Semantic Hypernetwork 학습에 의한 자연언어 텍스트의 의미 구분

이은석[12] 김준식[3] 신원진[2] 박찬훈[2] 장병탁[2]

[1]서울대학교 인지과학협동과정
[2]서울대학교 컴퓨터공학부 바이오지능 연구실
[3]서울대학교 신경정신과
eslee@bi.snu.ac.kr, jskim.ozmagi@gmail.com, {wjshin, chpark, btzhang}@bi.snu.ac.kr

Scaling Documents' Semantic Transparency Spectrum with Semantic Hypernetwork

Eun Seok Lee[12] Joon Shik Kim[3] Won-Jin Shin[2] Chan-Hoon Park[2] Byoung-Tak Zhang[2]

[1]Cognitive Science Program, Seoul National University
[2]BI Lab, School of Computer Sci. & Eng., Seoul National University
[3]The Department of Neuropsychiatry, Seoul National University

요 약

어떤 자연언어 문서가 전달하려는 의미는 그 텍스트의 성격에 따라 아주 명확할 수도(예: 뉴스 문서), 아주 불분명할 수도 있다(예: 시). 이 연구는 이러한 '의미의 명확성(semantic transparency)'을 정량적으로 측정할 수 있다고 가정하고, 이 의미의 명확성을 판단하는 데에 단어들의 연쇄(word association)의 확률통계적 성질들이 어떻게 기능하는지에 대해 논한다. 이를 위해 특정 단어가 연쇄체를 형성하면서 발생하는 neighboring frequency와 degeneracy를 중심으로 Markov chain Monte Carlo scheme을 적용하여 의미망('Semantic Hypernetwork')으로 학습시킨 후 문서의 구성 단어들과 그 집합들 간의 연결 상태를 파악하였다. 우리는 의미적으로 그 표상이 분명하게 나뉘는 문서들(뉴스와 시)을 대상으로 이 모델이 어떻게 이들의 의미적 명확성을 분류하는지 분석하였다. Neighboring frequency와 degeneracy, 이 두 속성이 언어구조에서의 의미망 기억과 학습 탐색 기제에 유의한 기질로서 제안될 수 있다. 본 연구의 주요 결과로 1) 텍스트의 의미론적 투명성을 구별하는 통계적 증거와, 2) 문서의 의미구조에 대한 새로운 기질 발견, 3) 기존의 문서의 카테고리 별 분류와는 다른 방식의 분류 방식 제안을 들 수 있다.

1. 서 론: 의미적 명확성Semantic Transparency

"Colorless green dreams sleep furiously." Noam Chomsky는 이 무의미(nonsense) 문장을 예로 들면서, 문장이 문법적으로 정확하게 만들어져 있다 하더라도 여전히 의미를 전달하지 못할 수 있음을 보였다.[1] 이 문장은 syntax 측면에서 보면 완벽한 문장이지만, 의미론적으로 보면 전혀 무의미하다는 것이다. 즉 syntax와 semantics의 처리 층위는 분명히 다르고, 거기에 따르는 규칙들은 분명히 따로 존재해야 언어구조가 확립된다는 것이다.

그러나 이 문장이 그의 말처럼 인간의 언어처리과정에서 의미를 전혀 불러일으키지 못하는 것은 아니다. 이 기념비적인 문장은 지금까지도 많은 논의를 불러일으키고 있다. 예를 들어 Chomsky의 문장을 토대로 조작된 다음의 문장들을 보라.

"Furious dog dreams about colorless green."
"Colorless green dreams sleep furiously."
"Dreams green furiously colorless sleep."
"Green green furiously colorless colorless."

이 문장들을 읽고 해석해보면, 아래로 내려갈수록 그 의미적 명확성이 점점 줄어들고, 그 해석(표상)은 점점 늘어난다. Chomsky는 생성문법 측면에서만 이 문장을 논했지, 독자가 특정 문장에 스스로 만들어 부착하는 표상의 구조적 측면에 대해서까지 논의한 것은 아니다.

문장의 의미적 유의성은 단순히 syntax와 semantics

의 규칙들을 적용하는 것으로 이루어지지 않는다. 대표적인 예로 무의미 시(nonsense poem)는 '의미없는 것들the meaningless'의 집합으로부터 의미가 떠오르는 현상을 가장 극명하게 보여준다. 반면 그 의미가 가장 뚜렷하게 정해져 있는 것은, 즉 문장과 그 문장이 표상하는 의미 사이의 사상mapping이 일대일인 문서는 뉴스일 것이다.

단어들은 여러 방식으로 상호작용한다. 어떤 단어들은 특정 단어들과 높은 확률로 동시에 나타난다. 각각의 단어들의 출현 빈도만으로는 단어들이 모인 텍스트의 성격을 파악할 수 없다. 만약 텍스트의 단어들의 위치가 아무렇게나 흐트러져 배치되어 있다면(scrambled) 그 텍스트 자체는 여전히 Zipf's law에 맞지만, 그 내용은 전혀 소통될 수 없다.[2]

2. Neighboring Frequency & Degeneracy

김준식 등(2007)은 뉴스 문서를 일종의 의미망으로 재구성하여 분석을 하고, 의미적으로 'well-structured'된 문서에서의 단어 연결의 확률통계적 양상에 대해 논했다.[3] 여기서 드러난 언어구조의 중요한 기질이 특정 단어가 가진 neighboring frequency와 degeneracy였다. 두 단어, 혹은 세 단어의 연쇄체들이 전체 문서 안에서 출현하고 짝지어지는 양상을 통계물리적 측면에서 보면 그 문서의 structure가 어느 정도 정합적(coherent)인지를 알 수 있었다. 여기서 우리는 단어간 결합이 높은 빈도로 일어나는 bigram의 경우 그렇지 않은 경우보다 더 좋은 문장구조를 이룰 수 있다는 것을 알 수 있다. 한편 너무 많은 동일한 bigram 연쇄체만으로 구성되는 문장이 되면 오히려 의미적으로 좋지 않은 문장구조를 형성하는 결과를 낳는다. 따라서 이 두 기질이 유비적으로 함의하는 것은, neighboring frequency는 단어간의 결합강도(affinity)를, degeneracy는 단어가 그 neighbor로 가질 수 있는 다양성(variety)이라는 것이다. 즉 이 둘은 의미판단의 기질이 될 수 있다.

이와 같이 growing 'semantic hypernetwork'로 의미망을 학습시켜 각기 의미적으로 그 투명성이 다른 문서들을 재구성한 후, 그 constituent들의 frequency와 degeneracy를 total group number, entropy, free energy, relative ranking 등을 그 변수로 이용하여 그 문서의 semantic transparency의 scaling measure로 사용할 수 있는지, 그 가능성을 연구하려 한다.

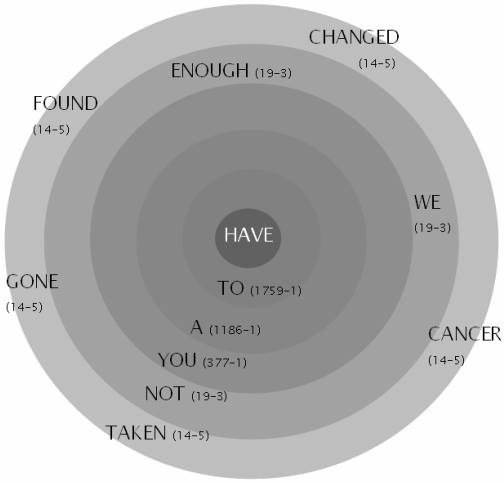

그림 1 단어 have와 연접 출현한 단어들의 neighboring frequency와 degeneracy. 괄호 안은 그 단어의 have와의 neighboring frequency와 degeneracy를 나타낸다. ('to'의 frequency는 1759, degeneracy는 1이고, 'changed' 'found' 'gone' 은 그 값들이 같다.) 중심 단어와 거리가 멀수록 frequency는 낮아지고 degeneracy는 높아진다.

3. 의미망으로서의 하이퍼네트워크

하이퍼네트워크 모델은 학습과 메모리에 관한 random graph 모델로서 가중치가 부여된 하이퍼그래프(hypergraph)로서 설명될 수 있다.[4] 하이퍼그래프를 구성하는 k개의 vertice로 이루어진 k-order 하이퍼에지는 그 edge의 가중치를 통해 하이퍼네트워크를 구성한 패턴과 그 성분(vertice)간의 k-order correlation을 나타낼 수 있다. 이미 하이퍼네트워크의 이러한 특성을 활용하여 다양한 분야에서 패턴 인식 및 완성에 관련된 실험[5,6,7]이 이루어진 바 있다.

부분문서(query)의 단어들을 따라가며 전체문서에서 그 단어 다음에 올 수 있는 모든 인접 단어를 조사하여 이를 그 빈도수(frequency)에 따라 분류한 그룹으로 나누었다. 우리는 다시 Markov Chain Monte Carlo 모델을 적용하여 부분문서의 단어 다음에 오는 인접 단어가 속한 그룹의 빈도수(frequency)와 그 축퇴도(degeneracy)를 살펴보았다. 여기서 축퇴도(degeneracy)란 같은 빈도수(frequency)를 가지는 모든 단어들의 개수를 의미한다. 빈도수(frequency)는 다음 단어와의 연결의 긴밀함을 나타내고 축퇴도(degeneracy)는 다양성을 나타낸다.

우리는 부분문서의 단어들의 전개에 따른 group들은 degeneracy와 frequency의 정보를 분석하여 보았고 이 다양성과 긴밀성을 나타내는 척도의 분포는 상당히 대칭적(symmetrical)임을 관찰하였다. 또한 단어 다음에 오는

그룹의 상대적 순위(relative ranking, group ranking/number of total groups)의 분포가 상위 부분과 하위 부분이 대칭적으로 높게 나옴을 관찰했다. 이는 상대적 순위(relative ranking)의 시계열 분포로부터 알 수 있듯이 중간 ranking에서 상대적으로 상위와 하위 ranking으로 주도적으로 전이되는 현상에 기인한다. 마지막으로 Gibbs free energy와 frequency 그리고 Gibbs free energy와 degeneracy 그래프로부터 frequency와 degeneracy가 부분적으로 깨짐(partial symmetry breaking) 현상을 관찰했다. 이는 높은 frequency를 가지는 단어들은 그 degeneracy가 1의 값만 가지지만 높은 degeneracy를 가지는 단어들은 1, 2, 3, 4, 5 등의 다섯 가지 이상의 frequency를 가지기 때문으로 분석된다. 이 현상으로부터 가능한 해석은, 이용(exploitation)은 한 가지 단어만을 계속하여 연결하는 것이 유리하고 탐색(exploration)은 여러 가지 경우를 가지는 것이 유리하기 때문이라는 점이다.

Hypernetwork 관점에서 분석한 인접한 단어들 간에도 관찰되었다. 즉, 빈도수(frequency)가 높은 단어의 축퇴도는 대체로 1이고 빈도수 (frequency)가 낮은 단어의 축퇴도(degeneracy)는 큰 값을 가진다. 그림 1에서 두 척도의 상관관계를 볼 수 있다.

여기서 degeneracy = exp(entropy/k) 와 frequency = exp(-enthalpy/kT) 의 열역학적 관계가 도입된다. k는 Boltzmann 상수이고 T는 절대 온도이다. 이로 볼 때, Gibbs free energy = enthalpy - T·entropy 이며 이는 $-kT$·ln(degeneracy · frequency) 로 주어진다. 위의 Gibbs free energy를 도입하여 부분문서(query)의 단어 간 진행을 Markov chain Monte Carlo (MCMC) 모델로 설명할 수 있다.

4. 실험 과정

실험 분석 과정은 김준식 등(2007)과 동일하다. 먼저 입력된 텍스트를 하이퍼네트워크로 재구성한다. 하이퍼네트워크는 order 2의 하이퍼에지(hyper edge)로 구성되어 있으며 이는 2개의 단어(feature)가 연결되었음을 표상한다. 하이퍼에지를 구성하는 2개의 단어는 문장을 기본 단위로 하여 연속되어 쪼개진다. 즉, "A B C"의 3개 단어로 이루어진 문장의 경우 "A B" 와 "B C" 두개의 하이퍼에지를 생성한다. 하이퍼에지(또는 library elements, 동일하게 사용)는 동일한 하이퍼에지가 많이 존재할수록 더 큰 weight를 갖게 된다. 즉 문장을 쪼개어 생성된 각각의 하이퍼에지는 기존에 동일한 하이퍼에지가 존재할 경우 그 하이퍼에지의 count를 하나(1)씩 증가시켜주며, 그렇지 않을 경우 새로운 하이퍼에지가 된다.

프로그램 상에서는 각 하이퍼에지마다 counter를 달아서 동일한 하이퍼에지가 입력될 경우 그 counter의 값을 1 증가시키도록 구현하였다. 전체문서(corpus)에 대하여 모든 문장을 하이퍼에지로 전환하면, 하이퍼네트워크가 구성된다. 구성된 하이퍼네트워크에 대하여 연속된 부분문서(query)를 입력하였다. query는 마찬가지로 문장 수준으로 들어오며, 이 문장을 initialization과 마찬가지로 order 2의 하이퍼에지로 변환하여 matching을 수행한다.

각 query 문장으로부터 생성된 하이퍼에지 각각은 만들어둔 하이퍼 네트워크와의 matching을 통해 corpus의 정보를 분석한다. query 하이퍼에지들은 자신의 첫 번째 vertex(단어)와 동일한 vertex를 포함하고 있는 하이퍼네트워크 상의 모든 하이퍼에지들을 검색하며, 그 하이퍼에지들의 다른 한쪽 vertex 정보를 각 query 별로 저장한다. 이때 저장되는 정보들은 각각의 query 하이퍼에지마다;

a. query 하이퍼에지와 첫 번째 vertex가 match되는 모든 하이퍼에지의 수 (total number of group)

b. query 하이퍼에지의 다른 한쪽 vertex를 포함하는 하이퍼에지가 속하는 group의 번호 (appointed group ranking)

* 각 그룹은 동일한 frequency (= count)를 갖는 하이퍼 네트워크상의 하이퍼에지의 집합으로 구성된다.

c. b group의 원소의 개수 (degeneracy)

d. b group의 frequency이다.

5. 실험 데이터

하이퍼넷에 입력되는 텍스트의 종류는 다음과 같다.

1. AP News 실험 문서로는 TIPSTER VOL1의 AP data를 썼으며 전체문서(corpus)는 2.2MByte이고 이중 처음의 56KByte 분량을 복사하여 부분문서(query)로 사용하였다.

2. 특정 작가의 무의미 시(nonsense poem). ('Tender Buttons' by Gertrude Stein) 전체문서의 분량은 약 100KByte이다.

6. 분석 결과

6.1. Degeneracy & Frequency

단어 간의 이웃 빈도수(neighboring frequency)와 축

퇴도 (degeneracy)의 분포를 살펴본다. 의미적으로 정합적인 문서인 경우, 이웃 빈도수와 축퇴도는 서로 대칭적인 분포를 가진다.(그림 2) 즉 자연언어로 만들어진 유의미한 텍스트라면 단어 간 연결 상태는 빈도수가 높거나 혹은 축퇴도가 높거나 하는 대칭적인 분포를 가짐을 알 수 있다. 즉 스케일을 무시하면 $y=x$ 직선에 대해서 대칭이다. 뉴스와 시 텍스트 둘 다 같은 양상을 보인다.

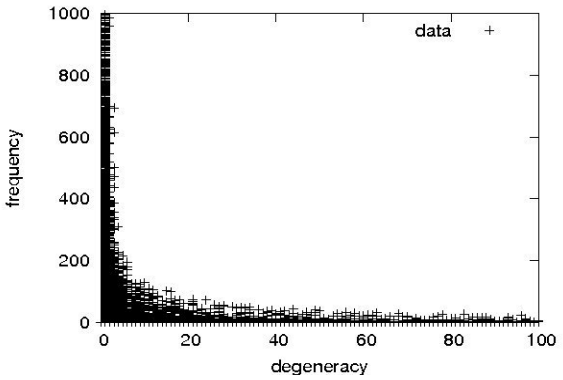

그림 2.1. AP 뉴스의 이웃빈도수와 축퇴도

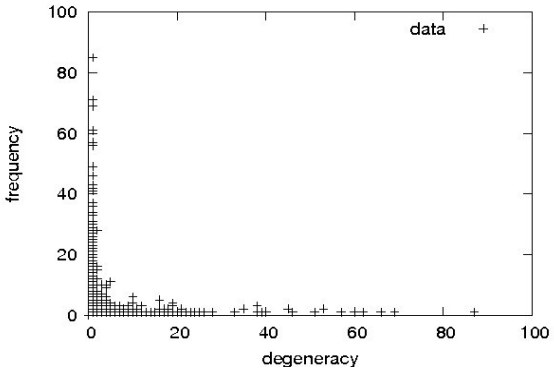

그림 2.2. *Tender Buttons*의 이웃빈도수와 축퇴도

6.2. Degeneracy와 Number (log-log)

그림 3은 data 들을 degeneracy 값들에 따른 도수 분포표이다. x, y 값은 자연로그를 취한 값들이다. 데이터들이 선형적 감소를 보인다. 이 결과는 degeneracy profile이 power law를 따름을 보인다.

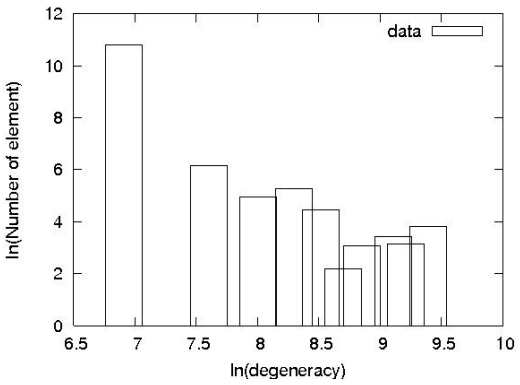

그림 3.1. AP 뉴스의 축퇴도 양상

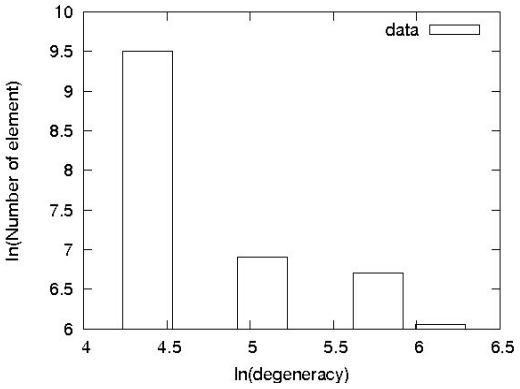

그림 3.2. *Tender Buttons*의 축퇴도 양상

6.3. Frequency & Number (log-log)

그림 4는 frequency 값들에 따른 도수 분포이다. 역시 x, y 값들은 자연로그를 취한 값들이다. 처음 4개의 데이터에서 선형적 감소를 볼 수 있고 그 옆에 두 개의 독립된 data들을 관찰할 수 있다.

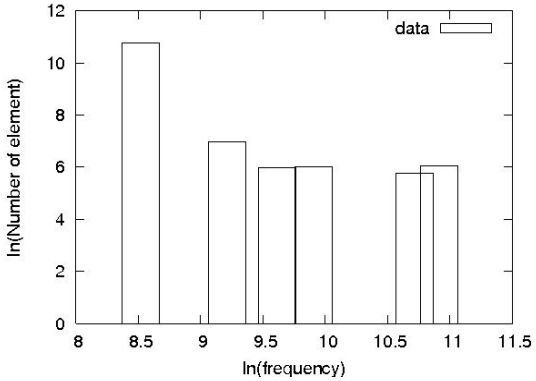

그림 4.1. AP 뉴스의 이웃빈도수 양상

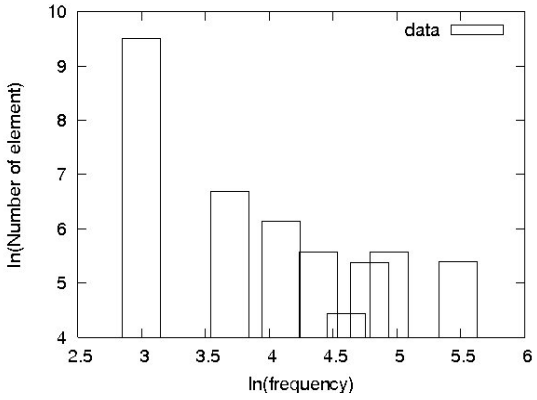

그림 4.2. *Tender Buttons*의 이웃 빈도수 양상

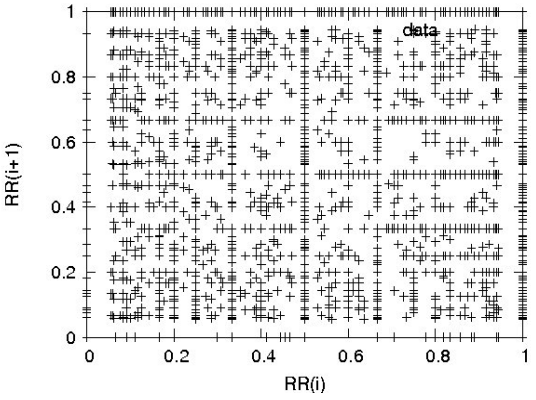

그림 5.2. *Tender Buttons*의 Relative Ranking

6.4. Relative Ranking dynamics

이웃하는 단어의 그룹(group)의 상대 순위(relative ranking)의 전이 그림. 그룹은 빈도수 (frequency)가 같은 모임을 뜻하며 앞 단어 다음에 올 수 있는 단어들을 그 빈도수 기준으로 그룹을 나누고 실제 부분문서(query)에서 선택되어지는 뒤 단어의 그룹 순위를 구한다. 그림 5에서 x 축은 i 번째 단어의 그룹순위를 전체 그룹의 개수로 나눈 상대 순위를 나타내며 y 축은 그 다음에 연결되는 (i+1) 번째 단어의 상대 그룹 순위이다. 이 그림에서 x 값이 0.1 이하이거나 0.9 이상이면 단어 연결은 모든 상대 순위 그룹으로 전이하지만 그 사이에서는 상위 그룹 (y 값이 0.2 이하) 혹은 하위 그룹 (y 값이 0.8 이상) 으로 우세하게 전이됨을 볼 수 있다. 실제로 다음 그림 5는 이를 잘 보여준다.

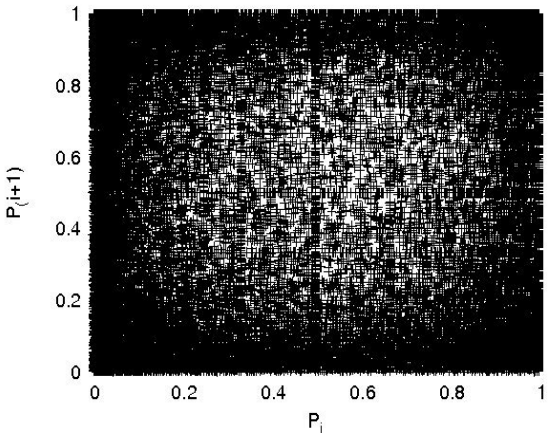

그림 5.1. AP 뉴스의 Relative Ranking

또한 이 Ranking dynamics로 두 데이터 사이의 차이를 볼 수 있다. 그룹의 상대 순위의 도수 분포를 나타낸다. 0.2 이하의 상위 그룹과 0.8 이상의 하위 그룹에 데이터가 몰려 있음을 볼 수 있다. 상위 그룹은 하위 그룹에 비해 이웃빈도수가 높고 축퇴도가 낮으며 하위 그룹은 그 반대이다. 이 그림에서 우리는 빈도수와 축퇴도사이의 대칭성을 볼 수 있다. 즉 x=0.5 직선을 중심으로 좌우 대칭이다. 그러나 무의미 시에서는 하위 그룹에만 집중되는 양상을 보인다.

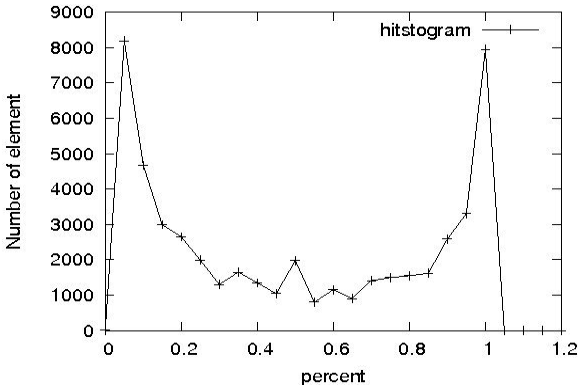

그림 6.1. AP 뉴스의 Relative Ranking histogram

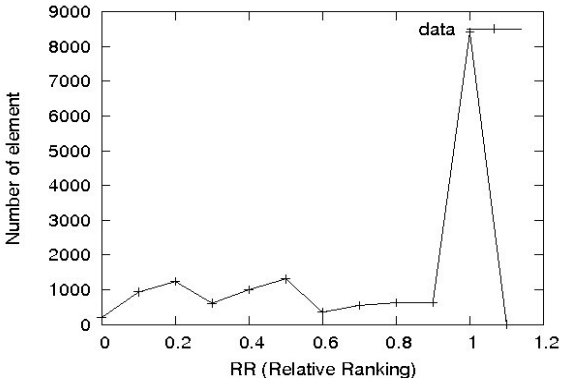

그림 6.2. *Tender Buttons*의 Relative Ranking histogram

7. 결론

어떤 자연언어 문서가 전달하려는 의미는 그 텍스트의 성격에 따라 다양하고, 이 다양성을 하나의 스펙트럼으로 정량화하여 표현할 수 있을 것이라는 것이 우리의 가정이다. 이러한 '의미의 명확성(semantic transparency)'은 단어들의 연쇄(word association)체들을 네트워크로 구조화시켜 형성된 망의 확률통계적 기질들을 보고 판단 가능하다는 것이다. 이것이 가능하다면 요약에서 언급한 1) 텍스트의 의미론적 명확성을 구별하는 통계적 증거와, 2) 문서의 의미구조에 대한 새로운 기질 발견, 3) 기존의 문서의 카테고리별 분류와는 다른 방식의 분류 방식 제안 등이 가능할 것이다.

이에 따라 본 연구는 자연언어 텍스트를 'Semantic Hypernetwork'로 growing시키고, 의미적으로 그 표상이 분명하게 나뉘는 문서들(뉴스와 시)을 대상으로 이 모델이 어떻게 이들의 의미적 명확성을 분류하는지 분석하였다. 이후 단어 연쇄체들의 neighboring frequency와 degeneracy, 이 두 속성을 중심으로 total group number, entropy, Gibb's free energy, relative ranking 등을 그 변수로 이용하였다. 그리고 이것들이 문서의 semantic transparency에 대한 scaling measure로 사용 가능한지 그 여부를 탐색했다. 현재는 미리 상정한 데이터들을 충분히 분석하지 못했으나, 각 결과 비교를 보면, 언어구조에서의 의미망 기억과 학습 탐색 기제에 유의한 기질로서 이 두 속성이 제안될 수 있다는 우리의 가정을 뒷받침한다. 차후 더 방대한 데이터를 분석하고 이에 따라 기존의 의미망 모델들과의 검증 및 비교를 수행한다면 본 연구의 지향점이 더 명확해질 것이라고 생각한다.

우리는 본 연구를 통해 자연언어 의미론(semantics for natural languages)을 표상하는 망 구조를 지배하는 보편적인 원칙들이 실제로 존재하며, 이 원칙들은 잠재적으로 언어 의미 구조의 변화, 발달, 학습 과정 등에 대해 중요한 함의를 갖고 있다는 점을 말하고자 한다. 물론 이 원칙들이 자연언어 의미론의 순수 이론을 제공하기 위해 존재하는 것이 아님은 분명하다. 단어들의 관계들로 이루어진 망 구조만으로는 의미론 구조의 가장 핵심적이고 중요한 단면을 모두 다 반영할 수 없는 것도 사실이다. 우리가 기대하는 것은, 기존의 언어 자료로부터 형성된 의미망이 언어구조에 대한 주요 이론들에 있어 어떤 특정한 역할을 한다는 것이다. 따라서 이 연구의 지향점은 의미망에 내재한 어떤 특성들을 고찰하고 이를 통해 의미론 연구의 기반작업에 기여하는 것이 될 것이다.

8. 참고문헌

1. N. Chomsky, *Syntactic Structures,* Mouton, 1957.
2. Wentian Li, *Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution*, IEEE Transactions on Information Theory, 38(6), 1842-1845 (1992).
3. 김준식, 박찬훈, 장병탁, 하이퍼네트워크 관점에서 본 문서에서의 단어간 긴밀성과 다양성의 대칭성, 한국컴퓨터종합학술대회 2007 논문집, 제34권 1(A), 31-32, 2007.06.
4. C. Berge, *Graphs and Hypergraphs,* North-Holland Publishing, Amsterdam, 1973.
5. B.-T. Zhang and J.-K Kim, DNA hypernetworks for information storage and retrieval, Lecture Notes in Computer Science, DNA12, 4287, 298--307, (2006).
6 S. Kim, M.-O. Heo, and B.-T. Zhang, Text classifier evolved on a simulated DNA computer, IEEE Congress on Evolutionary Computation (CEC 2006), 9196--9202, (2006).
7. J.-W. Ha, J.-H. Eom, S.-C. Kim. and B.-T. Zhang, Evolutionary hypernetwork models for aptamer-based cardiovascular disease diagnosis, The Genetic and Evolutionary Computation Conference (GECCO 2007), Vol. 4, 2709--2716, (2007).
8. E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabasi, Hierarchical organization of modularity in metabolic networks, Science 297, 1551--1555, (2002).
9. C. Furusawa, Zipf's law in gene expression, Physical Review Letters 90 (8), 088102, (2003).
10. A.-L. Barabasi and R. Albert, Emergence of scaling in random networks, Science 286, 509--512, (1999).
11. S. Goldwater, T.L. Griffiths, and M. Johnson, Interpolating between types and tokens by estimating power-law generators, Advances in Neural Information Processing Systems 18, 459--466, (2006).
12. K.E. Kechedzhi, O.V. Usatenko, and V.A. Yampol'skii, Rank distribution of words in correlated symbolic systems and the Zipf law, Physical Review E 72, 046138, (2005).
13. K.S. Krane, Introductory nuclear physics, John Wiley & Sons, Inc, 1988.
14. S. Maslov, M. Paczuski, and P. Bak, Avanlanches and 1/f noise in evolution and growth models, Physical Review Letters 73 (16), 2162--2165.