퍼지 및 지능시스템학회 논문지 2003, Vol. 13, No. 2, pp. 125-130

구문관계에 기반한 단서의 결정 리스트를 이용한 지도학습 어의 애매성 해결 방법

A Method of Supervised Word Sense Disambiguation Using Decision Lists Based on Syntactic Clues

김권양*
Kweon-Yang Kim

* 경일대학교 컴퓨터공학과

요 약

본 논문은 구문관계에 기반한 단서의 결정 리스트를 이용한 지도학습 어의 애매성 해결 방법을 제시한다. 이 방법은 주어진 단어의 어의 애매성을 해결하기 위해 애매한 의미를 가지는 단어와 문맥 내 주변 단어들 사이의 구문적 관계에 비중을 두며, 모든 단서들을 통합하는 대신에 주어진 문맥 내에서 애매성 해결에 최상이 되는 단일 증거를 규명하고 이용함으로써 올바른 의미를 결정한다. 10개의 한국어 동사에 대한 실험 결과 주변 문맥 단어 외에 구문적인 단서를 추가한 방법이 정확도 성능에 있어서 기준 정확도보다 33% 향상됨을 보였으며, 결정 리스트를 사용한 방법이 모든 애매성 해결에 대한 단서들을 통합하는 방법보다 3%의 정확도 성능 개선을 보였다.

Abstract

This paper presents a simple method of supervised word sense disambiguation using decision lists based on syntactic clues. This approach focuses on the syntactic relations between the given ambiguous word and surrounding words in context for resolving a given sense ambiguity. By identifying and utilizing only the single best disambiguation evidence in a given context instead of combining a set of clues, the algorithm decides the correct sense. Experiments with 10 Korean verbs show that adding syntactic clues to a basic set of surrounding context words improves 33% higher performance than baseline accuracy. In addition, our method using decision lists is 3% higher than a method using integration of all disambiguation evidences.

Key Words : 어의 애매성 해결, 결정 리스트, 구문적 단서.

1. 서 론

자연어 처리에서 어의 애매성 해결이란 주어진 단어가 가지는 여러 의미 중 문맥 내에서 사용된 올바른 의미를 구분하는 일을 말한다. 많은 단어들은 사전 정의에 따라 서로 다른 의미를 가지며, 이러한 단어들이 문맥 내에서 나타날 때 그 단어가 사용된 올바른 의미를 구분하기란 쉬운 일이 아니다. 어의 애매성 문제는 구문 분석과 같이 자연어 처리의 주요 영역으로 연구되어 왔으며, 자연어 이해와 관련한 응용분야에서 반드시 해결되어야 할 문제로 인식되어 왔다. 문맥 내의 주변 단어는 어의 애매성 해결에 중요한 단서를 제공한다. 따라서 많은 연구에서 보듯이 주변 단어를 어의 애매성 해결을 위한 주요 정보로 이용하고 있다[1].

최근에 Ng 등은 문맥 내의 주변 단어들 외에 이웃에 있는 주변 단어의 품사 정보와 어형 변화 정보, 국소적인 연어 정보, 동사-목적어 정보 등을 통합하여 의미 결정의 주요 단서로 사용하였고, 실험 결과 국소적인 연어 정보와 이웃 단어의 품사정보가 의미 구분에 신뢰성이 높은 증거임을 보였다[2,3]. 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.

Yarowsky는 주변 단어들에 대한 단서를 통합하여 의미를 구분하는 대신에 문맥 내에서 어의 애매성 해결에 최상이 되는 단일 증거만을 사용하는 결정 리스트 (decision lists) 방법을 제안하였다[4,5,6]. 결정 리스트 방법은 애매성 문제를 해결하는 기계 학습 방법으로 동작 원리가 비교적 단순하며 최근 어의 애매성 해결과 관련한 많은 실험에서 가장 우수한 성능 평가 결과를 보여준다[7].

본 논문에서 제안한 어의 애매성 해결 방법은 어의 애매성을 가지는 해당 단어에 대해 의미 구분에 증거가 되는 문맥 내의 모든 단서를 통합하는 대신에 결정 리스트를 사용하여 어의 애매성 해결에 최상이 되는 단일 증거만을 사용하며, Ng의 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-피

접수일자 : 2003년 2월 17일
완료일자 : 2003년 3월 24일

수식 관계 등의 구문 관계를 가지며, 이 구문 관계에 따른 문성분들은 그 동사의 의미와 밀접한 관계를 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.

2. 구문관계에 따른 어의 애매성 해결 단서

한국어는 어순이 비교적 자유로운 특성을 가지는 언어로서 명사 성분 뒤에 후치사가 붙어 해당 명사와 동사 사이에 술어-논항 관계에 따른 구문 관계를 나타내며, 또한 어미 성분이 동사에 활용하여 문맥 내에 다른 성분과 수식-피수식 관계를 형성한다.

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' 등이며, 이들 각각은 일반적으로 해당 동사와 목적어, 장소, 도구와 같은 구문 관계를 가지는 격 표지 정보를 나타낸다.

동사(관형형)-명사, 동사(타동사)-동사, 동사-동사(타동사), 부사-동사 등과 같은 단서는 주어진 동사와 직접적인 수식-피수식 관계를 가지는 형태를 표현하며, 이들 외의 구문 관계 표현을 위해 주어진 동사와 문맥 내의 왼쪽, 오른쪽과 같은 방향성과 출현 단서에 대한 표제어, 활용형 형태에 따라 주변 단어들과의 관계를 단서로 이용한다. 다음은 본 논문에서 타동사의 어의 애매성 해결을 위해 제시한 19개의 단서 유형이다.

- Verb(word form): 동사의 활용형태
- Verb-Object relation: 후치사구(명사+을/를) ... 동사
- Verb-Post_Obj: 후치사(을/를) ... 동사
- Verb-Locative relation:
  후치사구(명사+에/에는/에도/에만) ... 동사
- Verb-Post_Loc: 후치사(에/에는/에도/에만) ... 동사
- Verb-Instrument relation:
  후치사구(명사+로/으로/로는/로도/로만) ... 동사
- Verb-Post_Inst:
  후치사(로/으로/로는/로도/로만) ... 동사
- Verb-Arguments relation:
  후치사구(명사+가/이/는/은 등) ... 동사
- Verb(adjective)-Noun+postposition relation:
  동사(관형형)-명사+후치사 관계
- Adverb(or Adverbial)-Verb relation:
  부사/부사형-동사 관계
- L_Vt(Left)-Verb relation:
  (왼쪽 타동사 표제어)-동사 관계
- W_Vt(Left)-Verb relation:
  (왼쪽 타동사 활용형태)-동사 관계
- Verb-L_Vt(Right) relation:
  (오른쪽 타동사 표제어) 관계
- Verb-W_Vt(Right) relation:
  동사-(오른쪽 타동사 활용형태) 관계
- Verb-L_Vt(Left/Right) relation:
  동사-(왼쪽/오른쪽 타동사 표제어) 관계
- Verb-W_Vt(Left/Right) relation:
  동사-(왼쪽/오른쪽 타동사 활용형태) 관계
- L_Lcontext-Verb:
  (해당문장 내 왼쪽 주변 단어 표제어)-동사 관계
- W_Lcontext-Verb:
  (해당문장 내 왼쪽 주변 단어 활용형태)-동사 관계
- Verb-L_Rcontext:
  (해당문장 내 오른쪽 주변 단어 표제어)-동사 관계
- Verb-W_Rcontext:
  (해당문장 내 오른쪽 주변 단어 활용형태)-동사 관계
- Verb-L_LRcontext: (해당문장 내 왼쪽/오른쪽 주변 단어 표제어)-동사 관계
- Verb-W_LRcontext: (해당문장 내 왼쪽/오른쪽 주변 단어 활용형태)-동사 관계
- L_Pcontext-Verb:
  (이전 문장 내 주변 단어 표제어)-동사 관계
- L_Ncontext-Verb:
  (다음 문장 내 주변 단어 표제어)-동사 관계
- Verb-L_PNcontext:
  (이전/다음 문장 내 주변 단어 표제어)-동사 관계

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[8]. 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[9]. 이러한 방법들은 품사 구분을 위해 형태소 분석기가 요구되지만 구문 분석기보다는 비교적 해결 방법이 쉽고, 어느 정도 높은 정확도를 가지는 시스템이 제공되고 있다. 따라서 앞에서 제시한 각 단서들의 추출 방법은 주어진 동사의 위치에서 왼쪽 혹은 오른쪽으로 문맥의 처음 혹은 끝 위치로 이동하면서 해당 단서를 찾을 수 있다[10].

단서 Verb-Object relation은 주어진 동사와 목적어 관계를 가지는 단서로서 출현한 동사의 왼쪽으로 문맥을 따라가면서 처음 만나는 후치사구(명사+'을/를')에 해당하는 명사 성분이다. 이러한 단서는 문맥을 따라가며 해당 단서를 찾기 전에 다른 타동사가 출현하게 되면 목적어 성분이 문맥 내에서 생략된 것으로 본다. 그러나 다음 예문에서와 같이 주어진 동사 '쓰다'의 바로 왼쪽에 '골라', '뽑아', '덮어'와 같은 동사가 와서 부사형으로 사용될 경우에는 두 동사가 복합 동사적인 역할을 하게 되므로 '재료', '인재', '마스크'와 같은 명사 성분을 단서로 추출할 수 있게 하였다.

... 알맞은 재료를 골라 써야 한다.
... 유능한 인재를 뽑아 쓸 수 없다.
... 얼굴에 마스크를 덮어 쓰고 있다.

또한 목적어 성분이 내포문 형태인 "...것을/일을(ING)"일 경우에는 내포된 문장의 술어 동사가 목적어 성분으로 추출된다. 다음 예문에서 보듯이 동사 '막다', '쓰다'의 목적어 성분은 "...것을"과 같은 형태를 취하며, 따라서 '흐르는'과 '느낀' 이라는 동사에 대한 명사형인 '흐름', '느낌'이라는 명사

성분을 Verb-Object 단서로 추출한다. 이 경우에 해당 동사의 모든 활용형을 명사형 형태로 표현함으로써 다양한 활용형에 대한 적용도를 높이는 방법을 사용한다.

... 강물이 거꾸로 흐르는 것을 막기 위해 ...
... 보고 느낀 것을 써서 모아 책으로 ...

다음 예문과 같이 내포된 문장에서 술어가 "명사+하다", "명사+되다", "명사+시키다"와 같은 형태인 경우에는 해당 단서에 대한 적용도를 높이기 위해 '하다', '되다', '시키다' 성분을 제외한 명사 성분인 '부패', '발생', '누락' 만을 단서로 설정한다.

... 음식물이 부패하는(되는) 것을 막기 위해 ...
... 지역적인 분쟁이 발생하는(되는) 것을 막기 위해 ...
... 기업들이 매출액을 누락시키는 것(일)을 막을 수 ...

3. 어의 애매성 해결 알고리즘

본 논문에서 제안한 어의 애매성 해결 방법은 학습 시에 학습 자료로부터 주어진 단어와 구문 관계를 가지는 모든 단서를 추출하고, 주어진 단어의 특정 의미와 대응되는 단서의 가중치 값을 내림차순으로 정렬함으로써 결정 리스트를 작성한다. 시험 시에는 학습 시와 같은 방법으로 주어진 단어가 사용된 문맥으로부터 단서의 유형별로 단서를 추출한 다음 해당 단어에 대한 결정 리스트의 최상위 항목부터 비교하여 처음으로 부합하는 단서와 대응되는 의미를 선택함으로써 의미를 결정한다.

단계 1: 어의 애매성의 규명

한국어에 있어서 대부분의 단어들은 여러 의미를 가지는 다의어 특성을 가지며, 이런 단어가 특정 문맥 내에서 사용될 때 단어가 가질 수 있는 의미 중에서 하나의 의미를 가진다. 표 1은 동사 '쓰다'와 '막다'에 대해 각각의 사전 상의 의미 정의와 학습 자료 상에서 해당 의미로 사용된 예문의 출현 빈도수 분포를 보여준다.

표 1. 의미 정의와 빈도수 분포.
Table 1. Sense definition and frequency distribution.

표제어	의미 정의	빈도수
쓰다	글을 짓다.(S1-2)	373
	모자 따위를 머리 위에 얹어 덮다.(S2-1)	84
	어떤 일에 재료, 돈 따위를 들이다.(S3-1)	516
막다	버티어 지키다.(S3)	75
	어떤 현상이 일어나지 못하게 하다.(S4)	200

단계 2: 학습자료의 수집

단계 1에서 규명된 어의 애매성에 따라서 학습 말뭉치 내에서 주어진 단어가 출현하는 모든 문맥을 수집하고, 관찰된 출현에 대해 주어진 단어가 사용된 적절한 의미를 사전의 의미 정의에 따라 수작업으로 의미를 부여한다. 표 2는 동사 '쓰다'와 '막다'에 대해 학습 자료 상에서 수집한 의미 정의에 따른 문맥의 예를 보여준다.

표 2. 의미 정의에 따른 문맥의 예.
Table 2. Context examples for sense definitions.

표제어	의미 정의	문 맥
쓰다	S1-2	... 유럽의 전통 음악을 융합하여 독특한 작품을 썼다.
	S2-1	... 동곳으로 고정시킨 다음 대개 망건을 썼다.
	S3-1	... 니트로화 할 때는 질산을 써서 반응시킨다.
막다	S3	... 사나운 짐승의 습격을 막기 위해 동굴 속이나 ...
	S4	... 큰 도시에서 자동차 사고를 막고, 인도와 차도를 ..

단계 3: 구문/의미 단서의 출현 빈도수 측정

제안된 어의 애매성 해결 방법은 애매한 의미를 가지는 단어에 대하여 해당 단어가 문맥 내에서 특정한 의미로 사용될 때 이들 단어와 구문적인 관계를 가지는 단서들에 대한 빈도수 분포를 이용한다. 단계 3은 학습 시 애매한 의미를 가지는 단어에 대해 각 의미에 따른 단서의 빈도수 분포를 조사한 다음, 시험 시 애매한 단어가 출현하면 어떠한 단서가 해당 단어의 의미를 결정짓는데 가장 유용한 것인가를 결정하게 한다.

표 3은 의미에 따른 단서의 빈도수 분포를 보여준다. 표 3에서 단서 '머리'의 경우에 단서의 유형이 Verb-Object일 때는 동사 '쓰다'의 S3-4("마음이나 힘을 들이다") 의미로만 사용되었으며, Verb-Locative일 때는 S1-1("연필이나 붓으로 글씨를 적다") 의미와는 1회 S2-1("모자 따위를 머리 위에 얹어 덮다") 의미와는 28회에 걸쳐 나타났음을 보여준다. 또한 왼쪽 주변 단어들로 구성되는 단서 L_Lcontext-Verb에 대해서는 이들 단서들에 대해 3가지 의미로 분산됨을 알 수 있다. 이와 같이 구문관계를 고려하지 않은 단서 유형인 *L_Lcontext-Verb* 만을 어의 애매성 해결을 위한 단서로 사용하게 되는 경우에 상대적으로 빈도수가 작은 분포를 가지는 의미에 대해서는 정확한 의미 결정을 하기 어렵게 된다. 그러나 본 논문에서 제시한 방법은 주변단어와 같은 단서 외에 주어진 단어와 구문 관계를 가지는 여러 단서를 사용함으로써 단서 유형에 따른 분포만을 이용하여 정확한 의미 구분을 제시하고자 한다.

표 3. 의미에 따른 단서의 빈도수 분포.
Table 3. Frequency distribution of clues for each sense.

단서유형	단서	의미/빈도
Verb-Object	글씨	S1-1/32, S1-2/3
	머리	S3-4/3
Verb-Loc	머리	S1-1/1, S2-1/28
Verb-Inst	재료	S3-1/36
L_Lcontext	만들다	S1-1/5, S3-1/28, ...
	소설	S1-2/38,S1-1/1,S3-12/1
	머리	S1-1/1, S2-1/28, S3-4/3

애매한 의미를 가지는 주어진 단어 W_0의 각 의미에 대해 출현하는 해당 단서들의 빈도수 정보는 (clue-type, clue, {S_1:v_1, ..., S_n:v_n})의 형태로 표현되는데, clue-type은 해당 단서의 유형을 나타내며, clue는 해당 단서이다. n은 단어 W_0가 가지는 의미 개수이며, 또한 S_i:v_i은 W_0의 의미가 S_i일 때 단서 clue가 같이 나타나는 빈도수 v_i를 나타낸다.

예로서 동사 '쓰다'의 어의 애매성 해결을 위해 사용된 학습 자료 내에서 (L_Lcontext-Verb, 소설, {S1-1:1, S1-2:38, S3-12:1})는 단서 유형이 L_Lcontext-Verb인 단서 '소설'은 동사 '쓰다'의 S-1 의미와 S3-12 의미에서 1회 나타나며, S1-2 의미와는 38회 나타나지만 나머지 의미와는 같이 나타나지 않음을 보여준다. 학습 시에 S_1:v_1, ..., S_n:v_n 값은 다음 식(1) 값으로 변경된다.

$$\Pr_i = \frac{v_i}{\sum_{j=1}^{n} v_j} (1 \le i \le n) \quad (1)$$

따라서 위에서 예를 든 단서에 따른 빈도수 정보는 (L_Lcontext-Verb, 소설, {S1-1:0.025, S1-2:0.95, S3-12:0.025})로 변경되는데, 여기서 조건부 확률 Pr(Silclue) = v_i 는 주어진 단서 clue에 대하여 W_0의 의미가 S_i일 확률이 v_i이다. 즉 동사 '쓰다'의 출현에 대해서 단서 유형이 L_Lcontext-Verb(해당 문장 내 왼쪽 주변 단어 표제어)인 '소설'이라는 단서가 나타나면 W_0의 의미가 S1-2일 확률이 0.95임을 나타낸다.

단계 4: 결정 리스트

Rivest[11]에 의해 제안된 결정 리스트는 애매성 문제를 해결하기 위한 단순한 기계학습 방법으로서 이 후에 Yarowsky에 의해 엑센트 복원, 어의 애매성 해결에 적용되었다[4,5,6]. 결정 리스트가 다른 기계학습 방법에 비해 비교적 단순한 방법이지만 최근 어의 애매성 해결 문제에 대한 연구 결과[7]에서 높은 효과성을 보여주고 있다.

단계 3에서 학습 자료에 대해 단서의 유형에 따른 단서와 빈도수 정보가 추출되면 각 단서들은 단서 $Clue_k$에 대해 의미가 S_i일 가중치 Weight(S_i,$Clue_k$)가 부여된다. 이 가중치 값에 의해 내림차순으로 정렬된 단서의 리스트가 결정 리스트이다. 본 논문에서는 셋 이상의 의미를 가지는 어의 애매성 문제를 다루기 위해 다음과 같이 변형된 식(2)를 사용하였다.

$$Weight(S_i, Clue_k) = Log\left(\frac{\Pr(S_i|Clue_k)}{\sum_{i \neq j} \Pr(S_j|Clue_k)} \cdot W_m\right) \quad (2)$$

위 식에서 W_m는 주어진 단어의 어의 애매성 해결에 대한 각 단서의 유형에 따른 상대적인 중요도이다. 학습 시에 구축된 각 단서 유형의 단서들에 대해 여러 개의 분산된 동사 의미를 가지는 단서가 많을수록 해당 단서 유형은 낮은 중요도를 갖게 되어야 한다. 따라서 단서 유형에 따른 중요도 W_m는 다음 식(3)과 같이 정의된다.

$$W_m = \frac{1}{N_i} \sum_{clue \in clue-type} \left(\frac{1}{Num_{of}Senses_{clue}}\right) \quad (3)$$

위 식에서 N_i는 학습시 단서 유형별로 추출된 단서의 수를 나타낸다. 단서의 유형별 중요도 W_m에 대한 실험에서 동사와 술어-논항 관계를 표현하는 단서 유형이 가장 높은 값을 가졌다. 다음으로 수식-피수식 관계인 단서 유형이 높은 값을 가지며, 구문 관계를 고려하지 않은 주변 단어인 단서 유형에 대해서는 술어-논항 관계와 수식-피수식 관계에 비해 상대적으로 낮은 값을 가졌다.

주어진 단어에 대해서 가중치 값이 음수 값을 가지는 단서는 결정 리스트에서 제외하였으며 수식에서 분모 값이 0이 되는 경우에는 그 값을 0.1로 대치함으로써 평활화(smoothing)를 적용하였다.

시험 자료에 대해서 학습 시와 같은 방법으로 추출된 단서들은 결정 리스트에서 그 가중치 값의 크기 순서에 따라 각각 비교되어지고, 그 중 가장 큰 값을 가지는 단서가 문맥 내에서 단서로 나타날 때 학습 시 만들어진 결정 리스트에 있는 해당 단서와 관련된 의미를 올바른 의미로 선택하게 된다.

주어진 단어에 대한 특정 의미와 강한 연관성을 가지는 단서가 큰 가중치 값을 갖게 되며, 이 가중치 값의 크기에 따라 내림차순으로 정렬된 결정 리스트는 가장 강하면서 신뢰성이 높은 단서를 제일 먼저 나열한다. 단계 4에서 만들어진 결정 리스트는 시험 자료 각각에 대해서 주어진 애매한 의미를 가지는 단어에 대해 추출된 단서와 큰 가중치 값부터 비교되고, 이 때 해당 단서에 따른 의미를 선택함으로써 의미를 결정한다.

통계적 관점에서 볼 때 결정 리스트의 최상위에 있는 단서는 주어진 단어의 어의 애매성 해결에 가장 신뢰성 있는 단서가 된다. 그러나 시험 단계에서 추출된 단서들이 학습 시에 만들어진 결정 리스트 상에서 높은 가중치 값을 갖지 않는 경우에는 의미 결정에 대한 신뢰성을 갖기 어렵다. 따라서 다음과 같은 식(4)를 사용하여 기존 연구에서 제안한 방법[10]과 같이 이들 단서들에 대한 가중치 값의 합을 구함으로써 각 단서들에 대한 모든 증거를 통합하여 어의 애매성 해결을 시도할 수 있다.

$$Argmax \sum_{k=1}^{n} Log\left(\frac{\Pr(S_i|Clue_k)}{\sum_{i \neq j} \Pr(S_j|Clue_k)} \cdot W_m\right) \quad (4)$$

4. 실험 및 평가

본 논문에서 실험 평가 자료로 사용한 말뭉치는 계몽사에서 출판한 학생 대백과 사전으로 권당 500여 페이지이며 총 6권으로 구성되어 있다. 전체 말뭉치 자료의 크기는 약 1백4십만 단어이고 23,113개의 표제어로 구성되어 있다. 본 논문에서 제안한 어의 애매성 해결 방법에 대한 평가를 위해 말뭉치 내에서 비교적 출현 빈도수와 사전 정의에 따른 의미 구분의 수가 큰 10개의 타동사 '나누다', '막다', '만들다', '묻다', '받다', '세우다', '쓰다', '얻다', '잡다', '짓다'를 실험 대상으로 삼았다.

말뭉치 내에서 실험 대상인 10개 동사에 대한 모든 활용형태를 포함하는 문장과 이전, 다음 문장을 추출한 다음 어문각 우리말 큰사전의 의미 정의에 따라 문맥에 내에서 동사가 가지는 의미를 수작업으로 부여하였다. 실험 말뭉치 내에서 출현한 모든 용언의 기본형 수는 8,344개이고, 서로 다른 활용형의 수는 39,624 개에 이른다. 실험 말뭉치 내에서 10개 동사를 포함하는 문장과 이전, 이후 문장은 평균 17개의 단어들로 구성된다. 해당 동사의 활용형이 수동형인 경우에 동사가 요구하는 후치사구의 형태가 바뀌게 되므로 수동형의 활용 형태는 본 어의 애매성 실험에서 제외하였다.

어의 이매성 해결에 대한 실험의 성능 평가는 실험 대상인 모든 시험 자료에 대해 이미 수작업으로 부여된 의미와

제안된 애매성 해결 방법에 의해 선택된 의미가 동일한 경우의 비율인 정확도로 측정한다. 다음은 본 논문에서 적용한 어의 애매성 실험의 정확도와 적용도를 계산하는 식이다.

정확도 = 정확하게 의미를 구분한 시험 자료 수/시험 대상 자료의 수

적용도 = 의미 구분이 가능한 시험 자료 수/시험 대상 자료의 수

어의 애매성 해결에 대한 실험 결과 얻은 정확도 수치만 가지고는 해당 단어의 어의 애매성 문제가 얼마나 어려운지를 비교할 수 없기 때문에 기준점이 되는 기준 정확도(baseline)가 항상 제시되어야 한다. 이 기준점의 설정은 학습 시에 주어진 단어에 대해 말뭉치 내에서 가장 빈번히 출현하는 특정 의미에 모든 출현된 단어의 의미를 부여하는 방법으로 표준 평가 실험 자료가 없는 경우에 어의 애매성 해결 방법의 성능을 평가하는 방법이다.

제안된 어의 애매성 해결 방법의 성능 평가를 위해 두가지 실험이 실시되었다. 실험 1은 같은 학습 자료와 시험 자료를 사용하여 어의 애매성 해결을 시도한 후 정확도와 적용도에 대해 측정하였으며, 그 결과는 표 4와 같다. 표 4에서 10개의 타동사에 대한 기준 정확도는 평균 55%이며, 단서들을 통합한 방법인 Sum과 결정 리스트를 적용한 Dlist 방법 각각에 대해 측정된 정확도는 기준 정확도보다 우수함을 보인다.

또한 구문관계를 고려한지 않은 주변 단어만을 단서로 사용한 경우와 구문관계를 분리하여 고려한 방법에 대한 결과는 Sum 방법과 Dlist 방법에 있어서 각각 11%, 10.5%의 정확도 개선을 보여준다. 따라서 구문관계에 기반한 단서를 사용한 방법이 무순서적인 주변 단어 만을 단서로 사용한 방법보다 어의 애매성 해결에 더 나은 단서가 됨을 알 수 있다. 이 실험에서 사용한 구문관계를 고려하지 않은 주변 단어만 사용한 방법은 앞에서 제시한 여러 단서 유형 중 L_Lcontext, W_Lcontext, L_Rcontext, W_Rcontext, L_LRcontext, W_LRcontext, L_Pcontext, L_Ncontext, L_PNcontext 만을 사용한 실험 결과이다.

표 4. 동일한 학습 및 시험 자료 상에서 성능 평가.
Table 4. Performance evaluation for same training and test data.

동사	의미수	빈도수	기준%	Sum(단서 통합)				Dlist(결정 리스트)			
				주변단어		모든단서		주변단어		모든단서	
				정확도	적용도	정확도	적용도	정확도	적용도	정확도	적용도
나누다	6	734	87	88%	100%	99%	100%	85%	98%	99%	100%
막다	7	515	39	87%	100%	98%	100%	85%	97%	99%	100%
만들다	10	3573	75	89%	100%	99%	100%	89%	98%	99%	100%
묻다	7	585	55	85%	100%	98%	100%	86%	97%	97%	98%
받다	19	1722	68	87%	100%	99%	100%	87%	98%	99%	100%
세우다	10	933	48	88%	100%	98%	100%	89%	97%	97%	98%
쓰다	26	2084	21	88%	100%	99%	100%	90%	96%	99%	100%
얻다	11	881	67	86%	100%	97%	100%	85%	96%	96%	97%
잡다	18	560	33	90%	100%	99%	100%	92%	98%	98%	99%
짓다	13	926	55	85%	100%	98%	100%	87%	95%	97%	98%
평균	13	1251	55	87%	100%	98%	100%	87.5%	97%	98%	99%

표 5. 서로 다른 학습 및 시험 자료 상에서 성능 평가.
Table 5. Performance evaluation for different training and test data.

동사	의미수	빈도수	기준%	Sum		Dlist	
				정확도	적용도	정확도	적용도
나누다	6	734	87	93%	98%	95%	97%
막다	7	515	39	76%	96%	80%	96%
만들다	10	3573	75	85%	97%	88%	95%
묻다	7	585	55	75%	97%	81%	94%
받다	19	1722	68	87%	94%	89%	93%
세우다	10	933	48	85%	94%	88%	93%
쓰다	26	2084	21	80%	95%	85%	92%
얻다	11	881	67	91%	99%	94%	96%
잡다	18	560	33	93%	98%	94%	97%
짓다	13	926	55	90%	97%	92%	97%
평균	13	1251	55	85.5%	96.5%	88.6%	95%

실험 2는 말뭉치 상에 출현하는 10개 동사의 모든 출현을 미리 수작업으로 사전의 의미 정의에 따라 대응되는 의미를 부여한 다음, 이들 문맥 중에서 임의로 90%를 선택하고 이를 먼저 학습하여 해당 동사에 대한 단서들로 구성된 결정 리스트를 작성한다. 이어서 나머지 10%의 시험 자료에 대해 학습 시와 같은 방법으로 단서 유형별로 단서를 추출하고, 이들 추출된 단서들을 단서의 가중치 값에 의해 내림차순으로 정렬된 결정 리스트와 비교하여 가장 높은 가중치를 가지는 단서를 찾은 다음 이 단서와 연관된 의미를 올바른 의미로 선택한다. 임의 문장을 선택함에 의해 생기는 정확도의 편차를 줄이기 위해 임의로 분리한 학습 데이터와 시험 데이터 상에서 100회 실험한 후 얻어진 정확도의 평균값을 측정하였다. 실험 결과는 표 5와 같다.

실험 2는 실험 1과는 달리 학습 자료와 시험 자료가 다르기 때문에 서로 다른 자료 상에서 사용된 어휘의 차이로 인하여 실험 1에 비해 적용도 측면에서는 약간의 감소를 보이나 정확도 측면에서는 기준 정확도와 비교할 때 Sum, Dlist 방법 모두 약 30%의 높은 정확도 개선을 보여준다. 또한 Dlist 방법은 Sum 방법에 비해 적용도 측면에서는 약간 감소를 보이나 정확도 분석에서는 평균 3.1%의 개선을 보여준다. 따라서 다른 기존 연구[7]에서 보여주듯이 주어진 단어의 특정 의미와 연관된 모든 단서를 통합하는 방법보다 특정 의미에 대해 가장 신뢰성이 높은 단서들부터 비교하여 해당 단서가 포함된 경우에 그 단서와 관련된 의미를 부여하는 결정 리스트 방식이 정확도 성능에서 개선을 보임을 알 수 있다.

5. 결 론

본 논문은 애매한 의미를 가지는 단어가 특정 문맥 내에서 사용될 때 문맥 내의 주변 단어 외에 이들 해당 단어와 구문 관계를 가지는 주변 단서들을 이용함으로써 주어진 단어의 의미를 결정하는 지도학습을 이용한 어의 애매성 해결 방법을 제시하였다.

학습 시에 학습 자료로부터 주어진 단어와 구문 관계를 가지는 모든 단서를 추출하고, 주어진 단어의 특정 의미와 대응되는 단서의 가중치 값을 내림차순으로 정렬함으로써 결정 리스트를 작성한다. 시험 시에는 학습 시와 같은 방법으로 주어진 단어가 사용된 문맥으로부터 단서의 유형별로 단서를 추출한 다음 해당 단어에 대한 결정 리스트의 최상위

항목부터 비교하여 처음으로 부합하는 단서와 대응되는 의미를 선택함으로써 의미를 결정한다.

실험 결과 주변 단어 외에 구문 관계를 고려한 방법이 구문 관계를 고려하지 않고 문맥 내의 주변 단어만을 단서로 고려한 방법보다 정확도가 개선됨을 보였다. 구문 관계를 고려함이 없이 무순서적인 주변 단어만을 사용한 방법에 있어서 낮은 정확도 성능을 보인 가장 큰 요인은 제약된 문맥 표현에 있다. 사람의 경우에도 이와같이 주변 단어의 리스트만 가지고 의미를 부여하더라도 높은 정확도 성능을 기대하기 어렵다.

또한 주어진 단어에 대한 의미 결정에 단서가 되는 모든 증거를 통합한 방법보다 해당 문맥 내에서 의미 결정에 신뢰성이 가장 높은 단서부터 비교하여 부합되는 처음 단서와 대응되는 의미로 결정하는 결정 리스트 방법이 적용도 측면에서는 약간 떨어지지만 정확도에 있어서 성능이 개선됨을 보였다.

본 논문에서 어의 애매성 해결을 위해 선택한 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 동사와 명사들에 대한 실험이 추가적으로 진행되어야 할 것이다.

참 고 문 헌

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

[2] Ng, H.T. and Lee, H.B., "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, 1996.

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

[4] Yarowsky, D., "Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French", In Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics, pp. 88-95, 1994.

[5] Yarowsky, D., "Unsupervised Word Sense Disambiguation rivaling Supervised Methods", Proceedings of the 33rd Annual meeting of the Association for Computational Linguistics, pp. 189-196, 1995.

[6] Yarowsky, D., "Hierarchical Decision Lists for Word Sense Disambiguation." Computers and the Humanities, 34(2), pp. 179-186, 2000.

[7] Kilgarriff, A., and Palmer, M., Special doubt issue on SENSEVAL, Computers and Humanities, 34(1-2), 2000.

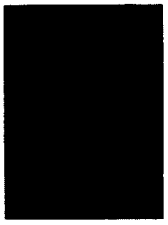
[8] Brown, P. F., Pietra, S. D., Della, V. J. and Mercer, R. L., "Word Sense Disambiguation Using Statistical Methods", Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, pp. 264-270, 1991.

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

[10] Kim, K.Y., Lee, J.H. and Choi, J., "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, 2002.

[11] Rivest, R., "Learning Decision Lists", Machine Learning, pp. 229-246, 1987.

저 자 소 개



김권양(Kweon-Yang Kim)

1983년 경북대학교 전자공학과 졸업.
1998년 경북대학교 컴퓨터공학과 공학박사
1999년~2000년 미국 University of Central Florida 방문교수
1991년~현재 경일대학교 컴퓨터공학과 부교수

관심분야 : 한국어정보처리, 자연어처리, 정보검색
Phone : 053) 850-7287
Fax : 053) 850-7609
E-mail : kykim@kiu.ac.kr