WordNet을 이용한 한국어 명사 의미지표 자동 구축

이지선[1], 전현경, 김남수, 이용석
전북대학교 컴퓨터과학과
{jslee,hgchon,nskim}@cypher.chonbuk.ac.kr, ylsee@moak.chonbuk.ac.kr

Automatic Construction of Korean Noun Semantic-Marker using WordNet

Ji-Sun Lee, Hyun-Gyung Chon, Nam-Su Kim, Yong-Seok Lee
Dept. of Computer Science, Chonbuk National University

요 약

컴퓨터는 자연언어로 된 문장을 올바르게 이해하기 위해 의미지식을 필요로 하며 이러한 의미지식을 정확하게 구축하기 위해서는 수작업을 필요로 한다. 그러나 수작업에 의한 의미지식 구축은 많은 비용과 시간을 필요로 하고, 작성자의 주관이 개입되며, 응용 도메인에 따라 의미지표 테이블이 수정되면 의미지표 사전의 재구축이 불가피하다. 이러한 문제점을 해결하기 위해 본 논문에서는 영어 WordNet과 한영 사전을 이용한 한국어 명사 의미지표 사전의 자동 구축 방법을 제안한다.

1. 서론

사람들은 자연언어로 된 문장을 이해하는데 거의 무의식적으로 의미분석을 하여 모호성 해결이나 개념 이해를 하는 반면 컴퓨터는 자연언어 처리를 효율적으로 수행하기 위해 의미지식을 참조한다[6]. 이러한 의미지식은 정보검색, 영한기계번역과 같은 자연언어처리 분야에 많이 이용된다. 예를 들어, 정보검색에서 사용자는 찾고자하는 정보를 얻기 위해 정보와 관련이 있다고 판단되는 키워드들을 정보 검색 질의어로 사용한다. 그러나 검색 시스템은 사용자가 원하는 정보를 찾기 위해 의미지표를 사용해서 사용자가 선택한 검색 질의어뿐만 아니라 선택한 질의어와 유사한 단어로까지 질의어를 확장한다. 또한 영한 기계번역 변환 과정 중, 대역어 선정방법으로 일반적으로 의미를 분명히 나타내기 위하여 습관적으로 동일한 형태로 사용되는 단어들의 연속을 의미하는 연어(collocation)정보를 사용하는 방법이 있다[10]. 이 방법을 사용할 경우 정확한 대역어 선정을 위해서는 영어 표현에 해당되는 모든 연어 정보를 사전에 입력해야하는데, 이것은 불가능한 작업이다. 이런 문제점을 해결하기 위해 사전에 들어있는 연어들의 공통적인 의미지표 구해서 사전에 입력하는 방법이 있다.

위와 같이 자연언어처리 분야에 사용되는 의미지표를 구축하는 방법 중에서 수동으로 구축하는 방법은 많은 시간과 비용이 소요되고, 작성자의 주관이 개입되며 한번 구축되어진 의미지표는 응용 도메인마다 의미지표의 수정이 불가피하다. 따라서 의미지표를 자동으로 구축하는 방법이 필요하다. 그러나 의미지표를 자동으로 구축하는 연구보다는 의미지표보다 더 광범위한 시소러스를 자동으로 구축하는 연구가 많이 시도된다.

시소러스를 자동으로 구축하는 연구 방법으로는, 크게 사전을 이용하는 방법과 대역어 사전(bilingual dictionary)과 이미 구축되어진 시소러스를 이용해서 새로운 언어의 시소러스를 구축하는 방법이 있다[3,4,5,6,7,8].

시소러스 자동 구축은 광범위한 작업이고 응용 도메인에 따라 시소러스를 직접 사용하기에는 불가능한 경우도 있다. 그래서 본 논문에서는 광범위한 시소러스 구축대신에, WordNet을 이용하여 의미지표 테이블을 구축한 후, 한영 사전을 이용하여 한국어 명사에 대한 영어 키워드를 추출하고, 영어 키워드를 WordNet과 의미지표 테이블, 확률을 이용해서 한국어 명사에 적합한 의미지표 사전을 자동 구축하는 방법을 제안한다.
2장에서는 WordNet과 의미지표 테이블 구축에 관하여 기술하고, 3장에서는 한영사전을 이용한 한국어 의미지표 사전 구축에 관하여 기술한다.

2. WordNet과 의미지표 테이블 구축

1990년 Princeton 대학에서는 단어의 의미, 상위 개념, 구성개념, 반대어, 관련어 등을 포함하는 영어의 WordNet 시스템을 구축했다[6]. 영어 WordNet은 전산 처리에 적합하도록 전산학자와 언어심리학자들이 공동으로 제작한 일종의 시소러스로서 기본 골격은 어휘개념인데, 동의어 집합(synonym set, synset)으로 표현되고 있으며, 이 동의어 집합들 간의 상·하위개념 관계는 계층 구조로서 표현된다[6]. WordNet은 명사뿐만 아니라 동사, 형용사, 부사의 엔트리에 Isa 계층구조를 가지고 있는데 동사는 크게 15개의 범주로 분류하고 명사는 크게 25개의 범주로 분류하고 있다[1].

WordNet의 명사 25개 범주 안에 대부분의 일반 명사들이 모두 포함될 수 있다는 분석 하에 본 연구에서는 의미지표 테이

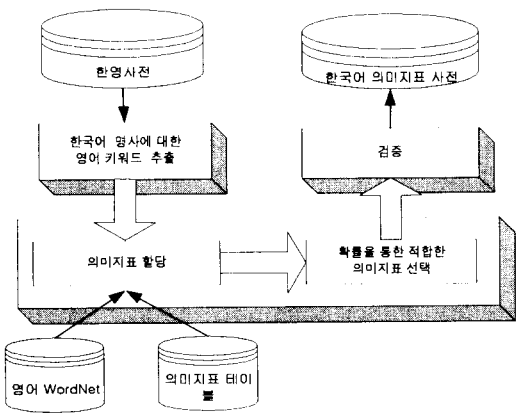
블을 구축 시 WordNet의 명사 25개 범주를 사용한다. 의미지표 테이블은 WordNet의 명사 25개 범주에서 각 응용 도메인에 따라 의미지표 레벨을 선택하여 사용할 수 있도록 2단계의 의미지표 레벨로 구축하였으며 1-level은 19개, 2-level은162개로 구성되어 있다.

- person#1	- time#6
- animal#1	- attribute#2
- plant#2	- relation#1
- artifact#1	- food#1
- substance#1	- natural_object#1
- location#1	- feeling#1
- event#1	- act#2
- group#1	- state#4
- cognition#1, knowledge#1	
- natural_phenomenon#1, nature#6	
- measure#3, quantity#1, amount#3	

[표 1] 1-level 19개 의미지표 테이블

3. 한국어 의미지표 사전 자동 구축

본 연구에서 한국어 의미지표 사전을 자동 구축하기 위해 [그림 1]과 같은 과정을 거친다. 첫 번째 과정으로는, WordNet을 이용하여 한국어 명사에 적합한 의미지표 테이블을 구축한다. 두 번째 단계로서 WordNet을 이용하여 구축한 영어 의미지표 테이블을 이용하기 위해, 한영사전을 이용하여 한국어에 해당하는 영어 키워드를 추출한다. 세 번째 단계로는, WordNet을 이용하여 한국어에 대한 영어 키워드들의 상위어들 중, 의미지표 테이블에서 매칭 되는 의미지표를 할당한다. 네 번째 단계로는 세 번째 단계에서 구한 의미지표들 중에서 적합하지 않은 의미지표들이 할당되는 경우가 발생하므로 의미지표에 확률을 적용하여 적합한 의미지표를 선택한다. 마지막 단계로 수작업에 의한 검증을 통하여 한국어 의미지표 사전을 자동 구축한다.



[그림 1] 한국어 명사 의미지표 할당 과정

3.1 한국어 명사에 대한 영어 키워드 추출

한국어 명사에 의미지표 할당 시, WordNet을 이용하려면 한국어에 해당하는 영어 대역어를 필요로 한다. 따라서 본 연구에서는 아래 예제와 같이 한국어 단어가 동음이의어인 경우 각각의 동음이의어들이 독립된 엔트리로 구성된 한영 사전을 이용한다.

말 1 : 말과에 속하는 짐승
(a colt/a horse..)
말 2 : 장기· 윷 등에서 사용되는 패
(a chessman/a piece....)
말 3 : 곡식· 가루 따위의 분량을 되는데 쓰이는 그릇
(a unit of measure)
말 4 : 사람의 생각이나 느낌을 표현하고 전달하는 음성기호
(a language/....)

한영 사전으로부터 영어 대역어를 추출 시, 몇 가지 문제점이 발생한다.
첫 번째, 한국어 단어에 영어 구(phrase)가 매칭 되는 경우, 구에 속한 여러 명사들 중에 어떤 명사를 키워드로 선택하는 것이다. 이런 경우 파싱을 통해 명사구의 head를 한국어 명사에 대한 영어 키워드로 선택한다. 그 결과 아래 예제의 *가래떡*의 경우 명사구의 헤드로서 *cake가* 선택된다.

가래떡 : a long and slender rice cake
rice cake in form of rounded stick

그러나 위의 방법을 아래 예제의 *가부장*에 적용했을 경우, 키워드 추출 시 단순히 명사구의 헤드를 추출할 수 없는 경우 즉, 어떤 명사를 키워드로 추출해야하는지 모호한 경우가 발생한다.

가부장 : a head of a family

두 번째, 한국어 단어에 대하여 추출된 영어 키워드가 다의어인 경우 한국어와 영어의 의미적 차이에 기인하여 잘못된 의미지표가 할당되는 문제점이 있다. 아래의 예제에서처럼 한국어 *계산서*에 해당하는 영어 키워드 account는 예금 구좌라는 의미로, *bill*은 *지폐, 법안* 등의 의미로도 쓰인다.
따라서 한국어 *계산서*의 의미지표로서 계산서에 해당되는 의미지표 뿐만 아니라 *지폐, 법안* 등의 의미지표도 할당되는데, 이런 경우의 해결 방법은 3.2절에서 논의한다.

계산서 : an account /a check /a bill /a tally card

세 번째, 한국어 단어에 대하여 영어 키워드가 하나인 경우 영어 키워드가 다의어인지 아닌지를 구분할 수 없는 문제가 발생한다. 아래의 예제에서처럼 한국어 *유아*에 해당하는 영어 키워드 *germ*은 *유아*라는 뜻과 *병원균*이라는 서로 다른 2가지의 의미를 갖는다.
이런 경우는 확률로서도 적절한 의미지표를 할당할 수 없기 때문에, 수작업을 통한 검증이 필요하다.

유아 : a germ

3.2 의미지표 할당과 적합한 의미지표 선택

WordNet을 이용하여 한영사전을 통해 추출된 영어 키워드들의 상위어들을 추출한 후, [표 1]의 한국어 의미지표테이블과 매

칭 되는 의미지표들을 추출하면 아래의 예제와 같은 결과를 얻는다.

```
말 1
horse : animal#1 artifact#1 group#1
colt : animal#1 artifact#1
filly : animal#1
foal : animal#1
mare : animal#1 location#1
pony : animal#1 relation#1 artifact#1
stallion : animal#1
stud : person#1 artifact#1 act#2
```

그러나 위의 예에서처럼, *말1*에 해당하는 영어 키워드가 여러 개일 경우, 영어 키워드 pony의 의미지표 중 relation#1처럼 한국어의 의미와 영어의 의미적 차이에 기인하여 적절하지 못한 의미지표가 할당된다. 이런 문제점을 해결하기 위해 영어 키워드에 할당된 의미지표 중, 공통적으로 포함된 의미지표를 우선적으로 선택하였다.
공통된 의미지표를 선택하기 위해 한영사전에 한국어에 대한 영어 키워드들이 여러 개 존재할 경우 가장 많이 사용되는 영어 키워드가 앞에 나온다는 가정 하에 아래의 수식을 이용한 확률을 사용한다.
아래의 수식을 통해 각 의미지표에 경험을 통해서 얻은 임계값 0.333이상을 갖는 의미지표만을 최종 의미지표로서 할당한다.

n : 키워드 수

SM_i : 각 키워드에 대한 의미지표 집합 $(0 < i < n)$

$$S = \bigcup_{i=1}^{n} SM_i$$

$$\forall w \in S$$

$$f(w) = \frac{2}{n+1} \cdot \sum_{i=1}^{n} \frac{n-(i-1)}{n}$$

말1에 대하여 각 의미지표에 더하여 계산된 확률은 다음과 같다.

```
말 1 :
[sem_prob]
animal#1 : 1.000        artifact#1 : 0.472
group#1 : 0.139         location#1 : 0.111
relation#1 : 0.083      person#1 : 0.028
act#2 : 0.028
```

위의 결과에서 임계값이하의 확률을 갖는 group#1, location#1, relation#1, person#1, act#2등의 의미지표는 최종 결과에서 제외되므로 다음과 같은 최종결과를 얻을 수 있다.

```
말 1
  : horse/colt/filly/foal/mare/pony/stallion/stud
  ==> animal#1 artifact#1
```

4. 결론 및 향후 연구

한국어 명사 36,000여 개에 대한 의미지표를 자동 구축한 후, 검증 결과 72%의 정확도를 가졌다. 그러나 한국어에만 있는 고유명사, 사자 성어 등을 제외할 경우에는 더 높은 정확도를 얻을 수 있다.
의미지표의 수동 구축은 많은 시간과 비용을 소요되며, 작성자의 주관이 개입되고 응용 도메인에 따라 수정이 불가피하다. 따라서 본 연구에서 제안한 한국어 명사 의미지표 자동 구축 방법을 각 응용 분야에 적용하기 위하여 적합한 의미지표 테이블을 구성한 후, 한국어 명사 의미지표 사전을 구축하여 적용하였다. 그 결과 한국어 명사 의미지표 사전을 한국어 처리에서의 모호성(ambiguity) 해결과 기계번역시 올바른 대역어 선택과 올바른 구문 분석 선택에 도움이 되었다.
한국어 명사에 대한 의미지표 사전을 구축시 한국어에 대한 영어 키워드 추출의 정확성이 중요하다. 따라서 향후 한국어에 대한 영어 키워드를 추출 시, 아래의 예제와 같이 한국어에 대한 영어 키워드로 복합 명사를 갖는 경우 어떤 명사를 한국어에 대한 영어 키워드로 추출해야 적합한 의미지표가 할당될 것인가에 대하여 고려해야 한다.

가루 치약 : tooth power
가속 전압 : acceleration voltage

[참고 문헌]

[1] Miller G.A, Beckwith R., Fellbaum C., Gross D. and Miller K., "Intorduction to WordNet : An On-Line Lexical Database." in Five Papers on WordNet, CSL report, Cognitive Science Laboratory, Princeton University, 1993

[2] Atserias J., Climent S., Farreras J., Rigau G. and Rodriguez H., Combining Multiple Methods for the Automatic Construction of Multilingual WordNets, In proceeding of the Conference on Recent Advances on NLP, 1997

[3] Benitez L., Cervell S., Escudero G., Lopez M., Rigau G. and Taule M. Methods and Tools for building the Catalan WordNet. In workshop Language Resources for European Minority Languages at LREC'98, 1998

[4] German R., Disambiguating biligual nominal entries against WordNet. Workshop On The computational Kexicon-ESSLLI95, 1995

[5] Bentiex L., Gervell S., Escudero G., Lopez M., Rigau G. and Taule M., Methods and Tool for building the Catalan WordNet. In Workshop Language Resources for European Minority Languages at LERC'98, 1998

[6] 문유진, 한국어 명사를 위한 WordNet의 설계와 구현, 정보과학회논문지(c) 제2권 제4호, 1996

[7] 이창기, 이근배, WordNet을 이용한 한국어 시소러스 자동 구축, 한글 및 한국어 정보처리, 1999

[8] 문유진, 김영택, 한국어 명사의 hypernym 자동 추출 방법, 한국 정보과학회 '94 가을 학술 발표 논문집 A, 제21권 2호

[9] 이호석, 김영택, 영어-한국어 기계번역을 위한 연어와 숙어 트랜스퍼 사전, 한국정보과학회 논문지, 제 20권 제 7호, 1993

[10] Concise 국어 사전 pp.422, 금성 출판사, 1996