구의 목표는 복합명사구를 구성하는 명사들 사이의 수식(의존)관계를 파악하는 것이다. 이를 위해서 최근에 많이 성행하는 통계에 기반을 둔 패러다임을 이용한다. 먼저 복합명사 분석을 위한 통계 모델(statistical model)을 제시하고 이에 의거한 분석

[1] 이것은 명사 사이에 수식어-피수식어 또는 의존소-지배소의 관계를 파악하는 것으로 볼 수 있다.

기법을 제안한다. 복합명사 분석을 위해서는 임의의 두 명사 사이에 수식 관계가 얼마나 잘 발생하는가에 대한 계 정보가 필요하다. 그러나 통계 기반 분석에서 항상 문제를 일으키는 것이 데이터 부족문제(data sparseness problem)이다. 데이터 부족문제에 대해 본 논문에서 제안한 방법은 통계 정보가 존재하지 않는 명사의 경우에 대해서는 이 명사와 유사한 명사들을 개념적 계층구조를 가진 워드넷(WordNet)에서 추출하고 이 유사한 명사들에 대한 통계 정보를 대신 이용하고자 하는 것이다.

본 실험의 목표인 복합명사를 이루는 구성 명사들간의 수식관계를 파악한 결과 통계 정보가 있는 복합명사의 경우에 대해서는 정확도 면에서 최대 87.5%의 정확도를 보였다.

통계정보가 있는 경우만 분석할 경우 분석률은 44.9%로 매우 낮은 수치이다. 통계정보가 없는 (즉 데이타 부족 문제가 생긴) 경우에 워드넷을 이용하여 분석한 결과 94.5%까지 분석될 수 있음을 볼 수 있었다.

2 복합명사 분석을 위한 통계 모델

2.1 구성 명사간의 수식관계

n 개의 명사로 이루어진 영어 복합 명사 <N_1 N_2 ⋯ N_n>가 있다고 하자. 이 복합 명사에 대해 가능한 구조(즉 수식 구조)를 결정하는 것이 복합명사 분석의 목표이다. 결정된 구조는 해당 복합 명사를 이루는 명사들간에 수식관계를 나타내고 있다. 명사 N_i 와 N_j 사이에 수식관계를 갖는다는 것은 N_i 와 N_j 사이에 의존소와 지배소의 관계(즉, N_i 가 N_j 를 수식하는 관계)가 있음을 의미한다.

n개의 명사로 이루어진 영어 복합명사 <N_1 N_2 ⋯ N_n>의 수식 원칙은 다음과 같다:

- 지배소 유일의 원칙 : 각 명사는 오직 하나의 명사를 반드시 수식한다 (단, 마지막 명사는 지배소가 없다.)
- 지배소 후위의 원칙 : 수식하는 명사는 수식 받는 명사보다 앞에 위치한다.
- 교차 금지의 원칙 : 의존소와 지배소를 연결한 아크들끼리 교차할 수 없다.

하나의 수식구조는 위의 3 원칙을 만족하는 <의존소,지배소>쌍의 집합이라고 볼 수 있다.

2.2 수식구조의 선택 및 확률값

임의의 복합명사에 대해 생성 가능한 수식구조는 복합명사를 이루는 명사의 개수가 늘어남에 따라 증가한다. 이와 같은 경우, 그 중 하나의 수식구조를 분석결과로 선택해야 하며, 이를 위해 여러 가지 가능한 수식구조의 생성확률을 구해야 한다. 이를 위해 명사쌍내의 두 명사가 <의존소, 지배소>관계를 갖는가에 관한 통계정보를 이용하여야 한다. 복합명사에 대해 가능한 여러 수식구조 중에서 생성 확률값을 최대로 가지는 수식구조를 해당 복합명사의 구조로 결정하는 기법을 사용하고 자 한다.

(1) 아크의 확률 $P(N_i \rightarrow N_j \mid N_j)$

이것은 명사 N_j가 주어졌을 때 명사 N_i가 N_j를 수식할 확률을 의미한다.

$$P(N_i \rightarrow N_j \mid N_j) = \frac{C_{pair}(N_i,\ N_j)}{C_{modified}(N_j)}$$

- $C_{pair}(N_i, N_j)$: 말뭉치에서 명사 N_i가 N_j와 수식관계를 갖는 경우의 카운트.
- $C_{modified}(N_j)$: 말뭉치에서 명사 N_j가 임의의 명사로부터 수식을 받는 경우의 카운트.

(2) 주어진 복합명사내에서의 아크의 확률

$$P(N_i \rightarrow N_j \mid NP = N_1 N_2 \cdots N_n) = \frac{P(N_i \rightarrow N_j \mid N_j)}{\sum_{k=i+1}^{n} P(N_i \rightarrow N_k \mid N_k)}$$

where $1 \le i < j \le n$

길이가 n인 복합명사 NP= <N_1 N_2 ⋯ N_n>내에서 N_i가 N_j의 의존소가 될 확률은 말뭉치에서 N_j가 주어졌을 때 N_i가 N_j의 의존소가 될 확률값을 길이가 n인 복합명사 내에서 $i < k \le n$ 인 임의의 명사 N_k가 주어졌을 때, N_i가 N_k와 <의존소 → 지배소> 관계를 가질 수 있는 확률값의 총합으로 나눈 값이 된다.

(3) 복합명사의 수식구조의 확률

$$P(DT_i \mid NP = N_1 N_2 \cdots N_n) = \prod_{j=1}^{n-1} P(N_j \rightarrow N_{h_i(j)} \mid NP)$$

where $h_i(j)$: the index of the noun modified by Nj in the structure DT_i

$P(DT_i \mid NP= N_1\ N_2 \cdots N_n)$는 하나의 수식구조 DT_i에 대한 확률값이다. 주어진 복합명사가 NP= <N_1 N_2 ⋯ N_n>이고 모든 가능한 구조들의 집합 DT_{NP} = <DT_1, DT_2, ⋯, DT_m> 일 때, 확률값이 가장 높은 구조 DT를 선택하여 해당 복합명사의 분석 결과로 한다.

$$DT = \mathbf{argmax}_{DT_i} P(DT_i \mid NP)$$

2.3 통계 정보의 추출

임의의 복합명사에 대한 수식구조의 확률값을 구하기 위해서는 <의존소, 지배소> 관계를 가지는 명사쌍에

대한 통계정보가 마련되어 있어야 한다. 길이가 n인 복합명사에 있어서 마지막 두 명사인 $n-1$번째 명사와 n번째 명사간에는 확실한 <의존소, 지배소> 관계를 가지고 있다. 말뭉치로부터 명사쌍 통계정보를 구할 때 위와 같이 확실한 수식관계를 가지는 명사쌍에 대한 공기(co-occurrence)정보를 구한다. 또한 임의의 명사로부터 수식을 받는 명사에 대한 통계 정보를 추출하는 작업은 확실한 수식관계 명사쌍의 공기 정보를 구할 때 병행하여 이루어진다.

복합명사를 분석하기 위해 복합명사를 구성하는 명사들 간의 수식관계 확률값 $P(N_i \rightarrow N_j \mid N_j)$을 구해야 한다. 이를 위해 수식관계를 갖는 명사쌍 통계정보와 임의의 명사로부터 수식을 받는 명사에 대한 통계정보가 필요하다. 학습 말뭉치로부터 수식관계를 갖는 명사쌍을 추출하면서 수식 관계를 갖는 명사쌍에 대한 통계정보와 임의의 명사로부터 수식을 받는 명사에 대한 통계 정보화일을 구축하여 이용한다.

3. 데이터 부족문제

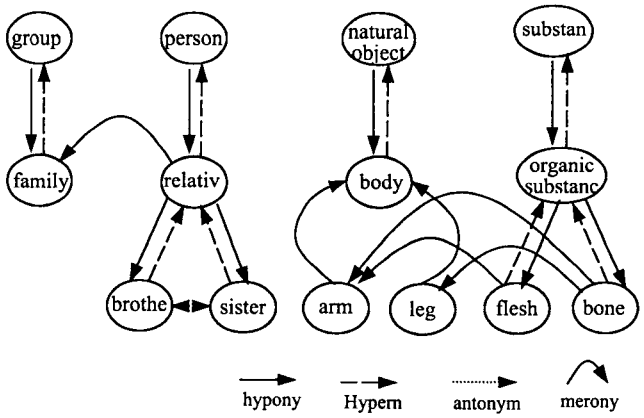
임의의 복합명사에 대해 생성 가능한 유도 트리의 생성 확률값을 알기 위해서는 말뭉치로부터 <의존소, 지배소> 관계를 갖는 명사쌍 통계정보를 가지고 있어야 한다. 그러나 통계 정보를 이용한 복합명사 분석은 학습 말뭉치의 크기가 충분히 크더라도 테스트 코퍼스 내의 명사쌍에 대한 통계 정보가 존재하지 않을 가능성은 항상 존재한다. 이와 같이 통계 정보가 존재하지 않는 경우를 데이터 부족문제(data sparseness problem)라 한다. 이를 해결하기 위해 워드넷 (WordNet) 이라는 계층구조를 사용함으로써 데이터 부족문제를 어느 정도 해소 시킬 수 있음을 보인다.

3.1 데이터 부족문제와 워드넷

임의의 복합명사 분석에 있어서 데이터 부족문제는 가능한 분석 구조의 생성 확률값이 모두 '0'인 경우에 데이터 부족문제가 발생했다고 가정하였다. 워드넷은 개념체계(ontology)의 하나로[5] 개념적 계층 구조를 이루고 있다. 워드넷의 계층구조는 신셋(synset) 이라는 노드들로 이루어져 있으며[2], 각각의 신셋들은 서로 hypernym[3], hyponym[4], antonym[5], meronym[6]과 같은 의미관계 포인터를 가지고 서로 연결 되어있다. 본 논문에서는 워드넷의 계층구조 중 명사 계층구조만을 사용하였으며, 여러 개의 의미관계 포인터 중 상위 의미관계 포인터 (hypernym) 와 하위 의미관계 포인터(hyponym)를 이용하였다. 워드넷의 명사 계층구조에는 총 94,473개의 명사들이 존재하고 있으며, 94,473개의 명사는 47,299개의 단일 명사(simple word)와 47,174개의 연어(collocation)로 이루어져 있다. 워드넷의 계층구조는 상위 계층에서 하위 계층으로 내려갈수록 포괄적 의미에서 구체적인 의미를 갖는다. 데이터 부족 문제가 발생한 경우, 본 논문에서의 해결 방법은 데이터 부족 문제를 발생시킨 명사를 대신할 대체 명사를 워드넷에서 충분히 많이 추출하여 대체 명사 집합간에 대체 명사쌍 조합을 만들어 데이터 부족문제를 발생시킨 명사쌍에 대한 통계정보 대신 이용하고자 하는 것이다.

워드넷은 최상위 신셋 노드에서 하위 신셋 노드로 내려갈 수록 신셋 노드는 포괄적 의미에서 세분화된 의미를 가지는 명사들을 포함하고 있다. [그림 1]과 같이 리프 신셋 노드를 제외한 신셋 노드들은 하위 포인터(hyponym)와 상위 포인터(hypernym)를 하나 이상 가지고 있다.[5]

데이터 부족문제가 발생했을 때 문제가 발생시킨 명사에 대해서 워드넷의 모든 신셋에서 해당 명사가 존재하는 신셋들을 찾아내고 신셋의 하위 포인터와 상위포인터를 적절히 이용함으로써 데이터 부족문제를 발생시킨 명사의 대체 명사들을 찾아냄으로써 데이터 부족문제를 해결하고자 한다.



[그림 1] 워드넷에서 synset간 의미관계

3.2 워드넷의 명사 계층구조와 개념 레벨

임의의 명사에 대한 유사어를 찾기 위해서는 워드넷의 명사계층을 검색하여야 하며 이를 위해서 워드넷을 이루는 파일들 중에서 noun.idx 화일과 noun.dat 화일을 이용하였다. noun.idx 화일에는 워드넷의 명사 계층구조에 존재하는 94,473개의 명사와 해당 명사가 다의어일 경우 다의어 정보를 함께 가지고 있다.

---

[2] 비슷한 의미를 갖는 단어들을 모아 놓은 유사어 집합
[3] 상위 의미관계 포인터
[4] 하위 의미관계 포인터
[5] 반대 의미관계 포인터
[6] 멤버 의미관계 포인터

noun.dat 화일은 명사 계층구조에 존재하는 모든 신셋에 대한 정보를 가지고 있다.

워드넷을 이용하여 데이터 부족문제에 대한 충분한 대체 명사들을 추출하기 위해서는 적절한 개념 레벨을 정하여, 각 개념 레벨에 존재하는 신셋들의 명사들을 중복 없이 추출하여야 한다. 실험에서는 워드넷에 적용하는 개념 레벨을 다음과 같이 5개의 레벨로 정하여 실험을 하였으며, 개념 레벨은 통계 정보가 존재하지 않는 명사쌍을 이루는 두개의 명사 모두에 대해서 동일 레벨을 적용하였다 (그림 2 참조).

• 개념 레벨 1

개념 레벨 1은 워드넷에서 데이터 부족문제를 발생시킨 명사가 존재하는 신셋을 모두 찾아내는 것이다. 임의의 명사가 존재하는 신셋 하나는 그 명사가 가지는 하나의 의미를 내포하고 있다. 두 개 이상의 의미를 가지는 다의 명사(polysemous noun)에 대한 의미 정보는 말뭉치로부터 알 수 없기 때문에 워드넷에서 대체 명사를 찾기 위해서는 명사가 가지는 의미의 수에 관계없이 해당 명사가 존재하는 신셋을 모두 찾게 하였다. 데이터 부족문제를 발생시킨 명사에 대해 워드넷에서 찾아낸 모든 신셋들에 존재하는 명사들을 하나의 대체 명사 집합으로 묶는다. 이 때 대체 명사집합을 이루는 각각의 명사들은 집합 내에서 유일하다.

• 개념 레벨 2

개념 레벨 2는 개념 레벨 1에서 찾아낸 신셋들이 가지고 있는 명사들과 신셋들이 가지는 하위 의미관계 포인터(hyponym)를 이용한다. 하위 의미관계 포인터는 해당 신셋이 리프 신셋 노드(leaf synset node)가 아닌 이상 신셋마다 하나 이상의 하위 의미관계 포인터를 갖는다. 따라서 워드넷의 개념 레벨 2에 존재하는 명사들의 수는 개념 레벨 1에 존재하는 명사들의 수보다 최소한 같거나 많다.

• 개념 레벨 3

개념 레벨 3는 개념 레벨 1에서 찾아낸 신셋들이 가지는 명사들과 상위 의미관계 포인터(hypernym)를 이용한다. 상위 의미관계 포인터는 해당 신셋(synet)이 루트(root)가 아닌 이상 신셋마다 하나 이상의 상위 의미관계 포인터를 갖는다. 개념 레벨 2와 마찬가지로 개념 레벨 3에 존재하는 명사들의 수는 개념 레벨 1에 존재하는 명사들의 수보다 최소한 같거나 많다.

• 개념 레벨 4

개념 레벨 4에서는 의미를 무시한 개념 레벨 1을 한층 더 개념적으로 확장시킨 것으로 레벨 1에서 가지는 상위 신셋의 하위 의미관계 포인터를 이용한 것이다. 또한 레벨 1에서 추출된 명사들과 레벨 3에서 추출된 명사들도 모두 이용하게 된다. 이는 한 개념 레벨을 이용하기 위해 거쳐온 개념 레벨의 대체 명사 정보를 잃지 않기 위해서다. 개념 레벨 4를 이용할 경우, 버리지 않고 이용되는 명사들은 개념 레벨 1과 3에서 추출된 명사들이다. 이 때 중복되는 단어들은 모두 제거된다.

• 개념 레벨 5

개념 레벨 5는 개념 레벨 4의 모든 하위 의미관계 포인터를 이용한 것으로 한층 더 구체적인 명사들이 추출될 수 있다. 개념 레벨 4에서와 마찬가지로 개념 레벨 5에 도달하기 위해 거친 레벨들의 명사 정보는 버리지 않고 이용한다.

통계정보가 존재하지 않는 명사쌍 $\langle N_i \rightarrow N_j \rangle$의 각각의 N_i와 N_j에 대해 [그림 4.2]와 같이 워드넷의 개념 레벨을 적용하여 추출한 대체 명사집합이 sw_set_i, sw_set_j 일 때,

sw_set_i : similar word set of N_i
sw_set_j : similar word set N_j
$sw\_set_i = \{sw_1\_N_i , sw_2\_N_i , \cdots sw_N\_N_i \}$
$sw\_set_j = \{sw_1\_N_j , sw_2\_N_j , \cdots sw_M\_N_j \}$

생성가능한 대체 명사쌍 조합은 다음과 같다.

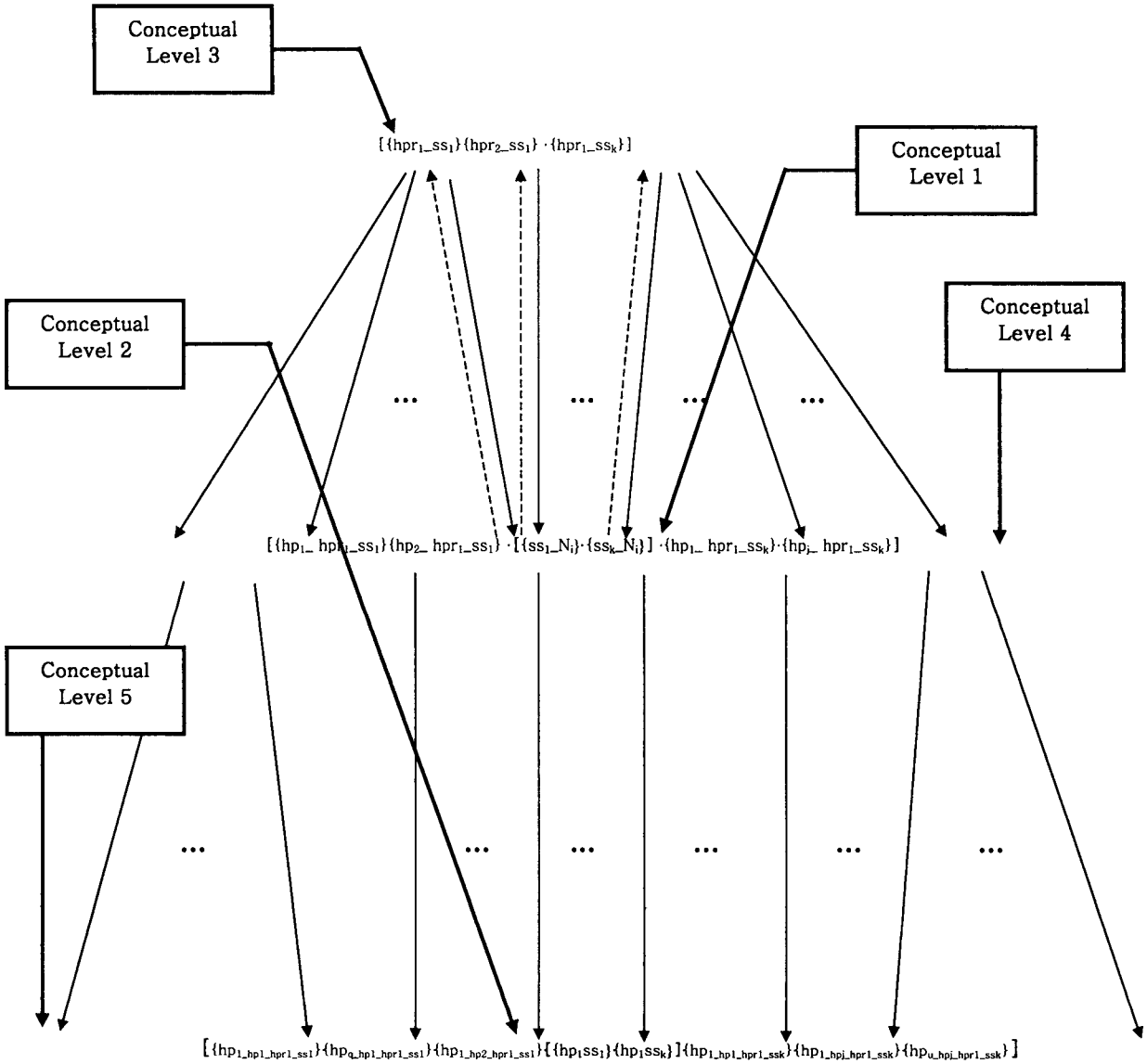
$pairset_{ij}$: possible noun pair set from N_i and N_j
$pairset_{ij} = \{\langle sw_1\_ N_i , sw_1\_N_j \rangle, \langle sw_1\_ N_i , sw_2\_N_j \rangle, \cdots \langle sw_1\_ N_i , sw_M\_N_j \rangle,$
$\langle sw_2\_ N_i , sw_1\_N_j \rangle, \langle sw_2\_ N_i , sw_2\_N_j \rangle, \cdots \langle sw_2\_ N_i , sw_M\_N_j \rangle,$
$\cdots$
$\langle sw_N\_ N_i , sw_1\_N_j \rangle, \langle sw_N\_ N_i , sw_2\_N_j \rangle, \cdots \langle sw_N\_N_i , sw_M\_N_j \rangle\}$

3.3 대체 확률값

수식관계 확률 $P(N_i \rightarrow N_j \mid N_j)$ 을 구해야 할 때 $C_{pair}(N_i, N_j)$, 즉 N_i가 N_j를 수식한 경우의 카운트가 필요하다. 그런데 $C_{pair}(N_i, N_j)$가 학습 말뭉치에서 마련한 수식관계 명사쌍 통계정보에 존재하지 않는다면 데이타 부족 문제가 생긴 것이다. 이 문제를 해결하기 위해서 우리는 워드넷을 이용한다. 문제가 되는 명사쌍 $\langle N_i \rightarrow N_j \rangle$에 대하여 앞에서 설명한 것처럼 워드넷의 개념레벨을 적용하여 대체 명사쌍을 구한다. 대체 명사쌍의 수가 n×m개 라고 가정하자. 그러면 대체 명사쌍을 이용한 수식관계 카운트는 다음과 같이 구한다.

$$RC_{pair}(N_i, N_j) = \frac{1}{n \times m} \times \sum_{k=1}^{n} \sum_{l=1}^{m} C_{pair}(sw_k\_N_i, sw_l\_N_j)$$

또한 수식을 받는 명사 N_j에 대한 대체 통계정보를 아래와 같은 식에 의해 구할 수 있다.



[그림 2] 통계 정보가 존재하지 않는 명사쌍 <N_i → N_j> 의 N_i에 대한 개념레벨 적용

ss	:	synset
$ss_1\_N_i$	:	1st synset of N_i
hp	:	hyponym
hpr	:	hypernym
$hpr_1\_ss_1$	:	1st hypernym synset of ss_1
$hp_1\_hpr_1\_ss_1$	:	1st hyponym synset of $hpr_1\_ss_1$
$hp_{1\text{-}hp1\text{-}hpr1\text{-}ss1}$	:	1st hyponym synset of $hp_1\_hpr_1\_ss_1$

$$RC_{modified}(N_j) = \frac{1}{m} \times \sum_{l=1}^{m} C_{modified}(sw_l\_N_j)$$

위 수식에서 이용된 텀의 의미는 다음과 같다:

$sw_k\_N_i$: k^{th} similar word of N_i
$RC_{pair}(N_i, N_j)$: replacement of $C_{pair}(N_i, N_j)$
$RC_{modified}(N_j)$: replacement of $C_{modified}(N_j)$

위의 두식에 의해 구해진 대체 통계정보를 명사 N_i가 명사 N_j를 수식할 확률 $P(N_i \rightarrow N_j \mid N_j)$을 구하는데 이용한다.

4. 실험 및 검토

실험을 위한 말뭉치로는 2,499개의 화일들로 이루어진 Wall Street Journal 말뭉치(총 1,289,201개의 단어들로 이루어짐)를 사용하였다. 통계 명사쌍 정보를 추출하기 위한 학습 말뭉치로는 2,499개의 WSJ 화일 중 2,249개의 화일(1,144,110개의 단어)을 사용하였고, 실험 말뭉치로는 250개의 화일(145,091개의 단어)을 사용하여 분석 대상인 명사 3개로 이루어진 복합명사를 추출하였다. 분석 대상인 복합명사는 총 222개가 추출되었으며, 222개의 실험 복합명사 중 실제 복합명사가 아닌 58개의 오류 복합명사는 제거되었다. 실험은 58개의 오류 복합명사를 제외한 164개의 실험 복합명사에 대해 이루어졌다.

데이터 부족문제가 발생했을 경우 워드넷을 사용하여 문제를 해결하고자 한다. 이 때, 의미적으로 불필요한 명사가 섞이는 것을 피하기 위해 [그림 2]의 5단계의 개념 레벨을 독립적 개념 레벨과 순차적 개념 레벨에 적용하였다.

시스템의 성능을 나타내는 척도로서 분석률은 분석된 복합명사에 대하여 올바른 분석을 했는가에 대한 판단없이 분석이 가능했는가에 대한 통계치를 말하며, 정확률은 분석된 복합명사 가운데 올바르게 분석된 복합명사의 통계치를 말한다.

$$\text{분석률} = \frac{\text{분석에 성공한 복합명사의 개수}}{\text{분석할 복합명사의 개수}}$$

$$\text{정확률} = \frac{\text{올바르게 분석한 복합명사의 개수}}{\text{분석에 성공한 복합명사의 개수}}$$

데이터 부족문제 극복률이란 분석에 실패한 복합명사들에 대하여 워드넷을 적용할 경우 분석에 성공한 통계치를 의미하며, 데이터 부족문제 정확률이란 워드넷을 적용하여 분석해낸 복합명사 가운데 올바르게 분석된 복합명사의 통계치를 의미한다.

$$\text{데이타부족문제 극복률} = \frac{\text{워드넷을 적용할 경우 분석에 성공한 복합명사의 갯수}}{\text{분석에 실패한 복합명사의 갯수}}$$

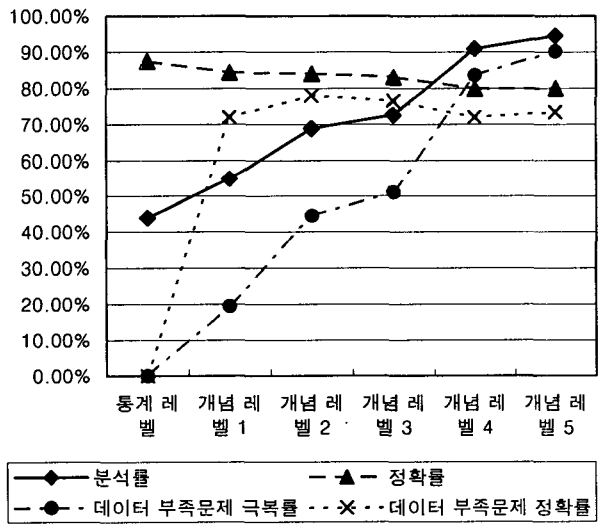
$$\text{데이타부족문제 정확률} = \frac{\text{올바르게 분석한 복합명사의 갯수}}{\text{워드넷을 적용할 경우 분석에 성공한 복합명사의 갯수}}$$

4.1 독립적 개념 레벨

독립적 개념 레벨은 복합명사를 분석할 때 데이터 부족문제가 발생하여 분석이 불가능할 경우, 워드넷의 5개의 개념 레벨들을 각각 독립적으로 적용하여 유사 명사를 추출하는 방법으로 각 개념 레벨에서의 분석률과 정확률, 데이터 부족문제 극복률과 데이터 부족문제 정확률을 측정하였다. 164개의 복합명사에 대한 독립적 개념 레벨을 적용하여 [표 5-1]과 같은 분석결과를 보이며 그래프는 [그림 3]과 같다[7].

[표 1] 독립적 개념 레벨에서의 분석결과

레벨	분석률	정확률	데이터 부족문제 극복률	데이터 부족문제 정확률
통계 레벨	43.9%	87.5%	0.00%	0.00%
개념 레벨 1	54.9%	84.4%	19.5%	72.0%
개념 레벨 2	68.9%	84.0%	44.5%	78.0%
개념 레벨 3	72.6%	83.1%	51.0%	76.5%
개념 레벨 4	90.9%	79.9%	83.6%	72.0%
개념 레벨 5	94.5%	80.0%	90.2%	73.4%



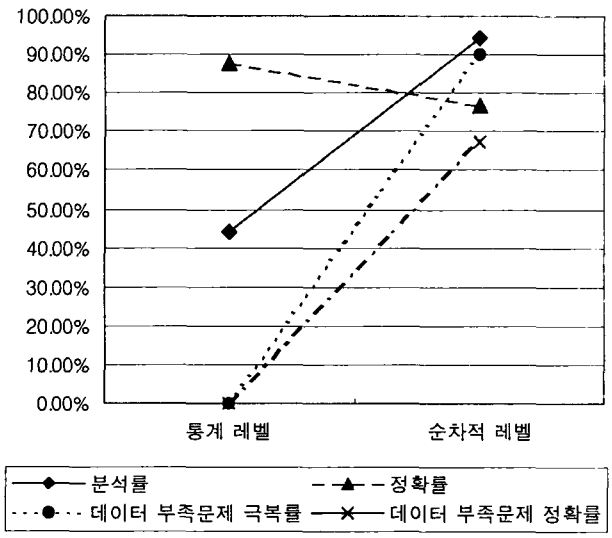
[그림 3] 독립적 개념 레벨에서의 분석률과 정확률

4.2 순차적 개념 레벨

순차적 개념 레벨은 독립적 개념 레벨모델과는 달리 각각의 개념 레벨에서 복합명사 분석을 시도하는 것이 아니라 한 레벨이 실패할 경우에 분석을 멈추지 않고 개념 레벨 1부터 개념 레벨 5까지 적용하는 것이다.

[7] 통계레벨이란 워드넷을 사용하지 않고 순수하게 통계정보만을 이용한 경우를 말한다.

개념 레벨 5까지 적용하고도 분석에 실패한 경우에만 실제 복합명사 분석에 있어서 데이터 부족문제가 해결될 수 없다고 보고 분석을 멈추게 된다.



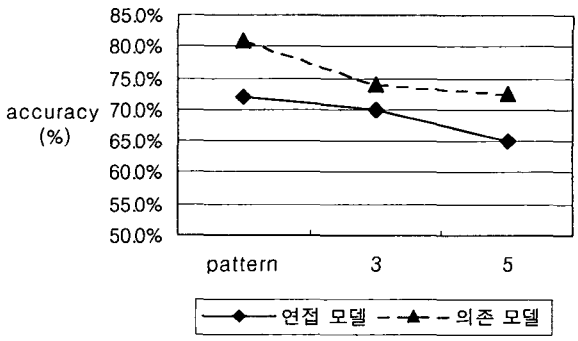
[그림 4] 순차적인 개념 레벨에서 분석률과 정확률

[표 2] 순차적 개념레벨 모델에서의 분석결과

레벨	분석률	정확률	데이터 부족문제 극복률	데이터 부족문제 정확률
통계 레벨	43.9%	87.5%	0.00%	0.00%
순차적 레벨	94.5%	76.8%	90.2%	67.4%

164개의 복합명사를 순차적 개념 레벨을 이용하여 분석한 경우, [표 2]에서와 같이 분석률이 94.5%까지 향상 되었고 정확률은 76.8%임을 보이며 [그림 4]와 같은 그래프로 정리된다.

Mark Lauer가 제안한 의존 모델은 로제 시소로스의 1043개의 카테고리 (category)간의 개념 결합도 (conceptual association)를 복합명사 분석에 이용하였다. Lauer는 1043개의 카테고리 중에서 임의의 두개 카테고리에 대한 개념 결합도를 Grolier's encyclopedia 말뭉치에 대해 먼저 구하고, 말뭉치에서 추출한 3개의 명사로 이루어진 복합명사들 중 로제 시소로스에 존재하는 명사들로만 이루어진 복합명사를 분석하는데 이용하였다.



[그림 5] 창의 크기에 따른 Lauer모델의 정확률

Lauer는 3개의 명사로 이루어진 244개의 복합명사에 적용한 경우 최대 77.5%의 정확률을 보였고 몇 가지 성능 향상을 통하여 최대 81%까지 정확률을 향상시켰다. Lauer가 제안한 의존 모델은 로제 시소로스의 카테고리간의 개념 결합도를 이용하여 결합도가 높은 카테고리들에 존재하는 명사들간의 결합도 또한 유사함을 보였다. 하지만 분석 대상 명사를 로제 시소로스에 존재하는 명사들로 제한함으로써 로제 시소로스에 존재하지 않는 명사들에 대한 분석에는 어떤 결과가 나올지 알 수 없다. 왜냐하면 로제 시소로스에 존재하지 않는 명사들간의 개념 결합도를 알지 못하기 때문에 새로운 카테고리를 추가하여 개념 결합도를 다시 구하거나 임의의 카테고리에 명사들을 포함시켜야 하는 제약이 존재하게 된다. [그림 5]에서 Lauer가 제안한 의존모델과 연접모델의 3개로 이루어진 복합명사에 대한 분석결과를 공기정보를 추출하는 창의 크기에 따라 나타내었다.

우리의 결과는 여러 면에서 Lauer 의 결과보다 향상된 것으로 판단된다.

5. 결론

본 논문은 영어 복합명사 분석을 위한 통계모델을 제안하였다. 통계정보를 기반으로 두 명사 사이에 수식관계가 존재할 확률을 구할 수 있다. 이 확률값을 기반으로 하여 주어진 복합 명사가 가지는 가능한 구조들 (즉 구성 명사들 사이의 수식관계)에 대한 확률을 구한다. 이 중에서 가장 높은 확률을 갖는 구조를 분석결과로 선택한다.

통계정보를 이용하는 언어 분석 기법에서 항상 등장하는 것이 데이터 부족 문제이다. 본 논문에서는 이 문제를 극복하기 위해 개념체계의 하나인 워드넷(WordNet)을 이용할 것을 제안하였다. 제안된 기법은 데이터 부족문제를 일으킨 명사쌍을 구성하는 두 명사에 유사 명사들을 워드넷으로부터 추출하여 유사 명사쌍들에 대한 통계정보를 구하고 이것을

원래의 명사쌍의 통계정보로 대체하여 사용하는 것이다.

본 논문에서 제안된 기법은 다음과 같은 특징을 가지고 있다. 첫째로, 우리가 제안한 기법은 데이타 부족 문제가 없는 복합명사에 대한 분석에 있어 정확도가 기존의 연구에 비해 더 높다. 둘째, 기존의 연구에서는 데이터 부족문제가 있는 복합명사를 심도 있게 다루지 않았지만 우리의 경우에는 워드넷을 이용하여 대체 통계정보를 구한 다음 분석을 시도한다. 여기에는 두 가지 방법이 있다. 독립적인 개념모델을 적용할 경우 분석률이 48%에서 94%까지 향상되었고, 정확도는 각각의 개념 레벨에서 최소 79.8%에서 최대 87.5%까지 높은 정확도를 보였다. 순차적인 개념모델 기법을 적용할 경우에는 분석률은 43.9%에서 최대 94.5%까지 향상되었으며, 76%의 정확도를 보였다.

이로써 통계정보를 사용한 복합명사 분석에 있어서 데이터 부족문제가 발생할 경우 워드넷과 같은 개념 체계를 이용하면 통계정보가 있는 경우와 유사한 정확도를 보였으며, 데이터 부족문제를 극복해 낼 수 있음을 보였다.

참고 문헌

[1] Y. Arens, J. Granacki, and A. Parker, "Phrasal Analysis of Long Noun Sequences," Proceedings of the 25th Annual Meeting of the ACL, pp.59-64, 1987.

[2] T. Finin, "the Semantic Interpretation of Nominal Compounds," Proceedings of First Annual National Conference, AAAI, pp. 310-321, 1980.

[3] M. Lauer, "Conceptual Association for Compound Noun Analysis," in Proceedings of 32nd Annual Meeting of the ACL, pp. 337-339, 1994.

[4] M. Lauer, "Corpus Statistics Meet the Noun Compound: Some Empirical Results," in Proceedings of 33rd Annual Meeting of the ACL, pp. 47-54, 1995.

[5] G. A. Miller, "WordNet: A Lexical Database for English," Communication of the ACM, Vol. 38. No 11. pp 39-41, November 1995.

[6] P. Resnik, and M. Hearst, "Structural Ambiguity and Conceptual Relations," Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, June 22, Ohio State University, pp. 58-64.

[7] J. Pustejovsky, S. Bergler and P. Anick, "Lexical Semantic Techniques for Corpus Analysis," Computational Linguistics, Vol. 19(2), Special Issue on Using Large Corpora pp. 331-358, 1993.

[8] 강승식, "한국어 복합명사 분해 알고리즘," 정보과학회지 논문(B), 제 25권 제 1호, 1998년 1월.

[9] 윤보현, 임희석, 임해창, "통계 정보를 이용한 한국어 복합명사의 분석 방법," 한국정보과학회 봄 학술발표 논문집, pp.925-928, 1995.

[10] 채영숙, 권혁철, "말뭉치로부터 추출된 통계 정보를 활용한 한국어 복합명사 분석," 인지과학, 제8권 2호, pp. 101-108, 1997.