평균 상호정보량에 기반한 동음이의어 중의성 해소

허정[0] 장명길
한국전자통신연구원 지식마이닝연구팀
{jeonghur[0], mgjang}@etri.re.kr

Homonym Disambiguation based on Average Mutual Information

Jeong Hur[0] , Myung-Gil Jang
Knowledge Mining Research Team, ETRI

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 연구에서는 평균 상호정보량에 기반한 동음이의어 의미 중의성 해소 기술을 소개한다. 사전 뜻풀이를 이용하는 기존 연구들은 어휘들간의 정확한 매칭에 의존하기 때문에 자료부족 현상이 심각하였다. 그러나, 본 연구에서는 어휘들간의 연관계수인 상호정보량을 이용함으로써 이 문제를 완화시켰다. 또한, 상호정보량을 가지는 어휘 쌍의 비율, 의미 별 빈도 정보와 뜻풀이의 길이를 가중치로 반영하였다.

본 시스템의 평가를 위해 질의응답 평가셋의 500여 개의 질의와 정답단락을 대상으로 동음이의어 의미 중의성 해소 평가셋을 구축하였다. 평가셋에 기반하여 두가지 유형의 실험을 수행하였다. 실험 결과는 평균 상호정보량만을 이용하였을 때 62.04%의 정확률을 보였고, 가중치를 활용하였을 때 83.42%의 정확률을 보였다.

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 태깅(Word Sense Tagging)에서부터 시작된다.

대부분의 어휘들은 쓰이는 상황과 문맥에 따라 다양한 의미로 이용된다. 하나의 어휘가 둘 이상의 의미로 이용되는 어휘들을 의미구분의 기준 [1] 에 따라 동음이의어(Homonym)와 다의어(Polysemy)로 구분할 수 있다.

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[2,6,9]. 다양한

[1] <표준국어대사전> 편찬 지침에서는 다음과 같이 정의하고 있다.
동음이의어로 처리할 것인가 다의어로 처리할 것인가는 표제어의 의미와 특성을 고려하여 결정한다. 의미적인 연관성이 있는 경우

언어처리 응용분야에서 어휘 의미 중의성 해소 기술의 효용성에 대한 연구가 활성화되면서, 어휘 의미 중의성 해소에 대한 다양한 연구가 진행되었다.

본 논문에서는 어휘들간의 상호정보량(Mutual Information : MI)정보[1]와 다양한 가중치 정보를 이용한 동음이의어 의미 중의성 해소 기술을 소개하고자 한다. 표준국어대사전의 의미체계를 따르는 ETRI 한국어 명사 개념망(ETRINET) 사전을 기준으로 중의성 어휘의 의미를 구분하고, 의미 별 뜻풀이와 중의성 어휘가 포함된 지역문맥(Local Context)내의 어휘들간의 평균 상호정보량과 가중치를 계산하여 어휘 의미 중의성을 해소한다.

본 논문은 다음과 같이 구성된다. 2장에서 어휘 의미 중의성 해소의 관련연구에 대해서 정리하고, 3장에서는 상호정보량 데이터베이스에 대하여 소개할 것이다. 4장과 5장에서는 어휘의 의미를 구별하는 방법에 대해서 상세히 언급하고, 6장과 7장에서는 실험 결과와 향후 연구방향에 대해서 언급한다.

2. 관련연구

사전에 기반한 방법은 Lesk[8]에 의해 제안된 방법에 기초하여 다양하게 변형된 기술들이 소개되고 있다. Lesk는 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 기반하여 자료부족 문제(Data Sparseness Problem)가 심각한 단점이 있다.

Cowie[7]는 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.

개념망에 기반한 방법으로는 David Yarowsky[3]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 클래스가 다르면, 서로 다른 문맥에 출현하는 경향이 있다. 그리고, 서로 다른 의미는 서로 다른 클래스에 속한다." 라는 가정에 기반하여 원시 코퍼스(Raw Corpus)로부터 각 범주별로 대표 문맥(Context)를 선정한다. 범주 별로 선정된 문맥들로부터 상호정보량에 기반하여 범주를 대표하는 어휘들(Salient Word)을 선정하고 이 어휘들은 범주에 대한 지표(Indicator)로 정의한다. 중의성 대상 어휘의 전역문맥[2](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 태깅(Word Sense Tagging)이라기 보다는 어휘 범주 태깅(Word Category Tagging)이라고 볼 수 있다.

Ganesh Ramakrishnan[4]은 워드넷의 Synset 의미 풀이말(Gloss)과 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.

Hee-Cheol Seo[5]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.

'다의어' 로 처리하고 의미의 연관성이 없는 경우 '동음이의어' 로 처리함을 원칙으로 한다

[2] 중의성 어휘를 기준으로 좌우로 50개의 어휘를 문맥(context)으로 본다.

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 각각의 어휘와 문맥 내 공기 어휘들간의 관계를 공기빈도 정보에 기반하여 확률적으로 계산하기 때문에 자료부족문제가 심각하게 발생할 수 있다.

본 논문에서는 관련 논문들 중 사전에 기반한 Lesk의 방법의 최대 단점인 자료부족문제를 완화시키기 위해서 어휘들간의 연관계수인 상호정보량을 이용하였고, 평균 상호정보량의 단점을 보완하고 사전에 구조적으로 내포되어 있는 의미 결정 단서를 반영하기 위해서 다양한 유형의 가중치를 적용하였다.

3. 상호정보량 데이터

컴퓨터에 의한 자연 언어 처리 기술이 발전하면서 어휘들간의 관계를 언어학적 이론이나 직관적 분석에 의존하지 않고 통계적인 분석을 통해 자동으로 파악하려는 연구가 활발해지고 있다[1]. 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.

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

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.

4. 상호정보량에 기반한 어휘 의미 중의성 해소

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

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

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

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

표 1. '고양이', '계곡', '범람', '다리' 의 뜻풀이

표제어	의미번호	뜻풀이
고양이		고양이과의 **동물**을 통틀어 이르는 말.
계곡		**물**이 흐르는 골짜기
범람		큰**물**이 흘러 넘침.
다리	01	**동물**의 몸통 아래 붙어 있는 신체의 부분. 서고 걷고 뛰는 일 따위를 맡아 한다.
	02	**물**을 **건너**거나 또는 한편의 높은 곳에서 다른 편의 높은 곳으로 건너다닐 수 있도록 만든 시설물.
	03	예전에, 여자들의 머리 숱이 많아 보이라고 덧넣었던 딴 머리.

예를 들어, "고양이가 다리를 통해 범람한 계곡을 건넜다." 라는 문장에서 '다리' 는 중의성 어휘이다. [표 1]에서는 문장에 출현한 명사('고양이', '계곡', '범람', '다리')의 뜻풀이를 제시하였다[3]. 중의성 어휘인 '다리' 의 의미 별 뜻풀이에 출현한 어휘와 문맥에 공기한 어휘인 '고양이', '계곡' 과 '범람' 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.

[3] [표 1]에서는 '다리' 의 동음이의어 번호만을 고려하였고, 각 동음이의어 별로 첫번째 다의어 뜻풀이만을 제시하였다.

이처럼, 문맥 내의 공기 어휘의 뜻풀이와 중의성 어휘의 의미 별 뜻풀이간에 공통적으로 사용되는 어휘에 대한 매칭은 자료부족이 심각하다. 그러나, 가정 3에 기반하여 문맥 내 공기한 어휘와 중의성 어휘의 의미 별 뜻풀이의 연관계수를 상호정보량으로 계산한다면, 자료부족문제를 극복할 수 있다.

표 2. 문맥 내 공기 어휘와 '다리' 의 의미 별 뜻풀이 어휘들 간의 상호정보량 샘플

공기 어휘	다리01		다리02		다리03	
	어휘	MI	어휘	MI	어휘	MI
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

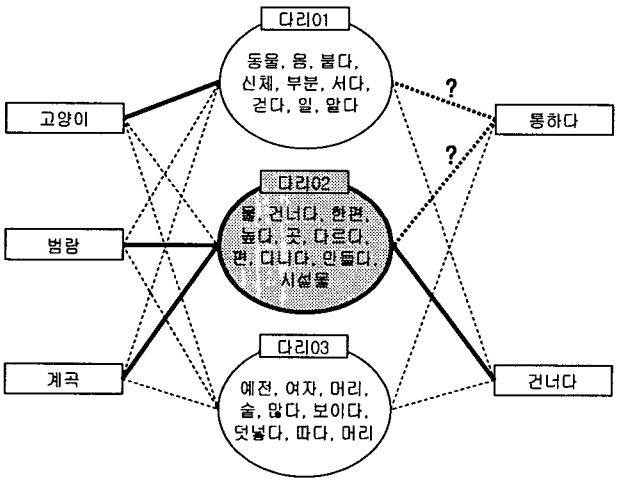

그림 1. 문맥 내 공기 어휘들과 '다리' 의 의미 별 뜻풀이와의 상호정보량에 따른 연관도

[표 2]은 문맥 내 공기 어휘와 '다리' 의 의미 별 뜻풀이에 출현한 어휘들과 문맥 내 어휘들간의 상호정보량의 일부를 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 2]에서는 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 1]은 문맥 내 공기 어휘들과 '다리' 의 의미 별 뜻풀이와의 상호정보량에 따른 연관도를 그림으로 표현한 것이다.

5. 알고리즘

5.1 평균 상호정보량 계산

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

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

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

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

C 는 중의성 어휘가 포함된 문맥이고, cw 는 문맥 내의 공기어휘들이다. E 는 뜻풀이이고, ew 는 뜻풀이를 구성하는 어휘들이다. 중의성 어휘 cw_{amb} 에 대한 어휘 의미 중의성 해소는 문맥 C 와 cw_{amb} 의 의미 별 뜻풀이인 E^{amb_i} 의 평균 상호정보량(AMI)으로 계산되고, 평균 상호정보량이 가장 큰 의미가 중의성 어휘의 의미로 결정된다.

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

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

5.2 가중치 계산

평균 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 비율을 반영하는 가중치를 고려하였다.

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

$MIF(cw,E)$ 는 어휘 cw 와 뜻풀이 E 를 구성하는 어휘들(ew) 중에 상호정보량이 0보다 큰 어휘 쌍의 비율로서 다음 수식과 같다.

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

또한, 문맥 내 공기 어휘와 뜻풀이간의 평균 상호정보량을 가지는 공기 어휘들의 비율을 가중치로 반영한다.

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

$AMIF(C,E)$ 는 문맥 내 전체 공기 어휘들(cw) 중 뜻풀이 E 와 평균 상호정보량을 가지는 공기 어휘의 비율로서 다음 수식과 같다.

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

일반적으로 실생활에서 많이 사용되는 어휘의 뜻풀이는 그렇지 않은 어휘에 비해 상대적으로 상세하게 기술된다. 따라서, 어휘 의미 별 뜻풀이의 길이를 가중치로 반영하였다.

또한, 의미 부착된 세종코퍼스로부터 의미의 빈도를 추출하여 이를 가중치로 반영하였다.

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

6. 실험

실험은 Lesk의 방법론과 평균 상호정보량에 기반한 방법론의 비교 실험과 평균 상호정보량에 다양한 가중치 정보값을 적용한 실험으로 구분하여 수행하였다. 실험을 위해서 질의응답에서 사용되는 500개의 질의와 정답 데이터에 의미를 부착하여 평가셋을 구축하였다.

6.1 Lesk방법론과의 비교

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.

[표 3]과 [그림 2] 에서 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 LESK실험에 비해서 상대적으로 윈도우 사이즈에 따른 정확률 향상이 완만함을 알 수 있다. 즉, 본 논문에서 제시한 평균 상호정보량이 자료부족문제를 많이 완화시킬 수 있음을 의미하는 것이다. 또한, LESK 실험에서 윈도우 사이즈에 따라 가장 낮은 정확률과 가장 높은 정확률의 차이가 13.37%이고, AMI

실험에서는 6.17%로 LESK실험의 차이보다는 많이 낮다는 것을 알 수 있다. 이 또한 AMI 방법론이 Lesk 방법론의 자료부족문제를 많이 완화시킨다는 것을 입증하는 결과이다.

표 3. Lesk 방법론과 평균 상호정보량에 기반한 방법론의 실험 결과

윈도우 사이즈	LESK	AMI
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

LESK:Lesk의 방법론
AMI : 평균 상호정보량을 이용한 WSD

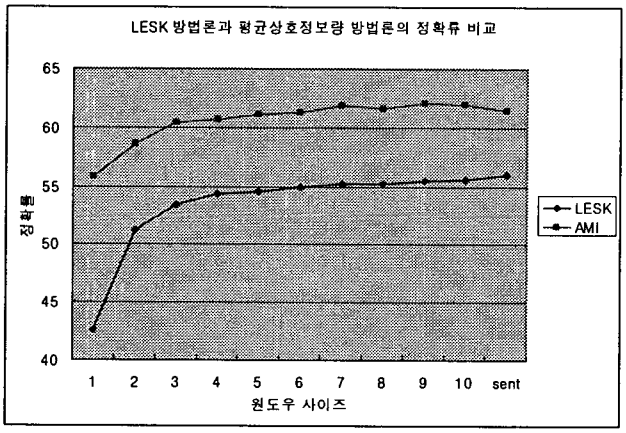

그림 2. Lesk 방법론과 평균 상호정보량에 기반한 방법론의 실험 결과 차트

LESK 실험에서는 문장전체를 봤을 때 정확률이 55.98%로 가장 높았고, AMI 실험에서는 윈도우 사이즈가 9일 때 정확률이 62.04%로 가장 높았다. 가장 높은 윈도우 사이즈를 기준으로 6.06%의 정확률이 향상되었다. 평가셋에서 중의성 어휘로 인식된 어휘들의 평균 의미 수는 5.67개였다.

6.2 가중치 부여 실험

본 논문에서는 평균 상호정보량의 단점을 극복하기 위하여, 세 종류의 가중치를 이용한다. 첫째, 상호정보량 값을 가지는 어휘 쌍의 비율(AMIF)을 가중치로 반영하였고, 둘째, 뜻풀이의 길이(DS)를 가중치로 반영하였고, 셋째, 의미 부착 코퍼스에서 추출한 의미 별 빈도 정보(SF)를 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 4. 가중치 적용에 따른 WSD의 정확률

윈도우 사이즈	AMI	AMI + AMIF	AMI + DS	AMI + SF
1	55.87	54.94	66.39	76.85
2	58.6	58.09	69.09	80.4
3	60.38	59.48	69.46	81.5
4	60.66	60.24	69.99	82.07
5	61.14	60.89	70.08	82.26
6	61.25	61.2	70.39	82.06
7	61.9	**61.56**	70.39	82.52
8	61.62	61.39	70.39	82.74
9	**62.04**	61.28	**70.67**	82.71
10	61.99	61.37	70.45	82.94
문장	61.42	61.53	70.19	**83**

[표 4]에서 알 수 있듯이, 정확률 향상에 가장 큰 영향을 미치는 가중치는 SF 가중치이다. 또한, DS 가중치도 정확률 향상에 많은 영향을 미쳤다. 그러나, AMIF 가중치는 정확률 향상에 영향을 미치지 않는 것으로 결과가 나왔다. DS 가중치가 적용되었을 때는 윈도우 사이즈가 9일 때 70.67%의 정확률로 가장 높았고, AMI만 이용했을 때의 최고 성능 보다 8.63%의 정확률 향상이 있었다. SF 가중치를 적용하였을 때는 문장 전체의 문맥 어휘를 대상으로 하였을 때 83%의 정확률을 보였다. 이는 AMI만 이용했을 때 보다는 20.96%의 정확률 향상이 있었고, DS 가중치가 적용되었을 때 보다는 12.33%의 정확률 향상이 있었다.

[표 5]는 세가지 가중치의 조합에 따른 어휘 의미 중의성 해소 정확률 변화를 분석한 결과이다. 가장 좋은 결과를 보인 조합은 모든 가중치를 다 적용한 경우와 AMIF + SF 가중치 조합을 사용한 경우이다. 가중치의 조합들 중에서는 DS + AMIF 가중치 조합이 윈도우 사이즈 9일 때 71.07%의 정확률로 가장 낮은 결과를 보였다. DS + SF 가중치 조합은 윈도우 사이즈 9일 때 78%의 정확률을 보였다. DS 가중치는 "일반적으로 자주 사용되는 의미의 뜻풀이는 상대적으로 길다" 라는 직관을 반영하기 위한 것이었다. DS 가중치가 개별적으로 적용되었을 때는 정확률 향상에 큰 영향을 미쳤다. 그러나, 모든 가중치를 적용한 결과와 SF + AMIF 가중치 조합을 적용한 결과가 비슷한 것을 봐서는 실질적인 의미 빈도 가중치인 SF 가중치가 DS 가중치가 담당한 역할을 포함하는 것으로 볼 수 있다. 즉, DS 가중치가 직관을 완벽하게는 반영하지 못하지만, 부분적으로 반영한다는 것으로 추정할 수 있을 것이다.

표 5. 가중치 조합에 따른 WSD 정확률

윈도우 사이즈	AMI+DS +AMIF	AMI+SF +AMIF	AMI+DS +SF	AMI +ALL
1	66.07	80.74	71.71	78.51
2	68.16	82.94	74.96	81.36
3	69.18	82.94	76.34	82.6
4	69.77	**83.42**	77.24	82.88
5	70.33	83.11	77.38	83.05
6	70.73	82.91	77.66	83.08
7	70.76	82.91	77.64	83.14
8	70.81	82.97	77.5	83.14
9	**71.07**	83.19	**78**	83.08
10	70.67	83.05	77.83	82.97
문장	70.75	82.94	77.95	**83.33**

ALL : 모든 가중치를 적용함

[그림 3]은 모든 가중치의 조합에 따른 정확률 변화에 대한 그래프이다. AMIF 가중치를 적용한 경우와 AMI만 이용한 경우를 비교했을 때, AMIF 가중치는 정확률 향상에 전혀 영향을 미치지 않는다. 또한 DS 가중치를 적용한 경우와 DS + AMIF 가중치 조합을 적용한 경우에도 AMIF 가중치가 정확률 향상에 영향을 미치지 못했다. 그러나, SF 가중치를 적용한 경우와 SF + AMIF 가중치 조합을 적용한 경우를 비교할 때, 윈도우 사이즈 7까지는 AMIF 가중치가 정확률 향상에 영향을 미치고 있다. 의미 중의성 해소를 위한 단서가 적은 경우(윈도우 사이즈가 작은 경우)에는 SF 가중치가 결과를 왜곡하는데, AMIF 가중치가 왜곡을 완화 시키는 것으로 분석되었다.

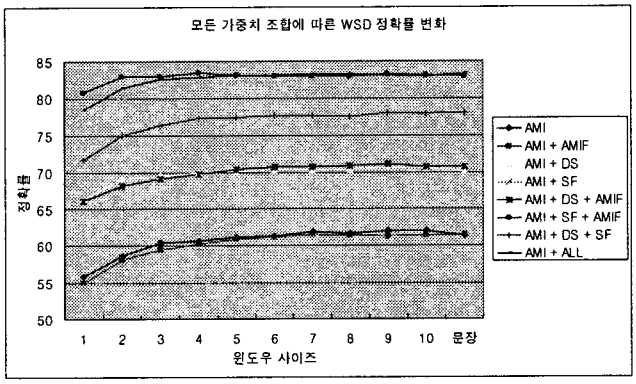

그림 3. 모든 가중치 조합에 따른 WSD 정확률 변화 그래프

6. 결론 및 향후 연구

본 논문에서는 평균 상호정보량을 이용한 동음이의어 의미 중의성 해소 기술을 소개하였다.

본 논문에서는 Lesk 방법론의 문제점인 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 평균 상호정보량을 계산하여 어휘의 의미 중의성을 해소하였다. 또한, 다양한 구조적 특징을 반영하기 위해서 상호정보량의 값을 가지는 어휘 쌍의 개수를 가중치로 적용하였고, 또한 사전 뜻풀이의 길이를 가중치로 반영하였으며, 의미 부착된 세종 코퍼스로부터 추출한 의미빈도를 가중치로 활용하였다.

본 논문에서 제안된 시스템의 성능에 대한 비교를 위해서, 크게 두가지 유형의 실험을 수행하였다. 첫째, Lesk의 방법론과 평균 상호정보량에 기반한 방법론의 성능 비교를 위한 실험으로써, 평균 상호정보량을 이용한 방법론이 Lesk의 방법론보다 6.40%의 정확률 향상이 있었다. 둘째, 다양한 가중치들이 어휘 의미 중의성에 미치는 영향을 분석하기 위한 실험으로써, 개별

가중치들 중 의미빈도 가중치가 가장 큰 영향을 미쳤다. 또한 가중치들의 조합에 따른 정확률 향상을 분석하기 위한 실험에서는 모든 가중치를 적용하였을 때와 의미빈도 가중치와 상호정보량을 가지는 어휘 쌍의 빈도 가중치를 조합하였을 때 가장 좋은 성능을 보였다.

본 연구의 의의를 간단히 기술한다면, 다음과 같이 정리할 수 있다.

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

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

앞으로 연구되어야 할 것을 정리하면 다음과 같다.

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

둘째, 본 논문에서는 개념망 사전만을 이용하였는데, 개념망의 다양한 의미관계들이 정확률 향상에 미치는 영향에 대해서 연구가 진행되어야 할 것이다.

셋째, 영어에 본 모델의 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 다의어에 기반의 어휘 의미 중의성 해소에서 요구되는 다양한 기술들에 대해서 분석하여야 할 것이다.

[참고문헌]

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

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

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

[4] Ganesh Ramakrishnan, B.Prithviraj, Pushpak Bhattacharyya, "A Gloss-centered Algorithm for Disambiguation", In Proceedings of SENSEVAL-3, 2004.

[5] Hee-Cheol Seo, Hae-Chang Rim, Soo-Hong Kim, "KUNLP System in SENSEVAL-3", In Proceedings of SENSEVAL-3, 2004.

[6] 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.

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

[8] 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.

[9] Mark Sanderson, "Word Sense Disambiguation and Information Retrieval", In Proceeding of ACM-SIGIR, 1994.