통계기반 의미중의성 해소를 이용한 정보검색

허정[0] 김현진 장명길
한국전자통신연구원 휴먼정보검색연구팀
(jeonghur[0], jini, mgjang)@etri.re.kr

Information Retrieval using Word Sense Disambiguation based on Statistical Method

Jeong Hur[0] ,Hyun-Jin Kim, Myung-Gil Jang
Human Information Retrieval Team, ETRI

요 약

인터넷의 발전과 더불어 기하급수적으로 늘어난 디지털 정보를 대상으로 사용자의 요구를 만족시키는 정보검색을 하기 위해 자연어처리 기술이 많이 응용되고 있다. 본 논문에서는 정보검색에 자연어 처리 기술 중, 의미중의성 해소(WSD) 기술을 적용하였다. HANTEC 12만 문서를 대상으로 9개의 중의성 단어를 실험한 결과 67.8%의 정확률을 보였다. 본 실험을 통해 WSD의 오분석이 정보검색의 정확률에 상당히 민감한 결과를 초래함을 알 수 있었다. 그리고, WSD 기술이 정보검색에 적용될 때 발생할 수 있는 여러 문제점들에 대하여 논의 하였고, 이 문제점의 근원적인 해결방안은 WSD 기술의 발전에 있다는 것을 알 수 있었다.

1. 서 론

정보 검색 시스템은 다양한 정보를 분석하여 일정한 형태로 저장한 후 정보에 대한 요구가 발생했을 때 적합한 정보를 제공하는 시스템을 말한다. 인터넷의 발전과 더불어 기하급수적으로 늘어난 정보를 대상으로 정보검색을 하기 위해서는 정보의 분석방법과 사용자의 요구에 적합한 정보를 검색하는 방법이 중요하다. 이와 같은 연구방향에 부합하여 최근 정보검색에 자연어 처리(NLP : Natural Language Processing) 기술이 많이 응용되고 있다.

의미 중의성 해소(WSD:Word Sense Disambiguation)는 단어의 의미를 의존관계에 있는 단어들과의 의미적 정합성으로부터 결정하는 것을 말한다. WSD는 NLP응용분야(정보검색, 기계번역 등)에서 성능저하의 원인 중에 하나이다. 최근 많은 연구의 결과로 다양한 방법의 WSD 기술들이 소개되고 있다.

정보검색에서 사용자 질의를 분석하여 제공되는 검색결과는 일반적으로 수 십 페이지에 이른다. 연관성 평가에 대한 기술의 발전으로 상위에 리스트되는 결과들이 비교적 사용자의 요구에 부합되는 결과들이나 아직 한계가 있다. 가장 큰 문제는 질의와 문서 내에 출현하는 단어들의 중의성 문제이다. 중의성을 지닌 단어가 질의어로 사용이 된다면, 검색된 결과에서 사용자가 요구하는 의미에 해당되는 키워드가 포함된 문서를 검색하는 것은 사용자의 몫이 된다. 이는 정보검색에서 사용자의 만족도를 저하시키는 주요한 요인이다.

본 논문에서는 사용자의 만족도를 높일 수 있도록 WSD기술을 정보검색에 적용하는 의미기반 정보검색 시스템을 제안[3]하고 그 문제점들에 대해서 논의한다.

2. 통계기반 단어 의미중의성 해소 방법

문장 중의 단어는 공기관계를 이루는 다른 단어에 의해 그 의미를 결정할 수 있다. 예를 들어, " 하늘에서 하얀 눈이 내린다." 에서는 " 하늘", " 하얀", " 내린다" 의 단어들과 " 눈" 의 의미적 정합성을 통해 " 눈" 의 의미가 " snow" 임을 알 수 있다.

본 시스템에서 사용되는 WSD 모델은 [1]의 엔진을 사용하였다. [1]에서는 사전의 의미기술에 사용된 단어를 의미정보 집합으로 추출한다. 자료부족(data sparseness)문제를 해결하기 위해 단어 의미기술의 패턴정보를 이용하여 단어의 상하관계를 파악하고 의미정보 집합의 대상을 하위어까지 확장하였다. WSD는 통계적인 방법에 의해 구현하였다. [그림 1]은 WSD의 전체 구성도이다.

[1]에서는 금성사전의 의미기술 문장만을 이용하여 의미정보집합을 구축하였으나, 본 시스템에서는 계몽대백과사전의 의미기술 문장과 소량의 WEB 데이터를 포함

하였다.

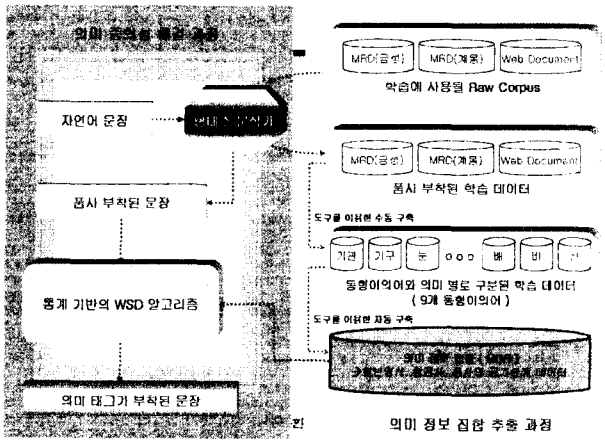

그림 1. WSD 시스템 구성도

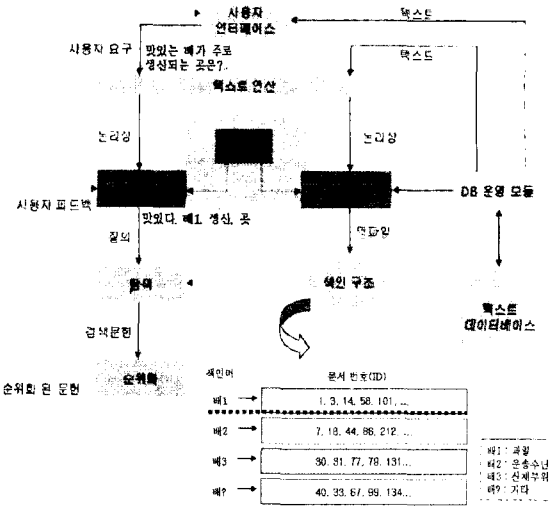

그림 2. WSD가 포함된 정보검색 구성도

3. 통계기반 WSD를 이용한 정보검색

정보 검색은 전통적으로 불리언 방식, 벡터방식과 확률 방식으로 나눌 수 있다[4]. 본 시스템에서는 벡터 방식을 이용한 [3] 의 정보검색엔진[1]을 이용하였다.

정보검색에서 WSD의 엔진이 포함되는 곳은 색인과 질의분석 부분이다. 문서를 색인 할 때, 문서 내에 출현하는 단어들 중, 중의성 단어에 대해서는 WSD 엔진을 이용하여 의미태그를 부착한다. 따라서, 색인 될 때는 색인어에 의미태그가 부착된 단어가 사용이 되는 것이다. 질의분석 시에도 색인 때와 동일한 방식으로 의미태그가 부착된다. 중의성 해소를 실패한 단어들을 포함한 문서를 고려하기 위해 중의성 해소 실패에 대한 의미태그를 추가한다.

[그림 2]는 정보검색에 WSD 엔진이 포함된 구성도이다.

[1] 위의 검색엔진은 5가지 의미색인단위(SIU:Semantic Index Unit)를 이용하여 색인한다.
SIU1 : 명사, SIU2 : SIU1 + 용언, SIU3 : SIU2 + 문장정보, SIU4 : 명사구, SIU5 : Head-Modifier .
본 논문에서는 SIU1의 색인 단위에 단어 의미중의성 해소 모듈을 적용하였다.
논문[3]과 관련된 자료는 다음 사이트를 참고하세요.
http://sir.etri.re.kr/paper.html

4. 실험 결과

실험은 HANTEC 12만 문서를 대상으로 9개의 중의성 단어(기관, 기구, 눈, 다리, 병, 배, 비, 신, 차)에 대해 실시되었다. 정답문서가 없기 때문에 평가기준을 다음과 같이 정했다.

o 평가기준 : 검색된 문서에 출현하는 중의성 단어의 의미들 중 질의문의 중의성 단어와 동일한 의미로 사용된 것이 있으면 정답으로 인정한다.

위의 평가기준을 바탕으로 9개의 중의성 단어에 대하여 상위 10위 안의 문서의 정확률은 평균 67.8%였다. [표 1]은 검색 결과이다.

표 1. 상위 10 내에서의 검색 정확률

중의성 단어	의미	정답수	정확률	중의성 단어	의미	정답수	정확률
기관	신체부위	6	60.0	병	그릇	10	100.0
	장치	9	90.0		사람	6	60.0
	조직	10	100.0		상태	10	100.0
기구	장치	7	70.0	비	도구	0	0.0
	조직	10	100.0		물	10	100.0
	신체부위	10	100.0		비석	1	10.0
눈	식물	1	10.0	신	신발	1	10.0
	기상현상	10	100.0		종교	5	50.0
다리	교각	9	90.0	차	운송수단	9	90.0
	발	9	90.0		음료	0	0.0
배	과일	10	100.0				
	운송수단	7	70.0				
	몸	5	50.0				

상대적으로 정확률이 낮은 의미들은 실험 대상 문서에 출현이 적은 것으로 대용량의 웹문서를 대상으로 실험하면 다소 정확률 향상이 있을 것으로 기대된다.

본 실험에서 문제점은 크게 WSD 와 정보검색에서의 한계에 따른 문제점으로 구분할 수 있다.

WSD의 문제점은 다음의 3가지로 볼 수 있다. 첫째, WSD의 오분석에 대한 대안이 없이 정보검색에 반영되었다는 것이다. 이 문제점은 질의분석과 색인 시에 큰 문제점으로 작용한다. 그러나 질의분석 시의 WSD 오분석은 사용자 피드백(user feedback)을 이용하여 해결할 수 있으나, 색인 시에 발생하는 WSD의 오분석은 정보검색의 정확률 저하의 중요한 요소이다. 그러므로, 색인 시의 WSD는 임계값을 정하여 의미분별을 하고, 임계값에 미치지 못하는 중의성 단어는 의미분별 실패로 의미태그 '?'를 부착하여 색인한다. 그리고 질의 분석 시의 WSD 결과를 가지고 해당 의미에 대한 색인 결과와 의미분별 실패에 해당하는 색인 결과를 동시에 고려하여 원하는 문서를 검색하는 방법(의미분별 실패에 해당하는 색인 결과의 가중치를 낮추어준다.)이 있을 수 있으나 연구가 진행되어야 할 부분이다. 둘째, 한 문장 내에 중의성 단어가 둘 이상 출현했을 때의 처리 방안이 없다는 것이다. 이 문제는 품사 태깅 문제와 유사하다. 그러므로 통계적 품사태깅 기술을 응용하여 해결이 가능할 것이다. 셋째, WSD가 적용된 단어가 9개로 다소 적다는 것이다. 이는 의미정보집합의 확장과 관련되는 문제로 의미정보집합의 자동 구축 방법과 연관이 된다. 현재 의미 태그가 부착된 전자사전이 구축되어 자동 의미정보집합 구축이 가능해졌으므로 조만간 WSD 적용단어의 확장이 이루어 질 것이다.

정보검색에서의 문제점은 크게 2가지로 볼 수 있다. 첫째, 질의문이 대체로 단문이거나 명사구 형태로 중의성 단어와의 공기관계를 이용한 중의성 해소에 한계가 있다는 것이다. 이 문제는 사용자 피드백을 이용하여 사용자가 원하는 의미의 정보를 검색할 수 있도록 할 수 있다. 둘째, 정보검색의 정확률이 WSD의 오분석에 민감하다는 것이다. WSD가 오분석되어 색인된 문서는 올바른 질의에 의해서는 절대로 검색할 수 없다는 것이다. 이 문제에 대한 해결방법은 WSD 시스템의 의미별 확률값 정보를 유지하면서 중의성단어의 모든 의미로 색인을 하는 방법이 있을 수 있다. 그리고, 검색을 수행할 때 의미별 확률값을 문서의 랭킹을 조정할 때 반영함으로써 WSD의 결과를 반영하는 방법이 있을 수 있으나, 상대적으로 모든 의미별로 색인을 하므로 색인양이 커지는 단점이 있다. 이 방법은 향후 연구할 대상이다.

결론적으로 WSD가 적용된 정보검색은 WSD의 결과에 상당히 민감하다. 따라서, 사용자가 요구하는 수준의 정보검색을 위해서는 WSD의 정확률이 크게 향상되는 것이 가장 근원적인 해결책이라는 것을 알 수 있었다.

5. 결론

본 논문에서는 통계기반 WSD 엔진을 이용한 정보검색 시스템을 구현하였다. 정보검색에서 WSD의 중요성은 누구나 인식하는 문제이다. 그러나, 아직 WSD 기술의 한계와 정보검색의 한계로 인해 괄목할 만한 결과를 보이지는 못했다. 실험은 9개의 중의성 단어를 대상으로 HANTEC 12만 문서에서 진행되었는데, 정확률이 67.8%이었다. 이번 실험을 통해 WSD의 정확률이 높은 수준에 이르지 않는다면 정보검색에서 그다지 좋은 성능을 기대하기는 힘들다는 것을 알 수 있었다. 그리고, WSD의 오분석이 정보검색에서의 정확률과 재현율에 민감한 결과를 초래함을 알 수 있었다.

참고 문헌

[1] 허정 외 1명, "사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 정보과학회논문지, 제 28권 제 9 호, 2001.09.

[2] Mark Sanderson, "Word Sense Disambiguation and Information Retrieval", ACM-SIGIR, 1994.

[3] 장명길 외 6명. "의미기반 정보검색", 정보과학회지, 2001.10.

[4] 김명철 외 5명 공역, "최신정보검색론", 홍릉과학출판사, 2001.

[5] 김영택 외 공저, "자연언어처리", 생능출판사. 2001.