어휘의미망을 이용한 중국어 비감독 어의 중의성 해소

롄광저, 김민호, 권혁철
부산대학교 컴퓨터공학과
e-mail:lgz3235943@gmail.com

Chinese Unsupervised Word Sense Disambiguation using WordNet

Guang-Zhe Lian, Minho Kim, Hyuk-Chul Kwon
Department of Computer Science, Pusan National University

요 약

어의 중의성 해소는 자연어처리에서 중요한 역할을 한다. 감독 중의성 해소 방법은 비감독 중의성 해소 방법보다 높은 성능을 나타내지만, 구축비용이 큰 대규모 의미부착 말뭉치가 필요하다. 본 논문에서는 중국어 어휘의미망(HowNet)과 의미 미부착 말뭉치를 이용한 중국어 비감독 어의 중의성 해소 방법을 제안한다. 의미 미부착 말뭉치에서 통계정보를 추출하고, 중국어 어휘 의미망에서 중의성 어휘의 의미별 형제어를 추출하여 중의성 어휘의 주변 문맥에 나타나는 어휘와 카이제곱검정(χ^2-test)에 의한 독립성 검정을 통해 어휘 간 연관성을 판단하고 중의성 해소를 한다. 본 논문에서 제안한 중의성 해소 방법의 성능을 SemEval-2007 평가데이터에서 측정한 결과 명사와 동사에서 각각 64.7%, 49.4%를 나타냈다. 이는 SemEval-2007 중국어 비감독 중의성 해소에서 가장 높은 성능을 나타낸 시스템보다 13.1%, 13.9% 높은 성능이다.

1. 서론

자연어처리에서 어의(어휘 의미) 중의성 해소(word sense disambiguation; WSD)란 한 개 이상의 의미가 있는 단어가 실제 문장에서 어떠한 의미로 사용되었는지를 주변 문맥 정보에 의해 정확하게 구분하는 방법을 말한다. 실제 조사에 의하면 중의성 어휘가 중국어 사전과 말뭉치에서 차지하는 비중은 각각 14.8%, 42%라고 한다. 중국어 어휘의미망인 HowNet에서는 21.3%의 단어가 2개 이상의 의미가 있다[1]. 어의 중의성 해소는 자연어처리와 연관된 기계번역, 정보검색, 문서분류 등 여러 응용분야에서 중요한 역할을 한다.

어의 중의성 해소 방법은 사용되는 언어 자원에 따라 크게 지식 기반 어의 중의성 해소와 말뭉치 기반 어의 중의성 해소로 구분할 수 있다. 지식 기반 어의 중의성 해소는 사용하는 자원에 따라 기계 가독형 사전의 뜻풀이를 이용한 방법, 시소러스나 기계 가독형 사전에 의하여 제공되는 의미 범주를 이용하는 방법, 2개 국어로 된 사전에서 단어 대응을 이용하는 방법으로 구분할 수 있다. 말뭉치 기반 어의 중의성 해소는 대량의 말뭉치에서 추출한 통계정보를 이용하는 방법이다. 이 방법에서 어의 중의성 문제는 기계학습에서의 통계적 분류 문제로 단순화 되어 여러 기계학습 기법을 적용하여 해결된다. 여기서 개별 의미부착 말뭉치를 사용하는지에 따라 감독 중의성 해소와 비감독 중의성 해소로 나누어진다. 실제로 감독 중의성 해소 방법은 비감독 중의성 해소 방법보다 성능이 높게 나오지만, 대규모 의미부착 말뭉치가 필요하다. 하지만 대규모 의미부착 말뭉치를 구축하는데 비용이 많이 들기 때문에 특정 어휘 몇 개만 대상으로 한 논문들이 주를 이루고 있다. 이러한 중의성 해소 방법은 실제 응용프로그램에 적용하기 어렵다. 본 논문은 실제 응용프로그램에 적용하고자 기존 비감독 어의 중의성 해소보다는 성능이 높은 방법을 제안하다. 논문의 구성은 다음과 같다. 2장에서는 HowNet에 대하여 간단한 소개를 하고, 3장에서는 어휘의미망을 이용한 어의 중의성 해소 방법을 제안한다. 4장에서는 3장에서 제안한 중의성 해소 방법의 성능에 대해 평가를 하고, 5장에서는 결론 및 향후 연구에 대하여 기술한다.

2. HowNet 소개

HowNet은 하나의 객체를 중국어와 영어로 묘사한 지식 베이스로서, 개념과 개념 사이 관계와 개념이 가지고 있는 속성 사이의 관계를 포함 한다[2]. 여기서 개념이란 어의를 말한다. 하나의 중의성 어휘는 여러 개의 개념으로 표현될 수 있으며 어휘의 각각의 의미는 서로 다른 개념과 대응된다. 개념보다 더 작은 단위로는 의의소(semem)가 있는데 여러 개의 의의소가 모여서 개념을 이룬다. HowNet에는 1,500여 개의 의의소가 있는데 이들 간에는 복잡한 관계들이 존재한다. 예를 들면 상, 하위 관계, 전체-부분 관계, 동의어 관계, 반의어 관계 등이 있다[3]. 본

논문에서 제안한 방법은 상, 하위 관계를 이용하도록 한다.

3. 어휘의미망을 이용한 어의 중의성 해소 방법

감독 중의성 해소 방법은 비감독 어의 중의성 해소에 비해 높은 성능을 보이지만, 대규모 의미 부착 말뭉치가 필요하다. 의미 부착 말뭉치가 있다면 중의성 어휘가 특정 의미로 사용될 때, 어떠한 어휘들과 연관성을 가지는지 알 수 있다. 하지만 본 연구에서는 구축비용이 큰 의미부착 말뭉치를 사용하지 않고 중국어 어휘의미망인 HowNet의 관계 정보와 자체적으로 수집한 의미 미부착 말뭉치를 이용하여 특정 의미가 어떠한 단어와 연관성을 가지는지를 측정하였다. 일반적으로 독립성 검정을 이용한 두 단어 사이의 연관성 측정은 t-test를 많이 사용한다. 하지만 t-test는 어휘의 분포 확률이 정규분포를 따른다는 가정에서 측정을 하는데 실제 언어 현상은 그렇지 않다[4]. 본 논문에서는 이러한 가정이 없이도 독립성 검정이 가능한 χ^2 test를 사용함으로써 t-test보다 높은 성능을 나타낸다.

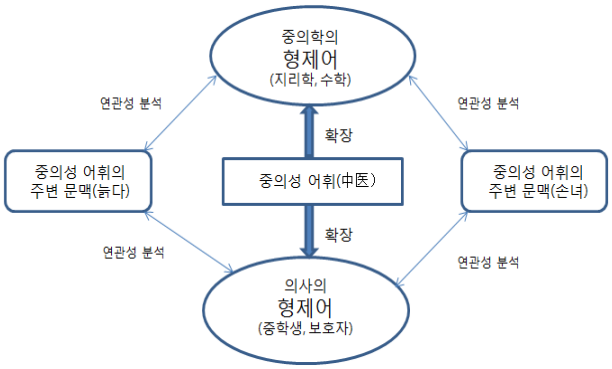
3.1 HowNet을 이용한 어휘 간 연관성 분석

중국어 어휘의미망 계층구조를 보면 형제어는 같은 성격을 지닌다. 예를 들면 중국어 中医는 의사(practitioner of Chinese medicine)와 중의학(traditional Chinese medical science) 두 가지 의미가 있다. 의사(practitioner of Chinese medicine)는 사람(human)의 하위어로 실제 문장에서 보면 孫女(손녀), 爺爺(할아버지)와 연관성이 있다. 의사(practitioner of Chinese medicine)의 형제어로는 학생, 보호자 등이 있는데 이러한 단어들도 孫女(손녀), 爺爺(할아버지)와 연관성이 있다. 반면에 중의학(traditional Chinese medical science)은 지식(knowledge)의 하위어로 孫女(손녀), 爺爺(할아버지)와 연관성이 없다. 마찬가지로 중의학(traditional Chinese medical science)의 형제어도 孫女(손녀), 爺爺(할아버지)와 연관성이 없다.

두 어휘의 연관성을 분석하려면 두 어휘의 공기 빈도를 알아야 한다. 독립성 검정(chi-square test)을 통하여 두 어휘 간의 연관성을 판단할 수는 있지만, 해당 의미는 판단할 수가 없다. 하지만 중의성 어휘 주변 공기 어휘와 의미별 형제어의 연관성을 판단하면 해당 의미를 구분할 수 있다. 아래와 같은 가정을 세워 문제를 해결하려고 한다.

> 가정: 중의성 어휘의 형제어와 주변 문맥 내 공기 어휘의 연관성은 중의성 어휘와 주변 문맥 내 공기 어휘와 정비례한다.

따라서 HowNet에서 중의성 어휘의 의미별 형제어를 추출하면 중의성을 해소 할 수 있다.
그림 1은 중의성 어휘의 의미별 형제어와 주변 문맥 내 어휘 간의 연관성을 표현한 것이다.



(그림 1) 중의성 어휘의 의미별 형제어를 이용한 연관성 분석

중의성 어휘 ω_{amb}의 의미가 S_k, 주변 문맥에 나타나는 공기어휘가 $v = \{v_1, v_2, \ldots, v_n\}$, r_i는 S_k의 형제어, l은 의미별 형제어 개수를 나타낸다. 아래 수식을 이용하여 두 어휘 간의 카이제곱(χ^2) 값을 구한다.

$$\chi^2(\omega_{abm} = S_k, v) = \frac{\sum_{i=1}^{l} \chi^2(r_i, v)}{l}$$

표 1은 위에 있는 수식을 이용하여 중의성 어휘의 의미와 주변 문맥 내 공기 어휘와의 연관성을 분석한 결과이다.

<표 1> '中医'의 형제어와 주변 문맥 내 공기 어휘와의 χ^2값

공기 어휘	中医(traditional Chinese medical science)와의 χ^2값	中医(practitioner of Chinese medicine)와의 χ^2값
走(가다)	26.72	20.83
遠(멀다)	8.96	14.93
遇到만나다)	4.13	5.62
老(늙은)	62.14	121.10
孫女(손녀)	1.76	233.24
說(말하다)	37.79	100.67
爺爺(할아버지)	0.003	100.32
知道(알다)	19.76	29.55
你們(너희)	4.41	41.13

3.2 중의성 어휘 의미 구분 방법

표 1에서 계산한 χ^2값을 이용하여 중의성 어휘 의미를 구분할 때 아래와 같은 몇 가지 방법을 사용할 수 있다. 첫 번째 방법은 χ^2값을 이용하여 공기 어휘가 해당 의미와 관련이 있는지를 판단하는 것이다. χ^2값이 일정한 수치(임계치)를 넘으면 관련 있다고 볼 수 있는데 관련 있는 공기 어휘의 개수가 많은 의미를 해당 의미로 판단한다. 예를 들면 임계치를 7.88이라고 했을 때 중의학(traditional Chinese medical science)과 관련 있는 공기 어휘가 5개 있고 의사(practitioner of Chinese medicine)와

관련 있는 공기 어휘가 8개 있다면 의사(practitioner of Chinese medicine)를 해당 의미로 결정한다. 두 번째로 중의성 어휘의 의미별로 주변 문맥 내 나타난 공기 어휘와 χ^2값의 비중을 이용하여 의미를 구분할 수가 있다. 아래 수식은 χ^2값의 비중을 곱으로 계산하여 값이 가장 큰 의미를 중의성 어휘의 의미로 결정한다.

$$WSD(\omega_{amb,}c) = \arg max_{s_k} \prod v_{j \in c} \frac{\chi^2(\omega_{amb} = s_k, v_j)}{\chi^2(\omega_{amb} \neq s_k, v_j)}$$

세 번째 방법은 두 번째 방법에서의 수식을 약간 변형한다. 위 수식에서 χ^2값의 비중을 곱하였다면 아래 수식에서는 χ^2값의 비중을 더한다.

$$WSD(\omega_{amb,}c) = \arg max_{s_k} \sum v_{j \in c} \frac{\chi^2(\omega_{amb} = s_k, v_j)}{\chi^2(\omega_{amb} \neq s_k, v_j)}$$

실험을 통하여 세 가지 방법을 비교해본 결과 세 번째 방법을 이용하였을 때의 결과가 가장 좋았다.

자료부족 때문에 χ^2값이 0이 나오는 경우가 있다. 이를 방지하기 위하여 smoothing method로 Laplace's Law를 사용하였다.

3.3 의미별 사용 정보 가중치

본 연구에서는 중의성 어휘 의미별 사용정보를 이용하기 위하여 아래와 같은 가정을 세웠다.

> 가정: 중의성 어휘의 형제어 빈도의 합에 대한 해당 의미별 형제어 빈도의 합의 비율은 의미별 사용 비율에 비례한다.

위 가정을 이용하여 계산한 의미별 사용 정보 가중치를 최종적으로 나온 χ^2값에 곱하면, 어느 정도 값을 바로 잡을 수 있다.

4. 실험 및 결과

4.1 실험 환경

본 논문에서는 통계정보를 추출하기 위한 학습 말뭉치를 자체적으로 수집하였다. 중국에서 가장 영향력이 있는 신문인 <<인민일보>>에서 2009, 2010년도 기사, 총 950만 단어를 수집하고 정제 및 형태 분석을 하였다. 실험에 사용된 태거는 중국과학기술원에서 개발한 ICTCLAS2012로써 98.45%의 성능을 나타낸다. 말뭉치에서 명사, 동사, 형용사 등 중의성 어휘 의미 결정에 영향을 주는 품사만 추출하여 단일 어휘 빈도 사전과 공기 어휘 빈도 사전을 각각 구축하였다.

어의 중의성 해소는 중의성 해소 대상 어휘에 따라 성능이 다르게 나올 수 있다. 본 논문에서 제안한 중의성 해소 방법과 타 시스템에서 사용한 방법을 비교하기 위하여 같은 평가 말뭉치인 SemEval-2007[5] Multilingual Chinese-English Lexical Sample Task[6]를 사용하였다. SemEval-2007은 ACL SIGLEX와 EURALEX의 후원으로 개최된 어휘 의미 중의성 해소 기술 평가 대회로 이전 명칭은 SensEval이다. SemEval-2007 Multilingual Chinese-English Lexical Sample Task의 평가 대상으로 되는 어휘는 명사 19개와 동사 21개로 구성되었다. 표 2는 중국어 평가 데이터에 구성에 대한 상세 내용이다.

<표 2> 중국어 평가 데이터 구성

	Average meaning	Training data	Testing data
nouns	2.45	1019	364
verbs	3.57	1667	571

어의 중의성 해소 평가 척도는 Macro average accuracy를 사용하였다.

$$P_{mac} = \frac{\sum_{i=1}^{N} p_i}{N}$$

여기서, N은 전체 어휘 수(The number of all test instances), $p_i = \frac{m_i}{n_i}$, m_i는 특정된 어휘에서 정확하게 의미를 구분한 어휘 수(correct labeled instances for one target word), n_i는 특정된 어휘 수(The number of test instances for one target word)를 의미한다.

4.2 실험 방법 및 타 시스템과의 비교

중의성 해소를 위해 중의성 어휘 주변 문맥의 어휘(윈도우 사이즈)를 이용해야 한다. 말뭉치의 규모나 성격에 따라 윈도우 사이즈가 다를 수 있다. 보통 윈도우 사이즈가 커짐에 따라 정확도가 점차 올라가다가 어느 순간부터 변화가 없거나 정확도가 떨어진다. 이는 윈도우 사이즈가 너무 커지게 되면 불필요한 정보들이 중의성 해소에 영향을 미치기 때문이다. 본 실험에서는 윈도우 사이즈가 6일 때 정확도가 가장 높게 나왔다.

중국어 어휘의미망인 HowNet에는 여러 가지 관계어가 존재하지만 본 논문에서는 형제어만 이용하였다. 문장에서 중의성 어휘를 찾은 다음 의미별 형제어 집합을 구축한다. 중의성 어휘의 주변 문맥의 어휘와 의미별 형제어 간의 연관성을 독립성 검정을 통하여 계산한 다음 그 값이 가장 큰 의미를 해당 의미로 선택한다.

본 논문에서 제안한 중의성 해소 방법의 성능을 평가하기 위해 SemEval-2007 중의성 해소 평가대회에 참가한 두 시스템을 Baseline으로 삼았다. TorMd[7]는 중국어 비감독 중의성 해소에서 가장 높은 성능을 나타 내였고 HIT는 Web을 이용한 비감독 중의성 해소 방법을 제안하

였다. 표 3은 명사를 대상으로 세 시스템 간 정확도를 비교한 것이다. 표 4는 동사를 대상으로 세 시스템 간 정확도를 비교한 것이다.

<표 3> 명사 실험 결과 비교

nouns	Meaning number	HIT	TorMd	CWSD
中医	2	0.500	0.438	0.813
儿女	2	0.500	0.500	0.850
單位	2	0.529	0.706	0.588
天地	3	0.440	0.560	0.68
旗幟	3	0.111	0.500	0.389
日子	3	0.344	0.281	0.406
本	3	0.320	0.720	0.800
机組	2	0.571	0.643	0.786
气息	2	0.571	0.857	0.646
气象	2	0.563	0.438	0.813
牌子	2	0.529	0.353	0.412
眼光	2	0.500	0.714	0.571
菜	2	0.632	0.474	0.579
表面	2	0.333	0.556	0.667
道	3	0.222	0.500	0.667
鏡頭	2	0.467	0.467	0.733
長城	3	0.619	0.429	0.524
隊伍	3	0.364	0.381	0.636
面	3	0.696	0.384	0.739
Ave P_{mac}	2.45	0.464	0.516	**0.647**

<표 4> 동사 실험 결과 비교

verbs	Meaning number	HIT	TorMd	CWSD
使	2	0.438	0.563	0.500
出	8	0.091	0.169	0.351
動	4	0.300	0.300	0.350
動搖	2	0.438	0.500	0.625
發	5	0.139	0250	0.389
叫	4	0.256	0.256	0.257
吃	4	0.174	0.174	0.348
帶	8	0.104	0.119	0.269
平息	2	0.500	0.375	0.750
開通	2	0.500	0.500	0.600
想	4	0.216	0.216	0.324
成立	3	0.407	0.481	0.370
挑	3	0.286	0.143	0.714
推翻	2	0.300	0.300	0.600
望	2	0.462	0.462	0.692
看	4	0.294	0.294	0.412
補	3	0.550	0.550	0.600
說明	2	0.556	0.444	0.667
赶	3	0.333	0.389	0.389
進	5	0.114	0.250	0.318
震惊	2	0.571	0.714	0.857
Ave P_{mac}	3.57	0.335	0.355	**0.494**

본 논문에서 사용한 중의성 해소 방법 CWSD는 인터넷에서 수집한 말뭉치와 HowNet에서 추출한 형제어 정보만 이용하여 중의성 해소를 하였음에도 불구하고 SemEval-2007 중의성 어휘 해소 평가대회에서 가장 높은 성능을 나타낸 TorMD보다 명사에서 13.1%, 동사에서 13.9% 높은 성능을 보였다. 반면에 TorMd는 여러 가지 언어학 자원을 이용하였다. 예를 들면 한, 영 번역 사전、Chinese News Magazine Parallel Text, Chinese News Translation Part1, Hong Kong Parallel Text, Chinese Treebank English Parallel News Text Version1.0 beta2 등을 포함한 각종 한, 영 Parallel corpus를 사용하였다. 만약 본 논문에서 제안한 방법에 이러한 자원을 적용한다면 더욱 높은 성능을 나타낼 것이다.

5. 결론 및 향후 연구

중국어 어휘 의미망에서 형제어는 비슷한 성격을 지니기 때문에 중의성 어휘의 의미별 형제어와 주변 문맥 내 공기 어휘와의 연관성을 분석하여 중의성 해소를 할 수 있다. 하지만 실험 결과로부터 보면 명사에 대한 중의성 해소는 비교적 만족스러운 결과를 얻었지만, 동사에 대한 결과는 만족스럽지 못하다. 원인으로는 아래와 같이 몇 가지로 볼 수 있다. 첫째, 동사의 평균 의미수가 명사에 비해 1.12개가 더 많다. 이것은 동사의 중의성 해소가 명사의 중의성 해소보다 난도가 높다는 것을 의미한다. 둘째, 동사가 가지고 있는 형제어 수가 명사에 비해 적다. 형제어 수가 적다는 것은 유의미한 통계 정보를 얻기 어렵다는 것을 의미한다.

앞으로 연구해야 할 문제는 다음과 같다. 첫째, HowNet에서 좀 더 다양한 관계 정보를 이용함으로써 시스템의 성능을 향상시킨다. 예를 들면 전체-부분 관계, 반의어 관계 등을 이용할 수 있다. 둘째, 자료 부족 문제 때문에 통계 정보로 해결되지 않는 어휘에 대해서 규칙을 이용하여 개선한다. 셋째, SemEval-2007 평가 데이터 외의 다른 다양한 데이터에 대해 평가함으로써 시스템의 신뢰성을 높인다.

참고문헌

[1] Chen Hao, CHENG Liang-lun, ZHANG Xiao-bo "Unsupervised approach to word sense disambiguation based on vetor space model" *Computer Engineering and Design*, Vol.28, No.5, pp.1215-1218, 2007 (in Chinese)
[2] Dong Zhendong, Dong Qiang, *Hownet and the Computation of Meaning*, 2006
[3] Xiaofeng Yu, Pengyuan Liu, Tiejun Zhao "A Method of Chinese Word Sense Disambiguation Based on Hownet" *Proceedings of The Second National Student Computational Linguistics Symposium*, 2004 (in Chinese)
[4] Church, Kenneth W, and Rober L.Mercer "Introduction to the special issue on computational linguistics using large corpora" *Computational Linguistics 19:1-24, 1993*
[5] SemEval-2007, http://nlp.cs.swarthmore.edu/semeval
[6] P. Jin, Y. Wu, S. Yu., "SemEval-2007 Task 5: Multilingual Chinese-English Lexical Sample," *Proceedings of the 4th International Workshop on Semantic Evaluations(SemEval)*, 2007
[7] S. Mohammad, G. Hirst, P. Resnik. Tor, "TorMd: Distributional profiles of concepts for unsupervised word sense disambiguation," *Proceedings of the 4th International Workshop on Semantic Evaluations(SemEval)*, 2007