어휘의 중의성 해소를 위한 의미 태깅

추 교 남, 우 요 섭

인천대학교 정보통신공학과

The Lexical Sence Tagging for Word Sense Disambiguation

Kyo-Nam Choo, Yo-Seop Woo

Dept. of Information and Telecommunication, University of Inchon

E-Mail : g971163@lion.inchon.ac.kr , yswooo@lion.inchon.ac.kr

요 약

한국어의 의미 분석을 위해서 의미소가 부여된 말뭉치(Sense-Tagged Corpus)의 구축은 필수적이다. 의미 태깅은 어휘의 다의적 특성으로 인해, 형태소나 구문 태깅에서와 같은 규칙 기반의 처리가 어려웠다. 기존의 연구에서 어휘의 의미는 형태소의 구문적 제약 등의 표층상에서 파악되어 왔으며, 이는 의미 데이터 기반으로 이루어진 것이 아니었기에, 실용적인 결과를 얻기가 힘들었다. 본 연구는 한국어의 구문과 의미적 특성을 고려하고, 용언과 보어 성분간의 의존 관계 및 의미 정보를 나타내는 하위범주화사전과 어휘의 계층적 의미 관계를 나타낸 의미사전(시소러스)을 이용하여, 반자동적인 방법으로 의미소가 부여된 말뭉치의 구축을 위한 기준과 알고리즘을 논하고자 한다

1. 서 론

자연어는 정보 전달의 매개체로, 이러한 정보는 어휘 각각의 의미가 모여 이루는 것(Compositionality)으로, 자연어 처리를 통한 궁극적인 목적은 언어 정보를 구조화시킴으로 객관적 처리가 가능하도록 하는 데 있다. 한국어는 용언을 중심어로 하여 여러 필수적 보어 성분의 결속으로 상황을 구성하기 때문에 정적인 구문을 이루지 않으며, 이는 한국어의 규칙 기반 처리를 어렵게 만들과 동시에, 여러 애매성의 원인이 된다. 이러한 애매성 해소(Disambiguation) 및, 사전 정보의 획득, 지식 기반 접근, 예제 기반 접근, 통계적 접근 등의 많은 자연어 처리 분야에서 말뭉치(Corpus) 활용에 대한 요구가 증가하고 있다. 의미적 관점에서 말뭉치는 많은 상황 정보를 담고 있으며, 이를 분석해 사전화된 정보로 사용함으로 애매성 등의 여러 근본적 문제를 해결할 수 있다. 자연어의 애매성은 의미적 관점에서의 다의적 특성으로 나타나게 된다. 예를 들어 문장상에 "눈" 이라는 어휘가 있을 때 뜻이 <EYE>인지 <SNOW>인지가 그것이다. 이러한 애매성은 형태소와 구문의 그것과는 다른 것으로 후보 제약 조건은 문장상의 주요 상황에 의존하는 수 밖에 없다. 의미 태깅은 말뭉치내 어휘의 다의성을 해소하여 의미소를 부여하는 작업으로, 우선 태그셋이 의미사전의 계층에 따라 정의되므로 방대하고, 태그셋을 결정하기 위한 구문적 의미적 특성을 반영한 하위범주화사전 등의 제약 규칙이 필요하다. 또한 이러한 어려움으로 인해 자동적 또는 반자동적인 방법으로 Sense-Tagged Copus를 구축하는 硏究는 드물었으며, 대부분 의미사전의 상위 계층을 범주로 하는 수작업의 태깅이 대부분이었다. 의미 태깅을 통한, 통계 기반의 어휘 중의성 해소(Word Sense Disambiguation)는 기계 번역의 역어 선택 및 정보 검색을 위한 개념 기반 인덱싱, 자동적인 하위범주화 패턴 또는 의미 기반 연어(Sense Co-occurrence) 추출 등에 활용될 수 있다. 본 연구는 ETRI-KONAN 그룹의 한국어 분석 시스템의 일환으로 용언 중심의 하위범주화사전과 의미사전을 지원으로 하는 반자동적인 태깅 방법과 알고리즘 구현을 논하고 의미소가 부여된 말뭉치를 구축한다.

2. Sense Tagging을 위한 데이터

태깅을 위해서 우선 정의된 태그셋이 필요하며, 이는 어휘의 의미간 계층적 관계를 이루고 있는 사전에 의해 결정된다. 본 연구에서는 의미소가 부여된 말뭉치 구축의 태그셋으로 한국어 의미사전[1]을 이용한다. 또한, 의미 제약 규칙을 위한 한국어 구문-의미 패턴은 본 연구실에서 구축한 용언 중심의 하위범주화사전을 이용한다. 또한 말뭉치는 한국전자통신연구소의 반자동적인 방법으로 형태소 분석이 이루어진 5만 문장 중 10000여 문장을 대상으로 삼았다.

2.1 태그셋을 위한 한국어 의미사전

한국어 의미사전은 의미 상위어 및 중간 노드가 12,702개를 기본 모델로 하고, BOTTOM-UP 방식의 의미 추출법을 사용하여 만든 어휘간 의미 상하위 계층 구조로 한 시소러스로 전체 어휘수는 약 25,000여개이다. 본 연구에서는 Sense-Tagged Corpus 구축을 위해서, 어휘의 계층적 특성을 반영하는 코드를 그림 1과 같이 제정하여 부여함으로서 어휘간의 의미적 거리 측정 및 하위범주화사전의 의미소와 매칭시 상하위 구조를 쉽게 파악할 수 있도록 하였다.

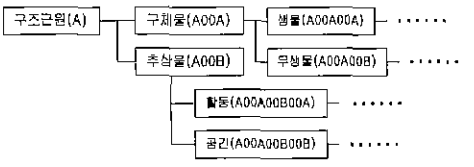

그림 1 한국어 의미사전의 의미 코드 정의

3.2 선택 제약을 위한 한국어 하위범주화사전

한국어 하위범주화사전은 본 연구실에서 구축한 것으로 말뭉치의 고빈도 용언 3000여개와, 32개의 의미역, 한국어의 조사 및 격 구문과 의미적 문형 패턴(동사 41개, 형용사 17개)을 정리하여, 그림 2와 같은 구조로 약 6000여개(동사 5435, 형용사 565)의 용언 중심의 패턴을 구축한 것으로 한국어 의미사전과 연동되도록 설계되어 있다

```
[용언] : 쥐다
[문형] : 1 이[AGT]   2 을[ACC]
[개념] 1. 사람   2. 무기류, 금권류, 도구
```

그림 2 한국어 하위범주화사전의 예

3. 의미 태깅을 위한 알고리즘 및 기준 정의

본 작업에서는 말뭉치의 대상 문장을 하위범주화사전과 의미사전과의 의미적 선택 제약을 바탕으로 일차적인 애매성 해소(Word Sense Disambiguation, WSD)를 함과 동시에 후보들을 제시하도록 하였으며, 자동적인 방법으로 커리된 분석 결과를 경험적인 면을 바탕으로 한 수작업 태깅을 하도록 하였다. 수작업 태깅에서는 자동적 방법으로는 해결하기 어려운 부분의 지침을 담고 있다

3.1 선택 제약을 이용한 자동 처리

일반적으로 하위범주화사전은 중심어의 보어와의 결속 특성상 구문 특질 뿐 아니라 의미 특질을 함께 담고 있다. 이는 곧 자연어 처리에 하위범주화 사전이 적용될 때, 각 보어나 절이 성분의 의미를 결정지어준다는 것이 된다. 이를 선택 제약(Selectional Restriction)이라 한다. 기동적인 방법에서는 그림 3에서와 같이 간단한 의존구조 해석기를 설계하고 의미사전, 그림 4와 같이 하위범주화사전과 매칭 모듈을 구현하여 의미를 찾는 방법을 이용하고 있다. 매칭 알고리즘은 우선 말뭉치의 대상 문장을 하위범주의 표층 상의 조사 매칭을 시도한 후 조사가 일치하면 어휘의 의미와 하위범주화사전의 의미 관련성 여부를 조사하고 Matched_Queue에서 여러 매칭 후보 중 최적으로 하위범주패턴을 만족하는 것을 선택하여, 그때의 의미 매칭 결과를 어휘의 의미로 본다. 최적의 후보 선택은 말뭉치의 문장의 조사와 어휘의 의미를 포함하는 패턴을 우선 순위로 하였다

3.2 경험적인 면을 바탕으로 한 수동 처리

수작업으로의 태깅은 자동적인 방법에서의 오류의 애매성이 해소되지 않은 부분을 위한 것으로 아래와 같은 몇가지 원칙하에서 수행하였다. 이래의 원칙들은 구문 분석상의 문제인 것도 있지만, 의존 관계 파악시 해결하기 쉽지 않은 점도 담고 있다

```
Matched_Queue = {},
 Loop for 한문장의 용언의 개수
    Loop for 용언에 의존 되는 조사의 개수
      If 용언 = 하위범주화사전용언 && 조사
            Matched_Queue = 말뭉치의 어휘와 해당
                            하위범주패턴정보,
    Return,
  If  Matched_Queue != {},
    Loop for Matched_Queue Item;
      If  용언 = 하위범주화사전의 용언 & 의존조사
          Matched_Queue = 말뭉치의 어휘와 해당
                          하위범주패턴정보 추가,
    Return,
 Return
```

그림 3 말뭉치의의 선택 제약을 위한 알고리즘

• 어휘의 중복 의존 – 예로 "철수는 점심을 먹고, 운동을 하였다" 라는 문장에서 '철수"는 "먹다"와 "하다" 에 모두 주격으로 쓰이고 있는 등위 접속 구문이다. 이러한 경우 두가지 모두의 의존 관계를 각성분과 함께 표시해 주었다. 또한 "철수는 수레를 끌고 있는 당나귀를 보았다" 에서 '당나귀"는 주격과 목적격을 동시에 갖고 있으므로 이들 모두의 격을 의존관계에 따라 표시하였다

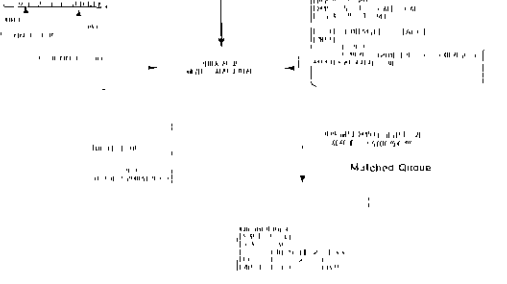

그림 4 Automatic Processing Modules Using Selectional Restriction

• Gap Filling - 다음 문장과 같이 소식언에 의존된 문장이 해당된다. "수레를 끄는 당나귀 ~" 라는 예문에서 용언에 따른 하위범주는 "당나귀가 수레를 끌다" 라는 문형 패턴을 형성하며 이는 완벽한

구문 분석이 아니라면 파악하기 어려운 문제에 대해 의존 관계를 정의하고 의미소를 부여하였다

• 피동 및 사동형 파악 – 본 연구실의 하위범주화사전은 능동형 용언을 기준으로 구축되었다 일부 중요한 피동형과 피동형이지만 능동형으로 전환되어 쓰이고 있는 몇가지를 담고 있으나 말뭉치의 형태소 분석에 의해 능동형 패턴이 제시되도록 되어 있다. "상자가 끈으로 묶이었다." 라는 문장은 "(누군가)가 상자를 끈으로 묶다"라는 능동형 패턴에서 문형 전환이 일어났다고 볼 수 있으며, 따라서 하위범주 패턴의 변환규칙(Meta-Rule)을 통해 능동형과 호환성을 갖도록 태깅하였다

3.3 Sense-Tagged Corpus의 Format

Sense Tagging 이 이루어진 말뭉치는 우선 해당 어휘의 의미와 의후보를 담고 있어야 하며, 하위범주화사전에 있는 패턴으로 태깅이 수행되었는지, 아니면 수행자가 패턴을 정의하였는지와 격성분도 담고 있어야 한다. 또한 능동, 피동, 사동형인지를 나타내도록 하였다

유역에는	1 1 0 0 0 0 (NN 유역)
	A00C00A00D00H00D00D00F
	2 0 0 0 0 0 (JO 에는)
평야가	1 2 1 4 1 CHD (NN 평야)
	A00A00A00B00G00B00U00B
	NO(개념없음)
	2 0 0 4 0 0 (JO 가)
적으니	3 0 0 5 2 3 4 0 0 (VJ 적)
	0 0 0 0 0 0 (EM 으나)

그림 5 Sense-Tagged Corpus의 예

위에 의미소가 부여되어 있는 말뭉치의 예이다. 그림 5에서 보면 품사와 해당 형태소가 나열되어 있고, 첫번째 번호는 명사와 용언(동사 형용사)과 조사를 나타내며 두번째 번호는 후보 개수, 세번째 번호는 해당 후보의 번호, 네번째는 의존 관계인데 용언에서는 태깅을 위해 이용한 하위범주화사전 번호이다. 다섯번째 번호는 자동처리와 수동처리를 구별하기 위해서이고, 끝으로는 격을 나타내고 있다

3.4 Sense Tagging 을 위한 도구의 구현

Sense-Tagged Corpus 구축을 위한 알고리즘을 실제적인 프로그램으로 구현하고, 수작업 태깅의 기준을 실행하기 위해 그림 6과 같이 태깅 도구를 설계, 제작하였다 Visual C++를 이용하여 윈도우 95/NT 환경하에서 운영되도록 하였다.

5. 실험 결과 및 검토

1차 구축 과정을 통해 10000여 문장에 의미소를 부여하였다 말뭉치의 용언 43279 개중 20500개의 하위범주는 의미사전과 하위범주화사전과의 매칭을 통한 자동적인 방법으로 얻었으며 22770여개의 의존 요소에 대해서는 자동적인 방법과 수작업을 병행하여 의미소를 부여하였다

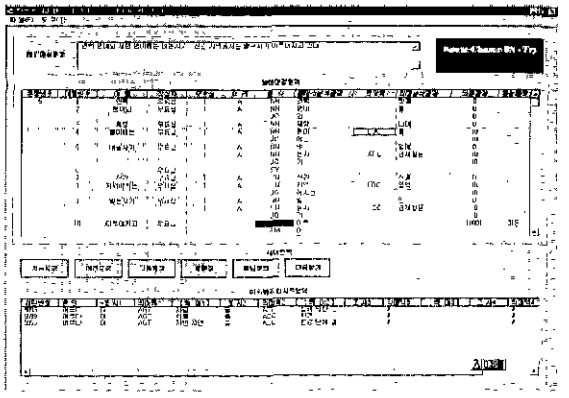

그림 6 Sense-Tagged Corpus 구축을 위한 도구

자동적인 방법으로 부여된 것 중에도 오류가 상당한 것으로 평가되었으니 이는 수작업으로 어휘의 비로 상위어를 부여하는 것에 비해 하위범주화사전의 의미소가 너무 높은 레벨에 존재했기 때문이다 또한 몇몇 용언에서는 자동적인 방법으로는 결과를 얻기가 어려웠는데, 이들은 어휘를 이고는 말아기도 하였으니 구나 결, 문장 전체를 받는 경우로 대부분의 하위범주화패턴이 조사에 있어 단어를 매임말로 하였기 때문이었다

6. 결 론

본 논문은 의미사전과 용언 중심의 하위범주화사전을 상황 제약 규칙으로 하여 의미소가 부여된 말뭉치를 구축하므로서, 한국어 의미 태깅의 방법과 규칙을 제시하였다 의미 태깅은 형태소 분석 결과의 근접의 의존관계 파악 그리고 여러 의미 데이터가 상호 관계를 갖으며 수행되는 작업으로, 앞으로 대량의 의미소가 부여된 말뭉치를 구축한다면 통계적 방법으로 어휘의 다의성을 해결 할 수 있으리라 생각한다

참고 문헌

[1] 토큰 기반 한국어 분석기 개발-한국어 의미 분석 사전 및 하위범주화 사전 구축, 한국전자통신연구원

[2] Yorick Wilks & Mark Stevenson, Sense Tagging Semantic Tagging with a Lexicon, Computational Linguistics On-Line e-print Achieve

[3] Llus Padro I Cirera A Hybrid Environment for Syntax-Semantic Tagging, de la Universitat Politecnica de Catalunya

[4] Atsushi Fujii Corpus-Based Word Sense Disambiguation, Tokyo Institute of Technology