상호 정보를 이용한 어의 모호성 해소에 관한 연구

°전 미선, 박 세영
한국전자통신연구소

A Study on Resolving Word Sense Ambiguity Using Mutual Information

°Meesun Jeon , SeYoung-Park
Electronics and Telecommunicaitons Research Institute

요 약

정보 검색 시스템의 정확성은 색인어의 정확성과 질의 해석의 정확성에 의존한다. 한국어 정보 검색 분야에서는 한국어의 특성을 고려하는 것이 무엇보다 중요하다. 한국어의 문서 색인과 질의 해석시 야기되는 어의 모호성(word sense ambiguity)을 가지는 단어에 대해서는 어의 모호성을 해소한 정확한 색인과 질의 해석이 전제되어야 정확한 문서를 검색해낼 수 있다. 본 논문은 한국어 문서 색인시 동음이의어(homonym)에 의해 발생하는 어의 모호성을 해소하기 위한 방안에 대해 다루고 있으며 의미적 관련 정보를 이용할 것을 제안하고 타당성을 보이는 실험 결과를 제시한다.

1. 서론

현대와 같은 정보화 사회에서는 사람이 관리하기에 불가능할 정도로 쏟아져 나오고 있는 수많은 정보들을 컴퓨터에 저장하여 색인을 비롯한 정보의 가공 작업을 거쳐 필요에 따라 사용자에게 서비스하는 정보 검색 시스템이 널리 이용되고 있다. 이러한 정보 검색 시스템(information retrieval system)의 성공 여부는 제공되는 정보의 정확성과 색인어의 정확성, 정보 검색의 속도, 질의 해석의 정확성 등에 의하여 결정된다. 색인이란 저장된 문서들을 미리 분석하여 그 문서의 주요 단어 또는 주요 어구를 추출한 후 찾기 편리한 형태로 저장하는 것을 말하며 검색시 문서의 내용 전부를 검색하지 않고 주요어만을 검색함으로써 빠른 시간에 사용자의 요구를 만족시킬수 있다.

색인어를 추출하는 방법으로는 우선 전문화된 사람에 의한 색인어 추출을 들 수 있으나 정보가 기하 급수적으로 늘어나고 있는 현실을 감안한다면 사람에 의한 색인어 추출 방법은 인력과 시간의 낭비를 초래한다. 그러므로 사람의 힘을 빌리지 않고 컴퓨터를 이용하여 입력 문서로부터 색인어를 자동으로 추출하는 방법에 대한 연구가 이루어지고 있다. 한국어 정보 검색 분야에서 한국어에 대한 자동 색인은 한국어의 특성을 고려하는 형태소 분석 및 구문 분석 등을 이용한 색인이 이루어져야 한다.

형태소 분석을 이용한 색인은 형태소 사전을 이용해 입력 문서의 각 단어에 대해 형태소들을 분리한후 단어의 원형을 복원하고 접두어, 접미어, 조사, 보조사 등을 분리해낸다. 그 다음 통계적 방법을 이용하여 가능한 명사를 모두 추출한 후 불용어 리스트(stop list)를 적용해 필요없는 명사를 제거하고 마지막으로 남는 명사들을 색인어로 취한다. 이 방법은 구현이 간단하고 한국어에도 쉽게 적용하여 사용할 수 있는 장점이 있으나 통계적 방법의 사용으로 인해 색인어 추출의 정확성이 떨어지고 형태소 분석을 각각의 단어(single word)에 대해 행한 후 빈도수 및 가중치를 계산하므로 구 단위의 색인어 추출이 어렵고 단어 자체에서 나오는 애매성이 있다.

구문 분석을 이용한 색인 방법은 형태소 분석 결과를 가지고 구문적 의미를 지니는 특정한 단어 및 단어구를 색인어로

추출한다. 이 방법은 형태소 분석 방법보다 색인어를 훨씬 잘 추출할 수 있을 뿐만아니라 구단위의 색인어도 훌륭히 추출할 수 있다. 그러나 구문 분석 결과에서 나오는 구문적 애매성(syntactic ambiguity)과 단어 자체에서 나오는 애매성(word sense ambiguity)이 있다.

한국어 문장을 분석하는데 있어서 형태소 분석과 구문 분석을 거치더라도 단어 자체에서 나오는 어의 모호성이 발생하며 이러한 모호성은 어떻게든 해소되어야 한다. 본 논문은 한국어 문장 분석시 동음이의어에서 나오는 어의 모호성을 해소하기 위해 필요한 의미 지식을 상호 정보 개념을 도입하여 백과 사전으로부터 습득하였다. 2장에서는 의미적 관련 정보를 획득하기 위한 입력 자료인 (주)계몽사 학생 백과 사전에서 사용된 마크업 기호들과 획득한 의미적 관련 정보에 상호 정보개념을 도입해 어의 모호성을 해소하는 과정을 보이며 3장에서는 실험 결과를 보인다.

2. 어의 모호성 해소

2.1 마크업

어의 모호성을 해소하기 위해 의미적 관련 정보 구축시 입력 자료가 된 학생 백과 사전의 내부에 의미를 부여하기 위해 사용된 마크업 기호들은 다음과 같다[1].

! : 백과 사전의 표제어를 뜻한다(약 22,000여개)
: 표제어가 2개 이상임을 뜻한다.
@ : 해설 옮김을 뜻한다.(See Also)
$: 참조어를 뜻한다. (Alternative)
% :대항목(약 2,000자 정도의 설명)의 부제목
\' : 의미 있는 최소 단위의 단어

\'는 표제어를 설명하는 단어들 중에서 표제어와 의미적으로 관련이 있다고 생각되는 단어들만 추려서 마크업하는데 이용되었다.

2.2 의미적 관련 정보

자연 언어(natural language)를 분석하는데 있어서 어의 모호성(word sense ambiguity) 해소는 매우 중요한 문제이다. 특히 정보 검색 시스템에서는 색인어 추출과 질의 해석시에 빈번하게 발생하며 이 문제를 제대로 해결하지 않고서는 실용적인 정보 검색 시스템을 개발한다는 것이 거의 불가능하다. 이 문제를 해결하기 위해 말뭉치로부터 습득한 지식을 사용하여 모호성을 해소하려는 방안이 연구되고 있다.

Jensen은 전자 사전(on-line dictionary)에 들어있는 뜻풀이나 예문으로부터 언어처리에 필요한 의미 지식을 습득하는 기법을 제시하였는데[8] 지식 습득 과정을 자동화했다는 점에 있어서는 가치가 있지만 습득된 지식이 너무나 일반적이어서 특정 분야(domain)에 사용되기에는 부족하다는 문제점이 있다. 이런 문제점때문에 말뭉치로부터 지식을 습득하는 방법에 대한 연구가 많이 진행되고 있다. 특정 분야에서 수집된 말뭉치는 그 분야의 전문 지식을 잘 반영하기 때문에 전자사전으로부터 지식을 습득할 때와 같은 문제가 생기지 않는다.

Yarowsky는 어의 모호성 해결을 위해 Roget's Thesaurus에서 분류된 개념 범주(conceptual category)를 이용한다[7]. 여기서는 말뭉치를 이용한 통계적 학습을 통해 각 범주를 지시하는 대표적인 단어들과 그 단어들의 가중치를 구한다음 이를 사용하여 문맥으로부터 그 안에 쓰인 특정 단어의 어의를 구분해내는데 이용하고 있다. 이 방안은 어의 모호성 해소에 도움을 주기는 하지만 잘 분류된 개념 범주 체계를 전제한다. 본 논문에서 고려하는 문서의 자동 색인시 발생하는 어의 모호성 해소 방법에는 백과 사전으로부터 얻은 통계적으로 얻은 의미적 관련 정보를 사용하는 방안을 도입한다.

본 논문에서는 전자 사전의 내용이 .일반적이라는 단점을 어느정도 해결하고자 백과 사전을 선택하였으며 백과 사전의 특성상 표제어와 표제어 풀이 문장에 쓰인 거의 모든 단어들이 의미적으로 밀접한 관계를 가지고 있으며 정확성을 기하기 위해 수동으로 의미적 관련이 있는 단어들의 앞뒤에 \'를 마킹하였으며 필요하다면 구단위의 마킹도 하였다. 예문(1)이 6가지 마크업 기호를 사용하여 마킹한 예이다. '가지'는 표제어 가부좌와 의미적으로 관계가 없으므로 마킹되지 않았으며 '앉는 법'은 구단위의 마킹이 되어있다. 그리고 예문 (1)에는 나타나 있지 않지만 말과 같은 동음이의어 표제어는 말[1], 말[2], 말[3]와 같이 표기되어 있다.

(1) ! 가부좌

\앉는 법\의 한 가지. 일반적으로 \책상다리\를 하고 \앉는 법\을 말한다. 원래는 \부처\의 앉는 모양을 가리키는 것으로 \결가부좌\의 준말이다.
\오른발\을 \왼편\ \넓적 다리\ 위에 올려 놓고, \왼발\을 \오른편\ \넓적다리\ 위에 올려 놓고 앉는 법을 이른다.

문장내에 쓰인 모든 단어는 그 자체로서 의미를 가진다기 보다는 다른 단어와 함께 쓰일 때 더욱 명확한 의미를 전달하게 되는 것이다. 따라서 한 단어의 의미는 그 자체에 내재되었다기 보다는 다른 단어에 의하여 정의된다고 볼 수 있다. 그러므로 한 문장내에서 함께 나타날 수 있는 단어들은 서로 의미적으로 관련이 있다. 의미적 관련이 없다면 그 문장은 자연스럽지 못한 문장이 되고 만다. 문맥내에서 자연스럽게 결합 가능한 단어들은 서로 연어 관계에 있다고 볼 수 있다. "세계를 간다(유럽 14개국)"이라는 관광 안내 서적에서 발췌한 예문 (2)는 의미적으로 관련 있는 정보들이 어의 모호성 해소에 사용될 수 있음을 보여준다.

(2) 프랑스는 풍부한 휴머니티와 자유 경쟁 원리, 평등, 박애의 정신에 입각한 프랑스 혁명으로 민주주의의 초석을 마련한 국가이다.

초석이란 단어는 단독으로 기초를 뜻할 수도 있고 질산칼륨이라는 화학 물질을 뜻할 수도 있는 2가지 뜻을 내포하는 동음이의어이다. 문장(2)에서 쓰인 초석이 가지는 어의 모호성은 한 문장내에 쓰인 프랑스, 휴머니티, 자유 경쟁, 원리, 평등,박애, 정신, 프랑스 혁명,민주주의와 같은 의미적 관련 정보에 의해 질산칼륨의 의미가 아닌 기반, 기초를 뜻함을 알 수 있다. 마찬가지로 국가가 가지는 모호성은 한 나라를 상징하는 노래가 아니라 통치 조직을 뜻함을 알 수 있다. 만약 한 문장내에서 의미적 관련 정보를 발견하지 못한다면 그문장이 속한 단락(paragraph)에서 찾아야한다. 이처럼 동음이의어 각각에 대해 충분한 의미적 관련 정보를 가지고 있다면 어의 모호성 해소가 가능하다.
예문(1)에서 표제어와 그 뜻풀이 문장들중에서 '\'로 마킹이 된 단어들만을 추출하여 표제어 명사x의 의미적 제약에 해당하는 정보를, 즉 명사x-명사 또는 명사x-명사구 쌍들의 빈도수에 관한 통계를 얻을 수 있다면 수식(1) 에 의해 상호 정보(Mutual Information) 값을 계산할 수 있다.

$$MI(x,y)=\log_2\frac{P(x,y)}{P(x)P(y)}$$

$$=\log_2\frac{N*f(x,y)}{f(x)f(y)}+1$$

$$(단 f(x)f(y)>0) \qquad (수식1)$$

상호 정보 수식(1)은 단어와 단어 사이의 연관성을 정량적으로 나타내기 위해 사용되어 왔다. 단어 x와 y가 밀접한 관계가 있다면 MI(x,y)의 값은 0보다 큰 값이 될 것이며 그다지 관계가 없다면 0에 가까운 값이 될것이며 전혀 관계가 없다면 0이 될 것이다. 예문 (1)의 표제어 '가부좌'를 예로 들어 공기 빈도수 f(x,y)를 구하는 방법을 설명하기로 한다. 예문 (1)에서 '가부좌'와 '앉는 법'이 의미적으로 연관이 있으므로 공기 빈도수 f(가부좌,앉는 법)을 하나 증가 시킨다. 마찬가지로 f(가부좌,책상다리)도 하나 증가시키면서 가부좌에 대한 설명이 끝날 때까지 반복한다. <표1>은 백과 사전으로부터 얻어진 상호 정보 값의 일부를 보인 것으로 단어x와 의미적으로 관련이 있다고 판명되어진 모든 단어 또는 구들중에서 빈도의 임계치(threshold)를 1로 설정하여 그 이하의 자료를 무시한후 상호 정보 값을 구하였다.

2.3 어의 모호성 해소

두 단어 x와 y가 어떤 의미 관계를 맺고 있다고 할 때 상호 정보 MI(x,y)는 단어 x와 y가 얼마나 긴밀한가 하는 것을 수량으로 나타낸 것으로 볼 수 있다고 했는데 이 정보를 이용하면 어의 모호성을 해소할 수 있다. 상호 정보를 이용하여 모호성을 해소하는 방법을 예를 들어 살펴보자.

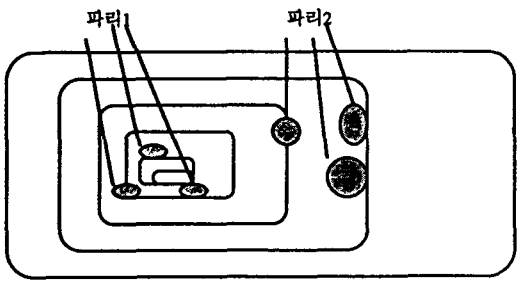
(3) '파리가 세계에서 가장 아름다운 도시라는 것에 이론을 제시할 여지가 없다...'라고 말한 빅토르 위고나 '모든 것, 그 추한 것까지도 매력으로 변하게 하는 파리'라고 한 [악의 꽃]의 시인 샤를 보들레르의 말과 같이 많은 예술가들이 파리에 매혹되고, 파리를 사랑하고, 파리를 찬미하는 언어를 계속 찾아왔다. 파리가 유럽에서 제일 큰 도시라고는 하지만 걸어서도 시내 구경을 할 수 있을 정도의 크기이다.

(4) 파리는 각종 병원균을 옮기는 해충으로 여름철에 많이 생겨나며, 전세계에 많은 종류가 퍼져있다. 파리의 뒷날개는 평형 감각을 느끼는 작은 기관(평균곤)으로 변형되어 있다.

<표1> 문장 (1)에 대한 상호 정보 값의 결과

x	y	I(x,y)
가부좌	앉는 법	17.2
가부좌	부처	10.7
가부좌	오른발	11.5
가부좌	왼발	11.8
가부좌	책상다리	16.2
가부좌	밥적다리	14.1
가부좌	결가부좌	15.2
가부좌	오른편	16.2
가부좌	왼편	16.2

예문 (3)과 (4)의 밑줄친 단어들은 형태소 분석과 구문 분석 과정 없이 단순히 어절 단위로 분리한 후 명사 사전과 조사 사전, 표제어 사전을 참조하여 명사 사전이나 표제어 사전에 있는 단어들로서 모호성 해소를 위한 의미적 관련 정보로 사용된다. 어의 모호성을 가지는 단어들은 22,000여개의 표제어를 갖는 학생 백과 사전에서 순수 표제어만을 뽑아 놓은 표제어 사전 참조시 판명된다. 문장(3)의 경우에서 상호 정보를 이용하여 어의 모호성을 가지는 단어 '파리'에 대하여 모호성을 해소해보자. 문장 (3)에 나타난 의미적 관련 정보들은 표(2)에 나타난 것처럼 파리^{1}보다 파리^{2}와 관련이 더 많다고 나타났다. <그림1>에서처럼 파리^{2}와 문장속의 단어들이 공통 부분을 만나지 못한다면 파리^{1}이나 파리^{2}의 의미적 관련 정보들을 한 단계 더 확장하여 나간다. 마찬가지로 문장 (4)에서 동음이의어 파리를 제외한 병원균, 해충, 여름철, 세계, 종류, 뒷날개, 평형 감각, 기관(평균곤)과 같은 단어들이 표(3)에 의해 파리를 파리^{1}으로 결정되도록 해준다.



(그림 1)단어간의 의미적 관련도

<표 2>파리^1 과 파리^2 의 예

x	y	MI	x	y	MI
파리^2	세계	0.003	파리^1	세계	0
파리^2	가장	0	파리^1	가장	0
파리^2	도시	0.439	파리^1	도시	0
파리^2	이론	0	파리^1	이론	0
파리^2	여지	0	파리^1	여지	0
파리^2	매력	0.03	파리^1	매력	0
파리^2	시인	0.03	파리^1	시인	0
파리^2	예술가	1.942	파리^1	예술가	0
파리^2	언어	0	파리^1	언어	0
파리^2	유럽	1.231	파리^1	유럽	0
파리^2	도시	0.439	파리^1	도시	0.1
파리^2	시대	0.013	파리^1	시대	0
파리^2	구경	0.102	파리^1	구경	0
파리^2	정도	0	파리^1	정도	0
파리^2	크기	0	파리^1	크기	0

<표 3> 파리^{1}의 상호 정보 값의 예

x	y	MI(x,y)
파리^1	병원균	0.401
파리^1	해충	0.251
파리^1	여름철	0.204
파리^1	세계	0.013
파리^1	종류	0.124
파리^1	뒷날개	0.310
파리^1	평형 감각	0.216
파리^1	기관	0.030
파리^1	평균곤	0.015
파리^2	병원균	0
파리^2	해충	0
파리^2	여름철	0
파리^2	세계	0
파리^2	종류	0
파리^2	뒷날개	0
파리^2	평형 감각	0
파리^2	기관	0
파리^2	평균곤	0

3. 실 험

본 논문에서는 형태소 분석이나 구문 분석 과정 없이 명사나 복합 명사 또는 표제어 사전에 나타나는 어의 모호성을 가지는 동음이의어에 대하여 상호 정보값으로 어의 모호성을 해소하는 실험을 해보았다. 빈도수 f(x,y)=1, 즉 명사-명사 쌍이 말뭉치내에 단 한번만 나타난 경우에는 잘못된 입력일 수도 있고 신빙성이 떨어지므로 제외하고 나머지 명사-명사 쌍에 대하여 의미적 관련 정보를 구하였다. 습득된 통계 정보가 실제로 모호성 해소에 사용되었을때 어느 정도 정확도를 보이는가 알아보기 위하여 100개의 샘플 문장으로 실험해보았다. 그 결과 63문장에 대해서는

모호성 해소가 올바르게 된 반면 37개 문장에 대해서는 모호성 해소가 잘못되었다. '가지'와 같이 백과 사전에 동음이의어로 등록이 안된 단어들은 백과 사전의 내용을 확장해 나간다면 해결될 수 있는 문제이고 '식물의 한 가지이다.'에서 처럼 '가지'가 식물때문에 먹을수 있는 '가지'를 뜻하게 되어 모호성 해소 결과가 잘못되었을 때에는 의미적 관련 정보를 이용한 의미론적인 방법 이외에 패턴 매칭으로 해결해야 할 것같다.

4. 결론

한국어 정보 검색 시스템에서 문서의 자동 색인시 동음이의어로 인해 발생되는 어의 모호성을 해소하기 위해 학생 백과 사전으로부터 자동적으로 습득된 의미적 관련 정보를 사용하여 구축한 키워드망으로 모호성을 해소할 것을 제안하였다. 백과 사전의 경우 표제어와 표제어 설명 단어들이 거의 대부분 의미적으로 밀접한 관계를 가지므로 어의 모호성 해소를 위한 정보 습득의 자료로는 안성 마춤이다. 기존의 정보검색시스템에서는 문장의 구조나 의미보다는 문서내의 단어에만 관심이 있었으므로 색인과 검색시에 찾고자하는 단어가 나타난 카드들만 검색해주므로 질의어내에 동음이의어가 포함되어 있을 경우 이 단어가 나타난 모든 카드들을 모두 찾아주게 되어 사용자에게 오히려 사용자에게 혼란을 야기시키게 된다. 본 논문에서 제안한 방법으로 질의어 분석시에도 동음이의어에 대한 어의 모호성을 해소한다면 사용자가 원하는 정확한 카드를 찾아줄 수 있을 것이다. 또한 질의어를 의미적으로 관련있는 다른 단어로 확장하여 폭넓은 내용의 검색도 가능하다.

참고문헌

[1] OH H.K. CHOI D.S. JUN M.S. PARK S.Y., "Design and Implementation of a Markup Editor for an Electronic Book System", Proc. of the 1994 International Confrence on Computer Processing Oriental Language", Seoul, pp. 383-388, 1994.

[2] KANG H.K., LEE C.Y., JANG H.W., PARK S.Y., "An Implementation of an automatic keyword extraction system", Proc. of the 1994 International Confrence on Computer Processing Oriental Language", Seoul, pp. 163-166, 1994.

[3] ROY RADA, JUDITH BARLOW., "Document Ranking Using an Enriched Thesaurus", Journal of Documentation, Vol. 47, No. 3, September 1991, pp. 240-253.

[4] D. Yarowsky, "Word-Sence Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora," Proc of COLING-92, pp.454~460, 1992.

[5] Karen Jensen and Jean-Louis Binot, "Disambiguating Prepositional Phrase Attachments On-line Dictionary Definitions," Computational Linguistics, Vol.13, pp. 251~260, 1987.

[6] GERARD SALTON, "Automatic Text Processing", Addison-Wesley Publishing Company, pp. 229-271,1989.

[7] 강현규, 이창렬, 박세영., "백과사전 검색 시스템의 설계 및 구현", 한국정보과학회 '93 가을 학술발표논문집, Seoul, pp. 1167-1170, 1994.