상호정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소

(Homonym Disambiguation based on Mutual Information and Sense-Tagged Compound Noun Dictionary)

허 정 † 서 희 철 † 장 명 길 †

(Jeong Heo) (Hee-Cheol Seo) (Myung-Gil Jang)

요 약 자연언어처리의 목적은 컴퓨터가 자연어를 이해할 수 있도록 하여, 인간에게 다양한 정보를 정확하고 빠르게 전달할 수 있도록 하고자 하는 것이다. 이를 위해서는 언어의 의미를 정확히 파악하여야 하는데, 어휘 의미 중의성 해소가 필수적인 기술이다.

본 연구는 상호정보량과 기 분석된 복합명사 의미사전에 기반한 동음이의어 의미 중의성 해소를 위한 기술을 소개한다. 사전 뜻풀이를 이용하는 기존 기술들은 어휘들간의 정확한 매칭에 의존하기 때문에 자료 부족 현상이 심각하였다. 그러나, 본 연구에서는 어휘들간의 연관계수인 상호정보량을 이용함으로써 이 문제를 완화시켰다. 또한, 언어적인 특징을 반영하기 위해서 상호정보량을 가지는 어휘 쌍의 비율 가중치, 의미 별 비율 가중치와 뜻풀이의 길이 가중치를 사용하였다. 그리고, 복합명사를 구성하는 단일명사들은 서로의 의미를 제약한다는 것에 기반하여 고빈도 복합명사에 대해서 의미를 부착한 의미사전을 구축하였고, 이를 동음이의어 중의성 해소에 활용하였다.

본 시스템의 평가를 위해 질의응답 평가셋의 200 여 개의 질의와 정답단락을 대상으로 동음이의어 의미 중의성 해소 평가셋을 구축하였다. 평가셋에 기반하여 네 유형의 실험을 수행하였다. 실험 결과는 상호정보량만을 이용하였을 때 65.06%의 정확률을 보였고, 가중치를 활용하였을 때 85.35%의 정확률을 보였다. 또한, 복합명사 의미분석 사전을 활용하였을 때는 88.82%의 정확률을 보였다.

키워드 : 상호정보량, 의미사전, 동음이의어, 중의성해소

Abstract The goal of Natural Language Processing (NLP) is to make a computer understand a natural language and to deliver the meanings of natural language to humans. Word sense Disambiguation (WSD) is a very important technology to achieve the goal of NLP.

In this paper, we describe a technology for automatic homonyms disambiguation using both Mutual Information (MI) and a Sense-Tagged Compound Noun Dictionary. Previous research work using word definitions in dictionary suffered from the problem of data sparseness because of the use of exact word matching. Our work overcomes this problem by using MI which is an association measure between words. To reflect language features, the rate of word-pairs with MI values, sense frequency and size of word definitions are used as weights in our system. We constructed a Sense-Tagged Compound Noun Dictionary for high frequency compound nouns and used it to resolve homonym sense disambiguation.

Experimental data for testing and evaluating our system is constructed from QA (Question Answering) test data which consisted of about 200 query sentences and answer paragraphs. We performed 4 types of experiments. In case of being used only MI, the result of experiment showed a precision of 65.06%. When we used the weighted values, we achieved a precision of 85.35% and when we used the Sense-Tagged Compound Noun Dictionary, we achieved a precision of 88.82%, respectively.

Key words : MI, Sense-Tagged Dictionary, Homonym, WSD

† 정 회 원 : 한국전자통신연구원 지식마이닝연구팀 연구원
jeonghur@etri.re.kr
hcseo@etri.re.kr
mgjang@etri.re.kr

논문접수 : 2006년 2월 24일
심사완료 : 2006년 11월 2일

표 1 '표준국어대사전'에서의 '다리'의 의미 구분

동음이의어	다의어 번호	뜻풀이
다리01	01	동물의 몸통 아래 붙어 있는 신체의 부분. 서고 걷고 뛰는 일 따위를 맡아 한다.
	02	물체의 아래쪽에 붙어서 그 물체를 받치거나 직접 땅에 닿지 아니하게 하거나 높이 있도록 버티어 놓은 부분
	03	오징어나 문어 따위의 동물의 머리에 여러 개 달려 있어, 헤엄을 치거나 먹이를 잡거나 촉각을 가지는 기관
	04	안경의 테에 붙어서 귀에 걸게 된 부분
다리02	01	물을 건너거나 또는 한편의 높은 곳에서 다른 편의 높은 곳으로 건너 다닐 수 있도록 만든 시설물
	02	두 사물이나 사람 사이를 이어 주는 역할을 하는 것
	03	중간에 거쳐야 할 단계나 과정
	04	지위의 등급
다리03	01	예전에, 여자들의 머리 숱이 많아 보이라고 덧넣었던 딴 머리

1. 서 론

자연언어처리의 목적은 디지털화된 다양한 자연어를 컴퓨터가 이해할 수 있도록 하여 인간에게 다양하고 정확한 정보를 빠르게 제공할 수 있도록 하는 것이다. 컴퓨터가 자연어를 이해한다는 것은 자연어가 표현하는 의미를 다양한 정보에 기반하여 파악하는 것으로 정의할 수 있다. 이를 위해서는 형태소분석, 구문분석, 의미분석, 담화분석 등과 같이 다양한 기술들이 요구된다. 특히 의미분석은 자연어가 표현하는 정확한 의미를 컴퓨터가 이해하기 위해서 반드시 요구되는 기술로써, 어휘의 의미를 파악하는 어휘 의미 분별(Word Sense Tagging)에서부터 시작된다.

대부분의 어휘들은 쓰이는 상황과 문맥에 따라 다양한 의미로 이용된다. 예를 들면, '다리'는 그 쓰임에 따라 표준국어대사전에서 표 1과 같이 의미를 구분하고 있다. 이처럼 하나의 어휘가 둘 이상의 의미로 이용되는 어휘들을 의미구분의 기준[1]에 따라 동음이의어(Homonym)와 다의어(Polysemy)로 구분할 수 있다.

표 1과 같이 동일한 형태의 어휘가 다양한 의미를 가지고 있을 때, 문맥 내의 정보를 기반으로 다양한 의미들 중 문맥에 적합한 하나의 의미를 선정하는 것이 어휘 의미 중의성 해소(Word Sense Disambiguation)이다. 어휘 의미 중의성 해소는 어휘 의미 분별의 핵심 기술로서 다양한 언어처리 응용분야들 (기계번역(Machine Translation), 정보검색(Information Retrieval), 질의응답(Question Answering), 음성처리(Speech Processing))에서 활용된다[1-3]. 다양한 언어처리 응용분야에서 어휘 의미 중의성 해소 기술의 효용성에 대한 연구가 활성화되면서, 어휘 의미 중의성 해소에 대한 다양한 연구가 진행되었다.

어휘 의미 중의성 해소에 대한 연구의 활성화로 다양한 어휘 의미 중의성 해소 기술에 대한 평가와 비교를 위한 체계적인 방법의 필요성이 요구되었다[4,5]. 1998년 ACL SIGLEX와 EURALEX의 후원 하에 어휘 의미 중의성 해소 기술 평가 대회인 SENSEVAL이 개최되어 2004년까지 4회에 이르렀다. SENSEVAL-2에서는 한국어에 대해서도 평가하였으나, 특정 샘플 어휘(Lexical Sample Task)에 대한 평가로서, 최고 성능을 보인 시스템은 고려대의 KUNLP 시스템으로 정확률(Precision)이 69.8%이고 재현율(Recall)이 74%였다[6-8]. 그러나, 국내에서 한국어에 대한 어휘 의미 중의성 해소에 대한 연구의 비활성화로 인해 한국어에 대한 평가는 SENSEVAL-3에서는 제외되었다.

어휘 의미 중의성 해소 기술은 의미 정보의 획득 유형에 따라 크게 지식에 기반한 중의성 해소 기술(Knowledge based WSD)과 코퍼스에 기반한 중의성 해소 기술(Corpus based WSD)로 구분할 수 있다[9].

지식에 기반한 중의성 해소 기술은 다양한 의미정보 지식을 기반으로 어휘의 중의성을 해소하는 기술로써 일반적으로 사전이나 워드넷(WordNet)[10]을 기반으로 한다. 사전에는 어휘에 대한 의미구분과 각각의 의미에 대한 뜻풀이를 비롯하여 문법정보, 용례정보, 상용어구 정보, 한자정보, 의미범주(Semantic Category)정보 등 다양한 정보를 포함하고 있다. 이러한 정보를 기반으로 선택제약(Selectional Restriction), 선택선호(Selectional Preference) 정보를 획득하여 어휘의 중의성을 해소한다. 워드넷은 심리언어학[2]을 기반으로 어휘들의 관계를 구조화한 언어자원으로써, 어휘들을 유의어 집합(Syno-

1) 〈표준국어대사전〉 편찬 지침에서는 다음과 같이 정의하고 있다.
동음이의어로 처리할 것인가 다의어로 처리할 것인가는 표제어의 의미와 특성을 고려하여 결정한다. 의미적인 연관성이 있는 경우 '다의어'로 처리하고 의미의 연관성이 없는 경우 '동음이의어'로 처리함을 원칙으로 한다. 의미의 연관성을 판단하기 어려울 경우 원어나 어원, 또는 '현대어/고어, 일반어/전문어'와 같은 기준에 따라 처리한다.
① 의미: 의미, 원어, 어원
② 특성: 현대어 / 고어, 일반어 / 전문어, 표준어 / 북한어 / 방언 / 비표준어, 고유어 / 한자어 / 외래어 / 특수어, 품사

2) 언어의 구조와 기능에 관한 언어이론을 심리학적으로 연구하는 학문

nym Set : Synset)으로 묶어서 표현하고 각 유의어 집합들을 의미적인 관계(Antonymy, Hypernymy, Hyponymy, Meronymy 등)로 연결하였다[10]. 표 2는 워드넷 2.1의 통계정보이다. 워드넷은 어휘들을 의미적 관계에 따라 구분하고 연결한 최고의 언어자원으로써, 다양한 언어처리 기술에서 중요하게 활용되고 있다.

코퍼스에 기반한 중의성 해소 기술에서는 일반적으로 기계학습(Machine Learning)을 위해 코퍼스를 이용하는데, 코퍼스의 유형에 따라 의미가 부착된 코퍼스를 이용하는 교사학습(Supervised Learning)과 원시코퍼스를 이용하는 비교사학습(Unsupervised Learning)으로 구분된다. 교사학습은 어휘 의미 중의성 해소를 어휘의미에 대한 분류문제(Classification Problem)로 보고, 중의성 대상 어휘(Target Word)의 의미 별로 학습데이타를 기반으로 학습하여 의미 별로 분류자(Classifier)를 획득하고 이를 이용하여 중의성 어휘를 구분한다. 그러나, 언어세계에 존재하는 모든 중의성 어휘에 대하여 의미 태그가 부착된 학습데이타를 구축하기란 현실적으로 어렵다. 또한, 중의성 어휘의 의미 수만큼의 분류자를 얻어야 하는 단점이 있다. 비교사학습에 의한 어휘 의미 중의성 해소는 일반적으로 어휘의미의 구분을 클러스터링 문제(Clustering Problem)로 보고, 중의성 대상 어휘들에 대해서 의미구별(Sense Discrimination)을 하고, 각각의 클러스터(Cluster)에 의미를 부여한다. 또는, 사전이나 워드넷의 언어자원을 활용하여 중의성 대상 어휘가 포함된 문맥에서 공기한 어휘들과 대상 어휘의 의미 별 의미정보(뜻풀이, 다양한 의미관계로 연결된 어휘들 등)들 간의 유사도 계산을 기반으로 중의성을 해소한다.

본 논문에서는 어휘들간의 상호정보량(Mutual Information : MI)정보[11], 다양한 가중치 정보와 복합명사 의미분석 사전을 이용한 동음이의어 의미 중의성 해소 기술을 소개하고자 한다. 표준국어대사전의 의미체계를 따르는 한국어 명사 개념망 사전을 기준으로 중의성 어휘의 의미를 구분하고, 의미 별 뜻풀이와 중의성 어휘가 포함된 지역문맥(Local Context)내의 어휘들간의 상호정보량과 가중치를 계산하여 어휘 의미 중의성을 해소한다. 또한, 본 논문에서는 어휘 의미 분별에 대한 평가를 위한 평가셋 구축에 대한 내용도 소개한다.

본 논문은 다음과 같이 구성된다. 2장에서 어휘 의미 중의성 해소의 관련연구에 대해서 정리하고, 3장에서는 어휘 의미 중의성 해소에 활용된 복합명사 의미분석 사전과 상호정보량 데이타베이스에 대해 소개할 것이다. 4장에서는 3장에서 언급된 자원을 바탕으로 어휘의 의미를 구별하는 방법에 대해서 상세히 언급하고, 5장에서는 평가에 활용될 평가셋에 대해서 설명한다. 6장과 7장은 실험 결과와 향후에 연구방향에 대한 언급으로 마무리한다.

2. 관련연구

사전에 기반한 방법은 LESK[3]에 의해 제안된 방법에 기초하여 다양하게 변형된 기술들이 소개되고 있다. LESK는 중의성 어휘의 의미 별 뜻풀이와 중의성 어휘가 출현한 문맥 내의 어휘들의 뜻풀이들 간에 공통된 어휘의 개수를 이용하여 중의성 어휘의 의미를 결정하였다. 고비용의 많은 자원을 요구하지 않고 구현이 쉬운 장점이 있으나, 어휘들간의 정확한 매칭에 기반하여 자료부족 문제(Data Sparseness Problem)가 심각한 단점이다.

Cowie[12]은 LESK의 방법을 이용하여, 한 문장 내의 모든 어휘에 대한 중의성을 동시에 해소하는 방법을 제시하였다. 문장 내의 모든 중의성 어휘들에 대한 의미 중의성 해소 계산의 최적화를 위해 시뮬레이티드 어닐링(Simulated annealing)방법을 이용하였다.

Andrew Harley[13]는 CIDE(Cambridge International Dictionary of English)을 이용하고, 네 종류의 하부태거(multi-word unit tagger, subject domain tagger, POS tagger, selectional preference pattern tagger)의 결과를 가중치 부여 규칙에 기반하여 가중치를 부여하여 결합된 의미값를 기준으로 의미를 선정한다.

개념망에 기반한 방법으로는 David Yarowsky[14]가 Roget 시소러스의 범주(Category)에 기반하여 통계적으로 어휘 의미 중의성을 해소하는 기술이 대표적이다. David Yarowsky는 Roget 시소러스의 1042개의 범주를 의미로 규정하고, 어휘 의미 중의성 해소를 1042개의 범주에 대한 분류문제로 정의하였다. "어휘의 개념적 클

표 2 워드넷(WORDNET) 2.1의 데이타 통계 자료

품사	어휘	비중의성 어휘 (Monosemous Words)	중의성 어휘 (Polysemous words)	중의성 어휘의 의미	중의성 어휘의 평균 의미
명사	117,097	101,321	15,776	43,783	2.77
동사	11,488	6,261	5,227	18,629	3.56
형용사	22,141	16,889	5,252	14,413	2.74
부사	4,601	3,850	751	1,870	2.49
총	155,327	128,321	27,006	78,695	

래스가 다르면, 서로 다른 문맥에 출현하는 경향이 있다. 그리고, 서로 다른 의미는 서로 다른 클래스에 속한다."라는 가정에 기반하여 원시 코퍼스(Raw Corpus)로부터 각 범주 별로 대표 문맥(Context)를 선정한다. 범주 별로 선정된 문맥들은 상호정보량에 기반하여 범주를 대표하는 어휘들(Salient Word)을 선정하고 이 어휘들은 범주에 대한 지표(Indicator)로 정의한다. 중의성 대상 어휘의 전역문맥[3](Global Context)에 포함된 어휘들을 범주결정의 단서로 보고 베이스 규칙(Bayes'Rule)을 이용하여 전역문맥 내의 모든 어휘들의 가중치 합을 계산하여 어휘의 의미를 결정한다. 그러나, 의미를 1042개의 범주로 결정함으로 인해서 어휘 의미 분별(Word Sense Tagging)이라기 보다는 어휘 범주 분별(Word Category Tagging)이라고 볼 수 있다.

Eneko Agirre[15]는 워드넷(WordNet)의 어휘 의미적 관계를 이용하여 어휘들간의 거리를 계산하는 개념적 밀도(Conceptual Density) 공식을 정의하고, 이를 기반으로 중의성 어휘를 포함한 문맥 내의 공기 어휘들과의 개념적 밀도를 계산하여 의미를 결정하였다. 그러나, 이 기술은 워드넷의 의미적 관계에 너무 의존되는 기술이라는 단점이 있다.

Philip Resnik[16]은 의미적으로 연관된 어휘들을 그룹화할 때 발생하는 어휘 의미 중의성 문제를 워드넷을 이용한 유사도 계산으로 해결하는 방법에 대해서 소개하고 있다.

Mauro Castillo[17]은 워드넷의 Synset 의미 풀이말(Gloss)을 의미 태깅하기 위해서 다양한 규칙(Heuristic)을 이용하는 방법에 대해서 소개하고 있다. SENSEVAL-3에서는 워드넷의 Synset 의미 풀이말에 대한 어휘 의미 중의성 해소(Task for Disambiguating WordNet Glosses) 기술 평가를 추가로 수행하였다. Mauro Castilo는 Synset의 의미 풀이말에 대한 어휘 의미 중의성 해소를 위해 워드넷의 의미 관계와 다양한 정보를 기반으로 한 규칙을 이용하여 의미 태깅을 수행하는 TALP 시스템을 소개하고 있다.

Ganesh Ramakrishnan[18]은 워드넷의 Synset 의미 풀이말과 중의성 어휘가 포함된 문맥의 단서들과의 유사도 계산을 이용하여 어휘 의미 중의성을 해소하는 방법을 소개하였다. Ganesh Ramakrishnan은 워드넷 Synset 의미 풀이말과 문맥의 단서들간의 유사도를 코사인 유사도(Cosine Similarity)와 자카드 유사도(Jaccard Similarity)를 이용하여 가장 유사도가 높은 풀이말의 Synset으로 의미를 결정한다. 또한 다양한 의미관계(Hypernyms, Holonyms)로의 확장을 통해 의미 분별에 미치는 영향을 분석하였다. 이 기술도 의미 풀이말과 문맥 단서 어휘의 정확한 매칭에 기반한 TF(Term Frequency)와 IGF(Inverse Gloss Frequency)를 이용하였으므로 자료부족 문제를 근원적으로 해결하지 못하고 있다.

Hee-Cheol Seo[19]는 중의성 어휘의 의미 별로 다양한 의미관계(Synonyms, Hypernyms, Hyponyms, Meronyms 등)로 연결된 어휘들 중 문맥 내 공기한 어휘들과 가장 확률적으로 밀접한 한 어휘를 선정하여 관련된 의미로 중의성을 해소한다. 중의성 어휘와 의미관계로 연결된 어휘들과 문맥 내 공기한 어휘들간의 확률값은 원시코퍼스로부터 추출된 공기빈도 정보(Co-occurrence Frequency Matrix)를 이용한다. 그러나, 이 방법론은 중의성 어휘와 의미적 관계에 있는 어휘 각각과 문맥 내 공기 어휘들간의 관계를 공기빈도 정보에 기반하여 확률적으로 계산하기 때문에 자료부족문제가 심각하게 발생할 수 있다.

교사학습에 기반한 어휘 의미 중의성 해소 기술은 기계학습의 분류기술(Classification)을 이용하여 학습데이타를 기반으로 의미 분류자(Classifier)을 획득하고, 중의성 어휘가 포함된 문맥의 정보를 기반으로 의미를 분류하는 방법이 대부분이다. 분류기술로는 주로 NB 모델(Naïve Bayes Model), ME 모델(Maximum Entropy Model)과 SVM 모델(Support Vector Machine)을 많이 활용하고 있다[14,20-25].

비교사학습에 기반한 어휘 의미 중의성 해소 기술은 기계학습의 최대 단점인 지식획득 병목현상(Knowledge Acquisition Bottleneck Problem)을 극복하기 위해서 소개되기 시작했다[18,19,26]. 일반적으로 사전이나 워드넷의 다양한 정보를 기반으로 중의성 어휘가 포함된 문맥 어휘들과의 유사도 측정을 통해 어휘의 의미를 결정한다.

최근의 어휘 의미 중의성 해소 기술들은 다양한 자원과 다양한 모델을 통합하여, 각 모델별로 제시되는 의미 구분 결과를 투표(Voting)하여 의미를 결정한다. 어휘 의미 중의성 해소 기술은 다양한 유형의 중의성 어휘를 모두 만족시키는 하나의 모델을 설계하기란 어렵기 때문에 기존의 다양한 모델을 통합하는 방법론이 많이 소개되고 있다[9].

본 논문에서는 관련 논문들 중 사전에 기반한 LESK의 방법의 최대 단점인 자료부족문제를 완화시키기 위해서 어휘들간의 연관계수인 상호정보량을 이용하였고, 상호정보량의 단점을 보완하고 사전에 구조적으로 내포되어 있는 의미 결정 단서를 반영하기 위해서 다양한 유형의 가중치를 적용하였다. 그리고, 기 구축된 복합명사 의미사전을 복합명사 어휘 중의성 해소에 이용하였다.

3) 중의성 어휘를 기준으로 좌우로 50개의 어휘를 문맥(context)으로 본다.

표 3 ETRINET 개념망 사전의 데이타 통계 자료

총 어휘	비중의성 어휘 (Monosemous Words)	중의성 어휘 (Polysemous words)	중의성 어휘의 의미	중의성 어휘의 평균 의미
131,347	94,791	36,556	131,790	3.61

표 4 ETRINET에서 IS-A관계로 연결된 어휘의 데이타 통계 자료

총 어휘	비중의성 어휘 (Monosemous Words)	중의성 어휘 (Polysemous words)	중의성 어휘의 의미	중의성 어휘의 평균 의미
81,322	65,459	15,863	46,358	2.92

3. 자원

본 논문에서 제시하는 어휘 의미 중의성 해소 시스템은 한국어 명사 개념망 사전, 복합명사 의미분석 사전, 원시코퍼스로부터 추출한 상호정보량을 활용한다. 이번 장에서는 언급된 세 종류의 자원에 대해서 상세히 기술한다.

3.1 ETRI 한국어 명사 개념망(ETRINET)

워드넷(WordNet)은 다양한 유형의 어휘 의미관계 정보를 포함하고 있기 때문에 의미분석에서는 가장 많이 활용되는 중요한 언어자원이다. 최근 발표된 버전 2.1은 155,327개의 어휘를 포함하고 있다. 그리고, 워드넷의 정보를 보다 확장하기 위한 다양한 프로젝트가 진행되고 있다. 그 중 XWN(eXtended WordNet) 프로젝트[4]는 기존 워드넷의 한계를 극복하기 위한 것으로 Synset 의미 풀이말(Gloss)에 대한 구문분석결과, 논리구조(Logic Form), 어휘 의미 분별 결과 등을 포함시키는 작업을 수행하고 있다. SENSEVAL-3에서도 Synset 의미 풀이말에 대한 어휘 의미 중의성 해소 기술 평가를 수행하였다[27]. 또한, 다양한 국가에서 워드넷을 자국의 언어로 변환하는 프로젝트를 수행하고 있다. 물론 워드넷을 한국어로 변환하고자 하는 노력과 연구도 많이 진행되었고, 변환된 워드넷을 활용한 언어처리 기술도 많이 소개되었다[28]. 그러나, 영어와 한국어의 근원적인 차이를 극복하기에는 한계가 있다.

워드넷과 같은 어휘의 의미를 계층적으로 연결한 언어자원의 요구가 급증하면서 다양한 유형의 한국어 워드넷과 시소러스가 구축되었다. 조평옥[29]은 국어사전의 뜻풀이와 표제어의 패턴을 분석하여 자동으로 명사 의미계층구조를 구축하는 방법을 소개하였다. [29]에서 소개된 방법에 기반하여 구축된 코난 시소러스(Konan Thesaurus)는 자연어 처리에서 요구하는 다양한 정보를 내포하고 있지 못하고, 어휘량도 부족하였다. 한국전자통신연구원(ETRI)에서 이와 같은 문제점을 극복하기 위하여 코난 시소러스를 기초로 한국어 명사 개념망(ETRINET)을 구축하였다[30,31].

4) http://xwn.hlt.utdallas.edu/index.html

ETRINET은 표준국어대사전에 기반한 어휘목록을 중심으로 개념망 사전을 구축하고, 개념망 사전에 포함된 어휘들 간에 의미적 관계를 설정하여 개념망을 구축하였다. 사전의 다양한 정보를 유지하면서 어휘의 개념적 관계를 설정함으로써, 사전 정보와 어휘 의미정보를 모두 가지는 효율적인 언어자원이다. 어휘목록은 표준국어대사전을 중심으로 다양한 국어사전, 전문용어 사전, 백과사전과 코퍼스를 기반으로 추출하였다. 추출된 어휘목록에 대해서는 표준국어사전에 기반한 의미번호(동음이의어 번호, 다의어 번호), 품사, 뜻풀이, 원어, 유의어, 반의어, 동의어 등의 정보를 지능형 워크벤치를 이용하여 개념망 사전 데이타베이스에 포함시켰다. 개념망 사전에 포함된 어휘들은 IS-A관계(상하관계)를 기준으로 연결하였다.

표 3은 개념망 사전에 포함된 어휘 목록에 대한 통계정보이고, 표 4는 개념망에 있는 어휘 목록 중 IS_A관계로 의미관계가 설정된 어휘들에 대한 통계정보이다. 표 1의 워드넷 2.1과 비교하면, 사전에 포함된 어휘목록(명사)은 ETRINET이 많으나, 의미관계가 설정된 어휘의 목록은 약 35,000개 정도 적다. 중의성 어휘의 평균 의미 수는 워드넷보다 조금 많다.

3.2 복합명사 의미분석 사전

복합명사들을 구성하는 단일명사들은 서로 의미적으로 제약을 한다. 따라서, 복합명사를 구성하는 단일명사들 중 중의성이 있는 어휘는 다른 단일명사들을 단서로 중의성을 해소할 수 있다.

예를 들면, "운동감각"의 경우에 '운동'과 '감각'이 서로의 의미를 제약하여 중의성을 해소 할 수 있다. 표 5는 '운동'과 '감각'의 의미 별 뜻풀이를 표준국어대사전에서 발췌한 것이다. 각각의 단일명사의 의미 별 뜻풀이를 봤을 경우, '운동'의 02번 의미와 '감각'의 02번 의미가 연결되어 '운동감각'의 복합명사를 생성한다는 것을 판단할 수 있다.

3.2.1 복합명사의 분포 분석

복합명사 의미분석 사전 구축의 효율성 판단을 위해서 원시 코퍼스를 대상으로 복합명사의 분포를 분석하였다.[5] 분석된 코퍼스로부터 추출된 전체명사 중 복합

표 5 복합명사 '운동감각'을 구성하는 단일명사들의 의미 별 뜻풀이(표준국어대사전)

동음이의어 번호	뜻풀이	
	운동	감각
01	1. 높이 솟아 있는 지붕의 용마루	1. 덜어 버림
02	1. 사람이 몸을 단련하거나 건강을 위하여 몸을 움직이는 일. 2. 어떤 목적을 이루려고 힘쓰는 일. 또는 그런 활동. 3. 일정한 규칙과 방법에 따라 신체의 기량이나 기술을 겨루는 일. 또는 그런 활동.	1. 눈, 코, 귀, 혀, 살갗을 통하여 바깥의 어떤 자극을 알아차림. 2. 사물에서 받는 인상이나 느낌.

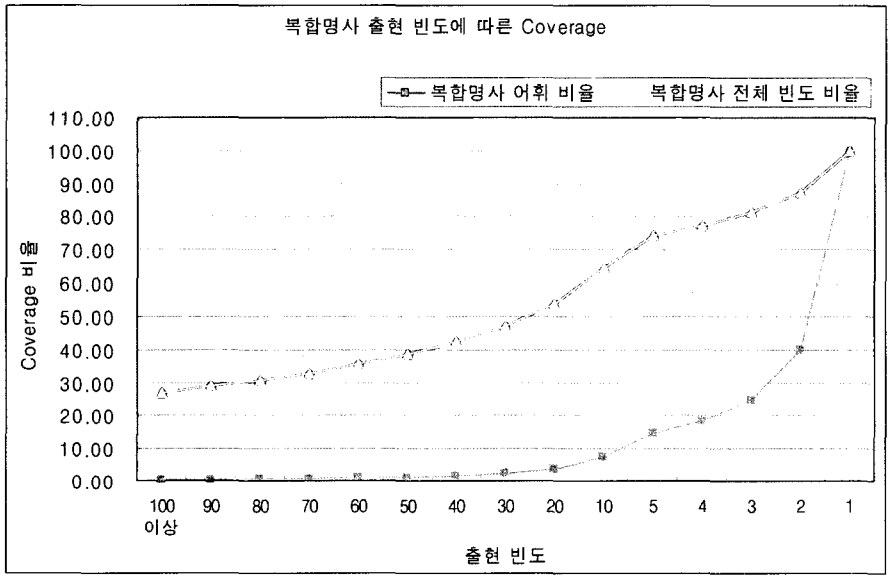


그림 1 복합명사의 빈도에 따른 전체 빈도에 대한 비율 변화

명사를 구성하는 명사의 비율이 약 30%정도였다. 또한 추출된 복합명사들의 빈도 별 분포와 비율을 계산하였다.

그림 1은 복합명사의 빈도 별로 전체 빈도에 대한 비율의 변화를 그래프로 표현한 것이다. 빈도 100이상의 복합명사가 전체 복합명사에서 차지하는 비율이 약 27%이다. 전체 복합명사 중 고빈도 복합명사 10%가 전체 복합명사 빈도의 60%이상을 커버한다. 따라서 고빈도로 출현하는 상위 10%의 복합명사들에 대해서 어휘의미 태깅을 한다면, 전체 복합명사의 60%이상을 커버할 수 있고, 전체 명사의 18% 정도를 커버할 수 있다.

3.2.2 복합명사 의미사전 구축

복합명사 의미사전은 고빈도의 복합명사 일부에 대해서 미리 의미태깅을 함으로써, 어휘 의미 분별의 정확률과 속도 향상을 목적으로 한다. 복합명사 의미사전을 위해 고빈도 복합명사 약 40,000어휘(출현빈도가 5이상인 복합명사 어휘)를 빈도 순으로 추출하였으며, 표준국어대사전의 의미체계에 기반하여 동음이의어와 다의어 번호를 부착하였다. 복합명사 의미사전의 구축지침은 다음과 같다.

A. 최소 단위 명사를 대상으로 의미번호를 부착하는 것을 원칙으로 한다.

B. 의미번호는 표준국어대사전의 의미체계를 따른다.

C. 의미번호는 동음이의어와 다의어[6]를 구분하여 부착한다.

기본 형태 : 명사_동음이의어번호_다의어번호

예) 가공/NN+법/NN[7] → 가공_1_1+법_3_0

D. 접두사와 접미사는 관련된 앞뒤의 명사와 묶어서 의

5) 다어절로 구성된 복합명사는 분석 대상에서 제외하였음.

6) 표준 국어 대사전의 동음이의어와 다의어의 구분은 다음과 같다. 해당 어휘를 검색하였을 경우, 여러 개의 어휘가 검색되면, 이 어휘들은 동음이의어의 관계에 있는 것이다. 각각의 어휘에 붙은 어깨번호가 동음이의어 번호이다. 그리고 해당 어휘의 대한 뜻풀이가 '「1」, 「2」, …'로 구분되어 있는데, 여기에 사용된 번호가 다의어 번호이다. 동음이의어와 다의어 번호가 없는 경우는 중의성이 없는 어휘로써 의미번호를 '0'으로 한다.

7) 품사태그는 다음과 같다.
NN : 일반명사, SN : 명사화 접미사, PF : 접두사

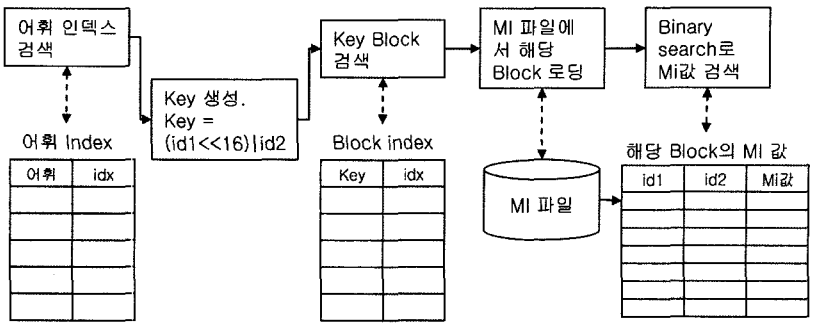


그림 2 메모리 기반 상호정보량 검색 방법

미를 부착하는 것을 원칙으로 한다.

예) 숙주/NN+염색/NN+체/SN → 숙주_2_2+염색체_0_0
시/PF+신경/NN+장애/NN → 시신경_0_0+장애_2_2

E. 접두사와 접미사를 포함한 어휘가 사전에 등록되어 있지 않을 경우, 분리하여 의미번호를 부착한다.

예) 국내/NN+총/PF+생산/NN → 국내_2_0+총+9_0+생산_1_1

F. 복합명사를 구성하는 단일명사들이 서로의 의미 중의성을 해소하지 못하는 경우에는 해당 복합명사는 복합명사 의미사전에서 제외한다.

예) 주요/NN+기관/NN : '주요'가 '기관'의 의미를 제약하지 못함.

표 6 구축된 복합명사 의미사전의 통계 자료

구분	원인	개수	비율
의미 부착한 복합명사		35,795	81.77%
의미를 부착하지 못한 복합명사	의미 모호성	5,410	12.36%
	사전 미등재	761	1.74%
	형태소 분석 오류	1,171	2.67%
	한글 외 언어 포함	639	1.46%
복합명사 총 개수		43,776	100%

표 6은 고빈도 43,776개의 복합명사를 대상으로 복합명사 의미사전 구축을 한 결과에 대한 통계자료를 정리한 것이다. 표 6에서 약 18%는 언급된 이유들로 인해서 의미를 부착할 수가 없었다. 구축된 어휘 수가 다소 적지만, 차후에 고빈도 순으로 더욱더 많은 복합명사에 대해서 의미사전을 구축할 예정이다.

3.3 상호정보량 데이타

컴퓨터에 의한 자연 언어 처리 기술이 발전하면서 어휘들간의 관계를 언어학적 이론이나 직관적 분석에 의존하지 않고 통계적인 분석을 통해 자동으로 파악하려는 연구가 활발해지고 있다[11]. 본 논문에서는 어휘 의미 중의성 해소를 위해 어휘들간의 연관성을 통계적으로 분석한 지식을 활용하는데, 다양한 연관계수 중, 일반적으로 가장 많이 활용되고 있는 상호정보량(Mutual Information)을 이용하였다. 상호정보량이란 두 독립사건의 확률변수 X와 Y사이의 의존관계를 정량적으로 나타낸 것이다.

$$MI(X,Y) = \log_2 \frac{P(X,Y)}{P(X) \times P(Y)} \quad (1)$$

상호정보량은 X와 Y의 연관성이 높을수록 높은 값을 가지고, 연관성이 적을수록 낮은 값을 가진다.

본 시스템에서는 명사, 동사, 형용사만을 대상으로 상호정보량을 추출하였다. 상호정보량추출을 위해서는 세종 코퍼스, 백과사전, ETRI 명사 개념망을 대상으로 하였다. 세종 코퍼스는 약 23,000,000 어절로 구성되어 있고, 백과사전은 약 12,000,000 어절로 구성되어 있다. 상호정보량에 대한 어휘 쌍의 수를 최적화하기 위해서 어휘들의 공기빈도가 10이상인 어휘만을 대상으로 하였다. 추출된 어휘 쌍은 약 20,000,000쌍 정도이다.

본 시스템에서는 대용량의 상호정보량에 대한 저장공간을 최소화하고 검색 시간을 줄이기 위하여 메모리 기반 상호정보량 검색 방법을 개발하였다. 검색을 위해서는 상호정보량을 일정한 블록단위로 인덱스를 생성하고, 어휘 쌍의 Key를 이용하여 해당 블록을 찾아서 파일로부터 블록을 메모리로 로딩하여 이진검색(Binary Search)으로 해당 어휘 쌍의 상호정보량을 얻어온다. 그림 2는 앞선 설명한 상호정보량의 검색 방법의 흐름을 보여준다. 블록의 사이즈는 내부실험을 통해 100으로 설정하였다.

대용량의 상호정보량에 기반한 동음이의어 의미 중의성 해소 기술의 실용적 측면에 대한 실험으로 속도측정을 하였다.[8] 약 92,400 개의 동음이의어를 포함하고 있는 1MByte의 텍스트 파일을 대상으로 실험을 수행한

8) 실험 컴퓨터 사양
- CPU : AMD Athlon 64 Processor 3800+ 2.41GHz
- Memory : RAM 2GB
- OS : Window XP

표 7 '고양이', '다리', '계곡'의 뜻풀이와 뜻풀이를 구성하는 어휘들의 집합

표제어	의미번호	뜻풀이	Bag of Words
고양이		고양이과의 동물을 통틀어 이르는 말.	고양이, **동물**, 통틀다, 이르다, 말
계곡		물이 흐르는 골짜기	**물**, 흐르다, 골짜기
범람		큰물이 흘러 넘침.	크다, **물**, 흐르다, 넘치다
다리	01	동물의 몸통 아래 붙어 있는 신체의 부분. 서고 걷고 뛰는 일 따위를 맡아 한다.	**동물**, 몸, 붙다. 신체, 부분, 서다, 걷다, 일, 맡다,
	02	물을 건너거나 또는 한편의 높은 곳에서 다른 편의 높은 곳으로 건너 다닐 수 있도록 만든 시설물.	**물**, ***건너다***. 한편, 높다, 곳, 다르다, 편, 다니다, 만들다, 시설물
	03	예전에, 여자들의 머리 숱이 많아 보이라고 덧넣었던 딴 머리.	예전, 여자, 머리, 숱, 많다, 보이다, 덧넣다, 따다, 머리

결과, 약 1,500초 가량이 소요되었다. 즉 평균 일초당 약 60개 가량의 동음이의어의 의미 중의성을 해소하였다. 실용적이 측면에서 속도 개선을 위한 다양한 연구가 진행되어야 할 것이다.

4. 어휘 의미 분별

LESK[3]는 중의성 어휘의 의미 별 뜻풀이와 문맥 내 공기 어휘들의 뜻풀이들 중 공통된 어휘가 많은 의미를 중의성 어휘의 의미로 결정하였다. 그러나, 사전의 뜻풀이 기술을 위해 사용되는 어휘는 제한된 일부 어휘들(Controlled Vocabulary)로써, 실세계의 문맥에서 출현하는 어휘들과는 다소 차이가 있고 LESK의 방법은 정확한 어휘 매칭에 기반하기 때문에 자료부족문제(Data Sparseness Problem)가 심각하다. 그러나, 어휘들간의 연관계수를 이용한다면, 이 문제를 상당히 극복할 수 있다.

4.1 알고리즘

본 논문에서는 다음의 가정을 기반으로 어휘들간의 연관계수를 이용하여 어휘 의미 중의성을 해소하려고 한다.

가정 1: 표제어와 뜻풀이에 출현한 어휘들은 의미적으로 밀접한 연관이 있다.

가정 2: 문맥 내에서 공기한 어휘들은 의미적으로 밀접한 연관이 있다.

가정 3: 특정 어휘의 뜻풀이에 출현한 어휘들은 문맥 내 공기 어휘들과 밀접한 연관이 있다.

예를 들어, "고양이가 다리를 통해 범람한 계곡을 건넜다."라는 문장에서 '다리'는 중의성 어휘이다. 표 7에서는 문장에 출현한 명사('고양이', '다리', '계곡')의 뜻풀이를 제시하였다.[9] 중의성 어휘인 '다리'의 의미 별 뜻풀이에 출현한 어휘와 문맥에 공기한 어휘인 '고양이'와 '계곡'에 출현한 어휘들간의 공통된 어휘는 '동물'과 '물'이다. 그런데, '동물'은 '다리01'의 뜻풀이와 공통되고, '물'은 '다리02'와 공통되고, 문맥에 공기한 동사 '건너다'는 '다리02'와 공통된다. 이처럼, 문맥 내의 공기 어휘의 뜻풀이와 중의성 어휘의 의미 별 뜻풀이간에 공통적으로 사용되는 어휘에 대해서는 자료부족이 심각하다. 그러나, 가정 3에 기반하여 문맥 내 공기한 어휘와 중의성 어휘의 의미 별 뜻풀이의 연관계수를 상호정보량으로 계산한다면, 자료부족문제를 극복할 수 있다.

표 8은 문맥 내 공기 어휘와 '다리'의 의미 별 뜻풀이에 출현한 어휘들간의 상호정보량을 보여주고 있다. 어휘들간의 연관계수를 이용함으로써 자료부족 문제가 상당히 완화되었다는 것을 알 수 있다. 또한, 문맥 내 공기 어휘들의 뜻풀이를 이용하지 않고 단지 공기 어휘만을 이용해서도 중의성 어휘의 의미를 분별할 수 있다는 것을 보여주고 있다. 표 8에서는 문맥 내 공기 어휘들 중 '고양이'는 '다리01'의 의미와 연관관계가 높은 어휘이고, '범람', '계곡'과 '건너다'는 '다리02'의 의미와 연관관계가 높은 어휘이다. '통하다'는 '다리01'와 '다리02'에 대해서 구분하기 힘들 정도로 비슷한 연관관계를 맺고 있다. 따라서, '다리'의 의미는 문맥 내 공기 어휘의 대부분과 연관관계를 가지는 '다리02'의 의미로 분별될 수 있다. 그림 3은 문맥 내 공기 어휘들과 '다리'의 의미 별 뜻풀이와의 상호정보량에 따른 연관도를 그림으로 표현한 것이다.

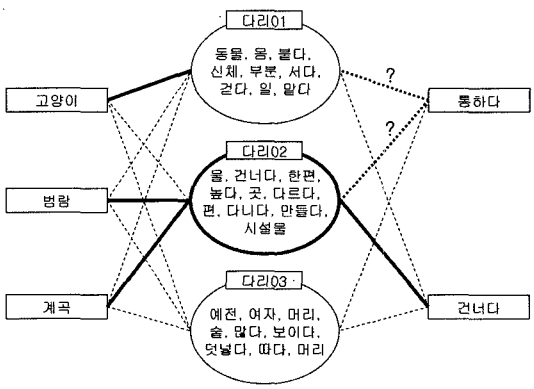


그림 3 문맥 내 공기 어휘들과 '다리'의 의미 별 뜻풀이와의 상호정보량에 따른 연관도

9) 표 7에서는 '다리'의 동음이의어 번호만을 고려하였고, 각 동음이의어 별로 첫번째 다의어 뜻풀이만을 제시하였다.

표 8 문맥 내 공기 어휘와 '다리'의 의미 별 뜻풀이 어휘들 간의 상호정보량 샘플

문맥 내 공기 어휘	다리01		다리02		다리03	
	어휘	MI 값	어휘	MI 값	어휘	MI 값
고양이	**동물**	**2.89**	물	0.85	여자	1.04
	걷다	**1.59**	만들다	0.24	머리	1.03
	몸	**1.50**	높다	0.16	예전	0.96
통하다	부분	0.88	만들다	0.83	많다	0.77
	동물	0.53	높다	0.68	보이다	0.71
	몸	0.42	곳	0.34	여자	0.09
범람	부분	1.43	**높다**	**1.96**	예전	1.52
	동물	0.61	**곳**	**1.46**	많다	1.46
			만들다	**0.64**	보이다	0.83
계곡	걷다	1.24	**건너다**	**3.37**	보이다	1.14
	부분	1.08	**물**	**3.17**	많다	1.11
	몸	0.8	**곳**	**2.27**	머리	1.00
건너다	걷다	2.58	**물**	**2.38**	보이다	1.15
	몸	0.76	**곳**	**1.69**	여자	0.88
			높다	**0.97**	머리	0.58

4.2 어휘 의미 중의성 해소 모델의 구성과 흐름도

본 모델은 ETRINET의 개념망 사전에 기반하여 명사의 의미를 분별하는 기술로 그림 4와 같이 구성된다.

4.2.1 명사관계 일반화와 복합명사 처리

문맥 내의 복합명사는 기 구축된 복합명사 의미사전을 참고하여 의미를 결정한다. 복합명사 의미사전을 이용하는 것은 어휘 중의성 해소의 정확률 향상과 처리속도의 개선에 도움을 준다.

문맥 내에서 다양한 조사로 연결된 명사들은 복합명사로 변환하여도 그 의미가 변질되지는 않는다. 이에 해당되는 관계 두 가지는 다음과 같다.

A. 속격조사 '의'에 의해서 연결된 관계

예) 정부의 정책 → 정부정책

한국의 교통문화 → 한국교통문화

B. 명사에서 파생된 동사의 목적어와 어근과의 관계

예) 역사를 연구하다 → 역사연구

평화를 조성하다 → 평화조성

위의 두 관계를 규칙에 의해서 복합명사로 변경하여 복합명사 의미사전에서 검색하고 복합명사가 존재하면 의미를 부착하고, 없으면 상호정보량을 이용하여 의미 분별한다.

4.2.2 단서 정보를 이용한 분별 대상 의미 축소

일반적으로 의미적으로 모호한 외래어는 의미 이해를 위해 괄호를 이용하여 원어정보를 제공한다. 특히 백과사전이나 사전과 같이 잘 정제된 언어자원의 경우, 다양한 유형의 괄호 기호를 이용하여 어휘들에 대한 정보를 기술하는데, 한자어와 영어 정보는 어휘의 중의성 해소를 위해 중요한 단서가 된다. 본 모델에서는 중의성 어휘에 대한 괄호 내 한자나 영어 정보를 기반으로 어휘의 의미를 해소하거나, 분별해야 할 의미 수를 축소한다.

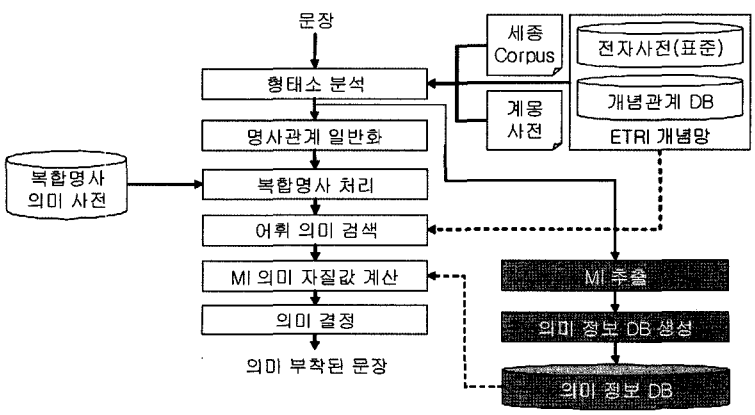


그림 4 어휘 의미 중의성 해소 모델의 흐름도

4.2.3 문맥 간 상호정보량 계산

중의성 어휘의 의미 분별은 지역문맥(Local Context) 내의 어휘들과 중의성 어휘의 의미 별 뜻풀이에 출현하는 어휘들간의 상호정보량을 기반으로 계산된다.

$$C = cw_1, cw_2, ..., cw_{amb}, ..., cw_{n-1}, cw_n$$

$$E = ew_1, ew_2, ..., ew_{m-1}, ew_m$$

$$WSD(C, cw_{amb}) = \arg\max_{amb_i} AMI(C, E^{amb_i}) \quad (2)$$

C는 중의성 어휘가 포함된 문맥이고, cw는 문맥 내의 공기어휘들이다. E는 뜻풀이이고, ew는 뜻풀이를 구성하는 어휘들이다. 따라서, 수식 (2)에서처럼, 중의성 어휘 cw_{amb}에 대한 어휘 의미 중의성 해소는 문맥 C와 cw_{amb}의 의미 별 뜻풀이인 E_{amb_i}의 상호정보량(AMI)으로 계산되고, 상호정보량이 가장 큰 의미가 중의성 어휘의 의미로 결정된다.

$$AMI(cw, E) = \frac{\sum_{y=1}^{m} MI(cw, ew_y)}{m} \quad (3)$$

$$AMI(C, E) = \frac{\sum_{x=1}^{n} AMI(cw_x, E)}{n} \quad (4)$$

$$AMI(C, E) = \frac{\sum_{x=1}^{n}\sum_{y=1}^{m} MI(cw_x, ew_y)}{n \times m} \quad (5)$$

수식 (3)은 문맥 내 공기어휘 cw와 뜻풀이 E와의 상호정보량을 계산하는 수식이다.

수식 (4)는 문맥 C와 뜻풀이 E와의 상호정보량이고 수식 (5)는 수식 (3)과 수식 (4)를 풀어서 표현한 것이다.

$$WSD(C, cw_{amb}) = \arg\max_{amb_i} \frac{\sum_{x=1}^{n}\sum_{y=1}^{m} MI\left(cw_x, ew_y^{amb_i}\right)}{n \times m} \quad (6)$$

수식 (5)를 수식 (2)에 적용하면, 수식 (6)이 된다. 결국, 최종적인 어휘 의미 중의성 해소는 수식 (6)에 기반한 상호정보량에 의해서 결정된다.

4.2.4 가중치 부여

상호정보량에 의한 어휘 의미 중의성 해소 방법은 특정하게 연관관계가 높은 소수의 어휘 쌍에 의해서 결과가 왜곡될 수 있다. 이와 같은 단점을 극복하기 위해서 본 논문에서는 상호정보 값을 가지는 어휘 쌍의 개수를 반영하는 가중치를 고려하였다.

수식 (3-1)은 수식 (3)에 가중치를 반영한 것이다.

$$MI(cw, E) = \frac{\sum_{y=1}^{m} MI(cw, ew_y)}{m} \times mif(cw, E) \quad (3\text{-}1)$$

$mif(cw, E)$는 어휘 cw와 뜻풀이 E를 구성하는 어휘들(ew) 중에 상호정보량을 가지는 어휘 쌍의 비율로서 수식 (7)과 같다.

$$mif(cw, E) = \frac{\sum_{y=1}^{m} \begin{cases} \text{if } MI(cw, ew_y) > 0, & 1 \\ \text{else}, & 0 \end{cases}}{m} \quad (7)$$

수식 (4-1)은 문맥 내 공기 어휘와 뜻풀이간의 상호정보량을 가지는 공기 어휘의 개수를 가중치로 반영하기 위해 수식 (4)를 수정한 것이다.

$$MI(C, E) = \frac{\sum_{x=1}^{n} MI(cw, E)}{n} \times amif(C, E) \quad (4\text{-}1)$$

$amif(C, E)$는 문맥 내 전체 공기 어휘들(cw) 중 뜻풀이 E와 상호정보량을 가지는 공기 어휘의 비율로서 수식 (8)과 같다.

$$amif(C, E) = \frac{\sum_{x=1}^{n} \begin{cases} \text{if } AMI(cw_x, E) > 0, & 1 \\ \text{else}, & 0 \end{cases}}{n} \quad (8)$$

일반적으로 실생활에서 많이 사용되는 어휘의 뜻풀이는 그렇지 않은 어휘에 비해 상대적으로 상세하게 기술된다. 따라서, 어휘 의미 별 뜻풀이의 길이를 가중치로 하여 $\log(m+1)$를 수식 (3-1)에 반영하면 수식 (3-2)와 같다.

$$MI(cw, E) = \frac{\sum_{y=1}^{m} MI(cw, ew_y)}{m} \times mif(cw, E) \times \log(m+1) \quad (3\text{-}2)$$

또한, 의미 부착된 세종코퍼스로부터 의미의 사용 빈도를 추출하여 이를 가중치로 반영하였다. 수식 (9)는 의미의 사용 비율이다. 코퍼스에서 중의성 어휘의 총 출현 횟수에 대한 해당 의미의 출현 횟수의 비율이다.

$$SF(cw_{amb_i}) = \frac{freq(cw_{amb_i})}{\sum_{i=1}^{n} freq(cw_{amb_i})} \quad (9)$$

즉, 어휘 의미 중의성 해소를 위해서 수식 (6)에 가중치들을 적용하면, 수식 (6-1)이 된다.

$$WSD(C, cw_{amb}) = \max_{amb_i} \frac{\left(\sum_{x=1}^{n}\left(\sum_{y=1}^{m} MI\left(cw_x, ew_y^{amb_i}\right)\right) \times mif(cw_x, E^{amb_i}) \times \log(m+1)\right) \times amif(C, E^{amb_i})}{n \times m} \times SF(cw_{amb_i}) \quad (6\text{-}1)$$

5. 평가데이타 구축

ETRI 평가 데이타는 표준국어대사전의 의미체계에

따라서 부착한다. ETRI 평가데이타 구축의 기본 지침은 다음과 같다.

A. 평가데이타의 의미체계는 표준국어대사전을 따른다.
B. 표준국어대사전에 의미가 없는 어휘는 태깅하지 않는다.
C. 복합명사에 대한 의미태깅은 복합명사 의미사전 구축 지침에 따른다.
D. 고유명사는 표준국어대사전에 등재된 어휘만을 대상으로 하여 태깅한다.

대상 데이타는 질의응답 시스템에서 사용된 약 200개의 질의문과 백과사전 문맥 데이타로서, 약 2014개의 명사에 의미태그를 부착하였다. 평가데이타의 구조는 일반 텍스트 형태로 그림 5와 같이 구성된다.

평가데이타 구축은 언어학을 전공한 2명이 수행하였다. 각각 동일 데이타에 대해서 평가데이타를 구축하고, 구축된 결과를 비교하여 일치하지 않는 의미태그에 대해서는 서로 의논하여 결정하도록 하였다. 2명의 작업자가 각자 의미를 부착한 평가데이타를 비교하였을 때, 동음이의어 수준에서 의미태그 일치 비율을 85.6%였고, 다의어 수준에서 의미태그 일치 비율은 78.6%였다. 오류 유형은 복합명사의 의미 태깅 범위 선정에서의 오류와 의미적 모호성에 따른 불일치가 대부분이었다.

6. 실험

실험은 크게 네 유형으로 수행하였다. 첫째, LESK의 방법론과 상호정보량에 기반한 방법론의 비교 실험이고, 둘째, 상호정보량에 다양한 가중치 정보값을 적용한 실험이고, 셋째, 복합명사 의미사전을 이용하였을 때의 정확률 향상에 대한 실험이고, 네번째, SENSEVAL-2에서 수행한 한국어 어휘 의미 중의성 평가에 사용된 학습데이타에 대한 실험이다.

6.1 LESK 방법론과 상호정보량에 기반한 방법론 비교 실험

기존의 연구에서 사전을 이용한 방법 중 대표적인 방법인 LESK의 방법론과 본 시스템의 비교를 위한 실험을 수행하였다.

LESK의 방법론은 앞선 2장에서 언급한 바와 같이 중의성 어휘의 의미 별 뜻풀이에 출현한 어휘들과 중의성 어휘와 공기한 문맥 내 어휘들의 뜻풀이에 출현한 어휘들 중, 공통되는 어휘가 많은 뜻풀이의 의미를 중의성 어휘의 의미로 선택하는 방법이다.

실험은 문맥의 윈도우 사이즈를 1에서 10까지 변경하면서 한 실험과 문장 단위로 실험한 결과를 비교하였다. 윈도우 사이즈는 중의성 어휘를 기준으로 좌우의 어휘 수를 의미하는 것으로써, 윈도우 사이즈가 1일 경우, 중의성 어휘의 좌우 한 어휘까지를 의미한다.

표 9 LESK 방법론과 상호정보량에 기반한 방법론의 실험 결과

윈도우 사이즈	LESK	AMI
1	46	57.48
2	51.74	60.96
3	53.89	62.09
4	55.64	62.6
5	55.33	63.32
6	56.15	63.22
7	56.25	63.73
8	56.46	64.86
9	56.56	64.96
10	56.35	65.06
문장	56.1	64.86

LESK:LESK의 방법
AMI :중의성 어휘의 뜻풀이에 기반한 상호정보량을 이용한 WSD

표 9에서 알 수 있듯이 LESK 실험은 윈도우 사이즈에 따라 가파르게 정확률이 상승하다가 윈도우 사이즈 4정도에서 안정적인 모습을 보인다. 이는 정확한 어휘의 매칭을 기반으로 하는 방법에서는 윈도우 사이즈가 작을 때, 자료부족 현상이 발생한다는 것을 의미한다. 그러나, 어휘들의 연관계수인 상호정보량을 이용한 AMI 실험에서는 LESK에 비해서 상대적으로 윈도우 사이즈에 따른 정확률 향상이 완만함을 알 수 있다. 즉, 본 논문에서 제시한 상호정보량이 자료부족문제를 많이 완화시킬 수 있음을 의미하는 것이다. 또한, LESK 방법론에서 윈도우 사이즈에 따라 가장 낮은 정확률과 가장 높은 정확률의 차이가 10.56%이고, AMI 방법론에서는 7.58%로 LESK의 차이보다는 다소 낮다는 것을 알 수 있다. 이 또한 AMI 방법론이 LSEK의 자료부족문제를 많이 완화시킨다는 것을 입증하는 결과이다.

LESK 실험에서는 윈도우 사이즈가 9일때 정확률이

```
구조 : <명사:명사_동음이의어번호:다의어번호>
예 : <연합:연합_3:1>
문장 : 현재 <세계:세계_2:1><노동:노동_3:1><연합:연합_3:1>에 <가입:가입_0:1>
       되어 있는 <나라:나라_1:1>는 모두 몇 개인가요?
```

그림 5 ETRI 평가 데이타의 형태

56.56%로 가장 높았고, AMI 실험에서는 윈도우 사이즈가 10일 때 정확률이 65.06%로 가장 높았다. 가장 높은 윈도우 사이즈를 기준으로 8.5%의 정확률이 향상되었다. 평가셋에서 중의성 어휘로 인식된 어휘들의 평균 의미 수는 5.45개였다.

6.2 가중치 부여 실험

본 논문에서는 상호정보량의 단점을 극복하기 위하여, 세 종류의 가중치를 이용한다. 첫째, 상호정보량 값을 가지는 어휘 쌍의 비율을 가중치로 반영하였고, 둘째, 뜻풀이의 길이를 가중치로 반영하였고, 셋째, 의미 부착 코퍼스에서 추출한 의미 별 사용 비율 정보를 가중치로 이용하였다. 각 가중치들이 어휘 의미 중의성 해소에 미치는 영향을 파악하기 위해 가중치 별로 실험을 수행하였다. 먼저, 각각의 가중치 별로 어휘 의미 중의성에 미치는 영향을 분석하기 위한 실험을 하였고, 둘째, 각 가중치들의 조합이 어휘 의미 중의성에 미치는 영향을 분석하기 위한 실험을 수행하였다.

표 10 가중치 적용에 따른 WSD의 정확률

윈도우 사이즈	AMI	AMI + MW	AMI + ESW	AMI + SFW
1	57.48	58.4	67.93	76.43
2	60.96	61.58	71.62	83.09
3	62.09	61.58	70.7	82.99
4	62.6	62.4	71.72	83.4
5	63.32	63.01	71.62	83.71
6	63.22	64.55	71.93	83.5
7	63.73	64.24	71.93	84.02
8	64.86	64.96	71.72	84.12
9	64.96	**65.27**	72.44	84.02
10	**65.06**	65.06	**72.75**	84.02
문장	64.86	**65.27**	72.34	**84.63**

MW : 상호정보량을 가지는 어휘 쌍의 비율 가중치
ESW : 뜻풀이 길이 가중치
SFW : 의미 사용 비율 가중치

표 10에서 알 수 있듯이, 정확률 향상에 가장 큰 영향을 미치는 가중치는 SFW 가중치이다. 또한, ESW 가중치도 정확률 향상에 많은 영향을 미쳤다. 그러나, MW 가중치는 정확률 향상에 영향을 미치지 않는 것으로 결과가 나왔다. ESW 가중치가 적용되었을 때는 윈도우 사이즈가 10일 때 72.75%의 정확률로 가장 높았고, AMI만 이용했을 때의 최고 성능 보다 7.69%의 정확률 향상이 있었다. SFW 가중치를 적용하였을 때는 문장 전체의 문맥 어휘를 대상으로 하였을 때 84.63%의 정확률을 보였다. 이는 AMI만 이용했을 때 보다는 19.57%의 정확률 향상이 있었고, ESW 가중치가 적용되었을 때 보다는 11.88%의 정확률 향상이 있었다.

표 11은 세가지 가중치의 조합에 따른 어휘 의미 중의성 해소 정확률 변화를 분석한 결과이다. 가장 좋은 결과를 보인 조합은 모든 가중치를 다 적용한 경우와 MW + SFW 가중치 조합을 사용한 경우이다. 가중치의 조합들 중에서는 MW + ESW 가중치 조합이 윈도우 사이즈 9일 때 74.69%의 정확률로 가장 낮은 결과를 보였다. ESW + SFW 가중치 조합은 문맥 전체 어휘를 이용할 때 79.92%의 정확률을 보였다. ESW 가중치는 "일반적으로 자주 사용되는 의미의 뜻풀이는 상대적으로 길다"라는 직관을 반영하기 위한 것이었다. ESW 가중치가 개별적으로 적용되었을 때는 정확률 향상에 큰 영향을 미쳤다. 그러나, 모든 가중치를 적용한 결과와 MW + SFW 가중치 조합을 적용한 결과가 비슷한 것을 봐서는 실질적인 의미 사용 비율 가중치인 SFW 가중치가 ESW 가중치가 담당한 역할을 포함하는 것으로 볼 수 있다. 즉, ESW 가중치가 직관을 완벽하게는 반영하지 못하지만, 부분적으로 반영한다는 것으로 추정할 수도 있을 것이다.

그림 6은 모든 가중치의 조합에 따른 정확률 변화에 대한 그래프이다. MW 가중치를 적용한 경우와 AMI만 이용한 경우를 비교했을 때, MW 가중치는 정확률 향상

표 11 가중치 조합에 따른 WSD 정확률

윈도우사이즈	AMI	AMI+MW+ESW	AMI+MW+SFW	AMI+ESW+SFW	AMI+ALL
1	57.48	68.13	81.35	71.52	80.84
2	60.96	71	84.02	76.95	83.61
3	62.09	71.62	83.81	77.05	84.22
4	62.6	72.85	84.43	77.97	84.43
5	63.32	72.95	84.02	77.66	83.71
6	63.22	73.05	84.12	79.1	84.43
7	63.73	74.18	84.22	79.1	84.63
8	64.86	74.08	84.84	79.51	85.04
9	64.96	74.08	85.35	79.71	85.35
10	65.06	74.69	85.14	79.41	85.35
문장	64.86	74.08	84.63	79.92	84.73

ALL : 모든 가중치를 적용함

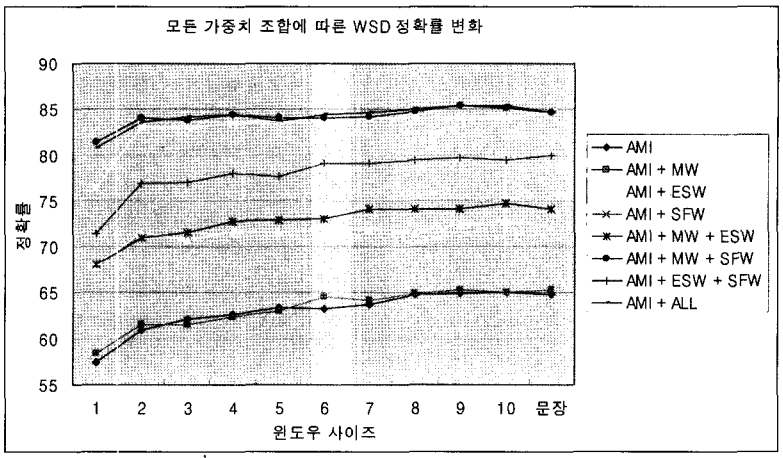


그림 6 모든 가중치 조합에 따른 WSD 정확률 변화 그래프

에 전혀 영향을 미치지 않는다. ESW 가중치를 적용한 경우와 MW + ESW 가중치 조합을 적용한 경우에는 MW 가중치가 윈도우 사이즈가 커질수록 정확률 향상에 영향을 미치는 것을 알 수 있었다. 또한, SFW 가중치를 적용한 경우와 MW + SFW 가중치 조합을 적용한 경우를 비교할 때도 MW 가중치가 정확률 향상에 영향을 미치고 있다. 즉, MW 가중치가 개별적으로는 정확률 향상에 큰 영향을 미치지는 않지만, 다양한 가중치와의 조합에서는 다른 가중치의 왜곡을 완화 시키는 역할을 수행하는 것으로 분석되었다.

ESW 가중치는 개별적으로 사용될 때와 MW 가중치와 조합하여 사용될 때에는 성능향상에 큰 도움이 된다. 그러나, SFW 가중치와 조합하여 사용될 때는 SFW 가중치만 사용했을 때에 비해 현저히 정확률이 하락한다. 이는 "일반적으로 자주 사용되는 의미의 뜻풀이는 상대적으로 길다"라는 직관에 의한 ESW 가중치과 실질적인 의미 사용 비율 정보인 SFW 가중치가 서로의 결과를 왜곡하는 것으로 보여진다.

6.3 복합명사 의미분석 사전의 활용에 따른 비교 실험

본 모델에서는 기 분석된 복합명사 의미사전을 이용하고 있다. 즉, 상호정보량에 의한 어휘 의미 중의성을 해소하기 전에 복합명사에 대해서는 기 분석된 복합명사 의미사전을 이용해 어휘 중의성을 해소한다. 복합명사 의미사전이 어휘 의미 중의성 해소에 미치는 영향을 분석하기 위해서 실험을 하였다.

표 12와 같이 복합명사 의미사전을 이용함으로써, 평균 3.44%의 정확률 향상이 있었다. 윈도우 사이즈가 10일 때 정확률이 88.82%로 3.47%의 정확률 향상이 있었다.

6.4 SENSEVAL-2의 한국어 평가 결과와의 비교 실험

SENSEVAL-2에서 한국어에 대한 어휘 의미 중의성 평가를 특정 샘플 어휘에 대해서 수행하였다. 한국에서

표 12 복합명사 의미분석 사전의 활용에 따른 실험 결과

WS	WAMI	WAMID	정확률 상승 폭
1	80.84	84.24	3.4
2	83.61	87.2	3.59
3	84.22	87.87	3.65
4	84.43	87.87	3.44
5	83.71	87.11	3.4
6	84.43	87.87	3.44
7	84.63	88.06	3.43
8	85.04	88.35	3.31
9	85.35	88.63	3.28
10	85.35	88.82	3.47
문장	84.73	88.16	3.43
평균			3.44

WAMI : AMI + ALL
WAMID : WAMI+복합명사 의미사전 활용

는 두 종류의 시스템이 평가에 참가하였다. 고려대와 KAIST의 시스템은 샘플에 대한 학습데이타를 기반으로 학습을 하고, 평가데이타에 대한 평가를 수행하는 교사학습에 기반한 시스템이었다. 평가에 사용된 샘플 어휘는 부록 A와 같다. 본 실험에서는 SENSEVAL-2의 학습데이타를 대상으로 실험하였다. 실험에 사용된 의미들이 개념망의 의미체계와 일치하지 않기 때문에 부록 A의 매핑 테이블을 기반으로 의미체계를 매핑하여 실험하였다.

표 13에서 알 수 있듯이 실험결과 KUNLP와 근소한

표 13 SENSEVAL-2 한국어 평가 시스템과의 비교

시스템	정확률
KUNLP	69.8%
WAMID	68.04%
KAIST	48.3%

차를 보이고 있다. KUNLP 시스템은 전통적인 교사학습에 기반한 시스템이고, WAMID는 최소한의 의미 코퍼스(의미 별 사용 비율 정보, 복합명사 의미사전)만을 이용하는 교사학습과 비교사학습의 중간적인 모델이다. 따라서, 본 모델의 성능이 나쁘다고 보기에는 무리가 있다. 또한, 본 실험에서는 평가데이타가 아닌 학습데이타를 대상으로 하였고, 의미적 체계에 차이가 있어 의미매핑을 하여 수행한 실험이라 절대적 비교를 할 수는 없다.

6.5 실험 결과 분석

앞서 수행된 네 유형의 실험에서 동음이의어 의미 중의성 해소에 있어 상호정보량에 기반한 방법론의 우수성, 다양한 가중치들이 정확률 향상에 미치는 영향과 복합명사 의미분석 사전 활용의 긍정적인 측면에 대해서 파악할 수 있었다.

첫째, LESK 방법론과 상호정보량에 기반한 방법론의 비교 실험에서는 어휘들의 정확한 매칭에 기반한 LESK 방법론에서는 상호정보량에 기반한 방법론에 비해, 윈도우 사이즈에 따른 정확률 상승폭이 크고, 정확률이 상당히 낮았다. 윈도우 사이즈에 따른 정확률의 상승폭이 큰 것은 어휘들간의 정확한 매칭에 따른 자료부족 현상이 원인인 것으로 분석된다. 반면, 상호정보량에 기반한 방법론은 상대적으로 윈도우 사이즈에 따른 정확률 상승폭이 완만하고 정확률이 높았다. 이는 어휘들간의 연관계수를 이용하여 정확한 매칭에 따른 자료부족 문제를 상당히 해소한 것으로 분석된다.

둘째, 3가지 가중치가 미치는 영향을 분석하기 위해서 실험을 수행하였다. "일반적으로 자주 사용되는 의미의 뜻풀이는 상대적으로 길다"라는 직관을 반영하기 위해서 뜻풀이의 길이를 가중치로 이용하여 실험한 결과, 상호정보량만을 이용한 실험에 비해 큰 폭의 정확률 향상이 있었다. 이는 뜻풀이의 길이가 의미의 사용 빈도를 반영한다는 것을 입증하는 것으로 분석된다. 또한, 의미 태깅된 코퍼스로부터 추출된 의미 사용 비율을 가중치로 적용한 경우에 가장 높은 정확률을 보였다. 이는 언어자원을 이용하는 확률 모델에서는 얼마나 실질적인 언어의 사용 패턴을 수치화하여 모델에 적용할 수 있는가가 연구의 핵심이라는 것을 다시 한번 확인하는 결과라고 볼 수 있다.

세째, 고빈도로 사용되는 복합명사에 대해서 의미를 부착한 복합명사 의미분석 사전이 동음이의어 중의성 해소에 미치는 영향을 분석하였다. 복합명사 의미분석 사전을 활용하였을 때, 정확률이 약 3.44%정도 향상되었다. 문맥에 표현되는 다양한 형태의 명사관계를 복합명사로 일반화할 수 있다면, 더욱더 많은 정확률 향상을 기대할 수 있을 것이다.

넷째, 국내에서 소개된 다양한 의미 중의성 해소 모델과 객관적인 비교를 위해서 SENSEVAL-2에서 소개된 모델과 비교실험을 수행하였다. 다양한 환경의 차로 인해 절대적 비교에는 무리가 있을 수 있으나, 본 논문에서 소개하는 방법론이 어느 정도의 수준인지 파악하는데 도움이 될 것이다. 실험결과, KAIST보다는 성능이 월등이 좋았고, KUNLP에 비해서는 조금 낮은 정확률을 보였다. 그러나, SEANVAL-2에서 소개된 두 모델은 교사학습에 기반한 모델이고, 본 논문에서 소개된 모델은 교사학습과 비교사학습의 중간적인 모델인 것을 감안한다면, 본 논문에서 소개한 방법론이 상당히 좋은 성능을 보였다고 분석할 수 있다.

7. 결론 및 향후 연구

본 논문에서는 상호정보량과 복합명사 의미사전을 이용한 동음이의어 의미 중의성 해소 기술을 소개하였다.

본 논문에서는 LESK[3]방법론의 문제 중, 자료부족 문제(Data Sparseness Problem)을 해결하기 위해서, 어휘들 간의 상호 연관계수를 이용하는 방법을 제시하였다. 본 논문에서 사용한 연관계수는 상호정보량으로써, 중의성 어휘의 문맥 내 공기 어휘들과 중의성 어휘의 의미 별 뜻풀이에 출현한 어휘들 간의 상호정보량을 계산하여 어휘의 의미 중의성을 해소하였다. 또한, 다양한 언어의 구조적 특징을 반영하기 위해서 상호정보량의 값을 가지는 어휘 쌍의 비율을 가중치로 적용하였고, 또한 사전 뜻풀이의 길이를 가중치로 반영하였으며, 의미 부착된 세종 코퍼스로부터 추출한 의미 사용 비율을 가중치로 활용하였다.

본 논문에서 제안된 시스템의 성능에 대한 비교를 위해서, 크게 네 유형의 실험을 수행하였다. 첫째, LESK의 방법론과 상호정보량에 기반한 방법론의 성능 비교를 위한 실험으로써, LESK의 방법론보다 8.5%의 정확률 향상이 있었다. 둘째, 다양한 가중치들이 어휘 의미 중의성 해소에 미치는 영향을 분석하기 위한 실험으로써, 개별 가중치들 중 의미 사용 비율 가중치가 가장 큰 영향을 미쳤다. 또한 가중치들의 조합에 따른 정확률 향상을 분석하기 위한 실험에서는 모든 가중치를 적용하였을 때와 의미 사용 비율 가중치와 상호정보량을 가지는 어휘 쌍의 비율 가중치를 조합하였을 때 가장 좋은 성능을 보였다. 셋째, 복합명사 의미사전이 정확률 향상에 미치는 영향을 분석하기 위한 실험으로 복합명사 의미사전을 활용하였을 때, 윈도사이즈가 10인 경우에 정확률이 88.82%로 가장 높았다. 윈도우 사이즈별 정확률 향상의 평균이 3.44%였다. 그러나, 정확률이 낮은 확률모델에 복합명사 의미사전을 활용한다면 정확률 향상의 폭이 더욱 클 것이다. 넷째, SENSEVAL-2의 한국어 평가에 활용된 학습데이타를 대상으로 한 실험에서는

정확률이 68.04%로 비교적 높은 정확률을 보였다.

본 연구의 의의를 간단히 기술한다면, 다음과 같이 정리할 수 있다.

첫째, 어휘들 간의 연관계수인 상호정보량을 이용함으로써, LESK의 방법에서 가장 큰 문제인 자료부족문제를 완화시킬 수 있다는 것이다.

둘째, 다양한 가중치들이 어휘 의미 중의성 해소에 미치는 영향을 분석하였고, 의미 사용비율 가중치가 가장 큰 영향을 미친다는 것을 알 수 있었다는 것이다.

셋째, 복합명사 의미분석 사전이 상대적으로 어휘 의미 분별에 큰 영향을 미친다는 것을 실험으로 알 수 있었다.

앞으로 연구되어야 할 것을 정리하면 다음과 같다.

첫째, 본 연구에서는 어휘들 간의 연관계수로써 상호정보량만을 이용하였는데, 향후 연구에서는 다양한 연관계수를 이용한 실험을 통해 의미 분별에 적합한 연관계수를 파악하여야 할 것이다.

둘째, 본 논문에서는 개념망 사전만을 이용하였는데, 개념망의 의미적 관계를 이용한 정확률 향상을 위해 어떠한 정보를 활용할 것인지 많은 연구가 진행되어야 할 것이다.

셋째, 영어에 본 시스템의 알고리즘을 적용하여 SENSEVAL-3에 참석한 시스템들의 연구결과와 비교해봐야 할 것이다.

넷째, 한국어의 특징을 최대한 고려한 다양한 형태의 가중치 부여 방안에 대한 연구가 추가되어야 할 것이다.

다섯째, 본 연구에서는 동음이의어에 기반한 어휘 의미 중의성 해소였는데, 의미적으로 유사한 다의어 기반의 어휘 의미 중의성 해소에서 요구되는 다양한 기술들에 대해서 분석하여야 할 것이다.

여섯째, 복합명사 의미사전이 어휘의 의미 중의성 해소의 정확률 향상에 많은 기여를 한다. 그러나, 복합명사 의미사전 구축은 상당히 노동집약적인 작업으로써 많은 비용이 소요된다. 따라서, 향후 복합명사 의미사전을 반자동으로 구축할 수 있는 방법에 대한 연구가 진행되어야 할 것이다. 또한, 복합명사 의미사전은 의미가 태깅된 코퍼스로써 폭넓게 활용할 방법에 대한 연구도 진행되어야 할 것이다.

마지막으로 동음이의어에 기반한 어휘 의미 중의성 해소 모듈이 실용화되기 위해서는 정확률과 함께 처리 속도도 상당히 중요한 요소이다. 그러나, 본 논문에서 제시한 방법론에서는 많은 계산량에 따른 속도 문제가 동음이의어 의미 중의성 해소의 실용화에 큰 걸림돌로 인식되고 있다. 향후에는 속도 개선을 통한 실용화 측면에 대한 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] Adam Kilgarriff, "What is word sense disambiguation good for?," In Proceedings of NLP Pacific Rim Symposium, 1997.

[2] Hyun-Kyu Kang, Se-Young Park, Key-Sun Choi, "A Word Sense Disambiguation Model Using Two-level Document Ranking with Mutual Information in Natural Language Information Retrieval," In Proceeding of ICCPOL, 1997.

[3] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries : how to tell a pine cone from an ice cream cone.," In Proceedings of ACM DIGDOC, 1986.

[4] Adam Kilgarriff, "SENSEVAL: An Exercise in Evaluating Word Sense Disambiguation Programs," In Proceedings LREC, 1998.

[5] Adam Kilgarriff, "Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs," Computer Speech and Language, 1998.

[6] Hee-Cheol Seo, Sang-Zoo Lee, Hae-Chang Rim, Ho Lee, KUNLP system using Classification Information Model at SESENVAL-2," In Proceedings of SENSEVAL-2, 2001.

[7] Philip Edmonds, Scott Cotton, "SENSEVAL-2: Overview," In Proceedings of SENSEVAL-2, 2001.

[8] Philip Edmonds, "SENSEVAL: The evaluation of word sense disambiguation systems," in the ELRA Newsletter, 2002.

[9] Mark Stevenson, "Word Sense Disambiguation : The Case for Combinations of Knowledge Sources," CSLI Publications, 2003.

[10] Christiane Fellbaum, "WORDNET:An Electronic Lexical Database," The MIT Press, 1998.

[11] 정영미, 이재윤, "한국어 텍스트 내 용어연관성 분석을 위한 기초 연구", 제5회 한국정보관리학회, 1998.

[12] Cowie, J., L. Guthrie, J. Guthrie, "Lexical disambiguation using simulated annealing," In Proceedings of COLING, 1992.

[13] Andrew Harley, Dominic Glennon "Sense Tagging in action: Combining different tests with additive weights," In Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics," 1997.

[14] David Yarowsky, "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora," In Proceeding of COLING, 1992.

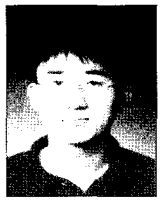
[15] Eneko Agirre, German Rigau, "Word Sense Disambiguation Using Conceptual Density," In proceedings of ACL, 1996.

[16] Philip Resnik, "Disambiguation Noun Groupings with Respect to WordNet Senses," In Proceedings of the Third Workshop on Very Large Corpora, 1995.

[17] Mauro Castillo, Real Francis, Jordi Asterias, Ger-

man Rigau," The TALP Systems for Disambiguating WordNet Glosses," In Proceedings of SENSEVAL-3, 2004.
[18] Ganesh Ramakrishnan, B.Prithviraj, Pushpak Bhattacharyya," A Gloss-centered Algorithm for Disambiguation," In Proceedings of SENSEVAL-3, 2004.
[19] Hee-Cheol Seo, Hae-Chang Rim, Soo-Hong Kim, "KUNLP System in SENSEVAL-3," In Proceedings of SENSEVAL-3, 2004.
[20] Armando Suarez, "A Maximum Entropy-based Word Sense Disambiguation system," In proceedings of COLING, 2002.
[21] Carlo Strapparava, Alfio Gliozzo, Claudio Giuliano, "Pattern Abstraction and Term Similarity for Word Sense Disambiguation: IRST at Senseval-3," In Proceedings of SENSEVAL-3, 2004.
[22] Eneko Agirre, David Martinez," The Basque Country University system: English and Basque tasks," In Proceedings of SENSEVAL-3, 2004.
[23] Gerard Escudero, Lluis Marquez, German Rigau," Naïve Bayes and Exemplar-Based approaches to Word Sense Disambiguation Revisited," In proceedings of ECAI, 2000.
[24] Namhee Kwon, Michael Fleischman, Eduard Hovy, "Senseval automatic labeling of semantic roles using Maximum Entropy models," In Proceedings of SENSEVAL-3, 2004.
[25] Yoong Keok Lee, Hwee Tou Ng, Tee Kiah Chia, "Supervised Word Sense Disambiguation with Support Vector Machine and Multiple Knowledge Sources," In Proceedings of SENSEVAL-3, 2004.
[26] David Yarowsky, "Unsupervised Word Sense Disambiguation Rivaling Supervised Mehtods," In proceedings of ACL, 1995.
[27] Kenneth C. Litkowski, "SENSEVAL-3 TASK: Word-Sense Disambiguation of WordNet Glosses," In Proceedings of SENSEVAL-3, 2004.
[28] 이창기, 이근배, "의미 애매서 해소를 이용한 WordNet 자동 매핑," 제12회 한글 및 한국어 정보처리 학술대회, 1997.
[29] 조평옥, 옥철영, "사전 뜻풀이에서 구축한 한국어 명사 의미계층구조," 인지과학회 논문지 제10권 제4호, 1999년.
[30] 왕지현, 장명길, "정보검색을 위한 한국어 명사 개념망 구축에 관한 연구," 제1회 한국시소러스연구회 국제학술포럼, 2003.
[31] Miran Choi, Jeong Hur, Myung-Gil Jang, "Constructing Korean Lexical Concept Network for Encyclopedia Question-Answering System," In proceedings of IECON, 2004.

허 정
1999년 울산대학교 전자계산학과 졸업(학사). 2001년 울산대학교 대학원 전자계산학과 졸업(석사). 2001년~현재 한국전자통신연구원(ETRI) 지식마이닝연구팀 선임연구원. 관심분야는 정보검색, 자연어처리, 시맨틱웹



서 희 철
1998년 고려대학교 컴퓨터공학과 졸업(학사). 2000년 고려대학교 대학원 컴퓨터공학과 졸업(석사). 2005년 고려대학교 대학원 컴퓨터공학과 졸업(박사). 2005년~현재 한국전자통신연구원(ETRI) 지식마이닝연구팀 선임연구원. 관심 분야는 시맨틱웹, 인공지능, 자연어처리, 정보검색

장 명 길
1988년 부산대학교 계산통계학과 졸업(학사). 1990년 부산대학교 대학원 계산통계학과 졸업(석사). 2002년 충남대학교 대학원 컴퓨터과학과 졸업(박사). 1990년~1997년 시스템공학연구소 연구원 1998년~1999년 한국전자통신연구원(ETRI) 선임연구원. 2000년~현재 한국전자통신연구원(ETRI) 지식마이닝연구팀 팀장(책임연구원). 관심분야는 자연어처리, 정보검색, 질의응답, 지식 및 대화 처리, 미디어 검색 및 처리, 시맨틱웹

부록 A

SENSEVAL-2의 한국어 평가데이타에 대한 정보와 ETRI 의미체계와의 매핑 관계

중의성 어휘	의미번호	의미	ETRI 의미번호[10]	데이타 수	
				학습	평가
거리	K00041	사람과 차가 다니는 길	**거리_01**	70	65
	K00042	음식 따위를 만들 감(물건). 행동이나 생각의 대상이 될만한 것(사물). 일~. 읽을~	**거리_02**	16	
	K00043	만드는데 주가 되는 물건		0	
	K00044	큰 이익	**거리_06**	0	
	K00045	음악이나 연극 또는 무속 음악 등에서 단락, 과장, 마당을 뜻하는 말	**거리_04**	0	

	K00046	두 곳 사이의 먼 정도	**거리_08**	45	
눈	K00011	보는 감각을 가진 사람, 동문의 기관	**눈_01**	125	66
	K00012	그물 같은 물건의 코와 코를 연결한 부분	**눈_03**	1	
	K00013	공중에 떠다니는 김이 찬 여섯 모가 난 결정체	**눈_04**	7	
말	K00001	사람의 생각과 느낌을 표현하는 수단	**말_01**	0	50
	K00002	끝	**말_11**	0	
	K00003	집짐승	**말_05**	11	
	K00004	곡식, 액체의 분량	**말_03**	57	
	K00005	고누, 윷 따위의 판의 군사	**말_07**	33	
목	K00071	머리와 몸의 잘록한 부분	**목_01**	99	49
	K00072	목과 비슷한 부분	**목_01**	1	
	K00073	통로 같은 곳으로 빠져나갈 수 없는 중요하고 좁은 곳	**목_01**	0	
	K00074	나무(木)		0	
바람	K00031	기압 변화로 일어나는 공기의 흐름	**바람_01**	1	49
	K00032	어떤 일이 이루어지기를 기다리는 간절한 마음	**바람_02**	0	
	K00033	어떤 일에 더불어 일어나는 기세	**바람_01**	97	
	K00034	몸에 차려야 할 것을 처리지 않고 나서는 차림 또는 그 행실	**바람_01**	0	
밤	K00091	밤나무의 열매	**밤_02**	29	50
	K00092	저녁 어두운 뒤부터 새벽 밝기까지의 동안	**밤_01**	72	
손	K00021	사람의 팔목 끝에 달린 부분	**손_01**	129	66
	K00022	손아랫사람	**손_04**	0	
	K00023	손해	**손_11**	0	
	K00024	성	**손_08**	1	
	K00025	후손	**손_09**	0	
	K00026	딴 데서 임시로 찾아오거나 가는 사람	**손_02**	1	
	K00027	자신의 힘이나 역량을 이르는 말	**손_01**	1	
의사	K00061	국사시험에 합격하여 면허를 받고 의료 활동에 종사하는 사람	**의사_12**	111	81
	K00062	외부로부터 위험을 느낀 동물이 움직이지 않고 죽은 체하는 일	**의사_09**	0	
	K00063	의를 위해 죽음	**의사_03**	0	
	K00064	마음먹은 생각	**의사_02**	12	
	K00065	의심스러운 말	**의사_07**	0	
	K00066	의논할 사항	**의사_14**	0	
	K00067	신라시대의 으뜸 벼슬	**의사_13**	0	
	K00068	의학이나 의료에 관한 일	**의사_11**	0	
	K00069	실제와 비슷함	**의사_10**	42	
자리	K00051	물체가 있거나 그것을 둘 수 있는 공간	**자리_01**	97	49
	K00052	어떤 일에 종사하여 활동하는 직위나 지위	**자리_01**	2	
	K00053	어떤 사람들이 모이도록 한 경우나 그러한 기회	**자리_01**	2	
	K00054	십진법에 따른 숫자의 자리	**자리_01**	0	
점	K00081	작고 둥글게 찍은 표	**점_10**	5	49
	K00082	어느 속성이나 측면의 개별적인 부분이나 요소	**점_10**	89	
	K00083	물품의 가지 수를 셀 때 쓰는 말	**점_10**	1	
	K00084	살코기 따위의 작은 조각들을 셀 때 쓰는 말	**점_10**	4	

10) ETRI 의미번호는 표준국어대사전의 동음이의어 번호이다. 그리고 굵은 글씨는 개념망에서 의미관계(IS_A)로 연결된 의미이고, 그렇지 않은 것은 개념망 사전에만 등록된 어휘이다.