국소 문맥과 공기 정보를 이용한 비교사 학습 방식의 명사 의미 중의성 해소

(Unsupervised Noun Sense Disambiguation using Local Context and Co-occurrence)

이 승 우 † 이 근 배 ††

(Seungwoo Lee) (Geunbae Lee)

요 약 본 논문에서는 한국어 명사의 중의성 해소를 위해, 원시 말뭉치로부터 얻을 수 있는 지식원으로서 국소문맥을 정의하고 추출하는 방법을 제시한다. 동일한 국소 문맥을 갖는 서로 다른 명사는 그 의미가 유사하다는 직관을 바탕으로 대상 명사의 중의성 해소를 위해 대상명사를 포함하는 국소문맥과 동일한 국소문맥을 갖는 단어를 단서로 사용함으로써 학습 자료의 활용도를 높일 수 있고 빈도수가 적은 단어의 의미 중의성도 해결할 수 있으며, 용언의 확장을 통해 자료 부족 현상을 줄일 수 있다.

대상 명사는 동일한 국소문맥에 의한 단서들과의 최대 유사도 계산을 통해 그 의미가 결정된다. 두 단어간의 유사도는 WordNet으로부터 차용한 의미 계층 구조에서 두 단어가 가지는 개념 사이의 거리에 의해 계산된다. 최대 유사도를 계산하는 과정에서는 단서들의 중의성을 점차 줄여 나감으로써 유사도 계산의 속도를 향상시킬 수 있다. 대상 명사가 둘 이상의 국소문맥을 가질 때에는 각 국소문맥의 종류에 따른 가중치를 부여하여 국소문맥의 종류에 따른 의미제약의 차이를 구현하였다. 또 하나의 지식원으로서 사전 정의와 예문으로부터 공기정보를 얻고, 이를 국소문맥을 보완하기 위한 지식으로 사용하여 최선의 의미를 선택할 수 있도록 하였다.

실험을 통해, 제안하는 방법은 국소 문맥의 적용률이 높고, 공기 정보는 국소 문맥과 상호 보완적으로 사용되어 정확도를 높일 수 있음을 보였다. 본 방법을 실험한 결과, 사용된 단어의 의미 중의성이 크면서도, 기존의 의미 부착 말뭉치를 이용한 교사 학습 방식의 성능보다도 높은 정확도(89.8%)를 얻을 수 있었다.

Abstract In this paper, in order to disambiguate Korean noun word sense, we define a local context and explain how to extract it from a raw corpus.

Following the intuition that two different nouns are likely to have similar meanings if they occur in the same local context, we use, as a clue, the word that occurs in the same local context where the target noun occurs. This method increases the usability of extracted knowledge and makes it possible to disambiguate the sense of infrequent words. And we can overcome the data sparseness problem by extending the verbs in a local context.

The sense of a target noun is decided by the maximum similarity to the clues learned previously. The similarity between two words is computed by their concept distance in the sense hierarchy borrowed from WordNet. By reducing the multiplicity of clues gradually in the process of computing maximum similarity, we can speed up for next time calculation.

When a target noun has more than two local contexts, we assign a weight according to the type of each local context to implement the differences according to the strength of semantic restriction of local contexts. As another knowledge source, we get a co-occurrence information from dictionary definitions and example sentences about the target noun. This is used to support local contexts and helps to select the most appropriate sense of the target noun.

Through experiments using the proposed method, we discovered that the applicability of local contexts is very high and the co-occurrence information can supplement the local context for the precision. In spite of the high multiplicity of the target nouns used in our experiments, we can achieve higher performance (89.8%) than the supervised methods which use a sense-tagged corpus.

† 비 회 원 : 포항공과대학교 정보통신연구소 연구원
pinesnow@nlp.postech.ac.kr
†† 종신회원 : 포항공과대학교 컴퓨터공학과 교수
gblee@postech.ac.kr
논문접수 : 1999년 5월 3일
심사완료 : 2000년 5월 24일

1. 서 론

인간의 언어 즉, 자연언어를 이해하는 데 있어서의 출발은 단어의 의미를 이해하는 데 있다. 그런데, 한 단어는 하나 이상의 의미를 가질 수 있기 때문에 각 단어의 의미 중의성을 해소(Word Sense Disambiguation; WSD)하기 위한 노력이 선행되어야 한다.

단어의 의미를 자동으로 구분하는 일은 처음으로 컴퓨터로 언어를 처리하려 했던 1950년대 이래 많은 자연언어 처리(Natural Language Processing; NLP) 연구자들의 관심의 대상이 되어왔다. 의미 중의성 해소는 그 자체가 최종 목표는 아니며, 대부분의 자연언어 처리 분야의 중간 과정으로서 중요하다. 메시지 이해와 인간-컴퓨터 상호작용과 같은 언어 이해 응용에서 필수적임은 말할 필요도 없고, 언어 이해가 목표가 아닌 다른 응용[1]에서도 반드시 필요하거나 적어도 유용한 도움이 될 수 있다.

일반적으로 단어 의미 중의성 해소(WSD)는 텍스트 내에서 주어진 단어와, 다른 의미와 구분 가능한 그 단어의 의미를 연관시키는 작업으로, 텍스트 내의 모든 단어의 서로 다른 모든 의미를 나열하는 단계와 그 중에서 적당한 의미를 선택하여 그 단어에 할당하는 단계로 나누어 볼 수 있다. 대부분의 최근 연구에서 첫번째 단계로서 일반 사전이나 유의어 사전, 번역 사전(transfer dictionary) 등에 정의된 의미를 사용하였다. 그러나 의미 정의에 대한 다양한 접근으로 인해 많은 관련 연구들의 상호 비교가 어렵게 되었다. 다만, 형태/구문적 중의성 해결(morpho-syntactic disambiguation)을 의미 중의성 해결과 분리시킴으로써, 동일한 구문 범주에 속하는 동형 이의어(homograph) 사이의 의미 구분에 주로 초점이 맞춰지게 되었다. 단어에 알맞은 의미를 할당하는 두번째 단계에서는 대상 단어의 넓은 의미에서의 문맥(context; 텍스트 내에 포함된 정보와 함께 상황(situation)과 같은 언어 외적 정보를 포함한다)과 사전이나 백과사전과 같은 외부 지식 자원(external knowledge sources)을 의미를 결정하는 주요 근거로 삼는다.

모든 WSD는 대상 단어의 문맥을 다른 지식 정보와 비교하는 과정을 포함하는데, 이때의 지식 정보로 사용하는 자원에 따라 현재까지의 WSD를 크게 3가지로 나누어 볼 수 있다. 초창기에는 컴퓨터의 연산 능력의 한계와 기억 공간의 부족 등의 이유로 자연언어에 대한 깊은 언어적 통찰에서 얻은 규칙을 이용하는 규칙 기반 방법이 주류를 이루었다. 그러나 점차 컴퓨터의 연산 능력의 급속한 향상과 기억 매체의 발달에 힘입어 기계 가독형 사전(Machine Readable Dictionary; MRD)과 같은 대량의 지식을 이용하는 지식 기반 방법(Knowledge-based method)이 등장하였고, 또한 컴퓨터의 보급과 인터넷의 도움으로 컴퓨터가 처리할 수 있는 문서가 많아졌고 대량의 지식에 대한 통계처리가 가능해짐에 따라 대량의 말뭉치(corpus)에서 얻은 통계적 정보를 이용하는 말뭉치 기반 방법(Corpus-based method)도 개발되었다. 최근에는 이러한 다양한 방법들을 혼용한 복합적인 방법에 관한 연구도 활발히 진행되고 있다.

말뭉치 기반 방법은 다시 크게 두 가지로 분류될 수 있는데, 하나는 교사 학습 방식(supervised method)으로 미리 의미 부착된 코퍼스로부터 통계 정보를 추출하는 방식이고, 나머지 하나는 비교사 학습 방식(unsupervised method)으로 의미 부착된 말뭉치(sense-tagged corpus)를 필요로 하지 않는 방식이다. 교사 학습 방식은 지금까지 비교적 좋은 성능을 보였으나 이용할 수 있는 의미 부착된 말뭉치가 극히 제한되어 있고, 또한 이를 만드는 작업이 어렵다는 근본적인 문제로 인하여 실용적인 측면에서 의미 태깅(sense tagging)에 적용하기 어려운 점이 있다. 이에 비하여 비교사 학습 방식은 원시 말뭉치(raw corpus)를 그대로 이용하기 때문에 세밀한 의미 구분이 어렵고 정확도를 보장하기 어렵다는 문제점이 있으나 의미 부착 말뭉치가 없는 경우 이를 마련하기 위한 기초 작업으로 이용될 수 있어 실용적인 측면이 강하다.

본 연구에서는, 한국어에 대한 원시 말뭉치를 이용하여, 기존의 비교사 학습 방식의 말뭉치 기반 방법들을 비교 분석하고, [1][2]에 제시된 다음의 가정을 기반으로 하여 학습 자료의 활용도를 높여 필요한 말뭉치의 크기를 줄일 수 있으며 다양한 지식원을 결합하는 실용적인 비교사 학습 방식의 한국어 명사의 의미 중의성 해소 방안을 제안하고자 한다.

> "서로 다른 명사가 동일한 문맥 내에 가지는 경향이 있다."

2장에서 WSD에 대한 관련 연구를 살펴보고, 구축하려는 시스템의 전체적인 구성을 3장에서 보인다. 의미 유사도 계산을 위해 필요한 의미 계층 구조를 4장에서 설명하고, 원시 말뭉치로부터 학습할 자료로서 국소 문맥과 사전으로부터 학습할 공기 정보를 5장에서 설명한다. 6장에서는 학습한 지식원을 이용한 최대 의미 유사

1) 기계 번역, 정보 검색(information retrieval), 내용과 주제 분석(content and thematic analysis), 음성 언어 처리, 텍스트 처리

도 계산 과정과 지식원의 결합을 포함한 전체 의미 결정 알고리즘을 설명한다. 7장에서는 실험을 통하여 제안한 방법을 평가하고, 끝으로 8장에서 본 연구의 결론을 맺는다.

2. 관련 연구

단어 의미 중의성 해소(WSD)를 위한 기존의 연구들은 대상 단어의 문맥을 다른 지식 정보와 비교하는 과정을 포함하는데, 이때의 지식 정보로 사용하는 자원에 따라 WSD를 위한 기존 연구들을 크게 3가지, 즉, 규칙 기반 방법(Rule-based method)과 지식 기반 방법(Knowledge-based method), 말뭉치 기반 방법(Corpus-based method)으로 분류할 수 있다.

자연언어 처리 연구의 초기에 주류를 이루었던 규칙 기반 방법은 자연언어에 대한 깊은 언어적 통찰에서 얻은 규칙을 수작업으로 구축하여 단어 의미 중의성 해소에 이용하는 방법이다[3]. 이 방식은 규칙을 수작업에 의하여 생성하기 때문에 근본적으로 지식 획득에 있어서 병목 현상을 겪게 된다. 뿐만 아니라, 실제의 모든 문장들에 적용될 수 있는 견고한 규칙을 수작업으로 구축하는 일은 쉽지 않으며, 적용되는 도메인(domain)이 바뀔 때마다 새로운 규칙들을 다시 구축해야 하는 어려운 점으로 인해 실용적이지 못하다.

지식 기반 방법(Knowledge-based method)은 1980년대에 들어 사전과 시소러스(thesaurus), 말뭉치와 같은 대량의 어휘 자원을 이용할 수 있게 되면서 지식 획득의 병목 현상을 극복하고자 제안된 방법이다. [4][5][6][7]는 기계 가독형 사전을 이용한 방법으로, [5]은 중의성을 가지는 단어와 문맥 내의 주변 단어의 사전 정의를 비교하는 방식을 사용하였고, [7]는 사전 정의에 나타난 단어의 공기 횟수를 계산하여 각 의미에 따른 지식을 개선하려 했다. [4]과 [6]는 LDOCE (Longman Dictionary of Contemporary English)의 각 의미별로 제공되는 박스 코드(box codes)와 주제 코드(subject codes)를 이용하였다. 시소러스를 이용한 방법에는 [8][9][10] 등이 있다. [10]은 Roget's Categories의 통계적 모델을 사용하여 무제한 텍스트에서 단어의 의미 중의성을 해소했다. Roget's index에서 각 단어에 대해 나열된 범주는 그 단어의 의미 구분에 해당하며, 1042개의 Roget's Categories에 대해 각각을 대표하는 문맥들을 모으고 여기서 현저히 두드러진 단어들을 찾아 가중치를 결정하고 이를 이용하여 중의성을 갖는 단어의 알맞은 범주를 예측하였다. [8]과 [9]는 WordNet[11]을 지식원으로 사용하여 의미 부착 말뭉치 없이 WordNet상에서 개념간의 유사성 관계를 이용하여 중의성 해소를 시도하였다.

이러한 사전 정보를 이용한 방법은 사전의 정의나 기술에 사용된 어휘가 제한적이고 너무 짧은 경우가 많아서 무제한의 어휘를 다루는 실제 문장에 적용하는 데에는 한계를 드러낸다. 또한 사전 정보에만 의존하기 때문에 문장내의 단어의 순서나 구문 정보, 품사 및 형태적 정보들을 고려하지 않고 있다. 따라서 단어 또는 개념 공기에 전적으로 의존하는 방법론적인 한계가 있다.

말뭉치 기반 방법(Corpus-based method)은 단어 의미 중의성 해소에 필요한 지식을 대량의 말뭉치로부터 습득하여 이용하는 방법이다. 이 방법은 실제 문장으로부터 지식을 얻는다는 점에서 문장의 성격을 잘 반영할 수 있다는 장점이 있다.

말뭉치 기반 방법은 사용하는 말뭉치의 성질에 따라 교사 학습 방식(supervised method)과 비교사 학습 방식(unsupervised method)으로 분류된다. 교사 학습 방식은 의미 부착된 말뭉치로부터 단어 의미 중의성 해소에 필요한 지식을 획득하는 방법이다[12][13][14][15][16][17][18][19]. 이 방법은 사전 정의에서와 같은 상세한 의미 구분이 가능하고 비교적 정확한 학습이 가능하다는 장점이 있는 반면에 의미 부착 말뭉치가 제한되어 있다는 점에서 실용적이지 못하다. 비교사 학습 방식은 의미가 부착되지 않은 원시 말뭉치를 그대로 사용하는 방법이다[1][2][20][21][22][23]. 이 방법은 이용할 수 있는 말뭉치는 제한되지 않지만 사전 정의에서와 같은 상세한 의미 구분은 어렵다.

교사 학습 방식의 말뭉치 기반 방법 중 [15]은 주변 단어의 품사 정보, 대상 단어의 어형 변화 정보, 주변 단어의 무순서 집합, 9가지의 국소적 연어 정보(local collocation), 동사-목적어 구문 관계 등의 다양한 지식원을 이용하였다. 예문들 사이의 유사성에 기반한 이 방법은 학습 시에 대상 단어의 출현에 대해 수작업으로 의미를 부여하고, 문맥 주변에 나타나는 단어들에 대해 추출된 지식을 이용하여 학습되지 않은 문맥에 나타나는 대상 단어의 의미를 구분한다. 기존의 말뭉치 기반 방법들에 비하여 상대적으로 높은 정확도(89.9%)를 보여 실용적이라 할 수 있으며, 특히 사용된 지식원 중에서 국소적 연어 정보에 의한 공헌도가 가장 높았다. 그러나 이 방법은 대량의 의미 부착된 말뭉치를 필요로 하기 때문에 특히 의미 부착된 말뭉치가 극히 부족한 한국어에 대해서는 이 방법을 사용하기가 곤란하다. WSD에서의 다양한 지식원의 필요성은 [14]에서도 제시되었다. [14]에서는 품사 정보, 단어 빈도수, 연어 정

보, 의미 문맥(semantic context), 선택 제약, 구문적 단서(syntactic cues) 등의 다양한 지식원이 단어 의미 중의성 해소에 필요하다고 주장하였다. [19]은 의미 부착 말뭉치의 부재를 극복하기 위해서 일-한 기계 번역으로부터 학습 말뭉치를 생성하는 새로운 방법을 제시하였다. 그러나 기계 번역 또한 의미 중의성 해소의 문제를 자체에 포함하고 있으므로 여기서 생성된 말뭉치의 의미 구분이 얼마나 정확한지가 역시 의문시된다.

비교사 학습 방식 중 [22]은 문맥 벡터를 기반으로 한다. 문맥 내의 각 주변 단어에 대해 그 단어들과 말뭉치 내에서 공기하는 단어들로 구성된 단어 벡터의 합으로 문맥 벡터를 구성하고, 의미적으로 유사한 문맥 벡터들을 단어에 대한 의미 표현으로 사용하는 방법을 제안하였다. 말뭉치 내에서 대상 단어와 같이 나타나는 단어들을 이용함으로써 다양한 어휘들과의 관련 정보를 포착할 수 있다. 하지만, 문맥 내의 단어들이 중의성을 가질 수 있고, 또한 반드시 대상단어의 단서로서 작용한다는 것을 보장할 수 없다. 즉, 문맥 내의 무순서적인 주변 단어들만을 문맥에 대한 표현으로 사용함으로써 단어들 사이의 술어-논항 관계나 수식-피수식 관계와 같은 의미적 제약이 강한 구문 관계에 관한 정보를 사용하지 않았다.

[21]에서는 WordNet을 이용한 단어 중의성 해소 방법을 제시하였다. 이 방법은 WordNet의 각 의미를 뜻하는 유의어 집합(synonym set) 내에서 중의성이 없는 유의어(monosemous synonym)를 대상 단어의 중의성 해소에 이용한다. 중의성을 가진 단어에 대해 WordNet에서 각 의미별로 중의성이 없는 유의어를 찾아 내고 이 유의어들을 포함하는 문장들을 대량의 말뭉치에서 뽑아낸다. 이 방법은 WordNet과 같이 단어들 간의 유의 관계를 잘 기술하고 있는 지식원이 있을 경우에는 가능한 방법이지만 아직 이런 지식원을 갖추지 못한 한국어에는 적용할 수 없다.

Yarowsky가 제시한 [23]은 의미 부착되지 않은 원시 말뭉치로부터 각 담화 내에서 단어는 하나의 의미로 사용된다는 성질(one sense per discourse)과 동일한 연어 내에서 단어는 하나의 의미를 갖는다는 사실(one sense per collocation)을 이용한 비교사 학습 방식의 단어 의미 중의성 해소 방법으로, 먼저 중의성을 해소하고자 하는 대상 단어를 포함하는 모든 예문을 수집한 다음, 대상 단어의 각 의미를 가장 잘 나타낼 수 있는 종자 연어(seed collocation)들을 정하고 이 종자 연어에 기반하여 예문들을 대상 단어의 각 의미로 분류한다. 그런 다음 각 의미별로 분류된 문장들에 대해 Decision List를 이용하는 교사 방식의 학습 알고리즘을 이용하여 새로운 종자 연어를 추출하고, 이 작업을 모든 예문을 분류할 수 있을 때까지 반복한다.

이 방법은 초기의 오류를 반복 과정에서 정정할 수 있으며, 기존의 교사 학습 방식의 성능과 비슷한 수준의 뛰어난 성능(96%)을 보였다. 그러나, 이 성능은 실험에 사용된 대상 단어의 의미 구분을 두 가지 의미로 제한하여 문제의 난이도를 상당히 낮췄기 때문에 가능했으며, 구분해야 할 대상 단어의 의미 수가 많은 경우에도 같은 성능을 발휘할 것이라고는 말할 수 없다. 즉, 단순히 실험 결과의 수치만으로는 다른 비교사 학습 방식에 비해 뛰어나다고 말할 수는 없다. 실제 중의성을 가진 단어는 그 의미 수가 7이 넘는다[15]. 또한 이 방법은 각각의 중의성을 가진 단어에 대해 Decision List를 학습시키기 위해 수천 개의 예제 문장을 사용하여 성능을 높일 수 있었다. 그러나, 중의성을 가진 단어 자체가 이미 수천 개[2]에 이르며, 따라서 이들 단어들을 이 방법으로 학습하기 위해서는 말뭉치가 수십억 단어를 포함해야 한다. 각각의 학습된 Decision List는 한 단어의 중의성 해소에만 사용될 수 있기 때문이다. 즉, 학습되지 못한 단어들은 처리할 수 없으므로 방법론의 일반성을 상실하고 있다.

[1]와 [2]는 의미 구분을 위해 WordNet의 계층 구조를 사용하고, 대상 단어와 동일한 문맥 내에 나타나는 서로 다른 단어들을 대상 단어의 의미를 구분하기 위한 단서로서 사용하기 때문에 의미 부착된 말뭉치를 필요로 하지 않는다. 이것은 동일한 문맥 내에서 나타나는 두 개의 서로 다른 단어는 유사한 의미를 가지는 경향이 있다는 직관에 바탕을 두고 있다. 우선 학습 문장들을 파싱하여 문장으로부터 주어-동사, 동사-목적어, 형용사-명사, 수식-피수식의 구문 관계를 추출하고 이러한 구문 관계를 이용하여 발생 빈도수를 이용하거나 개념 계층 구조에서의 유사도를 이용하여 단어의 의미 중의성을 해소한다.

이 방법은 동일한 지식을 다른 단어의 의미 중의성 해소에도 사용할 수 있으므로 자연언어의 일반성을 살려 학습 말뭉치에서 나타나지 않은 단어일지라도 그 중의성 해소가 가능하게 하였다. 의미 부착된 말뭉치를 이용하지 않기 때문에 교사 학습 방식에 비해 성능(68.5%)은 낮지만, 의미 태깅에 실용적으로 이용할 수 있고, 의미 부착 말뭉치를 마련하기 위한 기초 작업으로 활용될 수 있다. 그러나, 구문적 의존 관계만을 국소 문

2) WordNet의 명사 중 중의성을 가진 단어는 12,564개이다

맥이라는 지식원으로 사용하였기 때문에 기존 연구들 [14][15]에서 제시된 그 밖의 중요한 지식원들을 이용하지 못하여 성능이 낮을 수 밖에 없었으며, 또한 국소 문맥들을 동등하게 적용함으로써 구문 관계에 따른 의미 제약의 강도 차이를 이용하지 못하였다.

지금까지 살펴본 바와 같이 규칙 기반의 단어 의미 중의성 해소 방법은 실제 문장에 적용할 만한 견고한 규칙을 수작업으로 만들기가 어렵고 특정 영역이 정해지지 않은 경우에는 더욱 그렇다. 또 지식 기반 방법의 경우, 영어권의 LDOCE나 WordNet과 같은 다양한 언어 정보를 가진 기계 가독형 사전을 한국어의 경우에는 이용할 수 없고, 어순이나 구문 정보, 품사 정보나 단어의 어형 변화 정보 등의 다양한 정보를 활용하지 못하고 있다. 말뭉치 기반의 방법 중 교사 학습 방식은 좋은 성능을 보이지만, 미리 의미 부착된 말뭉치를 마련하여야 한다는 문제점으로 인해 의미 부착된 말뭉치가 극히 부족한 한국어에는 적용할 수 없으며, 또한 실제적인 의미 태깅에 적용하는 데에도 한계가 있다. 따라서, 본 연구에서는 의미 부착된 말뭉치를 필요로 하지 않는 비교사 학습 방식의 말뭉치 기반 방법을 채택하고 다양한 지식원을 의미 구분의 단서로 활용하여 말뭉치로부터 학습하고 사전의 정보를 보완적으로 활용하여 학습 자료의 활용도를 높이면서 성능을 향상시킬 수 있는 실용적인 한국어 명사 의미 중의성 해소 방법을 제안하고자 한다.

3. 본 WSD 시스템 구성

관련 연구에서 살펴본 바와 같이 기존의 비교사 학습 방식은 여러 가지의 문제점을 안고 있었다. 따라서, 본 연구에서는 이러한 문제점들을 해결할 수 있는 한국어의 명사 의미 중의성 해소 방안을 제시하고자 한다. 첫째, 한국어에 대하여 이용 가능한 의미 부착 말뭉치가 없으므로, 원시 말뭉치를 이용하는 비교사 학습 방식의 말뭉치 기반 방법을 채택한다. 둘째, 말뭉치 기반의 방식이 갖는 자료 부족 문제를 극복할 수 있도록 한국어 명사에 대한 의미 계층 구조를 이용하여 단어의 의미 사이의 유사도를 계산할 수 있도록 한다. 셋째, 적은 말뭉치를 사용하고도 효율을 높일 수 있고 학습되지 않은 단어에 대한 중의성 해소도 가능하도록 Lin이 [1]에서 제시한 직관을 적용하여 자연언어의 일반성을 살린다. 넷째, 국소 문맥으로서, 구문적 의존 관계 뿐만 아니라 명사의 대등관계, 명사 연어 정보 등을 중의성 해소를 위한 지식원으로 사용한다. 다섯째, 구문 관계의 종류에 따라 국소 문맥이 갖는 의미 제약의 강도 차이를 고려한다. 여섯째, 사전 정의와 예문을 통해 얻은 공기 정보를 국소 문맥을 보완하기 위한 지식원으로 사용하여 전체 성능을 향상시킨다.

한국어 명사의 중의성 해소를 위해 제안하는 시스템은 원시 말뭉치로부터 국소 문맥을 추출하고 사전 정의로부터 공기정보를 획득하여 국소 문맥 데이타베이스와 공기 정보 데이타베이스를 구축하고 의미 계층 구조로부터 의미 데이타베이스를 구축하는 학습 단계와 학습된 지식을 바탕으로 입력 문장의 중의성이 있는 명사의 의미를 결정하는 의미 결정 단계로 이루어 진다.

학습 단계와 의미 결정 단계의 시스템 구성은 각각 그림1과 그림2와 같다.

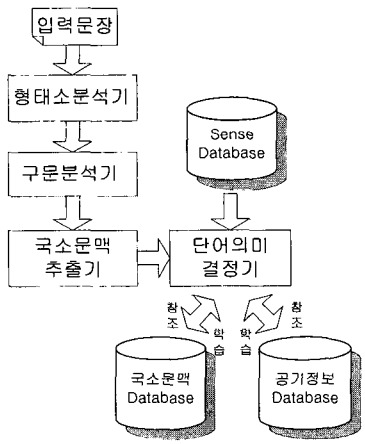

그림 1 학습 단계의 시스템 구성도

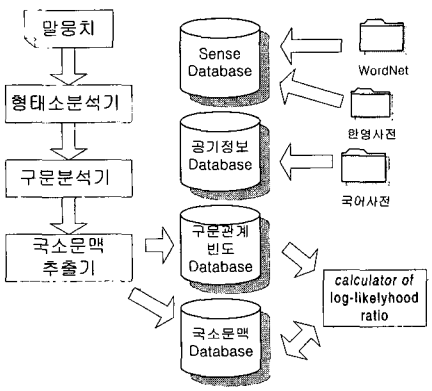

그림 2 의미 결정 단계의 시스템 구성도

4. WordNet을 이용한 의미 계층 구조

일반적으로 사람들은 자연언어로 된 문장을 이해하는 데 거의 무의식적으로 애매성을 해결하여 문장의 의미를 분석하고 그 개념을 이해한다. 그러나 비록 무의식적이라고는 하지만 실제로 사람이 자연어 문장을 이해하는 데에는 각자가 가진 상식이나 지식, 단어 개념 등의 지식 베이스를 이용한다. 이와 같이 일반적인 사람이 가진 지식 베이스를 컴퓨터에 도입하고 컴퓨터가 이를 참조하여 자연어 문장의 애매성을 해결하고 개념을 이해할 수 있도록 하려는 노력이 많이 있었다. 그러한 지식 베이스를 개념체계(ontology) 혹은 시소러스(thesaurus)라고 한다. 영어권의 경우 이에 대한 연구가 오래 전부터 있어 왔고, 그 결과로 현재 주로 사용되고 있는 개념체계에는 Roget's Thesaurus와 WordNet이 있다. 이러한 개념 체계들은 자연어 처리의 여러 응용 분야에서 중요한 역할을 담당하고 있지만, 아직 국내에서는 이에 대한 연구가 부족하여, 몇몇 특정 도메인(domain)에 한정되거나 상위 몇 단계의 개념만 분류하여 사용하고 있을 뿐, 두루 사용되고 있는 한국어에 대한 개념 체계가 없는 실정이다. 최근에 이러한 개념 체계에 대한 연구가 진행 중에 있지만[24][25], 아직 이용할 수 있는 단계는 아니다. 개념 체계를 구축하기 위해서는 많은 인력과 시간이 들기 때문에 새로이 개념 체계를 구축하기보다는, 비록 사용언어는 다르지만 이미 널리 인정되어 사용되고 있는 WordNet을 기반으로 하여 이를 한국어 어휘에 맞게 변환하는 방법을 택하였다.

WordNet은 인간이 갖고 있는 어휘 개념들에 대한 언어 심리학 이론들에 기반한 온라인 개념 체계로, 영어의 명사와 동사, 형용사, 부사들을 유의어 집합(synonym set; synset)이라는 어휘 개념과 이들 사이의 다양한 개념 관계(relation)로 구성되어 있다[11]. 현재 버전1.6이 제공되고 있으며, 대략 12만 단어와 10만 단어 의미(synset)를 담고 있으며, 명사만을 고려하면 94,000단어와 66,000단어 의미를 담고 있다. 다양한 개념관계 중에서 상하위 관계(IS-A relation; hyponymy/hypernymy)에 의해 각 개념들은 계층 구조를 이루게 되며 시소러스의 역할도 수행할 수 있다.

영어권과의 문화적, 생활 환경적 차이로 인하여 영어 어휘들이 커버하는 어휘 개념들의 분포와 한국어 어휘들이 커버하는 어휘 개념들의 분포에는 분명히 차이가 있으므로 이러한 개념 분포의 차이가 문제가 된다. 그러나, 일반적으로 영어의 어휘가 한국어 어휘에 비해 그 수가 많고 WordNet의 어휘 개념들이 상당히 세분되어 있기 때문에, 비록 부분적으로 한국어 어휘가 보다 세분된 영역들도 있지만, 많은 경우에 영어 어휘 개념들이 한국어 어휘 개념들을 커버할 수 있다. 물론 영어의 어휘 개념이 한국어의 어휘 개념보다 세분되어 있는 경우 영어 어휘 개념의 선택이 어려워질 수도 있으나, 이러한 경우 한국어 어휘 개념이 세분된 영어 어휘 개념 모두를 갖도록 하면 된다. 또한, 한국어 의미 해석뿐만 아니라 다른 대부분의 응용에서 어휘의 미묘한 의미 차이까지 구분하는 경우는 거의 없다. 이는 사람들 사이에도 그 미묘한 의미 구분에는 서로 다른 이견이 있을 수 있기 때문이다[26].

4.1 대역어 선정

먼저, WordNet을 한국어에 접목하기 위한 첫 단계로 한국어 어휘를 영어 어휘로 사상시키기 위해 한영 대역어 사전을 사용한다. 본 논문에서 사용한 한영 대역어 사전은 Internet에서 제공하는 한영 대역어 사전 서비스로부터 구축한 것으로 총 28178 단어로 이루어져 있다. 한영 대역어 사전은 한국어의 각 어휘에 대해 그 어휘가 가지는 의미별로 대역 가능한 영어 단어 또는 연어를 제공해 준다. 예를 들어, '가사'는 5개의 의미를 가지며 그 각각은 표 1과 같은 대역어를 갖는다. 표 1에서 대응하는 WordNet 개념(synset)이 둘 이상일 경우에는 슬래쉬(/)로 구분한다.

표 1 '가사'의 대역어와 WordNet synset

'가사'	의미	영어 대역어	synset
1	살림을 꾸려 나가는 일	household affairs, domestic affairs, housekeeping, housework	{housework#1, housekeeping#1}
2	집안의 사사로운 일	family affairs, household matters	{matter#1, affair#1, thing#8}
3	완전히 의식을 잃어 죽은 것처럼 보이는 상태	apparent death, suspended animation	{suspended animation#1}
4	노래의 내용이 되는 글	words, text, libretto	{words#3}/{libretto#1}
5	법의	surplice	{surplice#1}

4.2 대응 개념 결정과 의미 코드

다음으로, 이들 각각의 의미들이 WordNet 계층 구조의 어느 위치에 속하는지를 결정한다. 이를 위해 의미별로 사상된 대역어들을 WordNet에서 찾아 그 결과에서 공통으로 속해 있는 개념(synset)을 그 어휘의 해당 의

미가 속하는 개념으로 결정한다. 즉, '가사'가 '살림을 꾸려 나가는 일'을 의미할 때의 대역어는 네 가지(household affairs, domestic affairs, housekeeping, housework)가 가능하고 이 대역어들을 WordNet에서 검색해 보면 이 대역어들이 공통으로 synset ***{housework#1,housekeeping#1}***에 속함을 알 수 있다. 경우에 따라서는 '가사'의 네 번째 의미의 경우처럼 두 가지 이상의 synset이 가능한 경우도 있다.

한국어 어휘 중 명사에 대해 위와 같이 한영 대역어 사전을 이용하여 수동으로 한국어 명사 개념 체계를 구축하였다. 이것은 총 3897개의 단어와 7876개의 synset으로 이루어졌으며, 구조는 WordNet과 동일하다.

또한, 개념 유사도를 계산하는 데 사용될 계층 구조상에서 각 개념들간의 개념거리를 쉽게 얻을 수 있도록 각 개념에 synset외에 WordNet의 계층구조를 내포한 의미코드를 할당하였다. 예를 들면, 표 1에서 '가사'의 첫 번째 의미는 그림 3과 같이 계층 구조를 내포한 의미 코드를 할당받는다. 이 의미 코드는 synset과 함께 의미 태깅의 태그셋을 구성하는 데에도 사용될 수 있다.

```
[0B0F0N0H] {housework#1, housekeeping#1}
  => [0B0F0N] {work#1}
       => [0B0F] {activity#1}
            => [0B] {act#2, human action#1, human activity#1}
```

그림 3 '가사'의 1번 의미의 의미 코드

한국어 명사 개념 체계를 구축하기 위해서 본 논문에서 사용한 방법과 [24]를 비교해 보면, [24]에서는 영어권 WordNet의 영어 어휘를 한국어의 어휘로 변환하는 방법을 사용하였고, 본 논문에서는 한국어 어휘의 의미를 영어권 WordNet의 synset으로 매핑하는 방법을 사용하였다.

4.3 의미 데이타베이스

이와 같이 얻어진 명사에 대한 의미 코드들은 명사 의미 중의성의 해소 과정에서 의미 구분과 유사도를 계산하는 데 사용될 수 있도록 의미 데이타베이스로 구축된다. 의미 데이타베이스의 각 엔트리는 각 명사와 그에 대응되는 의미별 의미 코드들로 구성된다.

5. 중의성 해소에 이용할 지식원

5.1 국소 문맥

사람의 경우 비교적 좁은 영역의 주변 단어들이 주어질 때에도 단어의 의미 중의성을 해소할 수 있다는 사실이 언어 심리학 실험을 통해 밝혀졌다[27]. 이 사실은 단어의 의미 중의성을 해소하는 데 있어서, 그 단어를 둘러싼 좁은 범위의 주변 단어들만 잘 활용하여도 좋은 성능을 얻을 수 있다는 것을 간접적으로 보여준다.

지금까지의 단어 의미 중의성에 관한 연구들에서는 이러한 사실을 국소 문맥(local context) 또는 연어 정보(local collocation)라는 용어를 사용하여 다양한 방법으로 이용하였다. [28]에서는 대상 단어를 포함하는 문장에 나타나는 단어들의 무순서 집합과 이전 문장을 국소 문맥으로 사용하였고, [15]에서는 대상 단어의 주변 6개 단어의 품사와 어형 변화 정보, 9가지의 연어를 사용하였다. [23]에서는 대상 단어의 전후 k window내에 있는 주변 단어와 대상 단어로부터의 거리를 문맥 정보로 사용하였다. 또 [1]는 문장 내에서 대상 단어와 주변 단어 사이의 구문적 의존 관계를 국소 문맥이라 정의하였다.

단순히 주변 단어들의 무순서적인 나열을 이용하는 방법은 단어간의 순서, 구문 관계 정보, 연어 정보 등의 중요한 지식을 활용하지 못한다. 또 구문 관계 정보만을 이용하는 방법은 실제 문장에서 적용되지 못하는 경우도 발생할 수 있고 정확도가 낮을 수 있으므로, 대상 단어와 구문 관계로 연결된 주변 단어들을 국소 문맥으로 사용함과 동시에 대상 단어의 의미별 공기 정보도 활용할 수 있어야 한다. 따라서, 먼저 사용할 구문 관계에 따른 국소 문맥을 정의하고 이를 구문 관계의 종류에 따라 분류하였다. 그리고, 공기정보를 얻는 방법과 용언 확장을 통한 국소 문맥의 확장 적용에 대해 설명하도록 한다.

5.1.1 국소 문맥 정의

본 연구에서는 품사 태깅을 미리 수행하기 때문에 형태소의 중의성은 범위에서 제외하였다. 사용할 국소 문맥은 [1]에서의 정의를 확장하여 사용한다. 즉, 문장 내에서 명사와 다른 단어 사이의 의존 관계뿐만 아니라 명사들의 대등관계와 명사 연어(collocation), 명사-의존명사, 명사-접미사 관계를 구문 관계에 포함시켜 국소 문맥으로 정의한다. 주변 명사 W_c가 다른 주변 단어 W_a와 구문 관계 R로 연결되어 있을 때 명사 W_c의 국소 문맥은 다음과 같이 정의된다.

$$W_c : (R \quad \text{direction} \quad W_a)$$

direction은 W_c와 W_a사이의 방향성을 의미하는 것으로 W_a가 W_c를 지배하거나(Head) 반대로 W_a가 W_c에 지배되고(Modifier) 있음을 가리킨다. 때로는 W_c와 W_a 사이에 방향성이 없는 경우(Conjunct)도 가리킨다. 한

문장에서 한 단어가 하나 이상의 구문 관계를 가질 수 있기 때문에 각 단어는 다중의 국소 문맥을 가질 수 있다. 예를 들어, 표 2는 다음 문장에 나타난 명사들의 국소 문맥을 보여준다.

"한암은 가사와 장삼을 단정히 입고 앉아 있었다."

표 2 추출된 국소 문맥

중심 단어(Wc)	구문 관계 (R)	direction	주변 단어(Wa)
한암	주어-술어	Head	입-다
한암	주어-술어	Head	앉-다
가사	명사-조사-명사	Conjunct	장삼
가사	목적어-술어	Head	입-다
장삼	목적어-술어	Head	입-다

5.1.2 구문 관계에 따른 국소 문맥의 종류

한국어에 있어서 구문 관계는 주로 조사에 의해서 표현된다. 따라서 한국어의 조사들을 정리하여 각 조사별 명사와 용언 사이의 구문 관계를 설정하였다. 그밖에 수식-피수식 관계나 대등 관계, 명사 연어, 명사-의존 명사, 명사-접미사 등의 관계도 설정하였다.

본 연구에서 사용되는 구문 관계와 그에 따른 국소 문맥의 종류는 표 3과 같다.

표 3 국소 문맥의 종류

구분	중심단어	구문 관계 태 소	주변단어
조사/지정사에 의한 구문 관계 설정	명사	조사 (주격조사, 목적격조사, 여격조사, 부사격조사, 보조사)	용언
	명사	접속조사	명사
		관형격조사	
		지정사 (-이다, 아니다)	
형태소 쌍에 의한 구문 관계 설정	명사	-	관형사
	명사	-	명사
	명사	-	의존명사
	명사	-	`접미사

5.1.3 국소 문맥 추출

입력 문장으로부터 국소 문맥을 추출하기 위해서 입력 문장에 대해 형태소 분석과 구문 분석의 과정을 거친다. 입력된 문장은 형태소 분석기[29]에 의해 형태소 그래프가 생성되고 각 형태소는 알맞은 품사를 할당받는다. 이 형태소 그래프는 구문 분석기의 입력으로 사용된다. 구문 분석기[30]는 K-CCG(Korean-Combinatory Categorial Grammar)에 기반하여 형태소 그래프로부터 구문 트리를 생성해 낸다. 그림 4는 5.1.1의 예제 문장을 구문 분석한 결과로 생성된 구문 트리를 보여 주고 있다. 이 구문 트리로부터 5.1.1과 5.1.2에서 정의한 국소 문맥을 추출하는 알고리즘은 그림 5와 같다.

```
/complete/end---s<문장시작>(])
complete
\                /np---MPN<한암>(한암)
\              /v/(v\np[j는])
\              / \(v/(v\np[j는]))\np---jS<는>(은)
\           /v[D]\{np[j에]}
\           / \              /np---MCC<가사>(가사)
\           / \           /np/np
\           / \           / \(np/np)\np---jO<와>(와)
\           / \         /np
\           / \         / \np---MCC<장삼>(장삼)
\           / \       /v/(v\np[j를])
\           / \       / \(w/(v\np[j를]))\np---jC<을>(을)
\           / \    /v[D]\{np[j가]}
\           / \    / \ /v/v---BM<단정히>(단정히)
\           / \    / \v[D]\{np[j가],np[j를]}
\           / \    /    \v[D]\{np[j가],np[j를]}---DR<입>(입)
\           / \ /v/v
\           / \ / \(v/v)\v---eCC<고>(고)
\           / \v[D]\{np[j가],np[j에]}
\           /    \v[D]\{np[j가],np[j에]}---DR<앉>(앉)
\        /vp[E]
\        / \vp[E]\v[D]---bE<어#있>(아_있)
\      /vp[었]
\      / \vp[었]\vp---eGSt<었>(었)
\   /s[서술]
\   / \s[서술]\vp---eGEs<다>(다)
\ /s[X]
\ / \s[X]\s[X]---s.<.>(.)
\end
  \end\X---s<문장끝>(])
```

그림 4 구문 트리

Step 1. 구문 트리를 pre-order로 탐색한다.
Step 2. 국소 문맥을 포함하는 sub-tree의 root를 탐지한다.
노드가 np, cn, v[I], v[E], v[H], v[D] 중의 하나일 때, 이 노드의 category와 이 노드의 맨 오른 형태소의 품사, 그리고 왼쪽 노드의 category를 조사하여 판단한다.
Step 3. 국소 문맥의 구문 관계 종류를 결정한다.
조사가 있으면 그 조사에 따라 구문 관계가 결정되고, 그렇지 않은 경우에는 왼쪽 노드와 오른쪽 노드의 category에 따라 구문 관계가 결정된다.
Step 4. 구문 관계로 연결된 두 단어를 얻고 이들 사이의 방향성을 결정한다.

그림 5 구문 트리로부터 국소 문맥 추출

5.1.4 국소 문맥 데이타베이스

형태소 분석과 구문 분석을 통하여 학습 말뭉치로부터 추출된 국소 문맥은 통계 처리가 용이하도록 데이타베이스로 구축된다. 각 구문 관계별 빈도수를 기록하여 구문 관계 빈도수 데이타베이스에 저장하고 중심 단어에 대한 국소 문맥을 동일한 국소 문맥에 대한 중심 단어들의 리스트 형태로 국소 문맥 데이타베이스에 저장하였다. 구문 관계 빈도수 데이타베이스와 국소 문맥 데이타 베이스의 한 엔트리의 형식은 각각 다음과 같다.

[R : freq]

그림 6 구문 관계 빈도수 데이타베이스 (R:구문관계)

[(R, direction, W_a)
: {(W_1,$freq_1$,lh_1,$sense_1$),(W_2,$freq_2$,lh_2,$sense_2$),·,(W_n,$freq_n$,lh_n,$sense_n$)}]
여기서, $lh_i \geq lh_{i+1}$ 이고, lh_i는 log-likelihood ratio 이다.

그림 7 국소 문맥 데이타베이스

그림 7에서 W_i는 국소문맥의 중심 단어를 가리키고, $freq_i$는 W_i가 국소 문맥에 나타나는 횟수를 가리킨다. 또한 중심 단어가 국소 문맥에 어느 정도로 관련되는지를 측정하여 관련성이 높은 중심 단어를 우선적으로 고려할 수 있도록 log-likelihood ratio[31]를 계산하여 이 값에 의해 중심 단어들을 내림차순으로 정렬하였다. $sense_i$는 중심 단어 W_i가 국소 문맥하에서 결정된 의미를 뜻하며, 이 값은 초기에 NULL로 할당되지만, 유사도에 의한 단어 의미 결정 과정에서 국소 문맥하에서의 의미를 할당받게 된다. 이 중에서 술어-명사 구문 관계의 경우 명사들의 의미가 결정되면 이 명사들의 의미를 분류하여 술어의 의미를 구분하는 데 이용될 수 있다. 국소 문맥에 의한 의미 결정 과정은 6장의 알고리즘에서 자세히 다루고 있다.

그림 8은 국소 문맥 데이타베이스의 한 엔트리 예를 보여주고 있다.

[(목적어-술어, Head, 입-다)
: {(옷, 17, 81.97, *NULL*), (상처, 5, 31.82, *NULL*), (피해, 5, 31.82, *NULL*), ...}]

그림 8 국소 문맥 데이타베이스의 한 엔트리

5.2 공기 정보의 활용

2장에서 살펴 본 바와 같이 기존의 WSD 연구에서 이미 단어 의미 결정을 위한 지식원으로서 다양한 종류의 정보를 사용하고 있다. 구문적 의존 관계만을 지식원으로 사용하는 것은 실제 문장에 대한 정확도가 낮을 수 밖에 없다. 그리하여 본 연구에서는 5.1에서 설명한 것처럼 국소 문맥의 정의를 다양하게 확장하였다. 이 장에서는 또 하나의 지식원으로서 공기 정보를 활용하고자 한다. 본 연구에서 사용하는 공기 정보는 대상 단어가 특정 의미를 가질 때, 대상 단어를 포함하는 문장 내에 나타나는 명사들 중에서 대상 단어를 제외한 명사들로 정의한다. 다시 말해 공기 정보는 대상 단어의 의미와 대상 단어와 공기하는 다른 단어들 사이의 관계로 정의된다.

교사 학습 방식에서는 의미 부착된 말뭉치를 사용하기 때문에 이것으로부터 공기 정보를 학습할 수 있지만, 비교사 학습 방식의 경우 원시 말뭉치에서는 그러한 정보를 학습하기가 곤란하다. 말뭉치로부터 공기 정보를 얻기 위해서는 대상 단어가 하나의 의미로 결정되어 있어야 하고, 대상 단어의 의미를 결정하기 위해 다시 공기 정보가 필요한 순환 논리에 빠지게 된다. 이를 탈피하기 위해 본 연구에서는 최초의 공기 정보를 사전 정의와 예문으로부터 얻는 방법을 사용하였다. [32]에서 대상 단어의 각 의미 정의와 예문으로부터 대상 단어의 각 의미별 공기 정보를 얻어 공기 정보 데이타 베이스를 마련하였다. 또한 이 초기 공기 정보와 국소 문맥을 이용하여 입력 문장의 단어 의미 중의성을 해소하는 과정에서 대상 단어의 의미가 결정될 때 문장 내의 공기 단어들을 대상 단어의 선택된 의미에 대한 공기 정보에 추가하여 공기 정보 데이타베이스를 확장해 나갈 수 있다. 공기 정보 데이타베이스의 공기 단어들은 발생 빈도수에 따라 내림차순으로 정렬된다. 그림 9는 '가사'의 4번째 의미에 대한 최초 공기 정보를 보여주고 있다.

[(가사, sense#4)
: {(가곡,1), (가요,1), (곡,1), (오페라,1), (노래,1), (내용,1), (글,1), (말,1)}]

그림 9 최초 공기 정보 데이타베이스의 한 엔트리

5.3 용언의 확장

실제 문장을 처리할 때 국소 문맥의 적용률을 높이기 위한 방법으로 용언의 확장을 활용하였다. 한국어의 용언은 대체로 자동사-타동사, 피동사-사동사의 변환이 가능하며, 이때 조사에 의해 실현되는 구문 관계는 변하더라도 그 의미 제약은 그대로 유지된다. 따라서 용언의 변화와 그에 따른 조사의 변화를 파악하여 국소 문맥에 용언이 나타날 때 그 용언의 변화에 따라 변형된 국소 문맥을 얻고 이로부터 단서들을 얻을 수 있으므로 보다 충실한 단서의 획득이 가능하며, 국소 문맥의 활용도를

높이는데 기여한다. 국소 문맥 데이타베이스에서 동일한 국소 문맥을 찾을 때 용언 확장에 의한 국소 문맥도 동일한 국소 문맥에 포함된다. 용언의 변화와 그에 따른 조사의 변화는 [33]을 기반으로 하였으며, 그림 10에 용언 확장의 몇 가지 예를 보였다.

```
(-를, Head, 입-다) <=> (-를, Head, 입히-다)
(-가, Head, 가라앉-다) <=> (-를, Head, 가라앉히-다)
(-가, Head, 담기-다) <=> (-를, Head, 담-다)
(-에, Head, 담기-다) <=> (-에, Head, 담-다)
```

그림 10 용언 확장에 의한 국소문맥

6. 의미 결정 알고리즘

앞 장에서 설명한 국소 문맥과 공기 정보 등을 이용하여 입력된 문장에서 명사의 의미 중의성을 해소하여 하나의 의미를 결정하는 알고리즘을 소개한다. 먼저, 국소 문맥에 의한 단서들과 대상 단어의 유사도를 최대로 하는 대상 단어의 의미를 선택하는 과정을 설명하고, 다음에 다양한 지식원들을 결합하여 대상 명사의 최선의 의미를 결정하는 과정을 설명하겠다.

6.1 유사도 행렬을 이용한 최대 유사도

대상 단어의 의미 결정을 위해 자신이 포함되는 문맥들로부터 학습된 자료를 이용하는 방법과 달리 본 논문에서 제안하는 방법은 서로 다른 단어들을 단서로 활용할 수 있기 때문에 국소 문맥으로 학습된 단서들의 활용도를 높여 필요한 말뭉치의 크기를 줄일 수 있다. 대상 단어의 의미는 단서 단어들과의 유사도 계산을 통해 얻어질 수 있다. 두 단어의 유사도는 의미 계층 구조에 기반한 두 개념 사이의 의미적 거리를 이용하여 계산된다[34].

$$similarity(A,B)=\frac{2\times|A\cap B|}{|A|+|B|}$$

그런데, 국소 문맥에 의한 단서 단어들은 중의성을 내포하고 있을 뿐만 아니라 국소 문맥을 이루는 주변 단어 또한 중의성을 가질 수 있으므로 단서 단어들 중에는 대상 단어의 의미를 결정하는 데 오히려 잡음으로 작용하는 경우도 있다. 따라서, 단순히 대상 단어와 단서 단어들 사이의 유사도만 계산하는 것이 아니라, 단서 단어들 사이의 유사도를 함께 계산하여 서로간에 가장 강한 지지를 받는 단어의 의미부터 차례로 결정하여 잡음을 줄이면서 단서 단어들의 의미 중의성도 해소해 간다. 이 과정은 대상 단어의 의미가 결정될 때까지 반복된다. 이 과정에서 얻어진 단서 단어들의 의미 중의성 해소는 국소 문맥 데이타베이스에 반영되어 그 국소 문맥이 다음에 참조될 때에는 단서 단어들의 중의성이 많이 해소되었기 때문에 속도 향상을 기대할 수 있다. 다만 이 과정에서 발생할 수 있는 오류를 줄이기 위해 단서 단어의 의미가 다른 단서들로부터 받는 지지도의 평균이 임계치를 넘고, 가장 큰 지지도 값과 두 번째 큰 지지도 값의 차이가 클 때에만 이를 적용하였다. 이 과정은 Lin의 [1]와 Resnik의 [35]에서 사용한 방법과 유사하다.

대상 단어를 w_0, 단서 집합을 $\{w_1,w_2,,w_k\}$, w_i가 갖는 의미를 $\{s_{i1},,s_{ini}\}$라고 할 때 대상 단어와 단서 단어들 사이의 최대 유사도의 계산 과정은 그림 11과 같다.

Step A. (k+1) ×(k+1) SimBlocks로 이루어진 Similarity Matrix를 구성한다.

$$SimBlock_{ij}(l,m)=\begin{cases}similarity(s_{il},s_{jm}) & \text{if } i\neq j \;\&\; similariy(s_{il},s_{jm})\geq \grave{E}_1\\ 0 & \text{otherwise}\end{cases}$$

where, $l\in[1,n_i]$, $m\in[1,n_j]$

Step B. 중의성을 가진 단어의 목록을 기록한다.

$$IsAmbiguous[i]=\begin{cases}1 & \text{if } n_i>1\\ 0 & \text{otherwise}\end{cases}$$

Step C. 각 단어(w_i)의 각 의미(s_{il})가 다른 단어(w_j)로부터 얻는 지지도 값을 구하고 가장 큰 지지도 합계를 받는 단어(w_{imax})와 그 의미($s_{imaxlmax}$)를 구한다.

$$support(s_{il},w_j)=\max_{m\in[1,n_j]} SimBlock_{ij}(l,m) \quad \text{if } IsAmbiguou\,s[i]=1$$

$$(i_{\max},l_{\max})=\arg\max_{\substack{i\in[0,k]\\ l\in[1,n_i]}}\sum_{j=0}^{k} support(s_{il},w_j)$$

Step D. 다음 조건을 만족할 때 w_{imax}의 의미를 $s_{imaxlmax}$로 결정하여 국소 문맥 데이타 베이스(LCDB)에 반영한다.

$$\text{if } \min_{\substack{l\in[1,n_{i_{\max}}]\\ l\neq l_{\max}}}\left\{\frac{1}{k}\sum_{j=0}^{k} support(s_{i_{\max},w_j})-\frac{1}{k}\sum_{j=0}^{k} support(s_{l},w_j)\right\}\geq \grave{E}_2$$

then $LCDB[w_{i_{\max}}].sense \leftarrow s_{i_{\max}l_{\max}}$

$IsAmbiguous[i_{\max}]\leftarrow 0$

Step E. $i_{max}\neq 0$ 이면 Similarity Matrix를 갱신한다.

$$SimBlock_{i_{\max}j}(l,m)\leftarrow 0$$

$$\text{if } j\in[0,k]\;\; l\in[1,n_{i_{\max}}]\;\; l\neq l_{\max},\;\; m\in[1,n_j]$$

Step F. $i_{max}=0$일 때까지 Step C부터 반복한다.

그림 11 최대 유사도 계산 과정

6.2 다양한 지식원의 결합

중의성을 가진 명사가 문장에 나타날 때 그 명사는 둘 이상의 서로 다른 국소 문맥을 가질 수 있다. 그런

Step 1. 입력문장을 태깅하고 구문분석하여 구문트리를 얻는다.
Step 2. 구문 트리로부터 명사(중심어)의 국소 문맥을 추출하여 국소문맥 리스트(LClist)를 만든다.
LClist = {[W_{C1},LC_1],[W_{C2},LC_2],...} = {[W_{C1},(R_1,D_1,W_{art})],...}
W_C : 중심어, W_{ar} : 주변 단어,
R : 국소 문맥의 종류, D : 구문 관계의 방향
Step 3. SenseDB에서 각 중심어의 가능한 의미를 얻어 DecisionList를 만든다.
DecisionList = {[W_{C1},{($sense_{11}$,$support_{11}$), ($sense_{12}$,$support_{12}$),...},...}
Step 4. LClist의 모든 엔트리에 대해, 각 W_{Ci} 가 명사이면서 중의성이 있을 때
◎ 국소문맥데이타베이스(LCDB)에서 LC_i를 검색하고 그 결과를 단서리스트(SelectorList)에 저장한다.
◎ SelectorList[] <- SelectorList[] - W_{Ci}
◎ 중심어(W_{Ci})의 각 의미(s_{il})에 대한 단서들의 전체 지지도 값(TotalSupport$_{il}$)을 구한다(그림11)
◎ 전체 지지도를 평균하고 국소 문맥의 종류에 따라 가중치를 부여한다.
TotalSupport$_{il}$ <- ***weight***(Ri) (TotalSupport$_{il}$ / *# of Selectors*)
◎ 전체 지지도를 DecisionList의 해당 의미의 지지도에 누적한다.
DecisionList[i].support$_{il}$ DecisionList[i].support$_{il}$ + Total Support$_{il}$
Step 5. DecisionList[i]에서 가장 큰 support가 임계치 이상이면 그 의미를 최종 의미로 선택한다.

$$\text{support}_{il_{\max}} \leftarrow \max_{l \in [1,n_i]} \left(DecisionList[i].\text{support}_{il} \right)$$

if ($\text{support}_{il_{\max}} \geq \Theta_3$)
then $W_{c_i}.sense \leftarrow DecisionList[i].sense_{il_{\max}}$

Step 6. 그렇지 못한 경우, 공기 정보 데이타베이스(CODB)로부터 대상 단어의 공기 단어들과 비교하여 대상 단어 의미를 결정한다. 그러나 만약 적용되지 못한 경우에는 국소 문맥의 결과를 따른다.

$k_{\max} \leftarrow \arg\max_k (matchingCo\ unt(CooccurLis\ t, CODB[W_{c_i}.sense_k]))$
if $k_{\max} \neq$ NULL
then $W_{c_i}.sense \leftarrow DecisionLi\ st[i].sense_{ik_{\max}}$
else $W_{c_i}.sense \leftarrow DecisionLi\ st[i].sense_{il_{\max}}$

그림 12 단어 의미 결정 알고리즘

데, 5.1에서 정의한 국소 문맥들은 그 종류에 따라 중심 단어와 주변 단어 사이의 의미 제약의 강도가 다르다 [2]. 술어-목적어 관계는 술어-주어 관계보다 선택 제약(selectional restriction)이 더 강하고, 술어-논항 관계에 의한 제약은 수식-피수식 관계에 의한 제약보다 강하고, 또한 명사 연어 정보나 대등 관계에 의한 제약보다 강하다. 따라서, 대상 단어가 둘 이상의 국소 문맥에 속할 경우 [1]에서처럼 각 국소 문맥을 동등하게 취급하는 것보다는 가중치를 주어 각 국소 문맥의 기여도를 조정할 필요가 있다. 대상 단어는 국소 문맥의 단서들에 의해 각 의미별로 의미 유사도만큼의 지지(support)를 받는다. 즉, 둘 이상의 국소 문맥이 가능한 경우 각각의 평균 지지도에 국소 문맥에 따른 가중치를 곱한 후 이들을 합하여 가장 큰 지지도 값을 갖는 의미가 대상 단어의 의미로 선택된다. 여기서 가중치는 실험을 통해서 설정해 주었다.

그러나, 국소 문맥에 의한 지지도가 임계치(threshold)에 미치지 못할 경우 공기 정보가 고려된다. 적용될 국소 문맥이 없거나 관용적 표현으로 사용된 경우에는 국소 문맥에 의한 의미별 지지도는 불확실한 값을 갖게 된다. 따라서, 이런 경우에는 대상 단어의 각 의미별 공기 정보가 의미 결정의 근거로 사용된다.

말뭉치로부터 학습된 국소 문맥 데이타베이스(LCDB)와 사전에서 학습한 공기 정보 데이타베이스(CODB)를 이용하여 입력 문장에 나타나는 명사의 의미 중의성을 해소하면서 국소 문맥 데이타베이스와 공기 정보 데이타베이스의 내용을 학습해가는 전체 의미 결정 알고리즘은 그림 12에 보였다.

7. 실험 및 결과 분석

본 연구에서 제시한 명사 의미 중의성 해소 방법을 검증하기 위한 말뭉치로 국어 정보 베이스 CD-ROM (97.5.1)의 일부를 사용하였다. 국어 정보 베이스 CD-ROM은 ETRI 컴퓨터 소프트웨어 연구소에서 자연언어 처리 연구의 평가를 위해 제공된 말뭉치로 사전관리 시스템(TDMS)과 1,000만 어절 한국어 텍스트 말뭉치, 100만 어절 한국어 품사부착 말뭉치, 1만 어절 한국어 구문구조부착 말뭉치를 담고 있다. 실험에서는 이 중에서 한국어 텍스트 말뭉치의 일부를 사용하였다.

우선, 의미 중의성이 많은 두 개의 명사('가사'와 '공사')를 대상 단어로 선정하고, 이 단어들을 포함하고 있는 문장들을 말뭉치에서 추출하여 형태소 분석과 구문분석을 거쳐 대상 단어를 중심 단어로 하는 국소 문맥을 추출하여 국소 문맥 데이타 베이스를 구축하였다. 또한 이 국소 문맥이 나타나는 문장들을 말뭉치로부터 추출하고 국소 문맥과 단서 단어들을 국소 문맥 데이타베이스에 추가하였다. 다음으로, 대상 단어들과 단서 단어들의 가능한 의미들을 WordNet의 의미 계층 구조로부터 얻어 의미 데이타베이스를 마련하였다. 그리고, 초기 공기 정보로 이용될 지식을 얻기 위해 [32]에서 대상 단어의 정의와 예문들을 이용하여 초기 공기 정보 데이타베이스를 마련하였다.

표 4 '공사'의 의미

'공사'	의 미
1	[公使] 외교를 맡아보는 공무 원
2	[工事] 토목 · 건축 등의 일
3	[公私] 공공의 일과 사사로운 일
4	[公事] 공무
5	[公社] 공공 기업체의 하나

이 실험의 의미 결정 단계에에서는 앞에서 학습을 위해 마련한 대상 단어를 포함하는 문장들을 그대로 이용하였다. 본 연구에서 사용한 방법은 비교사 학습 방식으로 학습 단계에서는 단지 국소 문맥을 학습할 뿐이며 유사도를 계산하는 과정에서 대상 단어 자신은 단서 단어가 될 수 없기 때문에 학습에 사용한 문장들을 테스트에 그대로 사용하더라도 성능 향상에 아무런 영향을 미치지 않는다. 즉, 대상 단어가 학습된 국소 문맥의 중심 단어로 나타나지 않았다 하더라도 그 국소 문맥을 갖는 다른 단서 단어들을 이용하여 의미를 결정하기 때문이다. 최대 유사도 계산에서는 속도를 고려하여 사용되는 단서 단어의 개수를 49개로 제한하였다.

대상 단어 '가사'와 '공사'가 가질 수 있는 의미들은 각각 표 1과 표 4에 나와 있다. 대상 단어의 의미 중의성의 난이도를 알아보고 실험 결과에서 대상 단어 의미 결정의 성공과 실패를 판단하기 위해 5명의 사람이 실험에 사용된 문장들에서 대상 단어의 의미를 결정하게 하였다. 즉, 5명이 선택한 의미들 사이의 불일치 정도를 대상 단어의 중의성 해소의 난이도로 판단할 수 있고, 실험 결과에서 대상 단어의 결정된 의미를 5명의 사람 중에 선택한 사람이 있으면 그 결과를 성공으로 보고, 한 사람도 그런 의미를 선택하지 않은 경우에는 실패한 것으로 간주하였다. 이것은 문장에 나타난 중의성을 가진 단어의 의미에 대해 사람들 사이에 의견이 서로 다를 수 있다는 점을 고려한 것이다.

실험에 사용된 대상 단어의 난이도는 표 5와 같고 이를 기준으로 실험 결과를 분석한 결과는 표 6에 보였다. baseline은 말뭉치에서 가장 빈번한 의미를 선택했을 경우를 말한다. 두 단어 각각에서 각 의미별 값을 합한 값이 Tot.의 값과 다른 것은 바로 사람의 실험에서의 불일치를 고려하였기 때문이다.

표 5 대상 단어의 난이도

	'가사'	'공사'
난이도	7.8%	9.4%

표 6 실험 결과

		발생 빈도	Base line (%)	국소문맥		공기정보		국소문맥 + 공기정보	
				적용률 (%)	정확도 (%)	적용률 (%)	정확도 (%)	적용률 (%)	정확도 (%)
가사	1	20	-	100.0	85.0	60.0	100.0	100.0	90.0
	2	27	-	96.3	46.2	77.8	76.2	100.0	77.8
	3	1	-	100.0	100.0	100.0	100.0	100.0	100.0
	4	47	-	97.9	91.3	46.8	100.0	97.9	95.7
	5	29	-	100.0	86.2	13.8	100.0	100.0	86.2
	Tot.	115	40.9	98.3	77.9	47.8	90.9	99.1	87.7
공사	1	122	-	100.0	96.7	9.8	83.3	100.0	96.7
	2	272	-	98.9	88.8	31.6	100.0	99.6	90.8
	3	9	-	100.0	88.9	22.2	100.0	100.0	88.9
	4	22	-	100.0	77.3	18.2	75.0	100.0	77.3
	5	69	-	98.6	79.4	17.4	75.0	98.6	79.4
	Tot.	446	61.0	99.3	89.2	24.2	94.4	99.8	90.3
Total Ave.		561	56.9	99.1	86.9	29.0	93.7	99.7	89.8

실험 결과에서 보는 바와 같이 동일한 국소 문맥에 나타나는 다른 명사들을 단서로 사용함으로써 국소 문맥에 의한 적용률을 높일 수 있음을 알 수 있다. 또한, 이 사실은 대상 단어의 의미 중의성 해소에 필요한 말뭉치의 크기를 줄일 수 있음을 뜻한다. 사전에서 얻은 공기 정보는 그 자체로는 적용률이 너무 낮지만, 국소 문맥과 함께 고려함으로써 국소 문맥을 보완하여 정확도를 상당히 높이는 역할을 할 수 있음을 알 수 있다.

본 연구에서 수행한 두 번째 실험은, 제안한 방법을 기존의 방법들과 비교하기 위한 것이다. [19]과의 비교를 위해 '경기', '보도', '보수', '지도'의 단어에 대해 실험을 수행하였고, [36]와의 비교를 위해 '전자'에 대해 실험을 수행하였다.

표 7 기존 연구 결과와의 비교

	의미 수	발생 빈도	적용율(%)	정확도(%)
경기	4	62	96.8	78.3
보도	4	49	100.0	81.6
보수	3	37	97.3	72.2
지도	2	46	95.7	75.0
전자	2	79	96.2	75.0

기존의 교사 학습 방식의 연구와 비교해 볼 때, [36]는 '배', '전자'의 두 단어에 대해 각각 82.2%와 92.2%의 정확도를 보였고, [19]에서는 10개의 명사에 대해 평균 84.5%의 정확도를 보였는데, 본 논문에서 제안한

방법은 '전자'에 대해 75.0%의 정확도를 보였고, '경기', '보도', '보수', '지도'의 단어에 대해 평균 77%의 정확도를 보였다. 이 결과는 기존의 교사 학습 방식의 연구 결과보다 낮은 정확도를 보였지만, 본 연구가 비교사 학습 방식이라는 점을 감안하면 우수한 결과라고 할 수 있다.

8. 결론 및 향후 계획

기존의 비교사 학습 방식의 말뭉치 기반 단어 의미 중의성 해소 방법들을 분석하고 각각의 장점들을 취하여 한국어의 명사 의미 중의성 해소에 적용하였다. 우리말에 대한 의미 부착 말뭉치가 거의 없는 실정을 감안하여 이를 필요로 하지 않는 비교사 학습 방식을 택하였고, 초기의 공기 정보를 사전으로부터 학습하고 원시 말뭉치로부터 국소 문맥을 학습하였다. 또한 단어의 의미를 결정하는 과정에서 국소 문맥 데이타베이스의 단서들의 의미를 학습하고 공기 정보를 추가적으로 학습하는 방법을 제시하였다.

이 방법은 대상 명사의 의미를 결정하기 위해 대상 명사가 갖는 국소 문맥과 동일한 국소 문맥을 갖는 다른 단어들을 단서로 사용함으로써 학습 자료의 활용도를 높일 수 있었다. 또한 자료 부족 문제를 극복하기 위해 용언의 확장을 이용하였다.

두 단어 사이의 의미 유사도 계산을 위해 의미 계층 구조를 WordNet으로부터 차용하였고, 학습된 국소 문맥 데이타베이스의 단서들은 입력 문장 내의 대상 명사의 의미를 결정하는 과정에서 함께 의미가 결정될 수 있으므로 반복 과정을 통해 최대 유사도 계산의 속도를 향상시킬 수 있다. 또 대상 단어에 적용될 국소 문맥이 둘 이상일 경우 국소 문맥의 종류에 따라 가중치를 부여하는 방식으로 지식원 결합을 사용하였고, 평균 지지도에 따라 공기 정보를 보완적으로 사용하였다.

평가 방법에서 대상 단어의 난이도를 파악하고 사람의 의미 중의성 해소 능력과 비교할 수 있는 방법을 제시하였고, 두 명사를 사용한 실험으로부터 국소 문맥의 적용률이 상당히 높음을 보였고 공기 정보는 국소 문맥을 보완하여 정확도를 높이는 역할을 할 수 있음을 보였다.

앞으로는 WordNet과 한영 대역어 사전을 이용한 한국어 명사의 의미 계층 구조를 얻기 위한 수작업을 줄이기 위해 자동화 할 수 있는 도구를 개발할 필요성이 있다. 그리고 이와 유사한 방법을 용언의 의미 중의성에도 적용해 볼 만하고 보다 많은 학습 말뭉치로부터 학습한 후 의미 해석 시스템의 단어 중의성 해소에 적용하여 의미 해석 시스템의 성능을 향상하기 위한 연구가 수행되어야 할 것이다.

참 고 문 헌

[1] Lin, Dekang, Using Syntactic Dependency as Local Context to Resolve Word-Sense Ambiguity., in Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics. Somerset, N.J : Association for Computational Linguistics, 1997

[2] Resnik, P., Selectional Preference and Sense Disambiguation. in Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? , pp. 52-57, Somerset, N.J.: Association for Computational Linguistics, 1997

[3] Kelly, Edward F. and Philip J. Stone, Computer Recognition of English Word Senses, North-Holland, Amsterdam, 1975

[4] Guthrie, Joe A., Louise Guthrie, Yorick Wilks, and Homa Aidinejad, Subject-dependent co-occurrence and word sense disambiguation., in Proceedings of the 29th Annual Meeting, pp. 146-152, Berkeley, CA, June. Association for Computational Linguistics, 1991

[5] Lesk, Michael, Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. in Proceedings of the 1986 SIGDOC Conference, pp. 24-26, Toronto, Canada, June, 1986

[6] Walker, Donald E. and Amsler Robert, The use of machine-readable dictionaries in sub-language analysis., in Ralph Grishman and Richard Kittredge. 1986

[7] Wilks, Yorick A., Dan Fass, Cheng-Ming Guo, James E. MacDonald, Tony Plate, and Brian A. Slator. Providing machine tractable dictionary tools. in James Pustejovsky, editor, Semantics and the Lexicon. MIT Press, Cambridge, MA, 1990

[8] Agirre, E., and Rigau, G., Word-Sense Disambiguation Using Conceptual Density, in Proceedings of the 16th International Conference on Computational Linguistics, Somerset, N.J.: Association for Computational Linguistics, 1996

[9] Li, Xiaobin, Stan Szpakowicz and Stan Matwin, A WordNet-based algorithm for word sense disambiguation, in IJCAI'95, pp. 1368-1374, 1995

[10] Yarowsky, David, Word sense disambiguation using statistical models of Roget's categories trained on large corpora., in Proceedings of the 14th International Conference on Computational Linguistics, COLING'92, pp. 454-460, Nantes,

France, August, 1992

[11] Miller, George A., Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas, WordNet: An on-line Lexical database, International Journal of Lexicography, 3(4), pp. 235-244, 1990

[12] Bruce, Rebecca and Janyce Wiebe. Word-sense disambiguation using decomposable models, in Proceedings of the 32nd Annual Meeting, pp. 139-145, Las Cruces, NM. Association for Computational Linguistics, 1994

[13] Leacock, Claudia, Geoffrey Towwell, and Ellen M. Voorhees. Corpus-based statistical sense resolution, in Proceedings of the ARPA Human Language Technology Workshop, San Francisco, Morgan Kaufmann, 1993

[14] McRoy, Susan W. Using multiple knowledge sources for word sense discrimination, Computational Linguistics, 18(1), pp.1-30, 1992

[15] Ng, Hwee Tou and Hian Beng Lee, Integrating multiple knowledge sources to disambiguation word sense: An examplar-based approach, in Proceedings of the 34th Annual Meeting, pp. 40-47, University of California, Santa Cruz, CA, June, Association for Computational Linguistics, 1996

[16] Ng, Hwee Tou and John Zelle, Corpus-Based Approaches to Semantic Interpretation in Natural Language Processing, in American Association for Artificial Intelligence, pp. 45-64, 1997

[17] Niwa, Yoshiki and Yoshihiko Nitta. Co-occurrence vectors from corpora vs distance vectors from dictionaries in Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, pp. 304-309, Kyoto, Japan, August, 1994

[18] Yarowsky, David. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French in Proceedings of the 32nd Annual Meeting, pp. 88-95, Las Cruces, NM. Association for Computational Linguistics, 1994

[19] 김봉섭. 한-일 기계번역에서 의미 태깅된 말뭉치의 자동 생성 및 이를 이용한 명사의 의미 중의성 해소, 포항공과대학교 석사학위 논문, 1998

[20] Dagan, Ido and Alon Itai. Word sense disambiguation using a second language monolingual corpus, in Computational Linguistics, 20(4), pp.563-596, 1994

[21] Leacock, Claudia and Martin Chodorow. Using Corpus Statistics and WordNet Relations for Sense Identification, Computational Linguistics, 24(1), pp.147-165, 1998

[22] Schtze, Hinrich. Ambiguity in Natural Language Learning Computational and Cognitive Models, Ph.D. Dissertation, Stanford University, 1995

[23] Yarowsky, David. Unsupervised word sense disambiguation rivaling supervised methods, in Proceedings of the 33rd Annual Meeting, pp. 189-196, Cambridge, MA, June. Association for Computational Linguistics, 1995

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

[25] 조평옥, 한국어 명사의 의미 계층 구조 구축, 울산대학교 박사학위 논문, 1996

[26] Kilgarriff, Adam. I don't believe in word senses, manuscript

[27] Choueka, Yaacov and Serge Lusignan. Disambiguation by short contexts, Computers and the Humanities, 19, pp. 147-158, 1985

[28] Leacock, Claudia, Geoffrey Towwell, and Ellen M. Voorhees. Towards building contextual representations of word senses using statistical models, in Corpus Processing for Lexical Acquisition. The MIT Press, chapter 6, pp. 97-113, 1996

[29] Jeongwon Cha, Geunbae Lee and Jong-Hyeok Lee. Generalized Unkown Morpheme Guessing for Hybrid POS Tagging of Korean, in Proceedings of the 6th Workshop on Very Large Corpora, COLING-ACL'98, pp. 85-93, 1998

[30] 이원일. 단일화 기반 범주 문법에 기반한 음성 한국어 처리, 포항공과대학교 박사학위 논문, 1998

[31] Dunning, Ted. Accurate methods for the statistics of surprise and coincidence, Computational Linguistics, 19(1), pp. 61-74, March, 1993

[32] 김민수. 그랜드 국어 사전;on-line version, 금성출판사, 1993

[33] 홍재성 외 9인, 현대 한국어 동사 구문 사전, 두산 동아, 1997

[34] 권혜진. 범주 문법과 논리 구조에 기반한 자연어 질의의 의미 분석, 포항공과대학교 석사학위 논문, 1997

[35] Resnik, P., Disambiguating noun groupings with respect to WordNet senses, in Third Workshop on Very Large Corpora. Association for Computational Linguistics, 1995

[36] 이호, 백대호, 임해창. 최소한의 코퍼스 정보를 이용한 단어 의미 중의성 해결 기법, 한국 정보 과학회 봄 학술발표논문집 24권 1호, 1997

이 승 우

1997년 2월 경북대학교 공과대학 컴퓨터공학과(학사). 1999년 2월 포항공과대학교 컴퓨터공학과(석사). 1999년 1월 ~ 현재 포항공과대학교 정보통신연구소 위촉연구원. 관심분야는 자연언어 처리, 정보 검색, SGML/XML

이 근 배

정보과학회논문지 : 소프트웨어 및 응용
제 27 권 제 1 호 참조