인과관계 문맥정보를 사용한 용어 군집화 연구

장두성
KT 서비스개발연구소 음성언어 연구팀,
한국과학기술원 전산학과
dschang@kt.co.kr

최기선
한국과학기술원 전산학과,
전문용어언어공학연구센터, 언어자원은행
kschoi@cs.kaist.ac.kr

Term Clustering based on Causal Context Information

Du-Seong Chang
Spoken Language Research Team KT,
Division of Computer Science KAIST

Key-Sun Choi
Division of Computer ScienceKAIST,
KORTERM, BOLA

요 약

단서구문 및 어휘 쌍 확률 등을 이용하면 일정한 영역의 문서에서 사용된 용어의 원인이 되거나 결과를 나타나는 관련어들을 찾을 수 있다. 본 논문에서는 이러한 각 용어의 선행 원인과 후행 결과를 인과관계 정보라고 정의한다. 인과관계 정보가 유사한 용어들은 서로 유사한 개념에 속한다고 가정한다면, 용어의 직/간접적 인과관계로서 용어 온톨로지에서 그 용어가 속할 집합을 결정하는 데 도움을 줄 수 있다. 본 논문에서는 각 용어의 인과관계가 용어 군집화를 위한 유용한 문맥 정보의 하나라는 것을 실험을 통해 증명하였다. 속성으로 사용된 인과관계는 대용량의 코퍼스로부터 비지도식 학습방법을 통해 자동 습득하였으며, 그 정확도는 74.84%를 보였다. 1659개 용어에 대한 군집화 실험 결과 70.02%의 정확도를 보였으며, 어휘 유사도만을 사용한 경우에 비해 32.9%의 적용도 향상을 보였다.

1. 서 론

인과관계(Causality)란 어떤 사실과 다른 사실 사이의 원인과 결과 관계를 말한다. 의료영역과 같은 일정한 영역에 사용되는 용어는 그 영역의 지식을 표현하는 기본 단위로서 많은 경우 어떠한 사건이나 상태, 혹은 사건을 일으키는 개체를 나타내며, 이 경우, 인과관계의 대상이 된다. 아래 문장 (1a~c)에서 'oral bacteria', 'cigarette', 'smokeless tobacco product'는 'gum disease'의 원인이 되는 용어로 쓰였으며, 'toothache'는 결과를 나타내기 위해 사용되었다.

(1a) "The oral bacteria that cause gum disease appear to be the culprit."

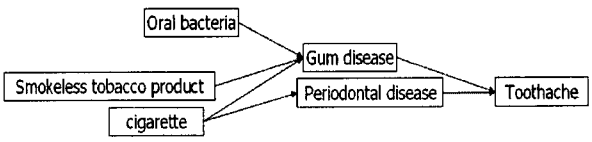
(1b) "Cigarette smoking and use of smokeless tobacco products may also cause gum disease"

(1c) "Gum disease is the second most common cause of toothache"

한 용어의 선행 원인과 후행 결과를 인과관계 정보라고 정의하며, 이는 코퍼스에서 인과관계를 가지는 명사구 쌍 혹은 용어 쌍을 추출하여 얻을 수 있다. 동일한 단어가 서로 다른 두 용어의 원인 혹은 결과를 표현할 때 이 두 용어는 인과관계의 일부를 공유한다고 정의한다. 아래의 문장 (2a~b)에서 'Periodontal disease'는 원인을 표현하는 용어로 'cigarette', 결과를 표현하는 용어로 'toothache'를 가진다. 이는 앞서 예를 든 'gum disease'의 인과관계 정보와 많은 부분은 공유하고 있으며, 실제 'Periodontal disease'는 'gum disease'의 동의어이다. 용어들의 인과관계 정보는 인과관계 용어 망으로 표현할 수 있다. 그림 1은 예제 (1), (2)에 나타난 용어들의 인과관계 용어 망이다.

(2a) "Periodontal disease can lead to toothache."

(2b) "Cigarette smoking is the number one environmental risk for periodontal disease."



[그림 1] 예제 (1), (2)의 인과관계 용어 망

실제 코퍼스에서 많은 수의 동의어들이 문장 내에서 유사한 쓰임새를 보이며, 이러한 현상에 비추어 두 용

이 연구의 일부분은 과학기술부와 과학재단의 지원에 의해 수행되었습니다

어가 많은 양의 인과관계를 공유하는 경우, 즉, 두 용어의 인과관계가 유사한 경우 이 두 용어는 서로 유사한 개념에 속한다고 가정할 수 있다. 실제 이러한 가정이 성립한다면, 용어의 직/간접적 인과관계로서 용어 온톨로지에서 그 용어가 속할 집합을 결정할 수 있다. 본 논문에서는 인과관계를 바탕으로 한 군집화 실험을 통해 유사한 인과관계를 가지는 용어는 유사한 개념에 속한다는 가정을 증명하고자 한다. 즉, 각 용어의 인과관계가 용어 군집화를 위한 유용하게 사용될 수 있는 문맥 정보라는 것을 보여서, 용어 온톨로지 구축을 위한 인과관계의 기여 가능성을 타진하고자 한다.

이를 위해 인과관계 분석 모델을 통해 생물의학 분야 코퍼스에 사용된 명사구 간 인과관계를 분석한다. 분석된 인과관계 명사구 쌍으로부터 인과관계 용어 쌍을 추출하고, 마지막으로 추출된 인과관계를 기반으로 용어 군집화 과정을 수행한다. 2장에서는 용어 군집화를 위해 기존 연구들에 대해 간략히 정리하고, 3장에서는 사용된 인과관계 분석 모델 및 용어 군집화 과정에 대해 소개한다. 4장에서는 평가를 위해 도입된 평가 함수들을 소개하고 군집화 결과를 보인다.

2. 관련 연구

용어 군집화는 유사한 용어들을 모아 몇 개의 군집으로 나누는 작업으로 용어의 개념화와 아주 밀접한 관계를 가지고 있다. 용어의 군집화를 위해 고려해야 할 부분은 용어의 유사도 속성 선택의 문제와 용어 간 유사도가 주어졌을 때 군집화 진행 알고리즘의 선택 문제이다.

용어 군집화를 위해서는 사용되고 있는 속성들은 크게 용어 내부 정보와 용어 외부 문맥 정보들이다. 용어 내부 정보로는 용어를 구성하는 주요 단어와 주변 단어 및 이들 간의 순서이며, 용어 외부 문맥 정보로는 용어가 문장에서 사용될 때 용어의 선후 위치에 존재하는 단어들, 문장 구조에서 용어의 상/하위에 존재하는 단어들로 구성된다. Bourigault와 Jacquemin(1999)은 용어를 구성하는 품사 나열 규칙만으로 군집화를 적용하여 93~98%의 정확도를 얻었다고 보고하고 있다[1]. Maynard와 Ananiadou(2000), Friedman 등(2001)과 같이 수동 작성된 시멘틱 프레임(semantic frame) 정보를 용어의 속성으로 사용한 시도도 있었다[3][6]. 여기에서는 용어의 어휘, 구조, 문맥 유사성을 측정하는 기준을 제시하고 이들 세 기준을 조합하여 용어 군집화를 수행한 Nenadic 등(2002)의 연구를 소개한다[9].

이 연구에서는 두 용어 간에 공유하고 있는 최상위 명사(head noun)와 수식어의 수에서 어휘 유사성을 추출하였으며, 문장 내에서 동격 구조가 사용되는 경우 여기에 포함된 용어들이 서로 구조 유사성이 있다고 판단하였다. 또한 용어를 추출하기 위해 사용된 주요한 패턴들의 가짓수를 파악하여 이들 간의 자카드 계수(Jaccard's coefficient)를 사용하여 문맥유사도를 계산하였다. 아래 수식 (1)은 두 용어, t1, t2간의 어휘 유사도로서, 식에서, Hi, Hj 는 두 용어 내 최상위 명사 집합, Mi, Mj 는 수식어 집합이다. a와 b는 가중치로서, a가 b보다 크게 설정되었다.

$$LS(t_i,t_j)=\frac{1}{a+b}\left(a\times|H_i\cap H_j|+b\times\frac{2|M_i\cap M_j|}{2|M_i\cap M_j|+|M_i\setminus M_j|+|M_j\setminus M_i|}\right) \quad (1)$$

예제 (3a)는 문장 내 동격구조의 예이며, (3b~c) 은 구조 유사성을 추출하기 위해 사용된 패턴의 일부이다.[1)]

(3a) steroid receptors such as estrogen receptor, glucocorticoid receptor, and progesterone receptor.

(3b) both 〈TERM〉 and 〈TERM〉

(3c) either 〈TERM〉 or 〈TERM〉

이 연구에서 군집화의 정확도는 63~71%로 보고되었다. 의료영역 코퍼스에서 세 유사도의 가중치를 지도식 학습한 결과 어휘유사도가 가장 큰 가중치(0.81)를 보였으며, 구문 유사도가 가장 적은 가중치(0.06)를 보였다[10].

용어의 외부 문맥 정보로 위치/구문적 주변 단어를 사용할 때 가장 근접한 한 두개의 단어들로서 용어의 정보를 대표하는 것이 일반적이다. 이러한 접근 방법은 구현과 계산의 단순성을 얻을 수 있는 장점이 있지만, 정보의 손실을 피할 수는 없다. 인과관계가 있는 두 용어 간의 위치/구문적 거리는 두 세 단어 이상이다. 인과관계 정보를 군집화의 속성으로 사용한다는 것은 문장 내에서 위치/구문적으로 비교적 먼 거리에 있는 관계 단어들로 표현되는 용어 정보를 군집화를 위해 사용할 수 있다는 의미이다. 문서의 지식을 용어간 관계의 합으로 표현할 수 있다면, 인과관계 이외의 지식을 표현하기 위해 사용될 수 있는 여러 관계들 (예를 들면, 시공간 정보, 조건관계 등) 역시 군집화를 위한 속성으로 시도될 수 있다. 본 논문에서는 이러한 시도의 첫 단계로 인과 관계를 문맥정보로 사용한 군집화를 시도

1) 여기에 사용된 예제는 (Nenadic 외, 2002)에서 인용하였다.

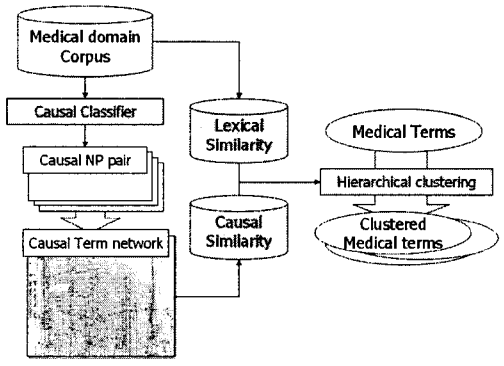
하고자 한다.

군집화 방법은 단계적 군집화 방법과 비단계적 군집화 방법으로 분류할 수 있으며, 비단계적 군집화 방법으로는 모든 용어에 대해 순서적으로 가장 가까운 용어와 군집화하는 Nearest Neighbor 방법이 대표적이다. Ward(1963)는 모든 용어를 각각 하나의 군집으로 초기화한 후 가장 유사한 두개의 군집부터 반복적으로 하나의 군집으로 통합하는 과정을 거쳐 원하는 개수의 군집을 얻는 방법을 제안하였다. 이러한 단계적 군집화 방법은 용어 간의 유사도로부터 군집간의 유사도를 얻는 방법에 따라 두 군집에 참여하는 용어 쌍의 평균 거리 측정 방법(Average link), 최소 거리 측정 방법(Single link), 최대 거리 측정 방법(Complete link) 등이 있다. Ward는 군집화 과정을 정보량 손실과정으로 보고 최소 정보량 손실을 유도하는 군집을 선택하였다[12].

3. 용어 군집화

본 논문에서는 (장두성 외, 2004)에 제시한 인과관계 분석 모델에 의해 분석된 인과관계 정보를 이용하여 용어 군집화를 시도하며, 여러 군집화 모델 중 가장 널리 사용되는 평균 거리 측정 방법을 이용하여 단계적 군집화를 시도한다. 그림 2는 제시하는 군집화 개요도이다. 군집화를 위해 용어의 어휘 유사도와 인과관계 유사도를 주요한 속성으로 사용하였으며, 이들 두 유사도의 조합으로부터 단계적 군집화를 위한 유사도 함수를 도출하였다.

용어간 인과관계 유사도는 인과관계 용어 망으로부터 얻어지며, 이를 위해 코퍼스에서 인과관계 명사구 쌍을 분석한다.

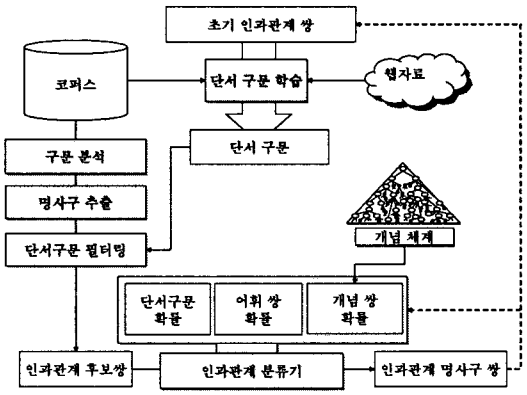


[그림 2] 군집화 방법

3.1. 인과관계 분석 모델

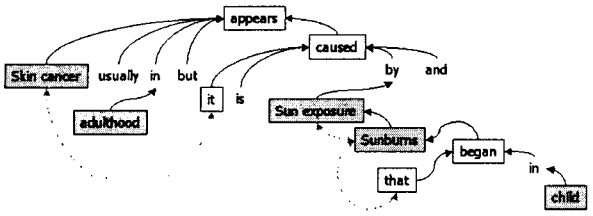
인과관계의 분석 문제는 두 인과관계 후보 명사구 쌍에 대하여 인과관계가 존재할 경우와 존재하지 않을 경우로 분류하는 문제로 해석할 수 있다. 인과관계의 분석 문제는 인과관계 후보 추출과 인과관계 분류의 두 단계로 이루어진다. 인과관계의 후보는 의존 구조에서 구문적으로 연결된 명사구 쌍을 추출하여 구성한다. 인과관계 분류를 위해 나이브 베이즈 분류기를 사용하며, 단서구문과 어휘 쌍 확률 등이 사용된다.

인과관계 분석기의 구성은 그림 3과 같다. 코퍼스에서 구문분석을 통해 사건으로 사용될 수 있는 명사구를 추출한다. 의존 구조에서 구조적으로 동사구에 의해 연결된 명사구는 후보 쌍으로 추출된다. 인과관계 분류기는 단서구문 및 어휘 쌍 확률을 이용하여 인과관계 분류를 하며, 분류된 인과관계 명사구 쌍은 단서구문과 어휘 쌍 확률을 재학습하기 위해 다시 사용된다.



[그림 3] 인과관계 분석기의 구성

'인과관계 후보'는 인과관계 분류의 입력이 되며, 원인-결과 후보 명사구 쌍과 두 후보 명사구를 연결하는 단서 구문의 3진 형태로 표현된다. 인과관계 후보는 〈원인 후보 명사구, 단서 구문, 결과 후보 명사구〉와 같이 표현된다. 그림 4는 사건이 명사구의 형태로 표현된 예문의 의존 구조이다. 이와 같은 의존 구조에서 인과관계의 후보를 추출하기 위해서는 동사구 및 명사구 추출, 명사 참조 해결, 동격 명사구 탐색, 3진 관계 추출 등의 단계를 거친다. 실험에서 의존구조는 Connexor 의존구조 파서[11]를 이용하여 생성하였다. 추출된 인과관계 후보의 명사구들은 인과관계 분류 후 분류 결과에 따라 원인 사건과 결과 사건, 혹은 인과관계에 관여하지 않는 경우로 결정된다.



[그림 4] 의존구조 예

인과관계 분류기는 인과관계 후보에 대해 인과관계 여부를 분류한다. 이러한 분류를 위해 단서구문 확률과 어휘 쌍 확률로 구성된 나이브 베이즈 분류기를 구성한다. 제안된 분류기는 코퍼스에서 비지도식 방식으로 학습된다. 단서구문 신뢰도 및 명사 의미 부류 순위 역시 인과관계 분류를 위해 사용 가능하다. 이들과 함께 사용하는 방법에 따라 세 가지의 분류 모델을 제안한다.

t_i가 문헌에서 추출된 인과관계 후보이고, y_i가 t_i의 인과관계 부류라 할 때, y_i의 값 c*는 수식 (2)와 같이 주어진 t_i에 대해 최대 확률을 갖는 c_j로 분류될 수 있다. c_j는 c_1(인과관계) 혹은 c_0(비인과관계)이다. 수식 (2)는 베이즈 규칙에 의해 수식 (3)과 같이 전개된다.

$$c^* = \arg\max_{c_j} P(y_i = c_j \mid t_i) \quad (2)$$

$$c^* = \arg\max_{c_j} \frac{P(c_j)P(t_i \mid c_j)}{P(t_i)} \quad (3)$$

인과관계 후보 t_i가 서로 독립적인 인과관계 속성들로 이루어져 있고, 인과관계 속성으로 단서구문 CP_{t_i}와 어휘 쌍 $LP_{t_i k}$만을 고려한다면, 수식 (3)에서 P(ti|cj)는 수식 (4)와 같이 쓰여질 수 있다.

$$P(t_i \mid c_j) = P(CP_{t_i} \mid c_j) \prod_{k=1}^{|t_i|} P(LP_{t_i k} \mid c_j) \quad (4)$$

수식 (4)에서 $P(CP_{t_i} \mid c_j)$와 $P(LP_{t_i k} \mid c_j)$는 각각 단서구문 확률과 어휘 쌍 확률이다. 이들은 인과관계의 여부 c_j가 표기된 코퍼스로부터 학습할 수 있다. 하지만, 대량의 코퍼스에 인과관계를 수작업으로 부착하는 일은 많은 시간과 노력이 필요하며, 현재 구축된 바도 없다. 본 연구에서는 인과관계의 여부가 표시되어 있지 않는 대량의 코퍼스로부터 인과관계 분류기의 학습과 인과관계 추출을 동시에 수행하는 비지도식 학습 방법을 사용하여 이를 해결하였다.

1단계는 '초기 인과관계 쌍 구축'이다. 학습이 용이한 초기 분류기를 구성하고, 이 초기 분류기로부터 초기 인과관계 쌍을 추출한다. 1단계에서 사용되는 초기 분류기로는 (Girju 외, 2002)에서 제안된 두 명사의 사전 의미에 기반한 인과관계 분류기를 사용하였다. 2단계는 기대치-최대화 (Expectation-Maximization) 단계이다. 이 단계에서는초기 인과관계 쌍으로부터 확률 파라미터들을 학습하고, 학습된 파라미터로 구성된 분류기로 인과관계 쌍을 다시 생성하는 과정을 반복한다. 2단계에서 학습되는 파라미터들은 사전 (prior) 확률, $P(c_j)$, 단서구문 확률, $P(CP_{t_i} \mid c_j)$, 어휘 쌍 확률, $P(LP_{t_i k} \mid c_j)$ 이다. 학습 코퍼스에 나타나지 않는 어휘 쌍을 위해 각 파라미터를 학습하는 과정에서 Laplace 평활화(smoothing)를 적용하였다.

본 논문에서는 인과관계 명사구 쌍 추출을 위해서 수식 (2)의 어휘쌍 및 단서구문 확률에 자동 학습된 코퍼스에서 추출할 수 있는 두 종류의 확률값을 같이 사용한 수식 (5)의 인과관계 분류 모델을 사용하였다. 추가로 사용된 확률값은 명사구의 의미분류($rank_{t_i}$)에 따른 인과관계 확률, $P(c_j \mid rank_{t_i})$와 사용된 단서구문에 따른 인과관계 확률, $P(c_j \mid CP_{t_i})$이다. 여기에서 가중치들, wnc, wcpc, wlp의 합계는 1이다.

$$P_{CP+NC+CPC+LP}(c_j \mid t_i) = w_{nc} \times P(c_j \mid rank_{t_i}) + w_{cpc} \times P(c_j \mid CP_{t_i}) + w_{lp} \times P(c_j \mid t_i) \quad (5)$$

3.2. 용어의 인과관계 정보 추출

용어의 인과관계 정보는 코퍼스에서 그 용어의 원인이나 결과로 사용된 다른 용어들과 명사들의 벡터표현으로 나타낼 수 있다. 수식 (6)은 용어 ti의 인과관계 벡터 표현으로서, 주어진 영역에서 사용된 용어와 명사의 수가 N개일 때, C_{ij}와 R_{ij}는 각각 이들 용어 및 명사가 용어 t_i의 원인과 결과 명사구에 포함된 경우의 수이다. C_{ij}와 R_{ij}가 명사인 경우 명사구의 최상위 명사로 사용된 경우만을 고려했다.

$$\vec{t_i} = \{C_{i,0 \wedge N}, R_{i,0 \wedge N}\} \quad (6)$$

코퍼스에서 인과관계를 가지는 명사구 쌍들이 추출되면 이들 명사구 쌍들로부터 주어진 용어의 인과관계 벡터를 추출할 수 있다. 모든 용어의 인과관계 정보가 명사구 쌍에만 존재한다고 볼 수는 없다. 하나의 사건은 하나의 동사구나 문장 혹은 여러 개의 문장으로도 표현이 가능하며, 이러한 사건들이 주어진 용어의 원인이나

결과로 문헌에서 사용가능하기 때문이다. 본 논문에서는 명사구로서 발현되고 있는 사건들로 인과관계 정보의 수집 창구를 단순화하고자 한다.

인과관계 명사구 쌍으로부터 용어의 인과관계 정보를 습득하고자 할 때 고려할 점이 몇 가지 있다. 그 첫째는 문장에서 인과관계의 대상이 용어 그 자체가 아닌 경우가 많다는 것이며, 두번째는 용어간 내포관계가 빈번히 존재한다는 것이며, 마지막으로 용어 자체에 의미 애매성이 존재하는 경우가 많다는 것이다.

우리는 인과관계 명사구 쌍을 분석하면서 실험 대상으로 삼은 용어 집합과 많은 경우 일치하지 않은 경우를 볼 수 있었다. 예문 (4a)의 'immunoregulatory protein'의 경우 용어 자체가 인과관계의 대상인 명사구로 사용되었으나, (4b~c)의 'IL-2R alpha transcripts', 'T cell', 'interleukin-2'는 각각 서로 다른 인과관계 대상 명사구에서 최상위 명사의 수식어나 하위 구조로 사용되었다. 본 논문에서는 이러한 용어들도 인과관계에 기여하고 있는 것으로 보고 용어의 인과관계 벡터 구성에 포함하였다.

(4a) T cell produced an important immunoregulatory protein.

(4b) IL-2 induces increase in IL-2R alpha transcripts.

(4c) T cell activation resulting in enhanced production of interleukin-2.

예제 (4b)에서 용어 'IL-2R alpha transcript'는 또 다른 용어 'IL-2R'을 구조적으로 품고 있다. 의료 영역 용어 사전인 Mesh[7]에 정의되어 있는 많은 수의 의료용어가 이러한 관계를 가지고 있으며, 이러한 용어가 사용될 때 내포된 용어에 대해서는 인과관계 벡터를 구성하지 않았다.

3.3. 유사도 및 군집화 모델

용어간 유사도는 수식 (7)과 같이 용어간 인과관계 유사도와 어휘 유사도의 조합으로 표현된다. 식에서 LS(t_i, t_j)는 두 어휘 사이의 어휘 유사도로서 수식 (1)에서 정의된 유사도 추출식을 인용하여 사용하였다. CA(t_i, t_j)는 인과관계 유사도로서 수식 (6)의 인과관계 벡터로 표현된 두 용어 사이의 코사인 유사도이다. wLS는 어휘 유사도의 가중치로서 0과 1사이의 값이다.

$$TS(t_i,t_j) = w_{LS} \times LS(t_i,t_j) + (1-w_{LS}) \times CA(t_i,t_j) \qquad (7)$$

군집화 과정은 여러 군집화 모델 중 가장 널리 사용되는 평균 거리 측정 방법을 이용하여 단계적 군집화를 시도한다.

4. 평 가

4.1. 인과관계 분석

인과관계 분석기는 Medline[8]에 등록되어 있는 의료 영역 논문 41만 건에서 비지도식으로 학습되었다. 학습된 인과관계 어휘 쌍 확률과 단서구문 확률을 이용한 인과관계 분석기의 성능은 표 1과 같이 측정되었다. 표에서 cTREC은 TREC에서 제공한 WSJ 1988년 기사 중 'cancer'를 포함하는 970문장이며, cADAM은 A.D. A.M Inc.에서 제공하는 의학백과사전 중 'cancer'를 포함하는 1147문장이다. 명사 부류 순위 확률(NC)을 위한 가중치 wnc와 단서구문 신뢰도(CPC)를 위한 가중치 wcpc는 각각 0.1을 사용하였다.

분류모델	실험집합정확도	재현률	F값
CP+NC+ CPC+LP	cTREC74.39	85.95	79.74
	cADAM75.32	84.67	79.73
	Total74.84	85.30	79.73

[표 1] 인과관계 분석 정확도

위에서 학습된 인과관계 분석기를 사용하여 자동 추출된 명사구 쌍 중 인과관계 확률이 비교적 높은 15만 5천 개의 명사구 쌍이 인과관계 용어 쌍 추출을 위해 선택되었다.

군집화 실험을 위해Mesh에 정의되어 있는 용어 중 1659개를 선택하였다. 이는 인과관계를 추출했던 코퍼스에서 50번 이상 등장했던 용어들 중 최상위 명사의 빈도수에 따라 선택한 결과이다. 이중 순수 빈도수만을 고려하여 상위 200개의 용어는 수식 (7)의 어휘 유사도 가중치 훈련을 위해 미리 검증자료로도 사용되었다.

4.2. 용어 군집화

일반적으로 군집화의 평가를 위해 정답세트를 만드는 것이 어렵고, 정형화된 평가함수를 제시하는 것 또한 쉽지 않다. 본 연구에서는 Mesh에서 정의하는 용어 온톨로지 최상위 계층 14개 부류와의 비교를 통해 군집화의 성능평가를 시도하였다. 또한 평가함수로 적용률과 정확도, 최대F값의 세 함수를 이용하여 평가하였다. 수식 (8)에서 적용률(Cr)은 주어진 용어 중 최종 결과에 제시한 군집들에 얼마나 많은 비율로 포함시켰나를 측정하는 것으로서, 군집화 모델이 사용될 수 있는 한계를 표현한다. (식에서, N은 전체 용어의 수, T_r은 군집 C_r로 분류된 용어의 수)

$$Cr = \sum_{r=1}^{c} \frac{T_r}{N} \quad (8)$$

수식 (9)에서 정확도(SP)는 군집으로 제시된 군집 Si이 부류, Lr로 분류될 수 있다고 가정할 때 각 부류별로 최대의 정확도를 가지는 군집-정확도 $P(L_r, S_i)$들을 구해 이들을 평균한 값이다. (식에서 N은 전체 용어의 수, nr은 정답 부류, Lr 에 속하는 용어의 수)

$$SP = \sum_{r=1}^{c} \frac{n_r}{N} P(L_r) \quad (9)$$

$$P(L_r) = \max_{S_i \in T} P(L_r, S_i) \quad (10)$$

정확도는 군집화 과정에서 적은 수의 용어로만 구성된 군집들이 많이 존재하게 되는 단계적 군집화의 중간 과정에서는 군집화의 성능을 제대로 표현하기 어렵다. 단계적 군집화의 경우 군집화의 성능 변화를 추적하기 위해 수식 (11)의 최대 F값[5]을 같이 사용하였다. 최대 F값은 각 부류별로 최대 군집-F값을 구해 이들을 평균한 값이다.

$$MFScore = \sum_{r=1}^{c} \frac{n_r}{N} FS(L_r) \quad (11)$$

$$FS(L_r) = \max_{S_i \in T} FScore(L_r, S_i) \quad (12)$$

그림 5는 어휘 유사도 가중치에 따른 전체 군집화 성능의 변화를 보인다. 가중치는 200개 용어 집합의 군집화 결과 가장 높은 최대 F값을 보인 0.3을 선택하였으며, 최대 F값은 낮지만 약간 높은 정확도를 보인 0.7을 후보 가중치로 선택하였다. 표2는 인과관계 정보만을 사용한 군집화 방법(CA only), 어휘 유사도만을 이용한 군집화 방법(LS only), 두 유사도를 조합한 두가지 방법(LS(0.3)+CA , LS(0.7)+CA) 등 총 4가지 군집화 모델에 대한 실험 결과이다. 1659개 용어 군집화 결과에서도 200개 용어에서 가장 높은 성능을 보인 가중치(0.3)에서 좋은 최대 F값을 보였다. 1659개 용어에 대한 군집화 실험 결과 70.02%의 정확도를 보였으며, 어휘 유사도만을 사용한 경우에 비해 32.9%의 적용률 향상을 보였다.

군집화 진행 중 각 성능평가 함수의 변화를 보면, 최소거리 측정방법에서 초기의 최대F값이 빨리 증가하나, 최종 결과는 평균거리 측정 방법에서 높은 정확률과 F값을 보였다. 이는 널리 알려져 있는대로, 최소거리 측정방법을 이용한 단계적 군집화의 편향된 군집나무(skewed tree)를 생성하는 경향이 반영된 결과로 보인다.

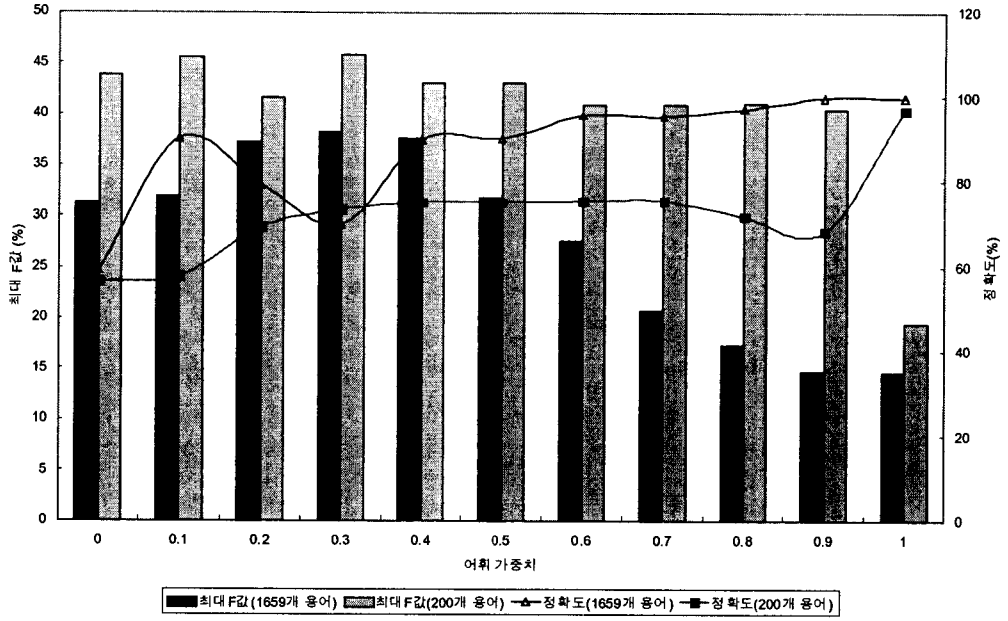
표2에서 어휘만을 사용한 방법(LSonly)에 비해 인과관계를 같이 사용한 모든 분류 방법에서 적용도 및 최대F값 모두 향상되었다. 이러한 결과를 통해 인과관계가 분류에 효과적인 속성으로 사용되었음을 알 수 있다. 또한, 인과관계를 어휘 정보와 함께 사용함으로서 인과관계만을 사용한 방법(CAonly)에 비해 비교적 높은 정확도를 확보하였다. 이러한 결과를 통해 인과관계는 용어 분류를 위해 긍정적인 속성의 하나로 사용될 수 있다고 해석된다. 하지만 이러한 사실로서 인과관계 정보가 유사한 모든 용어들이 서로 유사한 개념에 속한다고 말할 수는 없으며, 실험을 통해 증명한 사실은 인과관계가 유사한 용어들은 서로 유사한 개념에 속할 가능성이 그만큼 높다는 것이라고 하겠다. 이러한 내용은 인과관계 만을 이용한 군집화의 성능이 어휘 유사도와 결합된 경우에 비해 낮다는 실험 결과로서 뒷받침된다.

5. 결 론

본 논문에서는 용어의 선행 원인과 후행 결과를 인과관계 정보로 정의하고 이를 군집화의 속성으로 사용하고자 하는 시도를 하였다. 군집화 실험 결과 인과관계 속성이 사용된 경우 그렇지 않은 경우에 비해 32.9%의 적용률을 향상시켰으며, 어휘 유사도와의 조합을 통해 74.84%의 정확도를 보였다. 이러한 사실을 통해 각 용어의 인과관계가 용어 군집화를 위한 유용한 문맥 정보의 하나라는 것을 증명하였다. 이러한 결과를 바탕으로 용어의 직/간접적 인과관계로서 용어 온톨로지에서 그 용어가 속할 집합을 결정하는 데 도움을 줄 수 있을 것이다.

	실험용어세트(1659개)			검증용어세트(200개)		
	적용률	정확도	최대F값	적용률	정확도	최대F값
CA only	86.92	59.11	31.24	99.0	53.45	43.82
LS(0.3)+CA	99.10	70.02	38.18	100.0	73.79	45.79
LS(0.7+CA	99.22	95.43	20.74	100.0	72.40	40.90
LS only	66.18	99.78	14.56	28.0	96.92	19.41

[표 2] 집화 성능 평가 (종료 조건: 유사도<0.007적용)



[그림 5] 어휘 가중치 학습 및 실험 결과

참고 문헌

[1] Bourigault D. and Jacquemin D., 1999, "*Term Extraction + Term Clustering: An Integrated Platform for Computer-ided Terminology*,"In Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '99), pages 15--22, Bergen

[2] Chang, Du-Seong and Key-Sun Choi, 2004, "*Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities*,"in Proceedings of the first International Joint Conference on Natural Language Processing (IJCNLP-04)

[3] Friedman C., Kra P., Yu H., Krauthammer M. and Rzhetsky A., 2001, "*GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles,*" Bioinformatics, 17/1, pp. S74-S82.

[4] Girju, R. and D. Moldovan, 2002, "Mining Answers for Causation Questions," in *AAAI Symposium on Mining Answers from Texts and Knowledge Bases*

[5] Larsen B.and C. Aone, 1999, "*Fast and effective text mining using linear-time document clustering*" In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1622

[6] Maynard D. and Ananiadou S., 2000, "*Identifying Terms by Their Family and Friends,*" Proceedings of COLING 2000, Luxembourg, pp.530-536.

[7] Medical Subject Heading, 2004, http://www. nlm.nih. ov/mesh, U.S. National Library of Medicine

[8] Medline, 2004, http://www.nlm.nih.gov/PubM ed, U.S. National Library of Medicine

[9] Nenadic, G., Spasic, I., Ananiadou, S., 2002, "*Automatic Discovery of Term Similarities Using Pattern Mining*," in Proceedings of CompuTerm 2002, Taipei, Taiwan, 2002, pp. 43-49

[10] Spasi I., Nenadi G. and Ananiadou S., 2002, "*Supervised Learning of Term Similarities*," IDEAL 2002, LNAI series, Springer-Verlag, Berlin

[11] Tapanainen, Pasi and Timo Jarvinen, 1997, "*A non-rojective dependency parser*" in Proceedings of the 5th Conference on Applied Natural Language Processing, Association for Computational Linguistics, pages 64-71.

[12] Ward Jr. J.H., "*Hierarchical grouping to optimize an objective function,*" Journal of the Americal Statistical Assocation, 58(301):235-244, 1963.