코퍼스그람 실험과 개발에 대한 연구

이 호 석
호서대학교 공과대학 뉴미디어학과
hslee@office.hoseo.ac.kr

A Study on Corpusgram Experiment and Development

Ho Suk Lee
New Media Dept. College of Engineering, Hoseo University

요 약

코퍼스그람에서 실험이 필요한 부분은, 첫 번째는 변수 d와 $dist$의 정의 부분이다. 즉, 변수 d만을 이용한 경우, 변수 $dist$만을 이용한 경우, 그리고 변수 d와 $dist$를 모두 이용한 경우를 실험해 보아야 한다. 두 번째는 코퍼스그람에서 거리가 가까운 단어들의 조합, 예를 들어 명사와 명사, 동사와 명사, 형용사와 명사, 동사와 부사를, 조사하여 그 의미를 해석하여 보는 것이다. 세 번째로는 코퍼스그람의 단어들에 대하여 거리를 중심으로 단어 연결(connection) 네트워크를 구성하고 의미 네트워크와 비교하여 보는 것이다. 네 번째로는 연결 네트워크를 정보 검색 등의 응용에 적용하여 효과를 확인하는 것이다. 그리고 언어 처리, 온톨로지 등에 중요한 요소인 부분-전체 관계에 대하여 소개하였다.

주요 단어 : 코퍼스그람, 단어 연결 네트워크, 부분-전체 관계

1. 코퍼스그람

참고문헌 [1]에 코퍼스그람에 대한 논의가 다음과 같이 제시되어 있다. 예를 들어 v라는 동사의 빈도수가 c_v이고, n이라는 명사의 빈도수가 c_n 이라면, 동사 v와 명사 n의 거리 d는 다음과 같이 정의한다. 동사 v의 거리는 $d(v) = I(v) - I(v_{\max})$로 계산한다. 명사 n의 거리는 $d(n) = I(n) - I(n_{\max})$로 계산한다. 그리고 동사 v와 명사 n 의 거리 d는 다음과 같이 정의한다.

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

또한 단어들 간의 상호 정보(mutual information)를 다음의 식을 이용하여 계산할 수 있다.

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

그러면 앞의 수식 (1)에서 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.

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

단, 단어 사이의 거리를 위와 같이 정의할 경우에, $dist(x,x) = 0$으로 정의한다.

2. 코퍼스그람 실험

첫 번째로는 위의 코퍼스그람 설명에서 수식 (1), 수식 (2), 그리고 수식 (3)의 경우를 실제 코퍼스를 대상으로 실험하는 바람직한 경우를 찾는 것이 좋을 것이다.

두 번째로는 코퍼스그람에서 거리가 가까운 단어들의 조합을, 예를 들어 명사와 명사, 동사와 명사, 형용사와 명사, 동사와 부사, 조사하여 그 의미를 해석하여 보는 것이다. 명사와 명사의 경우에는 명사 관용어이거나 복합 명사일 가능성이 있을 것이다. 동사와 명사의 경우에는 관용어, 연속어(연어대신에 연속어라는 용어를 사용하기로 한다.), 혹은 현재 텍스트에서 주제어일 것이다. 형용사와 명사의 경우에도 관용어, 연속어, 혹은 현재의 상황을 나타내는 표현일 것이다. 동사와 부사의 경우에는 관용어 혹은 현재의 상황을 나타내기 위하여 자주 사용한 표현일 것이다. 단어들의 조합은 집합의 형태로 인식하여 처리할 수가 있

다. 즉, 코퍼스그람에서 단어들의 조합을 조사해 보면, 관용어, 연속어, 그리고 주제어를 파악할 수 있을 것이다.

세 번째로는 앞에서 파악된 단어들의 조합, WordNet 방식의 synset 개념, 그리고 원점에서 거리가 가까운 단어들을 계층별로 구분하거나 혹은 그룹별로 클러스터링(clustering)[3]을 수행하여 단어들의 연결 네트워크를 구축하는 것이다. 여기서 연결 네트워크라는 용어를 사용한 것은 기존의 WordNet 혹은 의미 네트워크에서는 사용하지 않는 클러스터 개념을 사용하여 단어들 간의 연결을 파악하기 때문이다.

네 번째로는 단어 연속어 집합과 단어 클러스터링은 정보 검색에 의한 텍스트 조사에 직접 적용될 수 있을 것이다. 참고 문헌 [2]에서 복합 명사의 처리에 대한 논의가 있는데, 이는 명사들의 연속어를 조사하면 용이하게 처리할 수 있을 것이다. 또한 의미 관계의 구성도 단어들의 연속어와 클러스터링를 조사하면 처리할 수 있을 것이다. 그 밖에 의미 네트워크의 구성, 의미의 과잉 일반화 해결 등도 코퍼스그람에 계층적으로 구축된 단어 연결 네트워크를 이용하면 용이하게 처리할 수 있을 것이다.

3. 관련 연구

자연언어처리에 대한 일반 연구로는 참고 문헌 [6]-[23]이 있다. 이 논문들은 영어에 대한 언어 처리에 대하여 논의하고 있다. 참고 문헌 [6]은 확률 파서에 대하여 논의한다. 우선 기본 확률 파서는 파싱 결과에 초기 확률을 붙여서 출력한다. 다음에 두 번째 모델이 초기 확률의 파싱 트리의 다른 정보를 이용하여 향상시켜 파싱 결과의 순위를 재조정한다. 이 방법의 강점은 파싱 트리가 피처(feature)들의 집합으로 나타난다는 것이다. 즉, 문법 전개(derivation)나 생성(generation)을 고려하지 않고 파싱 결과를 피처들의 집합으로 나타낼 수 있다는 것이다. 참고 문헌 [7]은 Penn Treebank 코퍼스에 의미 관계 정보 계층, 혹은 의미 역할 표시를 설정하여 의미 표현에 대하여 실제적인 접근을 한 연구이다. 연구의 결과는 coreference, quantification 등을 처리하지 않아서 정보 면에서 약한 면이 있으나, 코퍼스에 존재하는 모든 개별 동사의 모든 문장에 대하여 통계를 제시한다는 면에서는 포괄적인 장점이 있다. 참고 문헌 [8]은 기계 번역에 대한 것으로서, 평행하지 않은(non-parallel) 코퍼스로부터 평행한(parallel) 문장을 발견하는 새로운 방법을 논의하고 있다. 이 논문은 최대 엔트로피 분류기(maximum entropy classifier)를 사용하여 중국어, 아랍어, 그리고 영어 코퍼스에 대하여 번역 결과(sentence pair)를 찾을 수 있다고 한다. 이 방법의 또 다른 장점은 크기가 크지 않은 대략 100,000 단어정도로 구성된 코퍼스에도 적용할 수 있다는 것이다. 참고 문헌 [9]는 WordNet에 대한 것이다. 현재 WordNet 명사의 하위어(hyponym) 부분을 확장하여 클래스로서의 하위어와 인스턴스(instance)로서의 하위어를 구분할 수 있도록 하였다. 참고 문헌 [10]은 어휘간의 의미 관련성 연구에 대한 것으로서, WordNet을 사용하는 5가지의 어휘 의미 측정 방법을 스펠링 에러를 찾아서 수정할 수 있는 성능 면에서 평가하여 제시한다. 조사 결과는 Jiang & Conrath의 정보 내용 기반 방법이 Hirst, St-Onge, Leacock, Chodorow, Lin, Resnik의 방법에 비하여 우수한 것으로 판명되었다. 또한 분산적인 유사성은 어휘간의 의미 관련성을 나타내는데 적절한 것이 아니라는 것을 제시한다. 참고 문헌 [11]은 Hobbs가 제시한 담론 관계(discourse relation)를 이용하여, 사용하기 쉽고 적절한 자료 구조를 구성하기 쉬운 담론 구조 관계(discourse structure relation)를 논의한다. 즉, Wall Street Journal 신문과 AP 뉴스 기사 중에서 135개의 텍스트에 대하여 담론 결합성(coherence) 구조를 설정할 수 있는 표시 방법에 대하여 논의한다. 또한 트리 구조는 담론 구조를 표시하는데 적절한 구조가 아니라는 것을 제시한다. 참고 문헌 [12]는 기본적으로 단어 사이에서 파라메타를 사용한 분산 유사성(parameterized distributional similarity)을 계산하는 방법을 논의한다. 이 방법에서는 분산적으로 유사한 단어를 찾는 문제는 CR(Co-occurrence) 검색 문제로 전환되어서 정확성(precision)과 출력(recall)은 문서 검색에서 측정된 방법과 유사한 방법으로 측정된다. 이 방법은 자동 시소러스 생성(automatic thesaurus generation)과 유사-모호성 해소(pseudo-disambiguation) 응용에 적용되어 평가되었다.

참고 문헌 [13]은 텍스트로부터 부분-전체(part-whole) 관계(meronymy)를 자동으로 검출할 수 있는 도메인 독립적이며 의미적으로 집중적이고 그리고 조정을 받는(supervised) 텍

스 처리 방법을 제시한다. 부분-전체 관계는 오래전 그리스 시대부터 철학적으로 연구가 되었으며 근래에도 철학, 심리학, 언어학 등의 분야에서 연구가 계속 되고 있으며 온톨로지 관계 중에서도 가장 근본적인 것으로 인식되고 있다. 부분-전체 관계의 조사에서 어려운 문제는 패턴이 다른 의미 관계도 포함한다는 것이다. 따라서 패턴이 부분-전체 관계를 포함하는지 혹은 그 밖에 다른 관계도 포함하고 있는지 구분할 수 있는 학습 방법이 필요하다. 우선 명사구에 적용되어 ISS(Iterative Semantic Specialization) 방법을 통하여 부분-전체 관계가 학습된다. 다음에 소유격, 복합 명사, 그리고 전치사 절을 포함한 명사절에 적용되어 학습된다. 부분-전체 관계를 예를 들면, 복합 명사 door knob이 있다. 이것을 술어 형태로 표현하면 PART-WHOLE(knob, door)라고 할 수 있다. 부분-전체 관계는 일반 텍스트에서 여러 가지 형태로 나타난다. 예를 들면, 다음의 문장에는 여러 개의 부분-전체 관계가 있다.

The car's mail messenger is busy at work in the mail car as the train moves along. Through the open side door of the car, moving scenery can be seen. The worker is alarmed when he hears an unusual sound. He peeks through the door's keyhole leading to the tender and locomotive cab and sees the two bandits trying to break through the express car door.

즉, 1) mail car는 train의 부분, 2) side door는 car의 부분, 3) keyhole은 door의 부분, 4) cab은 locomotive의 부분, 5) tender는 train의 부분, 6) locomotive는 train의 부분, 7) door는 car의 부분, 그리고 8) car는 express train이 부분-전체 관계이다. 텍스트에서 이러한 부분-전체 관계는 텍스트의 의미 체계를 이해하는데 반드시 필요한 요소이다. 이 참고문헌에서는 WordNet, LA Times, Wall Street Journal, SemCor 1.7 텍스트 모음 등을 사용하여, 29,134 건의 긍정적인 예와 27,963 건의 부정적인 예로 구성되는 코퍼스를 구성하여 실험하였다. 분류 방법은 C4.5 판단 트리(decision tree)를 사용하였으며 형태는 if-then 규칙 형태로 표현된다. 일반적인 부분-전체 관계를 좀 더 구분하면 의미 적인 관점에서 다음과 같은 관계를 포함한다, 1) component-integral 관계, 2) member-collection 관계, 3) portion-mass 관계, 4) stuff-object 관계, 5) feature-activity 관계, 그리고 6) place-area 관계. 부분-전체 관계는 텍스트에서 의미 관계를 도출함으로써 텍스트의 지식 체계를 구성하는데 반드시 필요한 요소이다. 부분-전체 관계 중에서는 명확한 것도 있고, 애매한 것도 있다. 그리고 형태는 부분-전체 관계의 모양을 하고 있으나 전혀 아닌 것도 있다. 예를 들어서 1) The substance consists of three ingredients, 2) The cloud was made of dust, 3) Iceland is a member of NATO 등은 명확하게 부분-전체 관계를 보여준다. 반면에, 4) The lieutenant is part of the play 문장은 모양은 부분-전체 관계이나 의미적으로는 부분-전체 관계가 아니기 때문에, 부분-전체 관계라고 해석해서는 안된다.

참고 문헌 [14]는 의미 분석에 대한 것으로서 텍스트 분할의 정확성을 높이기 위하여 코퍼스로부터 의미 지식을 추출하는 방법을 논의한다. 사용하는 방법은 LSA(Latent Semantic Analysis)로서 다른 연구자들이 제안한 LSA 방법을 향상시켜 적용하였다. 참고 문헌 [15]는 자연 언어 생성에 대한 것으로 템플릿을 사용한 언어 생성이 다른 방법에 비하여 좋지 않다고 알려져 있으나, 이 논문은 그렇지 않은 경우를 보여 주고 있다. 참고 문헌 [16]은 Penn-II와 Penn-III Treebank 코퍼스를 위하여 자동 LFG(Lexical-Functional Grammar) f-구조 주석처리(annotation) 방법에 기반하여 서브카테고리(subcategory)를 추출하는 방법을 제시하고 있다. 논문은 구문-함수-기반 서브카테고리(LFG semantic form) 프레임, 전통적인 CFG 카테고리 기반 서브카테고리 프레임, 또한 함수/카테고리가 혼합된 프레임도 추출할 수 있다. 카테고리 프레임에는 확률이 부여되며, 능동과 수동 프레임이 구분되며, 원 자료 구조에서 긴 거리 의존 관계(long-distance dependency)를 모두 나타낸다. 다른 연구와 달리 본 연구에서는 추출될 서브카테고리 프레임을 미리 정의하지 않으며, 데이터에서 자동으로 학습하여 추출한다. 현재까지 본 연구가 영어에 있어서 가장 크고 가장 완벽한 서브카테고리 프레임 연구인 것으로 보인다. 참고 문헌 [17]은 문서 요약에 대한 것이다. 현재의 자동 문서 요약 연구는 문장 추출에 의한 요약과 단어 집합을 사용한 제목 생성이다. 이 논문에서는 문서와 요약 사이에서 자동으로 단어-대-단어 쌍과 구절-대-구절 쌍을 추출할 수 있는

방법을 개발하였다. 이 방법은 hidden Markov 모델을 확장하여 개발하였으며 감독받지 않고(unsupervised) 자동으로 수행한다. 참고 문헌 [18]은 other-anaphora 와 정확한 명사구 상호 참조(coreference) 해결을 위하여 선행 단어(antecedent)를 선택을 하는데 있어서 어휘 지식을 구하는 두 가지 방법을 비교한다. 비교하는 두 가지 방법은 WordNet을 이용한 방법과 직접 코퍼스를 조사하여 어휘 의미 패턴을 찾는 방법이다. 코퍼스로는 BNC(British National Corpus)를 사용하였으며 Web 텍스트도 사용하였다. 연구의 결과로는 (a) WordNet에 저장된 지식이 anaphora 해결을 위하여 부족한 것으로 보이며, (b) other-anaphora에 있어서는 웹기반 방법이 WordNet 기반 방법보다 좋은 결과를 내었다, (c) 정확한 NP 상호 참조 해결에 있어서는 웹 기반 방법이 WordNet 방법과 비슷한 결과를 내었으며 어떤 경우에는 WordNet 보다 좋은 결과를 내었다, (d) 두 가지 경우 모두에 있어서, BNC 방법이 데이터의 희소성으로 말미암아 더 좋지 않았다. 결론적으로 웹 기반 방법이 anaphora 해결에 있어서 어휘 지식의 부족을 경감시켜주는 좋은 자료가 되는 것을 알 수 있었다. 참고 문헌 [19]는 문장 요약에 대한 것이다. 즉, 이 논문에서는 문서에서 공통 정보를 합성하여 텍스트-텍스트 생성을 하는 새로운 방법을 제시한다. 우선, 유사한 정보를 나타내는 구절을 텍스트에서 찾아서 통계적으로 공통의 구절을 하나의 문장으로 묶는 방법을 사용한다. 참고 문헌 [20]은 감독되지 않는(unsupervised) 상태에서 부분적으로 파싱된 텍스트 코퍼스로부터 명사, 동사, 형용사에 대한 구문과 의미적인(syntactico-semantic) 근거를 찾는 방법을 제시한다. 첫째로, 의존 관계에 있는 두 개의 단어는 상호 필요한 존재라고 가정하고 상호 요구조건(corequirement) 라고 부른다. 둘째로, 유사한 장소에 나타나는 단어의 집합은 장소에 대한 단어의 요구사항이라고 정의한다. 이 연구의 학습 목적은 유사한 위치 나타나는 유사한 단어의 집단을 확인하는 것이다. 이것은 단어들의 구문적이고 의미적인 요구 사항을 학습하는 것을 의미한다. 이 연구의 결과는 접속(attachment)의 모호성(ambiguity)을 해결하는데 사용될 수 있다.

전체 연구 동향을 보면 텍스트 코퍼스와 웹 문서 등을 자동으로 감독 없이 조사하여 자연언어 처리에 필요한 언어적 지식과 정보를 얻어서 활용하려는 것으로 보인다. 특히 이 과정에서 많이 사용하는 방법이 통계적인 학습 방법이며, 텍스트 코퍼스나 웹 문서 혹은 처리 방법에 다른 가정이나 조건을 설정하지 않으려는 것도 특징이다.

4. 적용 분야

이 논문에서 논의한 방법은 자연언어 처리, 정보 검색, 온톨로지[4], 그리고 의미 웹 연구[5] 등에 폭 넓게 활용될 수 있을 것이다.

5. 결 론

이 논문에서는 코퍼스그람에 대한 실험 방법과 적용에 대하여 논의하였다. 앞으로의 연구로는 코퍼스를 구축하여 이 논문에서 논의한 사항들을 실험하여 실제 사용이 가능한 코퍼스그람 시스템을 구축하는 것이다.

또한 부분-전체 관계를 텍스트에서 자동으로 파악하여 텍스트에서 개념 구조를 생성할 수 있는 실용적이고 효율적인 방법을 찾는 것이 필요하다. 부분-전체 관계의 자동 조사 및 생성은 텍스트 개념 구조와 이해 구조의 생성, 온톨로지 체계의 구축 등을 위하여 반드시 필요할 것이다. 한국어는 한자어 처리가 한국어 단어의 개념 구조 구축에 필요할 것이다.

참고 문헌

[1] 이호석, 김영택, “단어들을 위한 새로운 메트릭 공간 : 코퍼스그람,” 2007 한국컴퓨터종합학술대회 논문집, 제34권, 제1(C)호, 185-188면, 2007.

[2] 이호석, 김영택, “WordNet과 텍스트 코퍼스에 기반한 의미 관계를 활용한 웹 텍스트 조사 기법,” 2007 한국컴퓨터종합학술대회, 제34권, 제1(C)호, 181-184면, 2007.

[3] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2006.

[4] Steffen Staab, Rudi Studer, Handbook of Ontologies, Springer-Verlag, 2004.

[5] John Davies, Rudi Studer, Paul Warren, Semantic Web Technologies, John Wiley & Sons, 2006.

[6] Michael Collins, Terry Koo, "Discriminative Reranking for Natural Language Parsing," Journal of Association of Computational Linguistics, 2005.
[7] Martha Palmer, Daniel Gildea, Paul Kingbuy, "The Proposition Bank : An Annotated Corpus of Semantic Roles," Journal of Association of Computational Linguistics, 2005.
[8] Dragos Stefan Munteanu, Daniel Marcu, "Improving Machine-Translation Performance by Exploiting Non-Parallel Corpora, Journal of Association of Computational Linguistics, 2006.
[9] George A. Miller, Florentina Hristea, "WordNet Nouns : Classes and Instances," Journal of Association of Computational Linguistics, 2006.
[10] Alexander Budanitsky, Graeme Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness, Journal of Association of Computational Linguistics, 2006.
[11] Florian Wolf, Edward Gibson, "Representing Discourse Cohesion : A Corpus-Based Study, Journal of Association of Computational Linguistics, 2005.
[12] Julie Weeds, David Weir, "Co-occurrence Retrieval: A Flexible Framework for Lexical Distribution Similarity, Journal of Association of Computational Linguistics, 2006.
[13] Roxana Girju, Adriana Badulescu, Dan Moldovan, "Automatic Discovery of Part-Whole Relations, Journal of Association of Computational Linguistics, 2006.
[14] Yves Bestgen, "Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore," Journal of Association of Computational Linguistics, 2006.
[15] Kees van Deemter, Emiel Krahmer, Mariet Theune, "Real versus Template-Based Natural Language Generation: A False Opposition?," Journal of Association of Computational Linguistics, 2005.
[16] Ruth O'Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, Andy Way, "Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks," Journal of Association of Computational Linguistics, 2005.
[17] Hal Daume III, Daniel Marcu, "Induction of Word and Phrase Alignments for Automatic Document Summarization," Journal of Association of Computational Linguistics, 2006.
[18] Katja Markert, Malvina Nissim, "Comparing Knowledge Sources for Nominal Anaphora Resolution," Journal of Association of Computational Linguistics, 2005.
[19] Regina Barzilay, Kathleen R. McKeown, "Sentence Fusion for Multidocument News Summarization, Journal of Association of Computational Linguistics, 2005.
[20] Pablo Gamallo, Alexandre Agustini, Gabriel P. Lopes, "Clustering Syntactic Positions with Similar Semantic Requirements," Journal of Association of Computational Linguistics, 2005.