단어들을 위한 새로운 메트릭 공간 : 코퍼스그람

이호석, 김영택
뉴미디어학과 공과대학 호서대학교, 컴퓨터신기술연구소 서울대학교
hslee@office.hoseo.ac.kr

A New Metric Space for Words : Corpusgram

Ho Suk Lee, Yung Taek Kim
New Media Dept. Hoseo University, Computer Technology Research Center SNU

요 약

본 논문은 코퍼스로부터 추출된 단어들을 빈도수에 따라서 적절하게 표시하고 거리를 계산할 수 있는 새로운 메트릭 공간(metric space)에 대하여 논의한다. 일반적인 Cartesian 좌표 평면은 단어와 빈도수를 표시하는데 불편한 점이 있다고 할 수 있다. 본 논문에서는 빈도수에 기반 한 새로운 좌표 평면과 정보 이론에 기반 한 새로운 거리 계산 방법을 제시하여, 코퍼스 기반 언어 처리에 필요한 계산을 더욱 적합하게 할 수 있도록 하였다.

1. 서 론

코퍼스 언어학 연구는 1960년대부터 시작되었다. 코퍼스 언어학은 기존의 이론적인 언어학이 아니고 코퍼스에 담겨져 있는 실제 문장을 언어학 연구의 대상으로 한 "경험적인" 언어학 연구의 분야이다. 참고 문헌 [1]에 있는 코퍼스에 대한 정의를 인용하면 "it can potentially contain any text type, including not only prose, newspaper, as well as poetry, drama, etc., but also word lists, dictionaries, etc." 이다. 이 정의를 보면 코퍼스는 매우 다양한 원천으로부터 텍스를 모아서 담고 있는 텍스트 자료 저장소라는 것을 알 수 있다. 그러나 일반적으로 코퍼스는 목적에 따라서 단일한 분야나 종류의 텍스트만을 모아서 구축하는 경우가 많다. 최초의 코퍼스에는 W. Nelson Francis와 Henry Kucera에 의하여 1963-1964년 사이에 Brown 대학에서 구축된 Brown 코퍼스가 있다[1]. Brown 코퍼스는 약 100만개의 단어를 포함하고 있으며 TAGGIT 이라는 프로그램을 사용하여 코퍼스의 단어들에 태그를 부여하여 단어의 사용에 대한 빈도수 조사를 실시하였다[2]. 가장 크기가 큰 코퍼스에는 영국에서 구축된 BNC(British National Corpus)가 있다. BNC에는 1억 개의 단어가 수록되어 있다고 한다. 미국에서도 이와 비슷한 ANC(American National Corpus)가 진행되고 있다. London-Lund 코퍼스는 많은 양의 발성된 영어 문장을 담고 있다. 1980년대에 들어서 코퍼스들은 자연언어처리 기술과 연계하여 크게 발전하기 시작하였다. 자연언어처리 기술과 컴퓨터 소프트웨어 기술에 의하여 코퍼스 편집기, 코퍼스 분석과 처리 도구 등이 개발되었으며, 그래픽 사용자 인터페이스 등도 개발되었다[1]. 또한 코퍼스에 대한 통계적이고 계량적인 연구가 많이 진행되었다[2]. 태깅 결과로 구축되는 코퍼스를 트리뱅크(treebank)라고 하였으며, 대표적인 트리뱅크에는 펜실바니아 대학의 트리뱅크와 LOB(Lancaster-Oslo/Bergen) 트리뱅크가 있다. 그리고 1990년대에는 ICE(International Corpus of English) 연구가 진행되어 전 세계 여러 지역에서 사용되는 영어를 수집하였다. ICE 연구에 참여한 국가는 영국, 미국, 호주, 캐나다, 뉴질랜드, 인도, 필리핀, 자마이카, 동아프리카(케냐, 탄자니아, 잠비아), 나이지리아, 싱가포르가 있다. ICE 코퍼스의 대표적인 것으로는 ICE-USA가 있다. 코퍼스 처리를 위한 소프트웨어 도구에는 PC Beta, ICECUP(ICE Corpus Utility Program), Sara 등이 있다[2].

전자 사전을 위한 대표적인 연구로는 George A. Miller의 WordNet 연구가 있다[3][4]. WordNet v2.1에는 명사에 117,097개의 단어, 81,426개의 synset(synonym set), 그리고 전체 명사-의미 쌍은 145,104개가 있다. 이 명사들 중 많은 것들은 연속어(collocation)를 구성한다. 동사에는 11,488개의 단어, 13,650개의 synset, 그리고 동사-의미 쌍은 24,890개가 분류되어 있다. 형용사에는 22,141개의 단어, 18,877개의 synset, 그리고 31,302개의 형용사-의미 쌍이 있다. 부사에는 4,601개의 단어, 3,644개의 synset, 그리고 5,720개의 부사-의미 쌍이 있다[5]. 동사의 경우에는 어형 변화를 나타내는 표기법을 사용하고 있다. WordNet의 목표는 일반 사전과 별로 다르지 않으며, WordNet의 시멘틱스는 사전을 제작할 때에 사용하는 단어의 의미(sense)를 나타낸다. 그러나 WordNet이 일반적인 사전과 다른 부분은 어휘 정보들이 조직되어 있는 방법이다. 일반 사전의 항목에는 일반적으로 철자정보, 발음, 원형과 변형, 어원, 품사, 주 정의와 부수 정의, 비슷한 단어, 반대 단어, 특수 용법 등이 등재된다. 그러나 사전을 기계가 읽을 수 있는 (machine-readable) 형태로 만들 때에 대부분의 이러한 정보들은 삭제된다. WordNet은 발음, 어형 변화, 어원 등의 정보를 제공하지 않는다. 반면에 WordNet은 단어 간의 의미 관계를 더욱 분명하고 사용하기 편리하게 만들어서 제공한다[5]. WordNet의 의미 관계를 구성하는 기본 단위는 비슷한 단어(synonymy)의 집합이다. WordNet에서는 이것을 synset이라고 한다. 비슷한 단어의 집합이 WordNet을 구성하는 기본 블록(building block)이 된다. 시소러스에서 비슷한 단어를 추출하는 연구에 진척이 있기는 하지만 기본적으로 이러한 정보는 수작업으로 진행되었다. 그리고 어휘 간에는 의미적인 면에서 상하 관계가 있다. 즉, 상위어(hypernym)가 있고 하위어(hyponym)도 있다. WordNet에서는 이것을 이용하여 어휘 간의 상하 관계를 구성한다. 그러면 제일 상위 집합의 형성을 가능하게 하는 최초의 유일한 단어는 어떻게 결정할까? WordNet에는 최초의 유일한 단어로서 25개의 단어를 지정하여 사용하고 있다. 이러한 단어에는 {act, activity}, {food}, ... {substance}, {time} 등이 있다. 또한 WordNet은 부분-전체 관계인 부분어(meronym)와 그 역인 전체어(holonym)를 지원하며 또한 반대말(antonym)을 지원한다. 그러나 WordNet은 is-not-a-(kind-of) 관계, is-used-as-a-kind-of 등의 어휘 관계는 다루지 않는다. 그 밖에 WordNet에는 데이터베이스와 검색 소프트웨어가 개

발되어 있으며 WordNet 어휘 관계를 자동으로 획득할 수 있는 시스템이 개발되어 있다. 따라서 결론적으로 WordNet은 전자 어휘 사전이며 어휘 간의 의미 네트워크를 구현한 어휘 의미 네트워크(lexical semantic network) 라고 할 수 있다. 현재에는 영상네트(ImageNet), 의미 분별(Sense Disambiguation), 지식 베이스(Knowledge Base), 의미 추출(Meaning Extraction)등의 연구가 진행되고 있다[5]. 그 밖에 WordNet에 대한 많은 연구가 진행되었다[6]~[22]. 유럽에서도 영어, 스페인어, 네덜란드어, 이탈리아어에 대하여 EuroWordNet 연구가 진행되었다[23].

의미 네트워크와 관련된 연구는 오래 전부터 있었다[24][25][26]. 참고 문헌 [24]에는 개념 그래프(conceptual graph)에 관한 다양한 내용이 자세하게 논의되어 있다. 현재의 관점에서 보면 개념 그래프는 온톨로지를 구축하기 위한 일종의 개념 도구라고도 할 수 있다. 참고 문헌 [25]는 지식을 표현하는데 사용되는 의미 네트워크(semantic network)에 대하여 논의 하고 있다. 의미 네트워크는 대상과 대상간의 의미 관계를 표현한 것인데, 온톨로지의 기본 형태라고 할 수 있다. 참고 문헌 [26]에는 논리, 온톨로지, 지식 표현, 지식 획득 등에 대한 중요한 논의가 제시되어 있다.

참고 문헌 [27]에는 온톨로지에 대한 전반적인 내용이 논의되어 있다. 여기에는 온톨로지에 대한 정의로 "Ontology is a philosophical discipline, a branch of philosophy that deals with the nature and the organization of being." 라고 설명하고 있다. 참고 문헌 [28]에는 온토로지에 대한 정의로서 "An ontology is an explicit and formal specification of a conceptualization of a domain of interest." 라고 제시하고 있다. 매우 다른 정의가 제시 되어 있는데, 컴퓨터 공학의 관점에서 보면 온톨로지라는 것은 사용자도 읽을 수 있고 컴퓨터도 읽고 처리할 수 있는 형태로 관심 영역에 대하여 필요한 개념, 정보, 지식을 표현하고 저장해 놓은 것으로 이해할 수 있다. 온톨로지는 3개의 계층으로 구성된 구조로 이해할 수 있다. 즉, 3개의 계층 중에서 제일 상위 계층에는 "기호/신택스(syntax) 구조"가 존재하고, 두 번째 계층에는 "개념/시맨틱(semantic) 구조"가 존재하고, 제일 하위 계층에는 "실제 세계의 대상"들이 존재하는 구조로 이해할 수 있다. 그리고 지식 베이스는 온톨로지를 사용하여 구축할 수 있다. 즉, 온톨로지는 구문과 의미를 나타내는 개념 구조로서 기호-개념-대상을 정확히 지칭하기 위한 것이고 지식 베이스는 온톨로지 구조를 기반으로 구축된 대상에 대한 지식과 정보의 집합체로 이해할 수 있다. 위의 기호-개념-대상을 "의미 삼각형"이라고도 한다.

근래에는 시멘틱 웹(semantic web)에 대한 논의와 연구가 활발히 진행되고 있다. W3C가 시멘틱 웹에 대한 기본 개념과 표준을 제공하였지만, 자연언어처리, 지식 개발과 온톨로지(ontology) 관리와 같은 다른 필요한 기술들도 지속적으로 발전하여 왔다. 온톨로지는 일반적으로 정의하면 관심 영역에 대한 형식적이고(formal) 명확한(explicit) 개념화라고 할 수 있다. 웹에서 온톨로지 작성을 위한 언어에는 OWL(Web Ontology Language)가 정의 되었다. 정보와 지식이 컴퓨터 속에 많이 저장되어 축적되면 컴퓨터는 점점 더 의미적이 되고 그리고 이것이 웹을 통하여 표현되고 전달되는 것을 시멘틱 웹이라고 할 수 있다. 이것은 컴퓨터 공학의 관련 분야, 즉, 인터넷 기술, 데이터베이스, 인공지능, 자연언어처리, 인터넷 문서(HTML, XML), 멀티미디어 기술 등을 다음 세대로 이끌어 가는 것이라고 할 수 있겠다.

본 논문에서는 코퍼스에 수록된 단어들을 빈도수에 따라 적합하게 표시할 수 있는 새로운 좌표계와 거리 계산 방법에 대하여 논의한다. 이 방법은 언어 처리에 계량적으로 유용하게 사용될 수 있을 것이다. 좌표계에 대하여서는 메트릭 공간(metric space)을 생각할 수 있다. 그런데 일반적인 Cartesian 좌표계는 단어와 빈도수를 표시하는데 불편한 점이 있다고 할 수 있다. 왜냐하면 영어 코퍼스의 경우에 영어 단어 table 이 10번 등장하고 pear가 0번 등장한다고 할 경우에, 10이라는 빈도수와 0이라는 빈도수를 원점을 기준으로 하는 일반적인 Cartesian 좌표계에 적절하게 표시할 수 있는 방법이 마땅치가 않기 때문이다. 예를 들어, pear의 빈도수가 0번이므로 이 단어를 원점에 위치시킬 수가 있고, table의 빈도수가 10이므로 이 단어를 원점에서 10만큼의 거리에 위치시킬 수가 있다. 그러면 원점에 위치한 영어 단어 pear가 현재 코퍼스 단어의 기준이 된다는 뜻인가? 이것은 적절하지 않다고 할 수 있다. 왜냐하면 pear는 코퍼스에 한 번도 등장하지 않았기 때문에 도리어 원점에 기준이 될 수도 없다. 도리어 코퍼스에 10번 등장한 table이 기준이 되는 것이 바람직하다고 할 수 있다. 이러한 모순을 어떻게 해결하여 코퍼스에서 추출한 단어와 빈도수들을 가장 적절하게 표시하여 여러 가지 유용한 계량적인 처리를 편리하게 할 수 있도록 할까? 이 논문은 이 질문에 대한 해답을 생각하는 과정이라고 할 수 있다.

2. 본 론

2.1 숫자를 위한 메트릭 공간

일반적인 Cartesian 좌표계에는1차원, 2차원, 3차원 좌표계가 있다. 예를 들어, 1차원 좌표계는 x를 변수로, 2차원 좌표계는 x와 y를 변수로, 그리고 3차원 좌표계는 x, y, 그리고 z를 변수로 사용하여 표시한다. 이러한 좌표계는 모두 원점을 0으로 하여 기준으로 하고 원점에서 멀어질수록 큰 좌표값을 부여하여 원점과의 거리를 표시한다. 2차원과 3차원의 경우에 각각의 변수는 원점에서 직각으로 교차한다. 그리고 변수들의 범위는 실수(real number)이다. 참고로 1차원 좌표계에서 두 점 x와 y사이의 거리는 d(x, y) = |x-y|로 표시할 수 있다. 다음은 2차원 메트릭 공간에 대한 정의이다[29].

[정의 1] 2차원 메트릭 공간은 (X, d)로 구성된다. 여기서 X는 집합이고 d는 X에 대한 메트릭으로서(혹은 X에 대한 거리 함수), $X\times X$에 정의된 함수이며, $x,y,z\in X$에 대하여 다음을 만족한다. 그리고 거리를 나타내는 d는 유클리디안 거리를 생각할 수 있다.

(1) d는 음이 아닌 실수값을 가지며 유한하다.
(2) $d(x,y)=0$ 오직(if and only if) $x=y$.
(3) $d(x,y)=d(y,x)$. // symmetry
(4) $d(x,y)\le d(x,z)+d(z,y)$. // triangle inequality

2차원과 3차원의 매트릭 공간에서 대부분의 거리 계산은 유클리디안 거리 공식을 사용한다[30].

2.2 단어를 위한 메트릭 공간

코퍼스에서 추출한 단어와 빈도수를 어떻게 적절한 방법으로 표시하여 단어에 대한 계량적인 처리를 적절하고 편리하게 할 수 있을까? 이 문제를 생각하는 과정에서 0값을 나타내는 원점을 중심으로 하는 기존의 좌표계를 거꾸로 생각하여 0값을 나타내는 점을 원점의 반대쪽에 위치시키는 방법으로 좌표계를 구성하는 새로운 좌표계를 생각해 보게 되었다. 예를 들어, x, y, z 축을 생각한다고 할 경우에, 코퍼스에서 단어

와 빈도수를 추출하여, 빈도수가 높은 단어는 점점 좌표계의 중심 부분에 위치시키고 빈도수가 낮은 단어는 점점 좌표계의 바깥쪽에 위치시키는 방법을 생각하게 되었다. 예를 들어, 빈도수가 더 높은 새로운 단어가 코퍼스에서 발견되면, 그 단어를 나타내기 위하여 좌표축은 양의 방향으로 확장되고 그 단어는 빈도수를 좌표로 하여 좌표계의 중심 부분에 표시되게 된다. 이러한 것을 코퍼스 전체를 조사하여 단어의 빈도수를 계산하고 새로 정의한 좌표 평면에 표시하였다는 의미에서 코퍼스그람(corpusgram) 이라고 부를 수 있을 것이다.

새로운 좌표계에서 예를 들어, x축의 양의 방향에 명사, y축의 양의 방향에 형용사, z축의 양의 방향에 동사, 그리고 x축의 음의 방향에 부사를 위치시키고, z축의 음의 방향에 복합 명사(compound noun)를 위치시키면 명사, 복합 명사, 동사, 형용사, 부사 등이 자연스럽게 자리를 잡게 된다. 또한 좌표계의 구조로부터 각 품사간의 관계가 일목요연하게 파악할 수 있다. 즉, 명사와 형용사와의 관계, 복합 명사와 형용사와의 관계, 명사와 복합 명사의 관계, 동사와 부사와의 관계, 형용사와 부사와의 관계, 동사와 명사와의 관계가 하나의 좌표계에 자연스럽게 표시가 되어 파악이 쉽게 되는 것이다. 더욱이, 코퍼스에서 빈도수가 높은 단어는 자연스럽게 좌표계의 중심부분에 집중하게 되어, 전체적으로 코퍼스가 포함하고 있는 단어의 빈도수와 분포 그리고 단어의 분포에 따른 코퍼스 전체의 개념의 구조를 파악할 수가 있을 것이다.

다음은 위에서 설명한 단어와 빈도수 표시를 위한 새로운 메트릭 공간에 대한 정의이다.

[정의 2] 단어들을 위한 3차원 메트릭 공간은 (X, d)로 구성된다. 여기서 X는 단어들이고 d는 X에 대한 메트릭으로서 (혹은 X에 대한 거리 함수), $X\times X$에 정의된 함수이며, $x,y\in X$에 대하여 다음을 만족한다. $X\times X$는 동사와 명사, 명사와 형용사, 그리고 동사와 부사를 의미한다. 필요한 경우에는 형용사와 부사를 의미할 수 있다.

(1) d는 음이 아닌 실수값을 가지며 유한하다.
(2) $d(x,y)=0$ 오직(if and only if) $x=y$.
(3) $d(x,y)=d(y,x)$. // symmetry
(4) $d(x,y)\le d(x,z)+d(z,y)$. // triangle inequality

이로서 단어들을 위한 새로운 메트릭 공간이 정의되었다. 그런데 여기서 거리함수 d는 일반적인 유클리디안 거리함수를 사용할 수는 없다. 왜냐하면 여기서 사용하는 좌표계는 Cartesian 좌표계가 아니고 단어의 빈도수에 의하여 정의된 새로운 좌표계이기 때문이다. 따라서 정보 이론 개념[31]을 활용하여 단어 사이의 거리를 계산하는 방법을 고안하였다. 즉, 개별 단어의 빈도수를 기반으로 자기 정보(self-information)를 계산하고 또한 단어와 단어 사이의 동시 발생 빈도수(co-occurrence)를 계산하여 상호 정보(mutual information)를 계산하여 단어와 단어 사이의 거리를 계산한다. 즉, 자기 정보양이 작은 것은 자주 발생하는 것이기 때문에 관계가 많아서 거리가 짧다고 해석하고, 자기 정보양이 큰 것은 자주 발생하는 것이 아니기 때문에 관계가 멀어서 거리가 멀다고 생각한다. 예를 들어 v라는 동사의 빈도수가 c_v이고, n이라는 명사의 빈도수가 c_n 이라면, 동사 v와 명사 n의 거리 d는 다음과 같이 계산할 수 있다. 여기서 코퍼스 전체의 단어수는 C , 동사 전체의 단어수는 V, 명사 전체의 단어수는 N으로 한다. 우선 동사 전체에 대한 동사 v의 확률을 계산하면 $P(v)=c_v/V$이고, 명사 전체에 대한 명사 n의 확률을 계산하면 $P(n)=c_n/N$이다. 따라서 동사 v의 자기 정보양은 $I(v)=-\log P(v)$가 된다. 여기서 빈도수가 가장 많은 동사의 자기 정보량을 $I(v_{\max})$이라고 할 경우에, 동사 v의 거리는 $d(v)=I(v)-I(v_{\max})$로 정의한다. 그러면 동사 v의 거리는 동사 중에서 가장 빈도수가 많은 동사를 기준으로 하여 거리가 정의된다. $I(v)$와 $I(v_{\max})$는 계산할 수 있는 값이기 때문에 이 들 값으로부터 간단하게 $d(v)$는 계산될 수 있다. 명사 n의 자기 정보양은 $I(n)=-\log P(n)$이 된다. 그리고 빈도수가 가장 많은 명사의 자기 정보량을 $I(n_{\max})$라고 할 경우에, 명사 n의 거리는 $d(n)=I(n)-I(n_{\max})$로 정의한다. 그러면 명사 n의 거리는 명사 중에서 가장 빈도수가 많은 명사를 원점으로 하여 거리가 정의된다. 그러면 일차적으로 동사 v와 명사 n 에 있어서 다음 식을 계산할 수가 있다. 이 식은 동사 v와 명사 n 사이의 정보량을 나타낸다고 할 수 있다.

$$d(v,n)=d(v)+d(n) \quad (1)$$

또한 단어들의 동시 발생 빈도수(co-occurrence)를 조사하여 단어들 간의 상호 정보(mutual information)를 계산할 수 있다. 단어의 동시 발생 빈도수는 조사하고자 하는 단어들이 하나의 문장에 동시에 존재하는 경우를 조사하여 구할 수 있다. 일반적으로 변량 A와 변량 B 사이의 상호 정보 계산식은 다음과 같이 정의된다.

$$MI(A,B)=\sum_{i=1}^{n}\sum_{j=1}^{m}P(a_i,b_j)\log\frac{P(a_i,b_j)}{P(a_i)P(b_j)} \quad (2)$$

그러면 앞에서 자기 정보를 이용하여 계산한 $d(v,n)$과 상호 정보를 이용하여 계산한 $MI(v,n)$을 기반으로 다음과 같이 동사 v와 명사 n 사이의 거리를 정의할 수 있다.

$$dist(v,n)=\min\{d(v,n),MI(v,n)RIGHT \quad (3)$$

단어 사이의 거리를 위와 같이 정의할 경우에, $dist(x,x)=0$으로 정의한다. 그리고 $dist(x,y)=dist(y,x)$가 성립한다. 왜냐하면 $d(v,n)=d(v)+d(n)$ 이고 $d(n,v)=d(n)+d(v)$이며, $MI(v,n)=MI(n,v)$이기 때문이다. 또한 $d(x,z)+d(z,y)=d(x)+d(y)+2d(z)$이므로 $d(x,y)\le d(x,z)+d(z,y)$도 성립한다. 그리고 $MI(a,c)\le MI(a,b)+MI(b,c)$도 성립한다. 따라서 자기 정보와 상호 정보에 의한 거리 정의인 [정의 2]는 단어들 위한 새로운 유효한 메트릭 공간을 만들어 낼 수 있다. 형용사와 명사, 동사와 부사 사이에도 동일한 방법으로 거리를 정의할 수 있다. 그리고 혹시 비슷한 부분에 단어의 빈도수가 몰리는 경우에는 단어 빈도수에 대하여 log 변환을 사용하여 단어 빈도수를 넓게 분포시켜서 거리를 계산하는 것도 좋을 것이다. 그리고 상호 정보 값이 계산되지 않는 경우에는 $d(x,y)$ 값을 거리로 간주하여야 할 것이다. 코퍼스 전체의 단어수는 정규화를 할 때 사용할 수 있다.

그리고 메트릭 공간에 단어들을 표시할 경우에 문서에서 단어들이 텍스트에서 나타난 위치를 함께 기록하면 좋을 것이다. 문서에서 단어들이 나타난 위치는 장(chapter), 절(section), 문단(paragraph), 문장을 의미한다. 즉, 단어에 장, 절, 문단, 그리고 문단 내의 문장의 위치를 기록하는 것이다.

2.3 단어 메트릭 공간의 활용

코퍼스를 기반으로 구성된 단어들의 메트릭 공간에 대한 응용으로는 다음을 생각할 수 있다. 예를 들어, 일차적으로 동사와 명사의 경우를 고려하면 명사가 주어인 경우에는 동사와 명사의 거리를 계산하여 우선 순위가 계산된 의미 관계(semantic relation)를 구성할 수가 있을 것이며, 명사가 목적어인 경우에도 역시 우선 순위가 계산된 술어(predicate)를 구성할 수가 있을 것이다. 이러한 의미 관계는 정보 검색, 의미 네트워크 구성, 그리고 분석의 다중성을

해결하는데 유용하게 사용될 수 있을 것이다. 여기에 상호 정보도 함께 고려하면 더욱 의미가 중요한 의미 관계나 술어를 구성할 수가 있을 것이다. 그리고 이 정보는 문장의 요약, 문장을 기반으로 한 온토롤지와 지식 베이스 구축, 그리고 정보 검색에 매우 효과적으로 활용할 수가 있을 것이다.

또한 명사들을 각 장이나 절마다 빈도수를 중심으로 정렬을 하면 텍스트에서 나타내고자 하는 대상, 주제, 그리고 의미의 흐름을 알 수가 있을 것이다. 이것은 텍스트 전체의 의미 구조와 논의의 흐름 과정을 나타낼 수 있으며, 온톨로지와 지식 베이스 구성의 기본 자료가 될 수 있을 것이다. 동사들도 빈도수에 따라서 정렬을 하면 행위의 흐름을 알 수가 있을 것이다. 형용사와 부사에 대하여서도 같은 조사를 수행하면 유용한 정보를 얻을 수가 있을 것이다.

그리고 명사의 경우에 단어의 빈도수에 대하여 동일화(equalization)를 수행하여 그 결과를 해석하여 보는 것도 의미가 있을 것이다. 동일화의 결과에 중요한 의미가 있을 수가 있다. 즉, 동일화의 결과가 전체 텍스트에서 명사의 고른 분포를 나타낼 수가 있을 수 있기 때문이다. 그리고 전체 명사의 고른 분포는 텍스트의 어떤 부분에서 있을 수 있는 명사의 집중 현상이나 희소 현상과 대비하여 그 근거를 설명할 수 있는 기본 자료가 될 수도 있을 것이다. 왜냐하면 일반적으로 명사가 전체 텍스트에서 균등하게 분포되어 있는 것이 바람직하다고 생각할 수 있기 때문이다. 또한 전체 텍스트의 각 장이나 절마다 명사의 개수와 빈도수에 대하여 평균과 분산을 계산하여 보는 것도 통계적인 관점에서 의미가 있을 것이다. 그리고 이를 바탕으로 빈도수가 높은 중심 명사와 그렇지 않은 명사들 간의 거리 혹은 거리의 역수를 변수로 하여 확률 분포 함수를 예측하여 보는 것도 의미가 있을 것이다. 동사의 경우에 대하여서도 동일한 과정을 생각해 볼 수 있다.

그리고 명사의 분포를 생각하는 경우에 하나의 중심 명사에서 다른 명사로 이어지는 의미의 진행과 연결 과정을 좀 더 세밀하고 균질하게 파악하기 위하여 명사와 명사사이에 의미적으로 비슷한 명사(synonym)들을 WordNet의 어휘 시멘틱 네트워크를 활용하여 찾아서 점진적으로 배치시키는 것을 수행하면 어쩌면 중간에 생략된 명사를 찾아서 분포시키는 결과를 얻을 수가 있을 것이며 그리고 이것은 중간에 생략된 개념을 찾는 결과를 얻을 수도 있을 것으로 생각한다. 동사의 경우에 대하여서도 동일한 과정을 수행하면 중간에 생략된 동사를 찾을 수도 있을 것이다. 그러면 새로 찾게 된 명사와 동사를 중심으로 문장을 형성하여 새로운 문단과 절을 구성할 수가 있을 것이다. 그리고 어쩌면 이것은 코퍼스에서 논의의 흐름 과정 혹은 추론 과정을 보간하여 나타내는 것일 수도 있을 것이며, 생략된 의미를 찾는 효과도 얻을 수 있을 것이다.

그리고 구성된 코퍼스그람과 정보들을 문장 생성 기능과 연결하여, 자연언어 문장을 생성하는 실험을 실시하는 것도 의미가 있을 것이다. 이때에 문장을 질의어 형태로 생성하여 자연언어 분석이나 이해 과정에서 발생하는 문제에 대한 질문을 사용자에게 할 수도 있을 것이다.

그리고 관심 영역에 대한 지식과 정보를 온톨로지에 의하여 지식 베이스를 구축하여 자연 언어처리에 유용하게 활용하는 방안을 찾아야 할 것이다. 이러한 지식 베이스를 구축할 때에 사용할 어휘는 WordNet의 어휘 시멘틱 네트워크를 활용하여 대상 응용 영역의 어휘 구조를 구축하면 좋은 효과를 얻을 수 있을 것이다.

그 밖에 이러한 코퍼스그람, WordNet, 온톨로지, 시멘틱 네트워크, 지식 베이스 등을 활용하면 어휘 분석의 모호성(ambiguity), 구문 분석의 모호성, 의미 분석의 모호성을 거의 해결할 수 있을 것이다. 그리고 대명사 지칭 문제, 관계 대명사 지칭 문제, 전치사 접속 문제 등 전통적인 어려운 문제들을 거의 대부분 해결할 수 있을 것으로 생각한다. 그리고 이러한 도구들을 활용하면 문장 재작성(paraphrasing), 텍스트 요약, 텍스트 이해, 텍스트 생성 등도 대부분 가능할 것이다.

3. 결 론

본 논문에서는 코퍼스, WordNet, 시멘틱 네트워크와 지식 베이스, 온톨로지 등을 논의하였다. 그리고 기존의 Cartesian 좌표 평면 이외에 코퍼스로부터 단어와 빈도수를 추출하여 구성할 수 있는 새로운 단어들을 위한 메트릭 공간에 대하여 논의하였다. 단어 메트릭 공간인 코퍼스그람은 언어 처리, 온톨로지, 지식 베이스, 시멘틱 네트워크 연구에 유용한 자원과 도구가 될 수 있을 것으로 생각한다. 앞으로의 연구로는 가까운 시일 안에 다량의 좋은 코퍼스를 확보하고 논문에서 논의한 코퍼스그람을 구축하여 언어처리와 온톨로지 구축에 적용하여 실험해 보는 것이다.

참고 문헌

[1] Charles F. Meyer, English Corpus Linguistics, Cambridge University Press, 2002.
[2] Stig Johansson, Anna-Brita Stenstroem(editor), English Computer Corpora, Mouton de Gruyter, 1991.
[3] George Miller, "WordNet : A Lexical Database for English," CACM, Vol. 38, No. 11, November 1995.
[4] Christiane Fellbaum(editor), WordNet - An Electronic Lexical Database, The MIT Press, 1998.
[5] http://wordnet.princeton.edu/ WordNet Documentation.
[6] Nitin Verma, Pushpak Bhattacharya, "Automatic Lexicon Generation through WordNet," GWC 2004
[7] B. A. Sharada, P. M. Girish, "WordNet Has No 'Recycle Bin'," http://wordnet.princeton.edu/.
[8] Anna Sinopalnikova, "Word Association Thesaurus As a Resource for Building WordNet," http://wordnet.princeton.edu/.
[9] Sa-Im Shin, Key-Sun Choi, "Automatic Word Sense Clustering Using Collocation for Sense Adaptation," http://wordnet.princeton.edu/.
[10] Adam Pease, Christiane Fellbaum, "Language to Logic Translation with PhraseBank," http://wordnet.princeton.edu/.
[11] Mauro Castillo, Francis Real, German Rigau, "Automatic Assignment of Domain Labels to WordNet," http://wordnet.princeton.edu/.
[12] Fernando Gomez, "Building Verb Predicates : A Computational View," http://wordnet.princeton.edu/.
[13] Nuno Seco, Tony Veale, Jer Hayes, "An Intrinsic Information Content Metric for Semantic Similarity in WordNet," http://wordnet.princeton.edu/.
[14] Philip Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," http://wordnet.princeton.edu/.
[15] Jiangsheng Yu, Zhenshan Wen, Yang Liu, Zhihui Jin, "Statistical Overview of WordNet from 1.6 to 2.0," http://wordnet.princeton.edu/.
[16] Elke Teich, Peter Fankhauser, "WordNet for Lexical Cohesion Analysis," http://wordnet.princeton.edu/.
[17] Jorge Morato, et al., "WordNet Applications," http://wordnet.princeton.edu/.
[18] Ann Devitt, Carl Vogel, "The Topology of WordNet : Some Metrics," http://wordnet.princeton.edu/
[19] Eneko Agirre, "Clustering of Word Senses," http://wordnet.princeton.edu/.
[20] Ganesh Ramakrishnan, et al., "Soft Word Sense Disambiguation," http://wordnet.princeton.edu/.
[21] Ivan Obradovic, et al., "Corpus Based Validation of WordNet Using Frequency parameters," http://wordnet.princeton.edu/.
[22] Roberto Navigli, et al., "Ontology Learning and Its Application to Automated Terminology Translation," http://wordnet.princeton.edu/.
[23]Piek Vossen, "EuroWordNet : a multilingual database for information retrieval," Proc. of the DELOS workshop on Cross-language Information retrieval, 1997.
[24] John F. Sowa, Conceptual Structures, Addison-Wesley Pub., 1984.
[25] John F. Sowa(editor), Principles of Semantic Networks, Morgan Kaufmann Publishers, Inc., 1991.
[26] John F. Sowa, Knowledge Representation, Brooks/Cole, 2000.
[27] Alexander Maedche, Ontology Learning for the Semantic Web, Kluwer Academic Publishers, 2002.
[28] John Davies, Rudi Studer, Paul Warren(editor), Semantic Web Technologies, John Wiley & Sons, Ltd, 2006.
[29] Erwin Kreyszig, Introductory Functional Analysis with Applications, John Wiley & Sons, 1978.
[30] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2006.
[31] Thomas M. Cover, Joy A. Thomas, Elements of Information Theory(2nd ed.), Wiley-interscience, 2006.