개념분류기법을 적용한 한국어 명사분류

정연수*○ 조정미* 김길창*
* 한국과학기술원 전산학과
전자우편 : {yeonsu|jmcho} @csone.kaist.ac.kr
전자우편 : gckim@csking.kaist.ac.kr

Korean Noun Clustering Via Incremental Conceptual Clustering

Yeon Su Jung*○ Jeong Mi Cho* Gil Chang Kim*
* Dept. of Computer Science, KAIST

요 약

많은 언어관계들이 의미적으로 유사한 단어들의 집합에 적용된다. 그러므로 단어들을 의미가 비슷한 것들의 집합으로 분류하는 것은 아주 유용한 일이다.

본 논문에서는 말뭉치로부터의 동사와 명사의 분포정보를 이용하여 명사들을 분류하고자 한다.

한국어에서는 명사마다 문장에서 그 명사를 특정한 격으로 사용할 수 있는 동사들이 제한되어 있다. 그러므로 본 논문에서는 말뭉치에서 나타나는 명사와 그 명사를 특정한 격으로 사용하는 동사들의 분포정보로부터 명사들을 분류하는 방법을 제시한다.

형태소 해석된 50만 단어 말뭉치에서 가장 빈도수가 높은 명사 85단어를 대상으로 실험하였다. 명사와 동사의 구문정보를 사용하므로 의미적으로는 다르지만 쓰임이 비슷한 단어들도 같은 부류로 분류되었다. 의미적으로 애매성을 가지는 명사들의 경우도 실험결과를 나쁘게하는 요인이 되었다. 그리고, 좀더 좋은 결과를 얻기 위해서는 동사들도 의미가 유사한 것들로 분류한 후, 명사와 동사의 분포정보가 아닌 명사와 동사들의 집합의 분포정보를 이용하는 것도 좋은 방법이 될 것이다.

1. 서론

자연언어처리에서는 단어자체보다 의미적으로 유사한 단어들의 집합을 사용하는 것이 유용한 경우가 많다. 예를 들면, 수식어들은 의미적으로 유사한 명사들을 선택하고, 의미해석에서 사용되는 선택제한도 단어들의 의미 클래스에 의해서 표현되며, 또한 의미 타입은 명사 합성의 가능성을 제한한다.[Hindle 1990]

특히, 통계적인 자연언어 처리에서는 말뭉치에 나타나는 분포정보를 이용하는 경우가 많은데, 미리 유사한 단어끼리 분류해 놓으면 단어자체에 대한 분포정보 대신 단어의 클래스에 대한 분포정보를 이용할 수 있으므로 처리해야하는 자료의 양을 엄청나게 줄일 수 있다.

그러므로, 단어들을 의미가 유사한 것들의 집합으로 분류하는 것은 많은 유용성을 가진다.

말뭉치에 나타나는 단어들의 분포 정보는 통계적인 자연언어 처리에 있어서 매우 유용하다. 말뭉치내에서의 단어의 분포정보로는 단순히 위치적으로 인접해서 나타나는 단어들에 대한 정보만을 사용하기도 하고, 특정 구문 관계로 나타나는 단어들에 대한 정보를 사용하기도 한다.

자연언어에서는 문장구조내에 특정한 구문관계로 나타나는 단어들에 대해서 제한이 있다. 예를 들어, 이동의 의미를 가진 동사들은 항상 위치나 방향을 나타내는 명사들과 같이 나타나야한다는 등의 제한이 있다.[Hindle 1990]

또한, 한국어 문장은 영어와 달리 조사에 의해서 구문적인 성질이 많이 결정된다. 특히, 격조사(가/이,을/를 등)의 경우는 조사만 보고도 동사와 명사사이의 구문관계를 알 수 있다.[윤길배 1986]

본 연구* 에서는 형태소 해석된 말뭉치로부터 조사들을 통해서 여러가지 구문관계로 나타나는 각 명사와 동사들의 분포정보를 구하고, 이 정보들을 이용하여 개념분류기법으로 명사들을 분류하고자 한다.

2. 관련연구

단어의 분포정보를 이용하여 단어를 분류한 연구에 대해서 간략하게 살펴보자. [Brown 1992]은 말뭉치내에서 서로 인접하여 나타나는 단어들의 Mutual Information을 이용하여 단어를 분류하였다.

[Hindle 1990]은 말뭉치내의 predicate-argument 구조로부터 명사들을 분류하였다. 여기서는 먼저 구문분석기를 이용하여 말뭉치에서부터 서술어-주어 관계와 서술어-목적어 관계에 있는 명사와 동사들의 Mutual Information을 구한 후, 이 정보를 바탕으로 명사들간의

*본 연구는 과학재단의 목적 기초 과제 "한국어 이해에 나타나는 중의성 문제 처리 모델 에 관한 연구"의 부분 지원을 받은 것입니다.

유사정도를 파악하여 명사를 분류하였다.

[Pereira 1993]도 역시 말뭉치로부터 동사와 목적어 관계에 있는 명사들을 추출하여 명사를 분류하였다. 여기서는 각 동사가 어떤 명사를 직접 목적어로 취하는지의 여부로써 그 명사의 context vector를 구한 다음, 이 정보로부터 명사들간의 상대 엔트로피를 구하여 명사들을 분류하였다.

[Grefenstette 1993]도 비슷한 실험을 했으나 그는 직접목적어 뿐만 아니라 주어, 간접 목적어 등 각 명사와 구문관계를 가지는 모든 다른 단어와의 정보를 이용하여 유사 명사를 추출하는 실험을 하였다.

본 논문에서는 형태소해석된 말뭉치내에서 주어-서술어, 목적어-서술어등의 여러가지 구문관계에 있는 명사-동사의 분포정보를 이용하여 명사들을 분류하였다.

3. COBWEB/3

개념분류기법[Michalski 1983]에는 여러가지가 있지만, 본 논문에서는 COBWEB/3라는 개념분류기법을 사용한 시스템을 이용하기로 한다.

COBWEB/3는 여러가지 속성들과 그 값의 집합으로써 표현가능한 실례(instance)들을 비슷한 속성을 공유하는 클래스들로 분류해주는 시스템이다.

COBWEB/3는 다음과 같은 특징을 가진다.

1. **계층적(hierarchical) 개념구성** 단순히 유사한 실례들의 클래스만 형성하는 것이 아니라 개념계층을 만든다.
2. **하향식(top-down) 분류** 새로운 실례를 분류할 때 제일 먼저 최상위계층으로 분류한 후 점점 하위계층으로 분류한다. 즉, 가장 일반적인 개념으로 분류한 후에 조금씩 특수한 개념으로 분류한다.
3. **비교사(unsupervised) 학습** 클래스에 관한 정보는 전혀 입력받지 않고 단지 입력으로 들어온 실례들의 특징으로 분류한다.
4. **점층적(incremental) 학습** 처음부터 여러 실례를 한꺼번에 분류하거나 이미 분류한 실례를 다시 분류하지 않고, 한번에 하나의 실례만 분류한다.
5. **hill climbing** 새로운 실례를 기반으로 현재 가지고 있는 개념계층에 수정을 가하기 때문에 hill-climing 검색의 특성을 갖는다.

3.1 자료의 표현과 연산자

COBWEB/3에서 사용하는 입력과 개념계층은 각각 그림 1과 그림 2와 같이 표현된다.

그림 1은 테니스공을 (지름, 색, 튀어오르는 정도)의 속성들로 표현한 예이며, 여기서 보인 실례는 각각의 속성에 대하여 (중, 녹색, 높다) (가), (2.52, 3, 0.71) (나)의 값을 가진다.

지름	중
색	녹색
튀어오르는정도	높다

(가) 명목적 속성

지름	2.52
색	3
튀어오르는 정도	0.71

(나) 숫적 속성

그림 1: 테니스공에 대한 입력자료

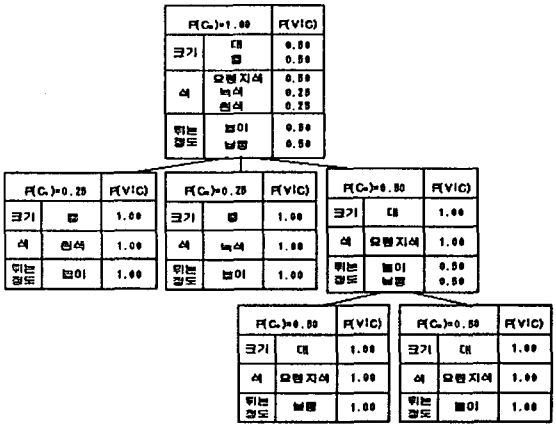

그림 2: 테니스공에 대한 개념계층의 예

COBWEB/3에서는 개념계층에 새로운 실례을 추가하기 위해서 4가지의 연산자를 사용한다.

그림 3는 이러한 4가지 연산자를 설명한다.

새로운 실례가 기존의 개념과 잘 맞을 때는 **병합연산자**를 적용한다. 병합연산자는 실례를 현재노드의 자녀노드에 추가한다. 그림 3 (가).

만약 이 자녀노드가 말단노드(terminal node)일 경우는 기존의 실례와 새로운 실례를 모두 포함하는 노드를 만든 후 두 실례를 새로운 노드의 자녀노드로 나타낸다. 그림 3 (나).

새로운 실례가 기존의 개념과 아주 다를 경우에는 **생성연산자**를 적용한다. 생성연산자는 현재노드에 새로운 새로운 실례를 표현하는 새로운 자녀노드를 만든다. 그림 3 (다).

COBWEB/3는 시스템에 기존의 개념계층의 구조를 어느 정도 수정할 수 있도록 하기 위해서 다음의 두 연산자를 추가로 가진다.

만약 기존의 계층이 과분기(overly branched)되었고, 두개의 클래스를 합친 개념이 새로운 실례와 더 잘 맞을 경우는 **융합연산자**를 적용한다. 융합연산자는 두 개의 자녀노드의 개념을 모두 포함하는 새로운 부모노드를 만든 후 실례를 이 노드에 병합한다. 그림 3 (라).

너무 일반적인 개념을 여러 개념으로 나누는 것이 새로운 실례와 더 잘 맞을 때는 **분할연산자**를 적용한다. 분할연산자는 현재노드를 제거하고 이것을 자녀노드들로 바꾼

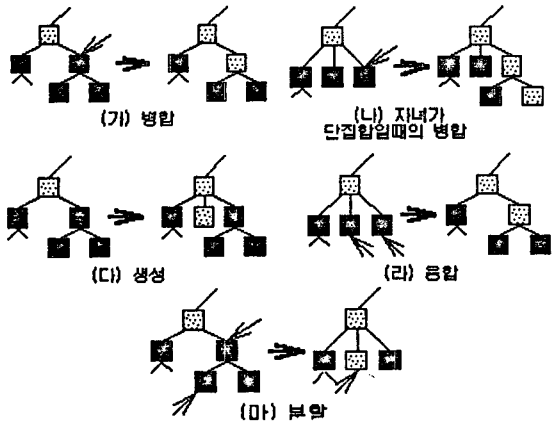

그림 3: COBWEB/3의 연산자

후 실례를 이 노드들 중 하나에 병합한다. 그림 3 (마).

3.2 평가함수

COBWEB/3는 4개의 연산자 중 가장 좋은 결과를 만드는 연산자를 선택하기 위해 category utility [Gluck 1985]라는 평가함수를 사용한다.

category utility는 생성된 계층이 잘 분류되어 있는지를 평가하기 위해서 클래스내유사도(intra-class similarities)와 클래스간상이도(inter-class differences)를 사용한다. 클래스내유사도란 같은 클래스내에서 구성원들이 가지는 속성들간의 유사한 정도($P(A_i = V_{ij}|C_k)$)를 말하고, 클래스간상이도란 서로 다른 클래스의 구성원들이 가지는 속성들간의 차이의 정도($P(C_k|A_i = V_{ij})$)를 말한다.†

클래스내유사도와 클래스간상이도가 클수록 좋은 계층이라고 말할 수 있으므로 category utility는 클래스내유사도와 클래스간상이도를 곱한 값의 기대값을 클래스로 나눈 값을 취한다.

이를 수식으로 나타내면, 먼저 기대값은

$$\sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i = V_{ij})P(C_k|A_i = V_{ij})P(A_i = V_{ij}|C_k) \quad (1)$$

이다. 여기서 $P(A_i = V_{ij})P(C_k|A_i = V_{ij}) = P(C_k)P(A_i = V_{ij}|C_k)$이므로, 식 (1)은

$$\sum_{k=1}^{K} P(C_k)\sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i = V_{ij}|C_k)^2 \quad (2)$$

와 같이 된다. 그런데, 이 기대값 중에는 클래스에 관계없이 일정한 값이 존재한다. 이 값은

$$\sum_{i=1}^{I}\sum_{j=1}^{J} P(A_i = V_{ij})^2 \quad (3)$$

이다.

†여기서 A_i는 i번째 속성, V_{ij}는 i번째 속성의 j번째 속성값, C_k는 k번째 클래스를 말한다.

그러므로, 클래스정보에 의존하는 기대값은 식 (2)에서 식 (3)을 뺀 값이므로, category utility는

$$\frac{\sum_k P(C_k)\sum_i\sum_j P(A_i = V_{ij}|C_k)^2 - \sum_i\sum_j P(A_i = V_{ij})^2}{K} \quad (4)$$

이다.

3.3 알고리즘

COBWEB/3의 알고리즘은 4단계로 이루어진다.

하향식 학습을 하므로 제일 먼저 루트노드에서 시작하여 말단노드에 이를 때까지 같은 과정을 반복하여 수행한다.

그 과정을 살펴보면 다음과 같다.

시작. 실례를 루트노드에 병합한 후 루트노드를 현재노드로 한다.

단계1. 현재노드에서 어떤 연산을 수행하는 것이 가장 좋은 개념계층을 얻을 수 있는 지를 평가함수를 사용하여 조사한다.

단계2. 단계1에서 얻어진 연산자를 실제로 개념계층에 수행하여 개념계층에 실례를 추가한다.

단계3. 단계2에서 새로운 실례를 표현한 노드가 말단노드이면 수행을 마치고 그렇지 않으면 그 노드를 현재노드로하여 단계1부터 다서 반복한다.

이상으로 COBWEB/3에 대한 간단한 소개를 마친다. 좀더 자세한 내용은 [McKusick 1990]와 [Fisher 1987]를 참조하기 바란다.

4. 개념분류기법의 적용

자연언어에서는 같은 문장내에 어떤 단어들이 함께 나타날 수 있는가에 대해 제한이 있다. 특히, 어떤 서술어의 성분들로는 어떤 명사가 올 수 있고, 어떤 명사가 올 수 없는지의 제한이 존재한다. 즉, 모든 명사들에 대해 그 명사를 목적어로 가질 수 있는 동사, 혹은 주어로 가질 수 있는 동사들의 제한된 집합이 있다. 예를 들면, 명사 '미소'를 목적어로 가지는 동사들은 '짓다','띄우다','보내다'등이 있으며, 명사 '모습'을 주어로 가지는 동사들은 '나타나다','보이다', '나오다'등이 있다.

또한, 한국어 문장은 영어와 달리 조사에 의해서 구문적인 성질이 많이 결정된다. 특히, 격조사(가/이,을/를 등)의 경우는 조사만 보고도 동사와 명사사이에 나타나는 구문관계를 어느정도 예측할 수 있다.

본 논문에서는 형태소 해석된 말뭉치로부터 '을/를', '이/가', '에', '으로/로', '에게', '에서' 관계에 있는 명사-동사들의 분포정보를 이용하여 명사의 속성을 표현한 후 개념분류기법으로 명사들을 분류하였다. 표 1은 명사 '말'과 '일'의 속성의 예이다.

본 논문에서 명사를 분류한 알고리즘은 다음과 같다.

명사	을/를			이/가		
	보다	받다	듣다	놀리다	오다	들다
말	-20.00	-2.91	3.82	-3.14	-0.42	-20.00
일	-0.05	-20.00	-1.92	-20.00	-20.00	-0.66

표 1: 명사 '말'과 '일'의 속성

1. 형태소 해석된 말뭉치로부터 '을/를', '이/가', '에', '으로/로', '에게', '에서' 관계에 있는 명사-동사의 분포정보를 추출한다.
2. 명사-동사의 분포정보로 부터 명사들의 속성을 표현한다.
3. COBWEB/3로 명사들을 분류한다.

4.1 분포정보의 추출

본 논문에서 사용하는 말뭉치는 구문분석된 결과가 아니라 형태소해석만 되어 있다. 그래서, 말뭉치로부터 분포정보를 추출할 때 몇가지 가정을 하고 간단한 구문분석기를 이용해서 명사와 동사사이의 구문 관계를 분석한다.

가정1. 모든 명사들은 가장 먼저 나오는 동사를 포함한 절에만 포함된다.

가정2. 형용사가 나오기 전까지 나온 명사들은 무시한다.

가정3. 명사전성어미 'ㅁ' 이나 '것' 을 포함한 절은 무시한다.

가정1.에 에 의하면 내포문을 포함한 문장일 경우 오류를 범하게 되는데, 예를 들면 "철수가 달리는 차에서 뛰어내렸다"와 같은 문장의 경우 '철수'가 '달리다'의 주어로 잘못 분석된다.

그러나, 형태소해석된 결과만으로 자동으로 모든 동사와 명사사이의 구문관계를 정확하게 분석하기는 사실 거의 불가능하다. 더우기 실제로 이와 같이 가정을 하고 실험을 했을 경우도 그러한 오류가 나타나는 확률이 매우 적었다. 그러므로, 본 논문에서는 이러한 오류는 무시하기로 한다.

본 논문에서는 명사의 속성을 표현하기 위해 이와 같은 방법으로 얻어진 동사와 명사사이의 구문관계들로부터 '을/를', '이/가', '에', '으로/로', '에게', '에서' 관계에 있는 명사-동사의 단어쌍들의 빈도수와 각 명사들의 빈도수, 각 동사들의 빈도수, 그리고 말뭉치에서 추출된 절의 수 등을 구한다.

4.2 명사의 속성 표현방법

본 논문에서는 명사의 속성을 표현하기위해 말뭉치에서 각 명사와 '을/를', '이/가', '에', '으로/로', '에게', '에서' 관계에 있는 동사들과의 Mutual Information(MI) [Fano 1961]을 명사의 속성으로 표현하였다.

명사와 $josa$관계에 있는 동사의 Mutual Information은 다음 식을 통해서 구한다.

$$I_{josa}(nv) = \log_2 \frac{\frac{f_{josa}(nv)}{N}}{\frac{f(n)}{N}\frac{f(v)}{N}} \quad (5)$$

식 (5)에서 $f_{josa}(nv)$는 말뭉치에서 명사 n이 동사 v와 $josa$ 관계로 나타나는 빈도수이며, $f(n)$과 $f(v)$는 각각 말뭉치에서의 명사 n과 동사 v의 빈도수이고, N은 말뭉치에서 나타난 절(clause)의 수이다.

여기서 명사와 동사가 동시에 나타난 빈도수를 직접 명사의 속성으로 하지 않고 굳이 MI로 한 근거는 다음과 같다.

동사 '하다'나 '되다'의 경우는 빈도수가 굉장히 높을 뿐 아니라 거의 모든 명사와 같이 쓰일 수 있으므로 명사와 동시에 나타나는 빈도수도 다른 동사에 비해서 굉장히 높을 것이고, 그러므로 당연히 명사들을 분류하는 데 큰 영향을 끼칠 것이다. 그러나, 이러한 동사들은 쓰임이 굉장히 일반적이므로 실제로 분류를 할 때는 오히려 명사와의 관계가 밀접한 동사와의 정보가 중요하다.[Church 1989] 예를 들어, 동사 '하다' 와 '걸다'를 명사 '말'과 목적어관계에 있는 경우로 비교해보면 '하다'와 같이 나오는 경우가 47번으로 '걸다'의 5번에 비해 월등히 크다. 하지만 명사 '말'의 속성으로는 '걸다'가 '하다'보다 더 중요하다. 그래서, 빈도수보다 두 개의 사건이 연관된 정보의 양을 나타내는 MI를 명사의 속성으로 사용하였다. 실제로 앞의 예에서 명사 '말'과의 MI값은 '걸다'가 1.29로 '하다'의 0.15로 더 높게 나왔다.

5. 실험 및 분석

한국어 명사 85개를 대상으로 하여 명사분류실험을 하였다. 실험에 사용한 말뭉치는 50만 단어의 형태소해석된 것으로, 교과서, 소설, 기사 등의 다양한 형태의 글로 구성되어 있다. 실험 대상으로 한 명사는 말뭉치로부터 빈도수가 높은 순으로 85개를 추출하였다. 그리고, 역시 말뭉치에서 명사들과 '을/를', '이/가', '에', '으로/로', '에게', '에서' 관계로 나타나는 빈도가 높은 동사를 각각의 조사에 대하여 15개씩 추출하였다. 이러한 명사 85개와 '을/를', '이/가', '에', '으로/로', '에게', '에서' 각각에 대한 동사 15개, 모두 90개 동사와의 Mutual Information을 구해 명사의 속성으로하여 실험하였다.

그림 4는 실험결과로 얻어진 단어들의 개념계층을 보여준다.

상위개념들간의 관계나 상위개념과 하위개념사이의 관계는 별로 의미를 가지고 있지 않다.

그러나, 대부분의 명사들이 의미가 유사한 것끼리 분류되었다. 특히, 표시된 것처럼 위치를 나타내는 명사와 사람을 나타내는 명사, 방향을 나타내는 명사, 소리정보를 나타내는 명사등은 아주 정확하게 분류되었다. 즉, 클래스간의 계층구조를 형성하기에는 조금 부족하지만 유사한

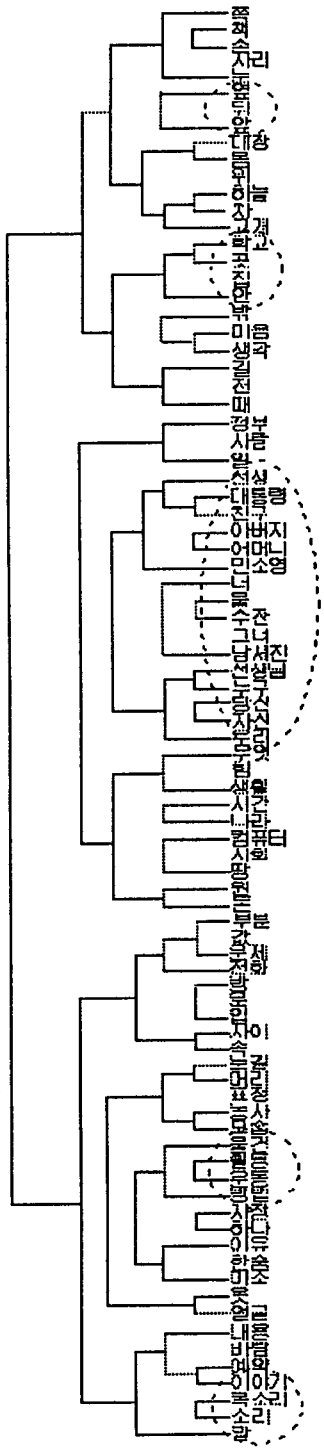

그림 4: 한국어 명사분류 실험결과

명사끼리의 클래스를 분류하는 데는 어느 정도 만족할 만한 결과를 보이고 있다.

6. 결론

본 연구는 말뭉치로부터의 분포정보에 개념분류기법을 적용하여 한국어 명사를 분류하였다.

분포정보로는 주어-서술어, 목적어-서술어등의 여러가지 관계로 나타나는 명사-동사쌍의 분포를 사용하였다.

본 논문에서는 명사와 특정동사와의 Mutual Information만을 이용했는데, 그 결과 빈도수가 작아 명사의 특성이 잘 나타나지 않는 부분들이 많았다. 그러므로, 특정동사뿐만이 아니라 동사들의 유사도를 측정하여 이와 유사한 동사와의 Mutual Information을 이용하는 방법도 생각해야할 필요가 있다.

본 논문은 약간의 개선해야 할 점이 있지만 명사를 의미에 따라 클래스를 나누는데 특정 구문관계로 나타나는 명사와 동사의 분포정보를 이용할 수 있음을 보여주고 있다.

앞으로 이렇게 분류한 명사를 실용적인 시스템에서 이용할 수 있게 하기 위해서 명사의 수를 좀더 확장하고, 자동분류된 명사들을 다른 전문가들이 쉽게 수정할 수 있도록 하는 도구를 만들 생각이다.

참고문헌

[윤길배 1986] 윤길배. 1986. 자연언어의 격분류에 관한 연구. 한국과학기술원 전산학과 석사학위논문.

[Brown 1992] Brown, P. F., Pietra, V. J. D., DeSouza, P. V., Lai, J. C., and Mercer, R.L. 1992. Class-based n-gram models of natural language. In *Computational Linguistics.*

[Church 1989] K. Church and P. Hanks. 1989. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Meeting of the Association for Computational Lin guistics.* Vancouver,B.C.

[Fano 1961] Fano, R. 1961. Transmission of Informationa. In *cambridge, Mass:MIT Press.*

[Fisher 1987] Fisher, D. 1987. Knowledge Acquisition Via Incremental Conceptual Clustering. In *Machine Learning (2): 139-172*

[Gluck 1985] Gluck, M., Corter, J. 1985. Information, uncertainty and the utility of categories. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society.*

[Grefenstette 1993] G. Grefenstette. 1993. SEXTANT: extracting semantics from raw text:

implementation details. In *Heuristics:The Journal of Knowledge Engineering.*

[Hindle 1990] Donald Hindle. 1990. Noun Classification From Predicate-Argument Structures. In *Proceedings of the 28st Meeting of the Association for Computational Lin guistics.*

[McKusick 1990] McKusick, K., K. Thompson. 1990. COBWEB/3: A Portable Implementation. In *Technical report FIA-90-6-18-2, NASA Ames Research Center*

[Michalski 1983] Michalski, R.S., Stepp, R. 1983. Learning from observation: Conceptual clustering. In *Machine learning: An artificial intelligence approach.*

[Pereira 1993] Fernando Pereira, Naftali Tishby and Lillian Lee. 1993. Distributional Clustering Of English Words. In *Proceedings of the 31st Meeting of the Association for Computational Lin guistics.* Columbus.