검색 엔진의 '색인 모듈'의 문제와 합성어 사전 및 구문 정보 사전의 필요성

Problems of Indexing Module in IR Systems and Lexicons of Complex Items and Syntactic Structures

남지순, 최기선

한국 과학 기술원 인공 지능 연구 센터 한글 공학 연구실

NAM Jee-Sun, CHOI Key-Sun

KAIST CAIR Language Engineering Laboratory

nam@world.kaist.ac.kr, kschoi@world.kaist.ac.kr

기존의 대부분의 정보 검색 시스템은 문서에 대한 '자동 색인 단계'를 거쳐 질의자의 요구에 적합한 문서들을 추출하도록 되어 있다. 이 과정에서 얼마나 적합한 문서를 빠짐없이 검색하였는가 하는 문제가, 검색 시스템의 효율성을 판단하는 데 가장 중요한 열쇠가 된다. 이 글에서는 '명사' 중심의 키워드 추출이 안고 있는 몇 가지 문제점들에 관해서 논의하였다. 즉, 합성어 키워드 구축의 필요성, 동사 구문 정보에 대한 필요성, 부사구 표현에 대한 기술 필요성, 그리고 발화 상황이 고려되어야 하는 점등이 검토되었고, 이에 관한 해결책으로, 어휘 정보 및 어절 정보, 나아가 구문 정보들을 담고 있는, 보다 체계적인 한국어 사전 시스템이 구축되어야 함을 강조하였다.

1. 머릿말

자동화된 정보 검색 시스템 (Information Retrieval System) 구현의 역사는 이미 1950년대를 전후하여 시작되었고 (사공철 1996), 자동 분류, 자동 색인에 관한 연구 및 시소러스의 개발 등이 특히 촛점이 되어 왔으며, 1970년대에 들어서서 온라인 서비스가 확대되면서 대규모의 데이터 베이스를 온라인으로 검색하기 위한 노력들이 점차 증가하였다. 또한 전문 데이터 베이스의 급격한 증가와 더불어 전문 검색, 멀티미디어 정보 검색 및 지능형 정보 검색 시스템에 대한 연구가 최근 들어 인터넷의 보급과 함께 더욱 활발하게 진행되고 있다.

기하급수적으로 증가하는 현대 사회의 정보의 양을 고려할 때 인간이 원하고자 하는 올바른 정보만을 찾아내어 제공할 수 있는 시스템의 필요성은 새삼스레 거듭 강조할 필요가 없다. 얼마만큼 적절한 정보를 컴퓨터가 제공하여 줄 수 있느냐 하는 문제는, 축적된 데이터의 양이 그리 많지 않을 때에는 그 문제의 심각성이 훨씬 완화되며, 이때 자동 검색 시스템은 단지 인간의 번거로운 작업을 보조적으로 도와주고 줄여주는 정도로써도

충분히 그 의미가 있다고 할 수 있다. 그러나 대형 도서관 전체에서, 또는 더 나아가 인터넷으로 연결된 전 세계 자료 베이스로부터 아주 특정한 정보를 얻고자 할 때, 자동 시스템이 얼마만큼 그 사용자의 의도를 이해하였는가, 그리고 얼마만큼 올바르게 그리고 신속하게 그 전체 데이터를 분석하여 제시하여 주는가 하는 문제는 실제로 매우 중요한 포인트가 된다.

예를 들어, 한국어 사용자가 '유령 회사'와 같은 질의어에 대한 정보를 얻고자 할 때, 검색 시스템이 자동 색인 (Automatic Indexing) 단계에서 '유령'과 '회사'라는 명사를 내포한 문서를 각각 분류해 놓고, 다시 그 두 자료집합의 교집합 (Intersection)을 추출해 내는 것으로 구성되어 있다면, 우리는 여기서, '유령'이라는 키워드를 단독으로 가진 문서는 실제 질의자가 요구하는 정보에 크게 도움이 되지 않을 것이라는 것을 쉽게 예측할 수 있다. 더구나, 만일 '유령'에 대한 문서의 양이 엄청나고, 다음과 같이

회사내에 유령이 출몰했다는 소문이 퍼졌다

'회사'라는 단어와 '유령'이라는 단어 사이의 공기 (Collocation) 정도율이 매우 높은 문장들이 문서 내에서 발견된다면, 이와 같은 유형의 자료들이 질의자의 의도와는 관계없이 검색 과정에서 출현할 것이다. 이와 같은 경우의 처리는, '유령 회사'라는 항목이 '유령', '회사'라는 단어와 별도로 하나의 색인어로서 존재할 때 비로소 만족할만한 결과를 가져올 것이다.

우리는 이 글에서 기존의 정보 검색 시스템에 보완해야 할 몇 가지 문제점들에 대해 논의하고, 그와 같은 문제점들에 대한 해결 방안으로 어떠한 점들이 더 연구되어야 하는지 살펴 보기로 한다. 효율적인 정보 검색 시스템을 구현하는 데 있어서 고려되고 개발되어야 할 분야는 무척 다양하다. 질의자와 시스템간의 인터페이스의 향상 및 그 질의문의 유형, 분석등에 대한 연구로부터 색인 및 검색 과정, 그리고 그 검색 결과물의 분류 및 추출 방법 등에 관한 연구에 이르기까지 종합적인 연구가 이루어져야 한다. 또한 이와 같은 과정을 거치는데 있어서 고려되어야 하는 메모리 및 처리 속도등에 관한 모델의 개발과 그 검증, 동시에 인간의 언어로 쓰여진 문서들의 기계 처리를 위한 필수적인 자연어 이해와 그 모델의 제시 등이 다 함께 이루어 질 때 비로소 이와 같은 시스템의 향상을 기대할 수 있을 것이다.

그러나, 정보 검색 시스템의 구현에 필요한 이와 같은 여러 영역에 있어서, '한국어 문서 분석 (Analysis of Korean Written Texts)'이 수행되는 '검색 엔진'을 단순한 하나의 모듈로 간주하기에는 실제로 다른 분야에 비해서 수십 배의 노력과 연구가 필요하다. 검색 시스템의 궁극적인 목적은 '얼마나 빠짐없이 (Recall Ratio), 그리고 얼마나 정확히 (Precision Ratio) 해당 문서들을 검색하느냐' 하는 것인데, 그것은 결국 한국어로 쓰여진 문서를 얼마나 올바르게 분석하느냐 하는 문제에 귀결된다. 한국어로 작성된 문서를 정확히 검색하고자 할 때, '명사위주의 색인어 목록'에 의존하여 문서를 재구성해놓은 시스템에서는 그 효율성의 향상에 한계가 있다. 질의문을 잘 이해하고 그 검색 결과들을 효과적으로 분류하고 추출하는 '지능형 에이전트 시스템 (Intelligent Agent System)'에 대한 연구들이 최근 활발하게 이루어지고 있는데, 정확하게 문서를 분석하는 검색 모듈이 제대로 구현되지 않는 한, 만족할 만한 성과를 기대하기 어려울 것이다.

2. 검색 엔진의 색인(Indexing) 모듈의 문제

현재 검색 시스템에서 색인에 필요한 키워드 (Key Word)의 구축은, 용언의 명사형을 포함한 '명사류'를 중심으로 이루어 진다. 가령 예를 들어, 다음과 같은 문서로부터,

(1) 그 지역 감나무밭에는 농약을 너무 사용해서 올해 오히려 감 생산량이 감소하였다

구축될 키워드 목록을 살펴보면 대체로 다음과 같다.

(2) 지역/ 감나무밭/ 농약/ 사용/ 올해/ 감/ 생산량/ 감소

이와 같은 색인 과정에서 발견되는 문제점들은 몇 가지로 분류될 수 있다. 다음에서 살펴 보자.

2.1. '합성어 (Complex Item)' 색인의 필요성

위 문서 (1)에서 나타난

(3) 감나무밭
생산량

의 경우를 보자. 이 명사들의 경우, 처음 것은 '단순어 (Simple Item)' 3개 (즉, 감/나무/밭) 가 결합하여 나타났고, 둘째 것은 '단순어' 1개 (생산)에 '접미사 (Suffix)' 1개 (량)가 결합하여 실현된 형태이다. 이들의 구조는 다음과 같이 분석될 수 있는데,

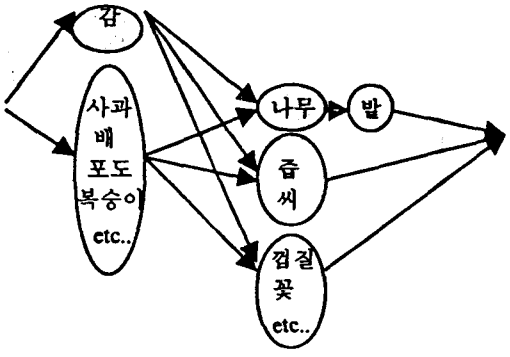
(4) *N-N-N*
N-SF

이들은 합성어가 갖는 여러 구조 유형중의 하나에 불과하다. 하나의 단순어 이상으로 이루어진 결합체들 (즉, 위의 복합어 및 파생어)을 모두 '합성어 (Complex Item)'라 부르면, 이와 같은 결합체들이 하나의 색인어로 등록되어야 할 때, 어떻게 이들을 인식할 수 있겠는가 하는 점이 문제이다.

색인어 등록을 위해서 사용되는 첫번째 단계는 우선 사전의 이용이다. 검색 문서를 '명사 사전'과 매칭을 하면서 판별되는 스트링들은 일단 '색인어' 후보로 등재될 수 있다. 이때 만일 위의 (3)에서 나타난 것과 같은 복합체들이 사전에 등재되어 있지 않다면, 이와 같은 명사들의 인식(Recognition) 을 위해 어떠한 과정을 거쳐야 하는가 ? 다음을 비교해 보자.

(5) 감나무밭, 감즙, 감씨,

(5)의 단어들은 기존의 대부분 사전들에 모두 등재되어 있을 가능성을 기대하기 어렵다. 왜냐하면, 이들은 그 생산성이 제한되어 있지 않고 따라서 다음과 같이 여러 조합의 경우 (다음 그래프는 25 개의 합성어를 보여 준다) 들을 기대할 수 있기 때문이다.



더우기 (5)와 같은 합성이들은, 다음과 같이 여백 (Typographical blank) 을 가지고 출현할 수 있는데,

포도 나무, 사과 껍질, 복숭아 꽃

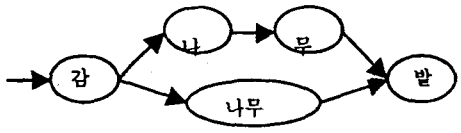
이때, 이들을 모두 사전에 등재시키기 위해서는 합성어 구성에 대한 더 세밀한 연구 결과들이 뒷받침되어야 한다. 현행 검색용 사전에는 따라서 이와 같은 유형의 합성어들이 체계적으로 수록되어 있지 않은 상태이며, 이때 이들을 색인어로 관별하기 위해서는, 우선 '단순 명사' 사전에 등재되어 있는 각 개별 단어들을 색인어로 따로 설정하여 문서상의 교집합을 구하거나, 또는 연이어 나타난 명사들의 관계를 예측하는 일정 알고리즘을 구상하여 이와 같은 유형의 복합 색인어를 구축하거나 하는 것이다. 두개 이상의 단순어가 함께 나타나서 그 의미 변화가 심하게 일어나지 않는 위와 같은 복합체들의 경우는 문제가 그렇게 심각하지 않다. 그런데, 만일 다음과 같이,

처녀 장가, 처녀 연설

'처녀'와 '장가', 또는 '처녀'와 '연설'의 의미로부터 단순히 유추가 가능한 합성어가 아닌 경우, 이와 같은 합성어가 하나의 색인어로 설정되지 않으면, 엄청난 양의 '비관련 문서 (Noise)'를 검색 결과로 제시하게 될 것이다.

그런데, 여기서 더 큰 문제는, 사전에 등재되지 않은 합성어들이 '여백' 없이 하나의 스트링으로 실현되는 경우이다. 이때 이들에 대한 색인어 처리를 위해서 다음 세 가지 유형의 프로세싱을 생각해 볼 수 있다.

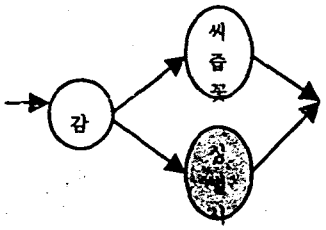
<1> 첫째는, 이러한 복합체를 이루고 있는 단순어들을 찾아내기 위해서, 모든 스트링들을 '분절 (Segmentation)' 하는 과정을 거치는 방법이다. 이때 분절이 가능한 모든 형태를 조사하기 위해서 일일이 사전과 '매칭 (Matching)'을 하는 작업이 수행된다. 그러면, 위에 제시되었던 '감나무밭'의 예를 보면, 분절의 가능성은 다음에서 보듯이 2 가지의 경로로 나타나게 된다.



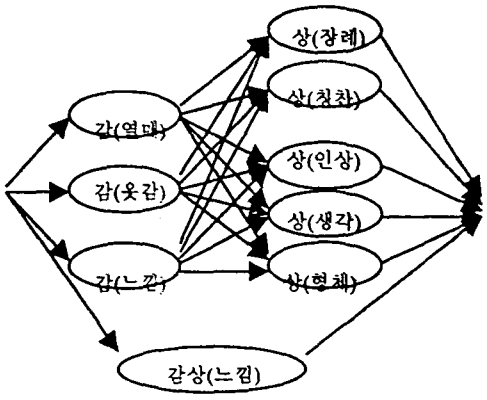
이것은 사전에 '감', '나', 나무', '무', '밭'의 5 개의 엔트리가 등재되어 있기 때문이다. 이와 같은 프로세싱을 도입할 때 문제는 매우 심각해 진다. 이것은 '감'으로 시작하는 스트링을 만날 때마다, '분절'의 가능성을 고려해야 하기 때문이다. 예를 들어, 다음과 같이, '감'뒤에 하나의 명사로 인식될 수 있는 '상'이나 '별', '시'와 같은 형태가 결합되면,

감상, 감별, 감시

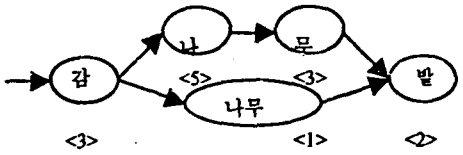
이들이 하나의 '단일어'인지, 또는 위의 (5)와 같이 '감'에 또 하나의 명사가 함께 실현된 '합성어' 형태인지에 대한 선택이 원칙상 불가능하다. '감'과 함께 결합되어 나타날 수 있는 모든 합성어의 목록이 구축되지 않는 한, 뒤에 나타난 명사 형태의 성분을 어떻게 분석할 것인가 하는 판단을 내릴 수 없다. 즉 다음에서,



둘째 셋트와 같은 분석이 불가능한 정보가 제공되어야 한다. 위에서 각 형태가, 많은 '동형어 (Homography)' 들을 갖게 되면 문제는 더욱 복잡해 져서, 다음과 같이 '감상'은 최소 3 개의 '감'과 5 개의 '상'으로부터 15 가지의 분석 중의성을 추가로 갖게 된다.



<2> 둘째는, 위와 같은 과분석의 문제에 부딪히지 않기 위해서, 사전에 존재하는 모든 형태들은 그대로 매칭시키고, 사전에 등재되지 않은 형태들만을 '분절'의 과정을 거치는 방법이다. 즉, 예를 들어, 위에서 '감상', '감별', '감시'와 같은 형태들이 사전에 등재되어 있으면, 이들은 그대로 분절없이 명사로 인식되고, '감나무밭'과 같이 사전에서 누락되어 있는 형태는 분절의 과정을 거치는 방법이다. 그러면 이때, '사전'에 등재된 '감상', '감별'등은 과분석의 부담을 덜게 된다. 그러나 '감나무밭'과 같은 합성어들의 상당수가 기존 사전에서 누락되어 있는 현단계에서 이와 같은 스트링들에 대한 분석 중의성은 여전히 해결되지 않은 채 남게 되어, 위의 '감나무밭'의 경우, 중의성은 다음에서처럼 96 (= (3x5x3x2) + (3x1x2))가지에 이르고 있다.



<3> 따라서 세번째로, 위의 합성어 '감나무밭' 과같이 사전 매칭에서 실패한 경우에도 분절 과정을 거치지 않고 '미등록어'로 직접 처리를 하는 방법을 생각해 볼 수 있다. 가령, '감나무밭'과 같이 사전에 누락되어 있는 경우, 이것을 일단 하나의 '명사'로 추정하여 색인하는 방법이다. 그런데, 미등록어 추정을 통해 명사로 인식되어야 하는 하나의 합성어는 대부분 '후치사'와 결합한 형태로 실현된다.

그러므로, 후치사를 인식하여 분리해 낼 수 있는 프로세싱을 거쳐야 한다. 미등록어 후보가 후치사와 나타났을 때, 후치사를 떼어내기 위해서, 뒤에서부터 후치사를 판별해 가는 방법을 가정할 수 있는데, 이때, 다음과 같이 후치사와 동일 형태를 취하는 부분이 미등록어의 일부를 이루게 되면 전혀 엉뚱한 색인어가 만들어 진다. 예를 들어,

뽕나무누에 다섯 마리가 ...

와 같은 문서에 나타난 '뽕나무누에'가 사전에 등재되지 않은 경우, 이것을 분절하는 과정을 거치지 않고 미등록어 추정 알고리즘으로 찾게 되면, '에'를 후치사로 분석하고 '뽕나무누'를 미등록어로 처리하는 오분석을 제거하기가 어렵다. 외국어로 된 고유명사들 중에는 이와 같은 형태적 중의성을 유발시킬 음절들이 상당수 들어 있어 미등록어 처리 방식은 모든 어절들을 분절하는 과정에서 수반되는 지나친 과분석의 수를 줄일 수는 있으나 원칙적인 해결 방안을 제시해 주기는 어렵다.

실제로 위의 <2>, <3> 유형의 프로세싱에서 사전에 등재되어 있어 매칭에 성공한 명사들의 경우에도, '분절 가능성'의 확인 절차를 완전히 배제시킬 수는 없다. 왜냐하면 가령,

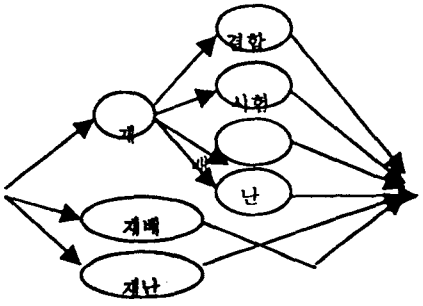
(6a) 놀이 기구는 모두 큰 방에 있다
(6b) 놀이 지고 있다

에서, (6a)의 '놀이'는 명사 사전에서 매칭되어 분절없이 그대로 색인어로 인식될 때 올바른 분석 결과를 제시해 주나, 분절 과정을 거쳐 'N-후치사 <Post>'라는 분석 결과를 가져야 하는 (6b)의 '놀이' 의 경우, '놀'이라는 색인어로 추출되어야 함에도 불구하고, (6a)와 마찬가지로 '놀이'라는 명사와 매칭되어서 전혀 다른 색인어 목록에 입력될 것이다. 이것은, 사전에 수록된 명사와 매칭될 수 있는 모든 형태들은, 분절없이 그대로 인식한 결과만을 보인다는 입장에서 안게 되는 문제점이다.

결국 한국어 문서의 경우, 색인 과정에서 명사 키워드들을 추출하기 위해서는 명사구 어절들에 대한 분절 (Segmentation) 가능성을 반드시 고려해야 하므로, 이 과정에서 생성되는 색인어 후보에 대한 오분석 내지는 과분석의 부담이 매우 커지게 된다. 합성어 사전은 이경우, 상당수의 과분석의 가능성을 제어해 주는데, 예를 들어,

재결합, 재시험, 재배, 재난

과같은 형태들이 나타났을 때, 접두사 '재'가 결합할 수 있는 모든 파생어들의 사전이 구축되어 있게 되면, 다음에서 회색 그림으로 표시된 분석 결과만이 출력되어 색인어 구축의 효율성을 향상시킬 것이다.



모든 접사들을 따로 분리하여 '접사 사전'을 구성하고, 단순 명사 사전을 명시적 기준에 의거하여 하나의 모듈 형태로 구축하면, '파생어 사전'은 단순어와 접사들 사이의 '어휘 형성'의 가능성을 '언어학적 조합 (Linguistic Combination)' 의 방식으로 검토하여 구축된다. '복합어 사전'은 단순어들 사이의 결합 가능성 및 그 유형 등이 체계적으로 연구되어야 하는 것으로, 띄어쓰기가 부자연스러운 '단음절 명사' 및 '뜻이 전성되어 다른 의미로 바뀌는 복합 구성 (관용어)'들을 우선 중심으로 이루어 져야 한다.

합성어 사전은 단시일내에 완성되지 않는다. 그러나 단순어와 접사들이 분리되어 하위 사전들을 이루고 있을 때, 파생어 사전의 구현이 우선 가능해지며, 궁극적으로, 고유 명사를 제외한 모든 색인어는 '사전' 검색을 통해서 자동으로 구성될 수 있을 것이다. 앞서 본 것과 같이 '유령 회사'라는 복합어가 사전에 수록될 것이므로, 이 경우 색인어 목록에서 '유령 회사'는 '유령' 또는 '회사'와는 별도의 색인어로서 존재하게 될 것이다.

2.2. 동사 구문(Verbal Construction) 기술의 필요성

2.2.1. 명사 중심 색인의 한계

위의 (1)에서 제시된 예를 다시 보자.

(1) 그 지역 감나무밭에는 농약을 너무 사용해서 올해 오히려 감 생산량이 감소하였다

여기서 추출된 키워드 목록 (2)에서 '사용', '감소'와 같은 명사들이 관찰되는데, 이들은 이 문서에서 '하다'와 결합해서 각각 '사용하다'와 '감소하다'의 동사 형태로 사용되고 있다. 그러면, 다음을 비교해 보자.

(7a) 그 지역 감나무밭에는 농약을 너무 사용하였다
(7b) 그 지역 감나무밭에는 농약을 너무 썼다

(8a) 올해 오히려 감 생산량이 감소하였다
(8b) 올해 오히려 감 생산량이 줄었다

많은 명사에 결합되어서 동사구를 유도하는 '하다'를 때로는 하나의 동사로 간주하기도 하고 (예를 들어, '사용을 자주 하였다'의 경우에서), 때로는 하나의 접사로 간주하기도 하는데 (예를 들어, 명사에 들러붙어 나타나는 '스럽다' 등과 같이 용언화 접미사로 간주하는 경우), 이와 같이 '하다'와 함께 나타난 명사는 하나의 색인어로 등재된다. 위의 (7a)-(7b)의 쌍과 (8a)-(8b)의 쌍을 비교해 보면, '사용'과 '감소'라는 키워드를 중심으로 (7a)와 (8a)의 문서는 검색이 이루어 질 것인데, 반면, (7b)와 (8b)의 문서는, 각각 (7a), (8a)의 문서들과 동의 관계를 이루고 있음에도 불구하고, 검색 과정에서 누락되어 버릴 것이다.

위와 같이 서술 명사 (Predicative Noun)가 '하다'와 같은 동사 유형 (Light Verb)과 결합하여 동사구를 이룬 경우에 한해서 그 명사가 색인어로 추출되는 방법에 의존할 때, '동의 관계'에 있는 단순 동사들에 관한 정보가 누락되는 한계를 넘어서기 어렵다.

2.2.2. 문맥을 떠난 의미 관계 연구의 한계

하나의 동사구를 이루는 '명사'가 키워드로 설정될 수 있는 경우, (7b), (8b)에서같이 '명사' 없이 단순 동사의 형태로 실현된 '동의 관계'의 서술어들이 검색 과정에서 모두 누락된다는 것은 몹시 심각한 문제를 가져온다. 이와 같은 문제의 해결을 위해서 '서술 성분'들 사이의 '의미 관계'를 일종의 '단어망 (Word Net)' 또는 '시소러스 (Thesaurus)'로 구축하려는 시도를 가정해 볼 수 있다. 그러나 '문맥 (Context)'을 벗어난 의미 관계의 연구는 우선 그 결과가 객관적 성격을 띄기 어렵고, 또한 구체적인 문맥 및 구조에 따라 그 관계가 매우 달라질 수 있다. 가령, 위의 (7a)-(7b)에서와 같은 문맥에서는 '사용하다'와 '쓰다'가 '뿌리다', '살포하다' 등과 동의어로 쓰일 수 있으나, 다음과 같은 문맥에서 위의 서술어들은 더이상 동의 관계를 가지고 있지 않다.

그는 늘 바른 말을 (사용한다 + 쓴다 + *뿌린다 + *살포한다)

또한 '쓰다'라는 서술어가 갖는 다음과 같은 여러 의미들이,

가면을 쓰다 / 글씨를 쓰다 / 먼지를 쓰다 / 힘을 쓰다 / 맛이 쓰다

모두 구별되지 않는 한, 단순히 '사용하다'와 '쓰다'의 의미관계를 설명하는 것은 그렇게 큰 유용성을 가지기 어렵다.

2.2.3. 구조적 동의성에 대한 분석의 필요성

다음 문장들을 비교해 보자.

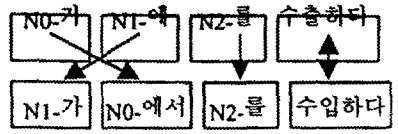
(9a) 한국은 미국에 자동차를 수출하였다
(9b) 미국은 한국에서 자동차를 수입하였다

위의 두 문장은 각각 '수출'과 '수입'을 키워드로 하여 검색될 수 있는 문서들이다. 이 두 문서는 동일한 정보를 내포하고 있는데, 이때, 단순히 '수출'과 '수입' 사이에 존재하고 있는 어떤 의미적인 관련성만을 보이는 것으로는 위와 같은 문서들의 동의성을 찾아낼 수 없다. 위의 문서 (9a)가, 반대 개념의 '수입'을 키워드로 가질 수 있는 (9b)와는 동의 관계에 있으나, 다음과 같이 동일 키워드 '수출'을 갖는 (10a)의 문장과는 같은 의미의 정보를 갖고 있지 않다는 것은,

(10a) =/= 미국은 한국에 자동차를 수출하였다

문장 구조에 대한 이해 없이는 분석이 불가능하다. 즉, 위의 (9a)-(9b)의 쌍은 다음과 같이 형식화 할 수 있는 구조 관계를 가지고 있다.

(11)



따라서, 다음과 같은 질의문이 입력된 경우,

한국이 미국에 수출한 품목에 대하여 알고 싶습니다

‹수출›과 ‹수입›이라는 단어사이의 의미 관계에 대한 정보가 제공되면, 관련 문서의 누락을 상당수 방지할 수 있으나, 여전히 (10a)와 같은 비관련 문서의 출현을 제어하기 어렵다. 따라서 ‹수출하다›를 술어로 취하는 문장의 그 논항 (예를 들어, 위의 N0, N1, N2 들) 구조와 ‹수입하다›를 술어로 취하는 문장의 그 논항 구조에 대한 관련 정보가 제공되어야 한다. 마찬가지로 ‹팔다›, ‹사다›, ‹매매하다›, 교류하다›, ‹교역하다›등의 술어에 대한 연구도 뒷받침되어야 하며, 이와 같은 연구는 단어 차원이 아닌 ‹문장› 차원에서 기술되어야 함을 전제로 한다.

(11)에서 보인 바와 같이, 주어진 하나의 어휘 성분 (여기서는 동사)에 대한 그 문장 구조의 유형을 형식화하여 나타내고, 그 문장 유형들 사이에 관찰되는 이와 같은 ‹동의 관계 (Synonymy)› 들을 체계적으로 기술하는 것을 목적으로 하는 것이 바로 ‹어휘 문법 (Lexicon-Grammar)›이다. 문장의 그 구조를 결정짓는 가장 중심되는 요소는 ‹술어 (Predicate)›이며, 각 술어를 중심으로 논항들 과의 관계를 기술할 수 있다. 예를 들어, 위의 (9a)에서 나타난 술어 ‹수출하다›는 원칙적으로 3 개의 논항을 필요로 한다 (행위가 이루어지는 두 주체와 그 행위의 대상물). 한국어의 기본 문장 구조에서 술어 성분으로 사용될 수 있는 품사로는 ‹동사›, ‹형용사›, ‹서술 명사›가 있다. 이와 같은 구문 정보의 구축을 위해서 ‹한국어 어휘 문법›이 계속 진행되어야 한다.

2.3. 부사 정보의 기술을 위한 ‹부분 문법›의 구축

2.3.1. 부정 표현의 부사 사용

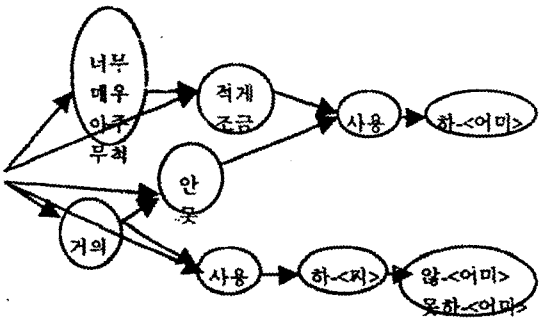
위의 (1)에서 제시된 예를 다시 들어 보면,

(1) 그 지역 감나무밭에는 농약을 너무 사용해서 올해 오히려 감 생산량이 감소하였다

‹너무›와 같은 ‹부사 (Adverb)›는 색인어 구축 과정에서 고려되지 않는다. 그런데, 다음 문서에서,

(12) 그 지역 감나무밭에는 농약을 너무 적게 사용해서 올해 오히려 감 생산량이 감소하였다

부사 ‹적게›의 출현은 문서 (1)과는 전혀 다른 의미를 가져온다. 그러므로, 만일 지나친 ‹농약 사용›이 문제가 된 지역 내지는 농장에 관한 정보를 얻고자 할 때, (12)는 비관련 문서가 된다. 이와 같이 문서의 의미를 바꾸어 놓을 수 있는 부사류에 대한 구문적인 기술이 이루어 져야 하는데, 그 대표적인 경우가, ‹부정형 (Negative)› 부사들이 사용된 문장 유형이다. 가령, 다음 그래프는 모두, (1)에서 나타난 ‹농약 사용›에 관한 정보와는 상반된 의미를 보이는 부사구 구조의 몇 가지 예를 보인다.



부정의 의미를 가진 부사들의 출현을 구조화하여 나타내기는 쉽지 않다. 예를 들어, 다음을 비교할 때,

(13) 한국과 중국 사이의 관계가 좋아지고 있다

(14a) 한국과 중국 사이의 관계가 좋아지지 않고 있다
(14b) 한국과 중국 사이의 관계가 악화되고 있다

(14a)에서 (13)에 대한 ‹부정›의 의미는 부사화 접미사 ‹지›에 ‹않다›라는 술어 성분이 동반되어 형성되었고, (14b)에서는 (13)의 동사에 어휘적으로 상반되는 동사가 실현되어 나타났다. 또한,

한국과 중국 사이의 외교 관계가
좋아지지 않고 있는 건 아니지만, 그래도...

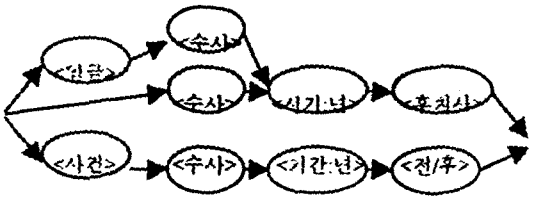
위와 같은 문서에서처럼 부정의 표현이 두 번 나타날 때, 그 의미가 '나빠지다', '악화하다'의 의미와는 상반된다는 사실을 올바르게 분석할 수 있기 위해서는 부정 표현의 논리 구조가 '어휘적으로' 낱낱이 기술되어야 한다는 사실을 보여준다.

2.3.2. 시간 부사구의 부분 문법

검색 시스템에서 '시간 표현'이 가지는 정보의 중요성은 때로 가장 먼저 고려되어야 할 사항이다. 질의자가 찾고자 하는 어떤 정보의 내용 (즉, 하나의 사건이나 상황등) 이 포함된 문서들이 검색되었을 때, 문서의 양이 증가하면 할수록, 얼마나 올바르게 '시간에 관련된 정보'를 추출하였는가 하는 문제는 검색 결과의 양과 질을 높이는 데 매우 중요한 관건이 된다. 가령 다음에서,

고종 34 년에 국호가 대한으로 바뀌었다
1897 년에 국호가 대한으로 바뀌었다
한일합방 13년전 국호가 대한으로 바뀌었다

여기 나타난 시간 부사구들은 모두 동일한 해 (1897년) 를 가르킨다. 1897년과 관련된 문서를 찾고자 하는 질의자에게 위의 문서들은 모두 검색 결과의 후보로 나타날 수 있어야 한다. 이때, 이와 같은 결과를 기대하기 위해서는 우선 '시간'을 나타내는 언어적 표현들이 올바르게 인식될 수 있어야 하며, 그 다음 그 언어적 표현이 갖고 있는 의미 지식 베이스가 함께 제공되어야 한다. 예를 들어, 위에서 나타난 언어적 구조들을 분석해 보면,



와 같은데, <임금> 이름과 함께 나타난 첫번째 <수사>의 경우 최대값 100을 넘기는 경우를 찾기 어렵고 (재위 기간동안만 가능하므로), 둘째 <수사>의 경우, '날짜'를 나타내는 단위 명사들 (년, 월, 일, 시 등) 과 공기할 수 있는 모든 수사들이 나타날 수 있다. 반면, 세째 <수사>의 경우, '기간'을 나타내는 단위 명사들 (년, 개월, 일, 시간 등) 과 함께 나타날 수 있는 수사들이 실현될 수 있으므로, 예를 들어 '20 개월전', '50일후'등과 같은 표현이 가능하다. 그것은 위의 '날짜'를 나타내는 표현에서 '1800년 (7+*20)월 (20+ *50)일'과 같이 12 개월, 31일을 넘어설 수 없는 경우와 대조적이다.

시간과 관련된 정보에는 위와 같이 정확한 한 시점을 나타내는 경우도 있지만, 찾고자 하는 시점을 포함하고 있는 시간 부사구의 표현으로 실현되는 경우도 많다. 가령, 아래의 문서들은 모두 위에서 얻고자 하는 1897 년과 관련된 정보를 가지고 있다.

1890 년대에는, ...
1880 년에서 1900 년 사이에, ...
19 세기에는, ...

따라서, 모든 시간적 정보를 가질 수 있는 언어적 표현 형식이 빠짐없이 연구되어야 하며, 이와 같은 표현들에 대한 형식적 기술이 이루어지면, 의미적으로 동일한 시점을 가르키는지의 여부가 언어외적 지식 베이스를 토대로 결정되어야 한다. 시간 부사구 유형에 대한 언어학적 연구는 '부분 문법 (Local Grammar)'의 형식으로 이루어질 수 있다. 여기서 말하는 부분 문법이란, 일정 어휘 요소가 실현될 수 있는 특정 구문을 규칙이나 변수의 설정없이 상수식의 개념으로 기술하는 방법이다.

'수사'가 나타난 시간 부사구를 인식하고 분석하는데 있어 어려움을 가져오는 요인중의 하나가, '수사'가 숫자가 아닌 자연어로 표기될 때 발생하는 현상이다. 가령, 다음과 같은 시간 표현이,

1777년 7월 7일 7시 7분 7초

자연어로 표기되면 다음과 같다.

천 칠백 칠십 칠년 칠월 칠일 일곱시 칠분 칠초
(천 *일곱백 *일곱십 *일곱년 *일곱월 *일곱일 *칠시 *일곱분 *일곱초)

위에서 보면 알 수 있듯이, '7'이라는 숫자가 모두 8번 나왔는데, 이때 '시각'을 나타내는 '7'의 경우만을 제외하고는 모두 '일곱' 대신 '칠'이라는 표현을 쓴다. 즉, 한국어 수사 표현에 있어서 '한글 체계 (하나, 둘, 셋, ...)'와 '한자 체계 (일, 이, 삼, ...)'가 어느 상황에서나 혼용되어 사용되기 어렵고, 오히려 거의 대부분의 경우, 함께 나타나는 '단위 명사'등, '문맥 (Context)'에 의해 결정되기 때문이다. 다음의 예에서 보듯이,

사과 (일곱 + *칠) 개
맥주 (일곱 + *칠) 병
설탕 (*일곱 + 칠) 킬로그램

셀 수 있는 명사뒤에 <수사> 표현이 오고 '단위 명사'가 뒤따르는,

이와 같은 구문에 대한 기술도 부분 문법'의 형태로 구현되어야 한다. 아라비아 숫자의 형태로 실현된 '수사 표현'을 자연어 표현으로 적합하게 대응시킬 수 있기 위해서는 함께 나타나는 '단위 명사'류 전체에 대한 어휘적인 연구가 체계적으로 이루어져야 하며, 이와 같이 부분 문법이 구축되면 검색 문서에서 '시간에 관련된 정보'들을 올바르게 추출하는 데 있어 큰 효율성을 가져올 것이다.

2.4. 문장외적 정보를 위한 담화론적 연구

2.4.1. 시간 부사구와 발화 시점

다시 위의 문서 (1)로 되돌아 가보자. 여기서, '올해'라는 색인어가 추출되었는데, 만일 이 문서가 작성된 시점이 1997년이었다면, 이것은 다음 문서에서 나타난 '1997년에는'의 시간 부사구와 동일한 '시간 정보'를 가진다.

(15) 그 지역 감나무밭에는 농약을 너무 사용해서
1997년에는 오히려 감 생산량이 감소하였다

즉, 시간을 나타내는 부사구중에는 위의 '올해'와 같이 '발화 시점'을 알아야 그 시간 정보를 구할 수 있는 경우들이 있다. 이러한 부사들은 시간에 대한 절대적인 지시성을 갖고 있지 못하며, 반드시 문서가 작성된, 또는 정보가 발화된 시점에 관련된 '상대적인 지시성'만을 갖고 있기 때문이다. 다음에서,

(16a) 작년에 홍콩이 중국에 반환되었다
(16b) 1년전에 홍콩이 중국에 반환되었다

위 두 문서는 현재 시점에서 작성된 문서들일 경우, 동일한 시간 정보를 가지고 있다. (16a)는 발화 상황에 관련된 '상대적' 표현이고 (16b)는 시간 전후를 비교하는 '기준 시점 (여기서는 '지금으로 부터'라는)' 이 생략되어 나타난 표현이다. 실제로, 정확히 명시되지 않은 시간 표현들이 검색 문서상에서 빈번히 나타나는 경우를 볼 수 있는데, 다음과 같은 신문 기사로부터,

이총재는 31일 총재직을 사퇴하면서, ...
신한국당은 다음 주 초까지 법률전문가와 ...
지난해 12월 최 지사의 탈당은 일부 ...
김 총재는 이날 <춘천 문화 방송> 주최...
<한겨레 신문 1997년 8월 2일자>

해당 날짜에 대한 올바른 정보를 추출할 수 있기 위해서는, 앞의 2.3.2.에서 논의한 바와 같이, '시간 정보'를 가지고 있는, 가능한 모든 언어적 표현 형식에 대한 연구가 우선적으로 이루어져야 한다. 그때 비로소 문서 작성 시점과 관련하여 올바른 '날짜'를 계산하여 그 결과로 제시할 수 있을 것이다.

2.4.2. 대명사와 지시 대상물 (Referent)

'발화 상황'에 대한 정보없이 그 정확한 정보를 검색하기 어려운 경우가 또 있는데, 가령, 다음과 같은 한 정치가의 발언을 보자.

비록 도지사가 탈당했지만 강원도민은 우리당을..

'우리당'이 어느 당을 지시하는지, 따라서 어느 당에 관련된 문서로 분류되어야 하는지의 여부는 발화자가 누구인지에 대한 정보가 제공되어야 한다 (위는 한겨레 신문에 개제된 '자민련' 총재의 발언). 검색 시스템은, 주어진 문서에 대한 단순한 통사적 구문 분석의 단계만으로는, 이와 같이 구하고자 하는 올바른 문서를 제공하여 줄 수 없다.

'우리', '나', '당신'등과 같이, '발화 상황'을 고려해야만 그 정확한 역할을 구분해낼 수 있는

언어 표현 형식들이 어떠한 것이 있으며, 또 어떠한 유형으로 분류될 수 있는지에 대한 연구가 이루어져야 하며, 그때 발화자 및 그 대화자에 대한 정보가 어떠한 형태로 문서에 삽입되어 나타날 수 있는지 고려되어야 한다. 가령 위의 문서에서, 발화자는 다음과 같은 형태로 실현되었다.

<N0-는 Loc-에서 "----"-고 말하다>

즉, 다음과 같다.

> 김총재는 이날 (...) 텔레비젼 토론회에서 "비록 도지사가 탈당했지만 강원도민은 우리당을 (...) 주었다" 고 말했다.

지시 대명사, 또는 관형사의 그 지시물을 파악하기 위해서, 위와 같이 '발화 상황'을 고려해야 하는 경우가 있다면, 다음과 같이 바로 '전 문장'에서 나타난 명사들을 검토해야만 알 수 있는 경우가 있다. 앞서 든 (1)의 예를 다시 고려해 보면,

---

(1) 그 지역 감나무밭에는 농약을 너무 사용해서 올해 오히려 감 생산량이 감소하였다

위에서 '그 지역'이 어디를 나타내는지는 이 하나의 문서만으로는 알아낼 수 없다. 이것은 그 앞의 문서에 나타난 '지명' 이름들을 인식하고, 그 다음 이와 같이 설정된 후보들로부터 올바른 매칭을 시킬 수 있어야 가능하다. 즉,

> 천안 지역의 올해 농산물 수확 현황을 살펴보면, 입장읍에서 포도의 생산량은 중부 지역권 내에서는 가장 큰 폭으로 증가하였다. 그러나, <그 지역 감나무밭에는 농약을 너무 사용해서 올해 오히려 감 생산량이 감소하였다>

에서와 같이, '그 지역'이 나타난 문장에 선행하는 바로 앞 문장으로부터 얻어진, '지명'을 가르키는 명사들은 모두 3 가지나 된다. '그 지역'의 올바른 지시물을 찾아내어 관련 문서에 연결시킬 수 있기 위해서는,

1. '지명'을 나타내는 고유 명사들을 인식하기 위한 어휘적 정보,
2. 어느 유형의 어절이 그 선행사가 될 수 있는지를 결정하기 위한 구문적 정보,
3. 어느 것이 논리적으로 더 연계성을 갖는지를 결정하기 위한 의미적 정보

등이 제공되어야 한다. 위에서 '그 지역'을 3 가지중의 하나의 명사와 올바르게 연결시킬 수 있기 위해서는 위와 같은 유형들의 정보를 획득하기 위한 체계적인 연구 방법이 더 제시되어야 한다. 현재 위에서 예시된 문서로부터 '그 지역'의 올바른 선행사를 찾아낼 수 있는 명시적인 방법은 아직 기대하기 어렵다. 지시 관형사 '그'뒤에 실현된 명사 '지역'과 동일 형태를 내포하고 있는 '천안 지역의'가 유력한 후보로 설정될 수 있으나, '중부 지역권'의 '지역'의 경우와 구별되기 어렵고, 또 '천안 지역의'가 단어 거리상 '그 지역'과는 가장 멀리 떨어져 있다는 점들이 해결하기 어려운 문제로 남는다.

발화 상황을 분석하기 위한 담화론적 모델은 자동 검색 시스템에서 이용되기에는 아직도 많은 연구를 필요로 한다.

3. 맺음말

대부분의 정보 검색 시스템은, 현재 검색 엔진에서 '명사 (Noun)' 중심의 '색인어 구축 (Indexing)' 단계를 거치고 있는데, 이때 발생되는 몇가지 문제점들을 살펴 보면 지금까지 논의한 바와 같다. 첫째, '합성어' 사전이 없이 명사 키워드를 추출할 때의 문제점, 둘째, 동의 관계의 동사를 고려하지 않고 술어를 이루는 명사만을 추출할 때의 문제점, 세째, '부사 정보의 올바른 분석이 이루어지지 않을 때의 문제점, 그리고 네째, 발화 상황 등을 고려해야 관련 문서를 구축할 수 있을 때의 문제점 등으로 나누어 검토하였다. 1, 2, 3과 같은 경우, 이러한 문제들을 해결하기 위하여 다음과 같은 방법론이 제시되었다. 첫째, 체계적인 '합성어 사전'을 구축하는 것, 둘째, 술어 체계를 중심으로 한 '한국어 어휘 문법 (Lexicon-Grammar)' 을 구축하는 것, 그리고 세째, 시간 부사구와 같은 구문에 대한 '부분 문법 (Local Grammar)'을 구현하는 것 등이다.

효율적인 정보 검색 시스템을 구현하기 위해 축적해야 할, 한국어에 대한 '형태적', '통사적' 정보의 양은 엄청나며, 실제로 다른 어느 모듈에 대한 연구보다도 선행되어야 하는 부분이다.

많은 연구들이 새로운 인터페이스 연구나 지능형 모델연구, 또는 새로운 알고리즘의 구현 등에 주력하고 있는 점은, 바로 이와 같은 어려움과 무관하지 않다.

자연어는 기계어와는 달라 어떠한 복잡한 수식이나 규칙만으로 그 원리를 가정하고 규명할 수 없다. 인간이 언어를 사용하고, 이해하고, 그리고 추론할 수 있는 '언어 능력(Competence)'이 어떠한 메카니즘에 의해서 이루어 지는지 그 실체가 밝혀지지 않는 한, 인간 언어를 100 % 이해하고 분석할 수 있는 시스템의 구현은 현재로는 분명히 불가능하다. 중요한 것은 자동 시스템의 효율성을 계속적으로 증가시킬 수 있기 위해서는 그 접근 방식의 모델이 초기 단계부터 올바르게 제시되어야 한다는 점이며, 그것이, 신속한 실용화 및 상용화를 위한 현실적 부담으로부터 어느 정도 자유로와 질 수 있을 때, 근본적인 향상을 위한 보다 진지하고 깊이있는 연구들이 진행될 수 있을 것이다.

참고 문헌

강현규, 장호욱, 전미선, 박세영, 1996, 인터넷 기반 멀티미디어 정보 검색 시스템 : 옥서 '95의 색인 및 검색, 제 8회 한글 및 한국어 정보 처리 학술 대회.

남영준, 1996, 코퍼스를 이용한 정보 검색용 전자 사전 구축에 관한 연구, 제 8회 한글 및 한국어 정보 처리 학술 대회.

김도완, 박재득, 박동인, 1996, 자연 언어 대화 Interface를 이용한 정보 검색 (WWW)에 있어서 사용자 모델 에이젼트, 제 8회 한글 및 한국어 정보 처리 학술 대회.

박영찬, 1997, 정보 검색을 위한 단어 지식의 통계적 구축, 박사 학위 논문, 한국 과학 기술원 전산학과.

박혁로, 1997, 한글 문헌을 위한 확률적 자동 색인 모델 연구, 박사 학위 논문, 한국 과학 기술원 전산학과.

사공철, 서경주, 1996, 정보 검색 발전사, 정보 관리학 회지 제 13권 2호.

이준호, 안정수, 박현주, 김명호, 1996, 한글 문서의 효과적인 검색을 위한 *n-Gram* 기반의 색인 방법, 정보 관리학 회지 제 13권 1호.

정영미, 1996, 국내 문자 정보 데이터 베이스의 색인에 관한 연구, 정보 관리학 회지 제 13권 1호.

최기선, 1991, 구문 및 의미 분석을 통한 한국어 자동 색인, 정보 관리 학회지 제 8권 2호.

Courtois, Blandine, 1987, Dictionnaire électronique du LADL, N-17, University Paris 7.

Frakes, William B.; Ricardo Baeza-Yates 1992, *Information Retrieval : Data Structures and Algorithms*, Prentice Hall, Englewood Cliffs, New Jersey.

Gross, Maurice, 1987, The use of finite automata in the lexical representation of natural language, *Lecture Notes in Computer Science* 377, Springer-Verlag.

Lancaster, F. W., 1979, *Information Retrieval Systems*, New York : John wiley & Sons.

Mohri, Mehryar, 1994, Application of Local Grammars Automata : an Efficient Algorithm, RT-IGM 94-16, University of Marne-la-Vallée.

Nam, Jee-Sun, 1995, *Constitution d'un lexique électronique des noms simples en coréen*, papers in LGC-1995, UQAM, Canada.

Nam, Jee-Sun, 1995, *Systèmes de numéraux et quelques grammaires locales en coréen*, LINX N° 35, Actes du Colloque RELEX : Lexique, syntaxe et analyse automatique des textes Université Paris X Nanterre.

Nam, Jee-Sun, 1996, *Construction of Korean Electronic Lexical System DECO*, Papers in Computational Lexicography Complex '96, ed. by F. Kiefer et al.: Linguistics Institute, Hungarian Academy of Sciences.

Nam, Jee-Sun, 1996, *Classification syntaxique des constructions adjectivales en coréen*, Amsterdam-Philadelphia : John Benjamins Publishing Company.

Salton, G., 1988, *Automatic Text Processing*, Addison Wesley Publishing Company.

Silberztein, Max, 1993, Dictionnaires électroniques et analyse automatique de textes, Le système INTEX, Paris : Masson.