구문 지시자를 통합한 통계적 어의애매성 해결

김권양† · 최재혁††

요 약

본 논문에서는 지도학습 알고리즘에 기반한 한국어 타동사의 어의 애매성 해결을 위한 통계적 방법을 제안한다. 본 논문에서 제안한 어의 애매성 해결 방법은 주어진 동사와 문맥 내에서 이들 동사의 주변 단어들과의 구문적 관계에 기반한 지시자들을 결합한 방법이다. 비교적 애매성이 심한 한국어 타동사 10개에 대한 어의 애매성 해결 실험 결과, 구문 관계에 기반한 지시자를 사용한 어의 애매성 해결 방법이 기준 정확도 성능 평가보다 27%의 정확도 성능 개선율 보였으며, 지시자 유형에 대해 가중치를 부여한 방법이 문맥 내에 무순서적인 주변 단어에 대한 정보만을 사용하는 방법에 비해 12% 정확도 성능 개선율 보였다.

A Statistical Word Sense Disambiguation Using Combinations of Syntactic Indicators

Kweonyang Kim† · Jaehuk Choi††

ABSTRACT

In this paper, we present a simple statistical method for performing word sense disambiguation(WSD), specially for Korean transitive verbs, based on a supervised learning algorithm. This approach combines a set of indicators based on syntactic relations between surrounding words and an ambiguous verb. Experiments with 10 Korean verbs show that accuracy performance of our WSD method using indicators based on syntactic relations is 27% higher than the baseline performance. Moreover, our method using weighting mechanism based on each indicator type is 12% higher than a method which uses only an unordered set of surrounding words in the context.

1. 서 론

컴퓨터와 컴퓨터 망의 보급으로 정보 습득과 정보 교환 방법에 일대 전환을 가져옴에 따라 종이 상에 작성된 문서를 우편에 의해 전달하던 전통적인 정보 전달 방법 대신에 컴퓨터 망을 통하여 엄청나게 많은 정보들을 빠른 시간 내에 접근할 수 있게 되었다. 따라서 대규모의 정보로부터 사용자에게 정확하게 요구되는 정보를 제공할 수 있는 기술에 대한 수요가 증가하게 되었다.

어의 애매성 해결(word sense disambiguation)이란 주어진 단어의 여러 의미들 중 문맥 내에서 사용된 올바른 의미를 구분하는 일을 말한다. 대부분의 단어들은 서로 다른 여러 개의 의미를 가지며, 이러한 단어들이 문장 내에서 나타날 때, 그 단어의 올바른 의미를 구분하는 것이 쉬운 일이 아니다. 어의 애매성 문제는 구문 분석과 마찬가지로 자연어 처리의 주요 연구 분야로서 기계 번역(machine translation)이나 정보 검색(information retrieval) 등과 같은 자연어 처리 응용 분야에서 반드시 해결되어야 할 문

† 정 회 원: 경일대학교 컴퓨터공학과 부교수
†† 종신회원: 신라대학교 컴퓨터교육과 교수
논문접수: 2002년 3월 24일, 심사완료: 2002년 4월 10일

제 중 하나로 인식되어 왔다.

애매한 의미를 가지는 단어의 문맥 내에서 올바른 의미는 그 단어가 사용된 문맥에 의존한다. 일반적으로 주제 분야를 제약함으로서 이러한 애매성 문제는 완화될 수 있지만[17], 제약된 주제 분야라 하더라도 많은 수의 어의 애매성이 발생하며, 특히 동사의 경우는 명사와 달리 제약된 주제 분야 내에서도 애매성이 발생한다[7][8].

문맥은 주어진 단어가 가지는 어의 애매성 문제를 해결하는 중요한 단서이다. 따라서 지금까지 대부분의 어의 애매성 해결에 관련한 연구에서 애매한 단어에 대한 지역적인 문맥인 주변 단어를 애매성 해결을 위한 주요 정보로서 사용하여 왔다[6]. 비록 문맥을 이용한 최근 어의 애매성 해결에 대한 연구 결과가 90%의 높은 정확도를 보여주고 있어나, 대부분의 연구에 있어서 실험 대상이 된 단어들은 그 의미가 비교적 분명하고 따라서 쉽게 구분될 수 있는 단어들에 대한 실험 결과이다. 그러나 일반적으로 동사의 경우는 가능한 의미들 간에 비교적 관련된 의미를 가지며, 따라서 이들 의미를 올바르게 구분하는 일은 쉬운 일이 아니다.

최근 연구는 이러한 사실을 감안하여 주어진 단어에 대한 문맥 내 주변 단어의 빈도수 정보만을 이용하는 대신에 연어(collocation)와 같은 정보에 더 높은 가중치를 부여하고 있다[11][12]. 또한 주어-동사 관계나 동사-목적어 관계와 같은 구문적인 정보 등을 빈번히 출현하는 연어 정보와 결합한 어의 애매성 해결과 관련한 연구들이 최근에 수행되고 있다[4][6][9][10].

본 논문에서는 지도학습(supervise learning) 방법에 기반하여 한국어 동사에 대한 어의 애매성을 해결하는 통계적인 방법을 제안한다. 이 방법은 애매한 의미를 가지는 동사와 주변 단어들 사이에 가지는 구문 관계인 지시자(indicator)들을 통합한 방법으로 각 지시자들에 대한 모든 증거들을 통합하여 어의 애매성 해결을 시도한다. 애매한 의미를 가지는 단어의 주변 단어에 대한 무순서적인 빈도수 정보 대신에 동사-목적어, 동사-장소, 동사-도구, 수식-피수식 그리고 이들의 복합 형태와 같은 구문 관계에 기반한 지시자들을 공기 정보로 사용하여 어의 애매성 해결을 시도한다.

2. 문제점 분석

본 논문에서 제시한 어의 애매성 해결 방법은 주어진 단어의 가능한 의미들과 문맥 사이에 의미적 관련성에 기반한다. 애매한 의미를 가지는 단어에 대하여 문맥 내의 올바른 의미는 해당 단어가 가지는 모든 가능한 의미들 중에서 문맥과 가장 의미적 관계가 높은 의미를 선택함으로서 결정된다.

<표 1> '쓰다'의 각 의미에 따른 주변단어

'쓰다'의 의미	주변단어/빈도수
(write)	글(writings)/33 소설(novel)/25 작품(product)/21문학(literature)/12 발표하다(announce)/10 지내다(spend)/9 책(book)/8 대학(university)/8 형식(form)/8 공부하다(study)/7 말(language)/7 신문(newspaper)/7 잇다(inherit)/7 논설(article)/5 문장(sentence)/5 수필(essay)/5 읽다(read)/5 ...
(put on)	머리(head)/28 모자(hat)/8 입다(put on)/7 남자(man)/6 갓(hat)/5 보호하다(protect)/5 막다(stop)/4 달다(hang)/4 햇볕(sunlight)/4 추위(coldness)/4 차리다(dress)/4 옷/(clothes)/4 어깨(should)/3 상투(topknot)/3 삿갓(rain-hat)/3 비(rain)/3 두르다(waer)/3 헬맷(helmet)/3 ...

<표 1>은 백과 사전 말뭉치 내에서 '쓰다:write'와 '쓰다:put on' 각각의 의미를 가지는 동사 '쓰다'와 함께 문맥 내에서 자주 출현하는 주변 단어들을 출현 빈도순으로 제시하였다. 문맥 내에서 '머리', '모자', '입다' 등과 같은 주변 단어들은 '쓰다:put on'의 의미와 의미적으로 밀접한 관계를 가지며, '쓰다'의 다른 의미보다는 '

쓰다:put on'의 의미를 가지는 동사 '쓰다'와 문맥 내에서 같이 출현함을 알 수 있다. 같은 이유로 '글', '소설', '작품' 등의 단어들은 '쓰다:write'의 의미를 가지는 동사 '쓰다'와 문맥 내에서 자주 나타나는 것을 알 수 있다. 따라서 동사 '쓰다'가 사용된 문장 내에서 '글'이나 '소설', '작품' 같은 단어들이 주변 단어로 나타난다면, 그 문장 내에서의 동사 '쓰다'의 의미는 '쓰다:write'로 사용되었음을 쉽게 결정할 수 있다.

이와 같이 동사 '쓰다'가 '쓰다:put on'이나 '쓰다:write'와 같이 두개의 의미가 명확하게 구분되는 동형이의어 의미를 가지는 경우는 동사 '쓰다'의 각 의미와 의미적으로 관련된 서로 다른 주변 단어들에 의해 '쓰다' 동사의 의미를 비교적 정확하게 구분할 수 있다. 그러나, 대부분 동사의 의미들이 이와 같이 항상 명확하게 구분되는 것이 아니라 서로 연관된 의미를 가지며, 따라서 이들 의미의 구분이 단순한 문제가 아니다.

이와 같이 의미들 사이에 밀접한 관계를 가지는 경우에는 구문 관계 정보없이 단순히 주변 단어들과 관련된 빈도수 정보만으로는 어의 애매성을 해결하기 어렵다. 동사 '쓰다'가 서로 관련된 의미를 가짐으로 인해 각 문장 내에서 같은 주변 단어를 가짐에도 그 의미가 서로 다르게 사용된 다음 예문을 보자.

예문 1-1 : 머리에 모자를 쓴다.
(put a hat on head)
예문 1-2 : 그 일에 머리를 쓴다.
(exercise brain to that work)
예문 2-1 : 영어를 유창하게 쓴다.
(speak English fluently)
예문 2-2 : 영어로 일기를 쓴다.
(write a diary in English)
예문 3-1 : 뿌리를 약으로 쓴다.
(use a root as a medicine)
예문 3-2 : 환자에게 새 약을 쓴다.
(administer a new medicine to a patient)
예문 3-3 : 약이 쓰다.
(medicine is bitter)

예문 1-1과 예문 1-2에서 명사 '머리'는 문장 내에서 동사 '쓰다'와 같이 공기하지만, 각 문장 내에서 후치사 '에'나 '를'이 명사 성분인 '머리'에 붙어서 동사 '쓰다'와 각각 장소나 목적어 같은 서로 다른 구문 관계를 이루고, 각각의 경우에 동사 '쓰다'의 의미는 서로 다른 의미인 '쓰다:put on'과 '쓰다:exercise'의 의미로 사용됨을 알 수 있다. 예문 2-1과 예문 2-2에서 명사 성분인 '영어'도 동사 '쓰다'와 같이 공기하지만, 동사 '쓰다'의 의미는 명사 '영어'에 붙은 '를'이나 '로'같은 후치사에 따라 목적어나 도구 같은 구문 관계를 이루고, 각각 '쓰다:speak'와 '쓰다:write' 의미로 사용됨을 알 수 있다. 또한 예문 3-1과 예문 3-2 경우에도 명사 성분인 '약'이 동사 '쓰다'와 같이 나타나지만 각각 동사 '쓰다'와 수단과 목적어의 구문 관계를 이루게 되고, 각각 '쓰다:use'와 '쓰다:administer' 의미로 사용됨을 알 수 있다. 특히 예문 3-3의 경우에는 명사 '약'이 주어 성분이 되어 '쓰다:bitter' 의미를 가지는 형용사로 쓰임을 알 수 있다.

문맥 내에 있는 주변 단어들에 대하여 구문 관계를 고려하지 않고 단지 각 의미에 따라 자주 출현하는 주변 단어들의 빈도수 정보만을 어의 애매성 해결에 적용할 경우, 앞에서 제시한 예문과 같은 경우에 대해서는 어의 애매성을 해결하기가 어려운 경우가 발생한다. 따라서 많은 수의 의미를 가지거나 이들 의미간에 밀접한 관련성을 가지는 단어의 어의 애매성을 해결하기 위해서는 주어진 단어와 문맥 내에 있는 주변 단어들간의 구문 관계의 차이에 중점을 가지고, 이러한 구문 관계에 기반한 단서들에 대한 정보가 요구된다. 이 방법은 같은 주변 단어들이 문장 내에서 나타나더라도 주어진 단어와 주변 단어들이 가지는 구문 관계에 따라 단어들이 서로 다른 의미적인 특성을 가진다는 사실에 근거를 가진다.

Ng[11] 등이 제안한 어의 애매성 해결 방법은 무순서적인 주변 단어 외에 이웃 단어의 품사 정보, 국소적인 연어 정보, 동사-목적어 구문 정보를 취합함으로써 구문 관계 정보의 중요성을 강조하였다. 실험 결과 가장 출현 빈도가 높은 의미에 모든 의미를 부여하는 방법보다 높은 정확도를 보였으며, 사용된 지식원 중에서 고정된 위치 정보를 가지는 국소적인 연어 정보 그리고 이

웃 단어의 품사 정보에 대해서 상대적으로 높은 정확도를 보였다. 그러나 이 두 정보는 주어진 단어의 특정 의미와 자주 나타나는 고정된 위치에 관한 정보로서 비교적 고정된 어순을 가지는 영어에 대해서는 의미 구분에 좋은 정보라고 할 수 있으나, 비교적 어순이 자유롭고 후치사에 의해 격이 실현되는 한국어인 경우에는 의미 구분에 대한 단서로서의 역할을 기대하기 어렵다.

3. 구문 지시자와 추출 알고리즘

본 논문에서는 주어진 동사의 어의 애매성을 해결하기 위해 해당 동사와 문맥 상의 주변 단어들 사이의 구문 관계에 기반한 구문 지시자를 사용한다.

먼저, 술어-논항 관계에 따른 정보는 목적어-동사, 장소-동사, 그리고 도구-동사 등과 같이 주요 명사-동사 관계를 타동사의 의미 구분을 위한 주요 단서로 이용한다. 명사에 붙은 후치사는 '을'/'를', '에', '로'/'으로'이며, 이들 각각은 일반적으로 해당 동사와 목적어, 장소, 도구와 같은 구문 관계를 가지는 격 표지 정보를 나타낸다. 한국어는 어순이 비교적 자유로운 특성을 가지는 언어로서 명사 성분 뒤에 후치사가 붙어 해당 명사와 동사 사이에 특정 구문 관계를 나타내며, 또한 어미 성분이 동사에 활용하여 문맥 내에 다른 성분과 수식-피수식 관계를 형성한다. 동사(관형형)-명사, 동사(타동사)-동사, 동사-동사(타동사), 부사-동사 등과 같은 지시자는 주어진 동사와 직접적인 수식-피수식 관계를 가지는 형태를 표현하며, 이들 외의 구문 관계 표현을 위해 주어진 동사와 문맥 내의 왼쪽, 오른쪽에 오는 주변 단어들과의 구문 관계를 구문 지시자로 이용한다.

타동사의 어의 애매성 해결에 단서가 되는 구문 관계 설정은 기존 규칙 기반 방식에서 이미 정의한 93개의 동사 하위 범주화(subcategorization) 패턴을 참고하여 술어-논항에 관한 단서를 정의하였다[1]. 또한 비교적 많은 수의 의미를 가지며 또한 이들 의미간에 관련성이 심한 두 동사 '쓰다'와 '막다'가 말뭉치 내에서 사용된 문장들을 분석하여 동사와 문맥 내에 있는 주변 단어 사이의 구문 관계에 따라 수식-피수식 구문에 관한 구문 지시자를 정의하였다. 다음은 본 논문에서 정의한 구문 관계에 따른 12개의 구문 지시자 유형을 제시한다.

지시자-1: Verb-Object relation: Noun+을/를…Verb
지시자-2: Verb-Locative relation: Noun+에…Verb
지시자-3: Verb-Instrument relation: Noun+로/으로…Verb
지시자-4: Verb-Arguments relation
지시자-5: Verb(adjective)-Noun+postposition relation
지시자-6: Adverb(or Adverbial)-Verb relation
지시자-7: Vt(Left)-Verb relation
지시자-8: Verb-Vt(Right) relation
지시자-9: Verb-Vt(Left/Right) relation
지시자-10: Lcontext-Verb
지시자-11: Verb-Rcontext
지시자-12: Verb-LRcontext

구문 관계에 대한 정보 이용은 강력한 구문 분석기를 요구하지만, 아직까지 구문 분석기의 정확도는 높지 않은 편이다. 따라서 문장 내의 제약된 범위 내에서 주어진 동사의 왼쪽 혹은 오른쪽으로 처음 만나는 특정 후치사가 붙은 명사와 같이 제약된 정보를 이용함으로써 강력한 구문 분석기 사용에 대한 대안을 제시하고자 한다.

Brown 등은 어의 애매성을 해결하기 위한 단서를 추출하기 위해 강력한 구문 분석기 대신에 왼쪽으로 처음 만나는 명사나 오른쪽으로 처음 만나는 동사와 같은 정보를 사용하였다[3]. 한국어의 경우에, Cho 등은 문장 내에 주어진 동사의 목적어 성분을 추출하기 위해 문장 내에서 동사와 가장 가깝게 위치하는 후치사 '을' 혹은 '를'이 붙은 명사와 같은 정보를 이용한 방법을 제안하였다[5]. 이러한 방법들은 품사 구분을 위해 형태소 분석기가 요구되지만 구문 분석기보다는 비교적 해결 방법이 쉽고, 어느 정도 높은 정확도를 가지는 시스템이 제공되고 있다[2].

앞에서 제시한 각 지시자들의 추출 방법은 주어진 동사의 위치에서 왼쪽 혹은 오른쪽으로 문장의 처음 혹은 끝 위치로 이동하면서 해당 지시

자를 찾을 수 있다. 다음 (그림 1)은 지시자-1의 추출 알고리즘이다.

```
/* input : a given sentence
        w[0] ... w[i-1] w[i] w[i+1] ... w[N]
   w[i] : a target verb for sense disambiguation
*/
/* output : Indicator-1 */
begin
/* extract the first noun ends with a
    postposition '을/를' */
while (preceding word of w[i] in the sentence) {
 if (w[i-1] is a transitive verb)  break;
 else if (w[i-1] == "것을")  {
  extract_previous_verb;
  break;
 }
 else if (w[i-1] ends with "을/를" && w[i-1] is
not a verb or an adjective) {
    extract noun substring of w[i-1] ;
    extract coordinate noun phrase;
    break;
 }
}
```

(그림 1) 지시자-1 추출 알고리즘

지시자 추출 알고리즘은 애매한 의미를 가지는 동사가 포함된 한 문장을 입력으로 받아 해당 동사의 왼쪽으로 진행함에 따라 목적어-동사 관계인 후치사 '을'/'를'을 가지는 첫 번째 명사를 지시자-1으로 추출한다. 그러나 내포된 문장의 경우는 이러한 지시자를 찾는 것이 어려운 경우가 발생한다. 예로서 문장 "강이 범람하는 것을 막다(keep the river from overflowing)"의 경우, 동사 '막다'에 대해서 '것을(ING형태)'을 첫 번째 명사로 인식한다. 이 경우에 본 알고리즘은 동사-목적어 관계로서 '것을(ING형태)'의 선행하는 동사인 '범람하는'의 명사형인 '범람'을 지시자로 추출한다. 다른 지시자의 추출 방법은 위 알고리즘과 비슷한 형태로 해당 지시자를 추출한다.

4. 어의 애매성 해결

본 논문에서 제안한 어의 애매성 해결기는 기존 지도학습 방법과 같이 학습 단계와 시험 단계로 구성된다. 먼저, 실험 대상인 말뭉치로부터 주어진 동사를 포함하는 모든 문장을 추출하고, 이들 문장 내에서 동사가 가지는 올바른 의미를 사전 정의를 기반으로 수작업에 의해 배정한다. 학습 단계에서는 주어진 동사가 포함된 학습 문장 각각에 대해 그 동사와 구문 관계를 가지는 모든 지시자들을 추출하고, 이들 지시자들은 각 지시자의 유형에 따라 이미 수작업에 의해 부여된 그 문장 내에서의 동사 의미와 함께 '지시자/동사의미/빈도수'의 형태로 각 해당 지시자 라이브러리에 기록된다. 이 지시자들은 해당 동사의 의미와 대응되는 문맥의 의미적 특성을 포착하게 함으로써 문맥 내에서 사용된 동사의 어의 애매성을 해결하게 한다.

시험 단계에서는 학습되지 않은 시험 문장에 대해 학습 시와 같은 방법으로 각 지시자의 유형에 따라 주어진 동사의 의미 구분에 필요한 모든 지시자들을 추출한다. 추출된 지시자들은 지시자의 유형별로 학습 단계에서 이미 추출된 지시자들과 비교된다. 시험 단계에서 추출된 지시자는 학습 시에 만들어진 지시자 라이브러리와 비교하여 그 지시자에 대한 해당 동사의 가능한 의미와 지시자의 출현 빈도수를 추출하고, 이 빈도수를 기반으로 각 지시자에 따른 동사의 의미와 관련된 점수를 계산한다. 시험 문장에서 추출한 모든 지시자에 대한 이러한 근거들을 통합하여 가장 높은 점수를 가지는 동사의 의미를 해당 문장 내에서의 올바른 의미로 결정한다.

각 지시자 라이브러리 내의 항목은 지시자-번호(지시자단어,n_1,n_2,...,n_i,...,n_N) 형태로 기록되고, 여기서 지시자 단어는 지시자에 해당하는 단어를 나타내며, n_i는 주어진 동사의 i번째 의미와 지시자 단어가 같은 문장 내에서 출현한 빈도수를, 그리고 N은 해당 동사가 가지는 최대 의미의 수를 나타낸다. 예를 들면, 지시자-1(모자,0,0,7,0,1)은 '모자(hat)'는 지시자-1에 속하며 학습 시 동사

'쓰다'의 세 번째 의미와 7번, 5번 째 의미와 1번, 그리고 나머지 의미와는 출현하지 않았음을 나타낸다.

시험 문장에서 추출된 한 지시자 단어 *inst*에 대해서 동사의 의미가 v_j일 확률 $\Pr(v_j|inst)$은 학습 시에 만들어진 해당 지시자 라이브러리로부터 다음의 조건부 확률에 의해 정의된다.

$$\Pr(v_j \mid inst) = \frac{n_j}{\sum_{j=1}^{N} n_j} (1 \le j \le N)$$

동사 v가 출현하는 시험 문장에서 추출된 모든 지시자들에 대해 주어진 동사의 의미는 각 지시자들에 대한 확률의 합을 구함으로써 다음 식과 같이 계산된다.

$$sense(v) = \underset{j \in \#\_of\_sense\ inst\in D}{Arg\max} \sum \Pr(v_j \mid inst) * W_i$$

위 식에서 W_i는 주어진 타동사의 어의 애매성 해결에 대한 각 지시자의 유형에 따른 상대적인 가중치이다. 학습 시에 구축된 각 지시자 라이브러리 내에 여러 개의 분산된 동사 의미를 가지는 단서가 많을수록 그 지시자 라이브러리는 낮은 가중치를 갖게 되어야 한다. 따라서 지시자 유형에 따른 가중치 W_i는 다음 식과 같이 정의된다.

$$W_i = \frac{1}{N_i} \sum_{inst\in Indicator-i} \left(\frac{1}{number\_of\_verb\_sense_{inst}}\right)$$

위 식에서 N_i는 지시자 라이브러리 내에 있는 서로 다른 지시자 단어의 총 수를 나타낸다. 지시자의 유형별 가중치 W_i에 대한 실험에서 동사와 술어-논항 관계를 표현하는 지시자-1, 지시자-2, 지시자-3이 가장 높은 값을 가졌다. 다음으로 수식-피수식 관계인 지시자-5, 지시자-6, 지시자-7, 지시자-8이 높은 값을 가지며, 구문 관계를 고려하지 않은 주변 단어인 지시자-10과 지시자-11, 지시자-12에 대해서는 상대적으로 낮은 값을 가졌다. 또한 지시자-4는 지시자-1,2,3을 지시자-9는 지시자-7,8을 보완하기 위한 지시자로 이들 각각 지시자보다는 낮은 가중치를 보였다.

본 논문에서 제안한 어의 애매성 해결 방법은 Ng[11]의 방법에 비해 주어진 단어와 그 단어가 포함된 문맥 내 주변 단어들 사이의 구문 관계에 더 비중을 둔 것이다. 동사와 문장 내 주변 단어들은 술어-논항 관계나 수식-피수식 관계 등의 구문 관계를 가지며, 이 구문 관계에 따른 문 성분들은 그 동사의 의미와 밀접한 관계를 가진다. 특히 타동사는 문장의 완전한 의미 표현을 위해 술어-논항 관계인 목적어 성분을 요구하며, 이 목적어 성분은 문장 내에서 주어진 타동사의 의미 구분에 직접적인 영향을 미친다. 또한 *명사-'에', 명사-'로'* 같은 술어-논항 관계나 부사나 부사구 같은 수식-피수식 관계도 주어진 동사의 특정 의미와 밀접한 연관을 가진다.

주어-동사, 동사-목적어, 관형어-명사 같은 구문 관계는 구문 관계를 고려하지 않은 주변 단어보다 더 직접적인 언어 정보를 제공한다. 이러한 구문 관계에 따른 단어간의 통계적 자료는 기존의 선택 제약과 같은 정보에 대한 통계적인 대안으로서 고려될 수 있다.

5. 실험 및 결과

본 논문에서 실험 평가 자료로 사용한 말뭉치는 계몽사에서 출판한 학생 대백과 사전이다. 학생 대백과 사전 자료의 크기는 약 1백4십만 단어이고 23,113개의 표제어로 구성되어 있다. 본 논문에서 제안한 어의 애매성 해결 방법을 평가하기 위해 말뭉치 내의 출현 빈도수와 사전 상의 의미 구분의 수가 비교적 큰 10개의 타동사 '나누다', '막다', '만들다', '묻다', '받다', '세우다', '쓰다', '얻다', '잡다', '짓다'를 실험 대상으로 삼았다.

어의 애매성 해결 실험을 위해 말뭉치 내의 10개 타동사의 모든 출현을 미리 수작업으로 사전의 의미와 대응되는 의미를 부여하고, 다음 두 가지 실험을 하였다. 실험 1은 같은 학습 자료와 시험 자료를 사용하여 어의 애매성 해결을 수행한 후 정확도를 측정하였다. 정확도는 시험 대상 문장 수에 대한 정확히 의미 구분된 문장 수의 비율로 계산된다.

<표 2>는 본 논문에서 지시한 구문 관계에 기반한 12개의 지시자를 사용한 어의 애매성 해결 방법, I_1~I_{12} 열이 가장 빈도수가 높은 의미를 올

바른 의미로 선택하는 기준 정확도(baseline)보다 우수하며, 또한 무순서적인 주변 단어 사용인 I_{12} 열의 정확도 결과보다 평균 11%의 정확도 향상을 보여준다.

실험 2는 말뭉치 내의 문장 중에서 임의로 90%의 문장을 선택하여 이를 먼저 학습한 후에 나머지 10%의 문장에 대하여 그 동사의 어의 애매성을 해결하는 실험을 수행하였다. 임의 문장을 선택함으로써 발생하는 정확도의 편차를 줄이기 위해, 임의로 분리된 학습 데이타와 시험 데이타 상에서 100회 실험을 수행한 후 그 정확도의 평균값을 측정하였고 또한 가중치를 사용하지 않았을 때와 가중치를 사용하였을 때의 평균 정확도의 개선 정도를 측정하였다.

<표 2> 같은 학습, 시험 자료에 대한 실험결과

동사	의미수	빈도수	Baseline	I_{12}	I_1~I_{12}
나누다	6	734	87%	88%	99%
막다	7	515	39%	87%	98%
만들다	10	3573	75%	89%	99%
묻다	7	585	55%	85%	98%
받다	19	1722	68%	87%	99%
세우다	10	933	48%	88%	98%
쓰다	26	2084	21%	88%	99%
얻다	11	881	67%	86%	97%
잡다	18	560	33%	90%	99%
짓다	13	926	55%	85%	98%
평균	13	1251	55%	87%	98%

<표 3>은 실험 2에 대한 정확도 성능의 개선 정도를 보여준다. 본 논문에서 제안한 구문 관계에 기반한 지시자를 이용한 I_1~I_{12} 열이 기준 정확도에 비해 평균 27%의 성능이 향상되었고, 무순서적인 주변 단어들을 사용한 I_{12} 열보다 평균 9%의 정확도가 향상됨을 보여준다. 또한 가중치 열은 가중치를 사용한 방법이 가중치를 사용하지 않은 방법보다 평균 3%의 정확도가 향상되었음을 보여주고 있다. 실험 결과 주변 단어에 대한 구문 관계를 기반으로 한 단서를 사용한 방법이 무순서적인 주변 단어를 이용한 방법보다 타동사의 어의 애매성 해결에 더 좋은 단서가 됨을 알 수 있고 또한 가중치를 사용한 방법이 더 높은 정확도를 얻을 수 있음을 알 수 있다.

<표 3> 서로 다른 학습, 시험 자료를 이용한 실험

동사	의미수	빈도수	Baseline	I_{12}	I_1~I_{12}	가중치
나누다	6	734	87%	89%	90%	93%
막다	7	515	39%	60%	70%	75%
만들다	10	3573	75%	77%	82%	85%
묻다	7	585	55%	59%	68%	72%
받다	19	1722	68%	75%	85%	87%
세우다	10	933	48%	74%	79%	85%
쓰다	26	2084	21%	62%	76%	79%
얻다	11	881	67%	79%	88%	91%
잡다	18	560	33%	79%	90%	93%
짓다	13	926	55%	72%	87%	90%
평균	13	1251	55%	73%	82%	85%

6. 결 론

단어의 의미를 올바르게 구분하는 일과 관련한 많은 최근 연구 결과에 따라 어의 애매성 해결 시스템은 제약되지 않은 형태의 문장을 입력으로 받아 합리적인 정확도와 효율성을 가지고 가장 적합한 의미를 부여할 수 있게 되었다. 어의 애매성 해결을 위한 단서로서 단순히 무순서적인 주변 단어들의 빈도수 정보 만을 사용하는 방법에 있어서 낮은 성능을 보인 가장 큰 요인은 제약된 문맥의 표현에 있다. 따라서 사람도 이러한 제약된 문맥 표현만 가지고서 어의 애매성을 시도한다면 정확한 의미를 결정할 수 없을 것이다.

본 논문에서는 애매한 의미를 가지는 동사와 문맥 내에 있는 주변 단어 간의 직접적인 구문 관계에 기반한 지시자들을 통합함으로써 한국어 타동사의 어의 애매성을 해결하는 통계적 방법을

제안하였다. 이 방법은 주어진 타동사와 구문 관계를 고려하지 않은 무순서적인 주변 단어들 외에 술어-논항, 수식어-피수식어, 인접어나 이들의 결합 형태 같은 12개의 구문 관계에 기반한 지시자들을 이용한다.

다른 동사에 비해 실험 대상인 말뭉치 내에서 출현 빈도수가 크고, 사전 내에서 정의된 의미 구분의 수가 커서 어의 애매성이 비교적 심한 10개 타동사 '나누다', '막다', '만들다', '묻다', '받다', '세우다', '쓰다', '얻다', '잡다', '짓다'에 대해 어의 애매성을 해결하는 실험을 수행한 결과, 본 논문에서 제안한 구문 관계에 기반한 지시자를 이용하는 방법이 기준 성능 평가 방법보다 평균 27%의 높은 정확도 개선를 보였다. 또한 본 논문에서 제안한 어의 애매성 해결 방법은 구문 관계를 고려하지 않고 단지 주변 단어만을 사용하는 어의 애매성 해결 방법보다 평균 12%의 정확도 개선을 보였다.

이와 같은 실험 결과를 통하여 주어진 동사와 문장 내에서 같이 출현하는 주변 단어들이 그 동사의 특정 의미와 밀접한 연관을 가지며, 이들 주변 단어들을 동사의 올바른 의미 구분의 단서로 사용할 경우에 기준 정확도에 비해 정확도에 대한 성능이 개선됨을 알 수 있다. 또한 구문 관계에 기반한 단서를 사용한 방법이 무순서적인 주변 단어만을 단서로 사용한 방법보다 타동사의 어의 애매성 해결에 더 좋은 단서가 됨을 알 수 있다.

어의 애매성 해결을 위한 대상이 목적어 성분을 취하는 타동사이고, 타동사의 의미가 주로 목적어 성분에 따라 결정되는 사실에 비추어 볼 때, 다른 타동사에 대해서도 본 논문에서 제안된 방법이 효과성을 가질 것으로 생각된다. 물론 이후에 특정 응용 시스템에서 사용되기 위해서는 나머지 동사들에 대한 실험이 추가적으로 진행되어야 할 것이다.

참 고 문 헌

[1] 신수송(1989). 한국어 문법의 정의와 파싱, 자연언어 처리의 기초 연구, KOSEF860207, 한국과학재단.

[2] 최재혁, 이상조(1993). 양방향 최장 일치법에 의한 한국어 형태소 분석에서의 사전 검색 횟수 감소 방안, 한국정보과학회 논문지, 제20권, 제10호, pp. 1497-1507.

[3] Brown, P. F., Pietra, S. D., Della, V. J. and Mercer, R. L.(1991). Word Sense Disambiguation Using Statistical Methods, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 264-270.

[4] Bruce, R. and Wiebe, J.(1994). Word Sense Disambiguation Using Decomposable Models, *Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics*, pp. 139-146.

[5] Cho, J. M. and Kim, G. C.(1995). Korean Verb Sense Disambiguation Using Distributional Information from Corpora, *Proceedings of Natural Language Processing Pacific Rim Symposium 95*, pp. 691-696.

[6] Ide, N. and Veronis, J.(1998). Introduction to Special Issue on Word Sense Disambiguation: the State of the Art, *Computational Linguistics*, Vol. 24, No 1, pp. 1-40.

[7] Kim, K., Park, S. and Lee, S.(1997). An Integrating Clues Method based on Syntactic Relations for Korean Verb Sense Disambiguation, *Proceedings of the 9th International Conference on Tools with Artificial Intelligence*, pp. 178-181.

[8] Kim, K.Y., Lee, J.H. and Choi, J.H.(2002). Combining Syntactic and Semantic Indicators for Word Sense Disambiguation, *Proceedings of International Conference on East-Asian Language Processing and Internet Information Technology, pp. 499-504.*

[9] Kwong, O.(2001). Word Sense Disambiguation with an Integrated Lexical Resource, *Proceedings of the NAACL Workshop on WordNet and Other Resources: Applications, Extensions and*

Customizations, pp. 11-16.

[10] Mcroy, S.W.(1992). Using Multiple Knowledge Sources for Word Sense Discrimination, *Computational Linguistics*, Vol. 18, No 1, pp. 1-30.

[11] Ng, H. T. and Lee, H. B.(1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Examplar-based Approach, *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 40-47.

[12] Ng, H. T. and Lee, H. B.(1997). Exemplar-based Word Sense Disambiguation: Some Recent Improvements, *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, pp. 208-213.

[13] Yarowsky, D.(1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *Proceedings of the 33rd Annual meeting of the Association for Computational Linguistics*, pp. 189-196.

김 권 양

1983 경북대학교 전자공학과 (공학사)
1990 경북대학교대학원 전자공학과(공학석사)
1998 경북대학교대학원 컴퓨터공학과(공학박사)
1983~1988 한국전자통신연구원(ETRI) 연구원
1999~2000 미국 University of Central Florida 방문교수
1991~현재 경일대학교 컴퓨터공학과 부교수
관심분야: 자연어처리, 인공지능, 정보검색, 한국어정보처리
E-Mail: kykim@kiu.ac.kr

최 재 혁

1984 경북대학교 전자공학과 (공학사)
1986 경북대학교대학원 전자공학과(공학석사)
1994 경북대학교대학원 전자공학과(공학박사)
1988 한국전자통신연구원(ETRI) 위촉연구원
1999~2000 미국 UCI 방문교수
1989~1994 신라대학교 전자계산학과 조교수
1995~현재 신라대학교 컴퓨터교육과 교수
관심분야: 한국어정보처리, 자연어처리, 정보검색, 컴파일러
E-Mail: jhchoi@silla.ac.kr