복합명사 의미사전을 이용한 동음이의어 중의성 해소

허정○ 장명길
한국전자통신연구원 지식마이닝연구팀
{jeonghur○, mgjang}@etri.re.kr

Homonym Disambiguation using Sense-Tagged Compound Noun Dictionary

Jeong Hur○ , Myung-Gil Jang
Knowledge Mining Research Team, ETRI

요 약

본 논문에서는 평균 상호정보량에 기반하고 복합명사 의미사전을 이용한 동음이의어 중의성 해소 기술에 대해서 소개한다. 평균 상호정보량을 이용한 방법은 사전의 뜻풀이를 이용하는 기존 방법의 자료부족문제를 완화시킨다. 복합명사 의미사전은 복합명사를 구성하는 단일명사들의 의미제약 관계를 이용하여 구축된다. 이 구축된 복합명사 의미사전은 어휘 의미 중의성의 정확률을 향상시키고, 연산 시간을 줄여 시스템의 효율성을 극대화시킨다. 평균 상호정보량을 이용한 실험에서는 62.04%의 정확률로 LESK의 방법에 비해 6.06%의 향상이 있었고, 복합명사 의미사전을 이용하였을 때는 68.13%의 정확률로 12.76%의 정확률 향상이 있었다.

1. 서 론

컴퓨터가 디지털화된 자연어를 이해하기 위해서는 다양한 기술이 요구된다. 그 중, 어휘의 의미를 이해하는 것은 자연어 의미분석을 위한 가장 기초되는 기술로서 중의적인 의미로 사용되는 어휘의 의미를 분별하는 중의성 해소 (Word Sense Disambiguation)기술이 핵심이다. 중의성 어휘는 의미적인 연관성에 따라 다의어(Polysemy)와 동음이의어(Homonym)으로 구분된다.

어휘 의미 중의성 해소는 다양한 언어처리 응용분야들(기계번역, 정보검색, 질의응답, 음성처리 등)에서 활용의 효용성이 입증되었지만, 기술적 완성도 문제로 인해 한계를 가지고 있었다. 그러나, 최근 SENSEVAL과 같이 어휘 의미 중의성 해소 기술에 대한 활성화로 인해 기술적 한계를 많이 극복한 상황이다.

Lesk[4]는 중의성 어휘의 의미 별 뜻풀이와 중의성 어휘가 출현한 문맥 내 어휘들의 뜻풀이들 간에 공통된 어휘의 개수를 이용하여 중의성 어휘의 의미를 결정하였다. 고비용의 많은 자원을 요구하지 않고 구현이 쉬운 장점이 있으나, 어휘들간의 정확한 매칭에 기반하기 때문에 자료부족 문제(Data Sparseness Problem)가 심각한 단점이다.

Ganesh Ramakrishnan[3]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 의미 관계(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 이용하여, 자료부족 문제를 근원적으로 해결하지 못하고 있다.

본 논문에서는 자료부족 문제를 완화하기 위해서 평균 상호정보량을 이용한 동음이의어 중의성 해소 기술에 대해서 간단히 언급하고, 복합명사 의미분석 사전이 정확률 향상에 미치는 영향을 분석하였다.

본 논문은 2장에서 평균상호정보량에 대해서 기술하고, 3장에서 복합명사 의미사전에 대해서 기술한다. 4장에서는 실험과 결과에 대한 분석이 언급되고, 5장에서 결론과 향후연구에 대해서 기술한다.

2. 평균 상호정보량

사전에 기반한 기존 연구들은 문맥 내 공기어휘와 사전의 뜻풀이를 구성하는 어휘들간의 정확한 매칭에 의한 유사도 계산에 기반한 방법들이었다[3,4]. 따라서, 정확하게 매칭되지 않는 어휘들에 의한 자료부족 문제가 심각하다. 이를 완화하기 위해서 본 논문에서는 어휘들간의 연관관계를 정량화한 상호정보량(Mutual Information)을 이용하였다. 상호정보량은 두 독립사건의 확률변수 X와 Y 사이의 의존관계를 정량적으로 나타낸 것이다[1].

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

평균 상호정보량에 기반한 어휘 의미 중의성 해소는 중의성 어휘의 의미 별 뜻풀이에 출현한 어휘들의 집합과 중의성 어휘와 공기한 문맥 내 어휘들 집합과의 상호정보량의 평균을 구하여 이를 기반으로 어휘의 의미를 결정하는 것으로 다음의 수식으로 표현된다.

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

cw_x 는 중의성 어휘와 공기한 어휘를 의미하고, $ew_y^{amb_i}$ 는 중의성 어휘의 i번째 의미 뜻풀이에 출현한 어휘들을 의미한다[2].

3. 복합명사 의미사전

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

표 1. 복합명사 '운동감각' 을 구성하는 단일명사들의 의미 별 뜻풀이

동음이의어 번호	뜻풀이	
	운동	감각
01	높이 솟아 있는 지붕의 용마루.	덜어 버림.
02	사람이 몸을 단련하거나 건강을 위하여 몸을 움직이는 일.	눈, 코, 귀, 혀, 살갗을 통하여 바깥의 어떤 자극을 알아차림.

예를 들면, "운동감각"의 경우에 '운동' 과 '감각' 이 서로의 의미를 제약하여 중의성을 해소 할 수 있다. [표 1]은 '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.

3.1 복합명사의 분포

복합명사 의미사전 구축의 효율성 판단을 위해서 먼저 원시 코퍼스를 대상으로 복합명사의 분포를 분석하였다[1]. 분석된 코퍼스로부터 추출된 전체명사 중 복합명사를 구성하는 명사의 비율이 약 30%정도였다. 또한 추출된 복합명사들의 빈도 별 분표와 비율을 계산하였다.

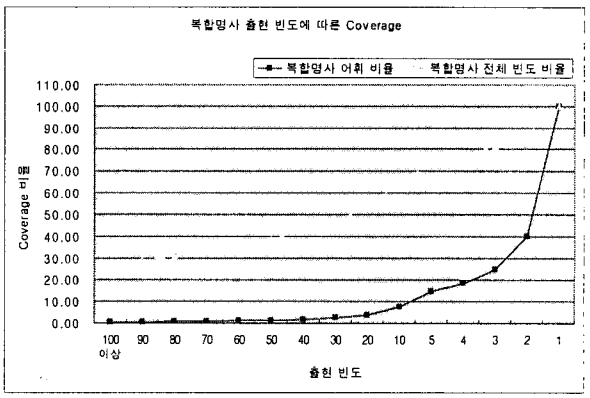

그림 1. 복합명사 출현 빈도에 따른 Coverage 변화

[그림 1]은 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 복합명사의 60%이상을 커버한다. 따라서 고빈도로 출현하는 상위 10%의 복합명사들에 대해서 어휘 의미 태깅을 한다면, 전체 복합명사의 60%이상을 커버할 수 있고, 전체 명사의 18% 정도를 커버할 수 있다.

3.2 복합명사 의미사전의 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 다의어를 구분하여 부착한다.
   기본 형태 : 명사_동음이의어번호_다의어번호
   예) 가공/NN+법/NN → 가공_1_1+법_3_0
D. 접두사와 접미사는 관련된 앞뒤의 명사와 묶어서 의미를 부착하는 것을 원칙으로 한다.
   예) 숙주/NN+영색/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

1 다어절로 구성된 복합명사는 분석 대상에서 제외하였음.

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.

고빈도 43,776개의 복합명사를 대상으로 복합명사 의미사전을 구축한 결과, 약 18%는 다양한 이유들(의미적 모호성, 형태소 분석 오류, 사전 미등재 등)로 인해서 의미를 부착할 수가 없었다.

3.3 명사관계 일반화

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 평균 상호정보량을 이용하여 의미 분별한다.

4. 실험 및 결과

실험은 크게 두 종류로 수행하였다. 사전 뜻풀이를 이용하는 가장 대표적인 방법인 LESK방법론과 평균 상호정보량에 기반한 방법의 비교와 복합명사 의미사전을 이용하였을 때의 정확률 향상에 대한 실험이다.

표 2. LESK 방법론과 평균 상호정보량에 기반한 방법론의 실험 결과

윈도우 사이즈	LESK	AMI(평균 상호정보량)
1	42.61	55.87
2	51.18	58.6
3	53.44	60.38
4	54.4	60.66
5	54.57	61.14
6	54.94	61.25
7	55.19	61.9
8	55.22	61.62
9	55.47	62.04
10	55.56	61.99
문장	55.98	61.42

실험은 문맥의 윈도우 사이즈를 1에서 10까지 변경하면서 한 실험과 문장 단위로 실험한 결과를 비교하였다. 일반적으로 윈도우 사이즈가 크면 좋은 결과를 보이지만, 연산량이 많아져 비효율적이다.

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.

복합명사 의미사전을 이용한 실험은 AMI에서 가장 좋은 성능을 보인 윈도우 사이즈 9에서 수행하였다. 실험 결과 복합명사 의미사전을 이용하였을 때, 정확률 68.13%로 약 6.7%의 정확률 향상이 있었다. 중의성 어휘의 평균 의미 수는 5.67개이다.

5. 결론 및 향후연구

본 논문에서는 평균 상호정보량에 기반한 동음이의어 중의성 해소 방법에 복합명사 의미사전을 추가로 이용하는 수정된 모델을 제시하였다. 복합명사는 단일명사들 간의 의미적 제약으로 인해, 사전에 의미를 부착할 수 있다. 또한 명사관계 일반화를 통해 구문 내의 명사들을 복합명사화 하여 복합명사 의미사전을 이용할 수 있다. 평균 상호정보량을 기반으로 한 방법은 LESK의 방법론에 비해 6.06% 높은 62.04%의 정확률을 보였다. 또한 복합명사 의미사전을 이용하였을 때는 68.13%의 정확률을 보여 평균 상호정보량을 이용하였을 때보다 약 6.7%의 정확률 향상이 있었다. 본 논문의 실험을 통해, 고빈도의 복합명사에 대해 기 구축된 의미사전은 어휘 중의성 해소에 큰 도움이 될 수 있는 것으로 분석되었다.

수작업에 의해 구축되는 복합명사 의미사전은 상당히 노동집약적으로 비용이 많이 드는 작업이다. 따라서, 앞으로 반자동을 복합명사 의미사전을 구축할 수 있는 지능형 워크벤치를 구축하기 위한 연구가 진행되어야 할 것이다. 또한, 기 구축된 복합명사 의미사전을 다양한 언어분석 기술에 어떻게 활용할 수 있을지 연구하여야 할 것이다.

참고문헌

[1] 정영미, 이재윤 "한국어 텍스트 내 용어연관성 분석을 위한 기초 연구", 제5회 한국정보관리학회, 1998.
[2] 허정, 장명길, "평균상호정보량에 기반한 동음이의어 중의성 해소", 제17회 한글 및 한국어 정보처리 학술대회, 2005
[3] Ganesh Ramakrishnan, B.Prithviraj, Pushpak Bhattacharyya, "A Gloss-centered Algorithm for Disambiguation", In Proceedings of SENSEVAL-3, 2004.
[4] 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.