EM 알고리즘을 이용한 전문용어 온톨로지 클래스간 관계 정의를 위한 동사 클러스터링

김미훈 남상협 이용훈 이종혁
포항공과대학교 전자컴퓨터공학부
첨단기술연구 정보센터
{meixunj,namsang,yhlee95,jhlee}@postech.ac.kr

Verb Clustering for Defining Relations between Ontology Classes of Technical Terms Using EM Algorithm

Meixun Jin, Sang-Hyob Nam, Yong-Hoon Lee, Jong-Hyeok Lee
Department of Computer Science and Engineering, Electrical and Computer Engineering Division
and Advanced Information Technology Research Center (AITrc)
Pohang University of Science and Technology (POSTECH)
{meixunj,namsang,yhlee95,jhlee}@postech.ac.kr

요 약

온톨로지 구축에서 클래스간 관계 설정은 중요한 부분이다. 본 논문에서는 클래스간 상 하위 관계 외의 관계 설정을 위한 클래스간 관계 자동 정의를 목적으로 의존구문분석의 (주어, 용언) (목적어, 용언) 쌍들을 추출하고, 이렇게 추출된 데이터를 이용하여 용언들을 클러스터링 하는 방법을 제안한다. 도메인 전문 코퍼스 데이터 희귀성 문제를 해결하고자, 웹검색을 결합한 방식을 선택하여 도메인 온톨로지 구축 클래스간 관계 자동 설정에 대한 방법론을 제시한다.

1. 서론

온톨로지 클래스 간 관계는 크게 상하위(is-a) 관계와 기타 관계로 나눌 수 있다. 상하위 관계가 온톨로지를 구성하는 골격 역할을 한다면, 기타 관계는 온톨로지가 내포하고 있는 지식의 양과 질을 나타낸다. 온톨로지의 각 클래스 간 기타 관계들을 어떻게 정의할 것인가 하는 것은 모든 온톨로지 구축자들이 고민하는 문제이다.

온톨로지의 사용 용도와 목적에 따라 정의될 클래스 간 관계도 다를 수 있다. 온톨로지를 구성하고 있는 클래스간의 다양한 지식을 나타내기 위하여, 이를 나타내는 다양한 관계 정의를 하는 것이 필요하다. 이러한 관계 정의를 위하여 NLP 커뮤니티에서 많이 사용하는 방법 중 하나가 의존구문분석결과를 이용하여, 온톨로지의 클래스를 나타내는 용어와 의존관계를 가지고 있는 용언으로 해당 클래스간 관계를 정의하는 것이다. 비슷한 의미를 지닌 용언들은 적절히 한가지 관계로 정의할 수 있다. 아래 예는 IT 전문용어 도메인에서 추출한 일부 문장이다. 데이터 방송과 디지털 방송이란 전문 용어는 같은 방송이라는 의미로 클러스터링이 가능한 단어이다. 이 두 전문용어가 사용하는 용언을 살펴 보면 이 용언들도 비슷한 정황에 사용되는 용언임을 볼 수 있다. 즉 용언을 클러스터링 한 결과가 전문용어 클러스터링에도 영향을 주게 된다.

예 :

데이터 방송이 시행되다
디지털 방송이 실시되다

본 논문에서는 EM 알고리즘을 이용하여, 코퍼스 중에 나타난 (주어, 용언) (목적어, 용언) 쌍들을 이용하여 클래스 간 관계를 자동으로 정의하는 방법을 제안한다. 구성은 아래와 같다. 2장에서는 온톨로지 자동 구축과 관련 기존 연구를 살펴보고, 3장에서는 EM 알고리즘을 이용한 클러스터링 방법에 대해서 소개한다. 4장에서는 실제로 어떠한 실험 과정을 거쳐서 전문 도메인의 용언을 클러스터링 하였는지 소개하고, 최종적으로 5장에서 결론을 맺는다.

2. 기존 연구

코퍼스로부터 클래스 간 관계를 추출하는 방법 중에서 많이 사용되는 방법은 구문분석의 (주어, 용언, 목적어)

트리플을 이용하는 것이다. 가장 공통적인 방법은 패턴 매칭을 통한 방법으로 두개의 용어와 용언을 포함하는 (주어,용언, 목적어) 트리플 사이의 문법적 관계를 추출하고 이 트리플에서 적절한 의미적 관계를 추출하는 방법이다. 기본적인 생각은 일반적인 패턴을 찾고 그것을 새로운 패턴에 적용시키는 것이다. 한국어 문장 중 자주 나타나는 생략으로, 이러한 방법은 일반 도메인 대용량 코퍼스가 가능할 경우에는 적용이 가능하지만, 특정 도메인 소량 코퍼스에 이 방법을 적용했을 경우 관계 추출 결과는 만족스럽지 못했다.

이렇게 추출된 패턴들은 관계 설정을 위하여 클러스터링 하게 되는데, 클래스터링 방법으로 상대 엔트로피를 Self-Organizing Map에 적용시켜서 동사를 클러스터링 하는 방법[1]과 동사와 명사 짝(pair)의 빈도수를 이용한 EM 알고리즘을 사용하여 동사와 명사 짝을 클러스터링 하는 방법[2]이 제시되었다.

[3]에서는 동사의 구문적인 종속 관계와 WordNet을 이용한 선택적인 선호도 등을 이용한 동사 클러스터링을 하는 방법 등이 제시 되었다.

[2]에서는 구문분석으로 얻은 (주어, 용언) (목적어, 용언) 쌍들을 추출하고 이들의 빈도수를 이용하여 EM 알고리즘을 이용한 클러스터링 방법을 제안하고 있다. 이 방법의 장점은 코퍼스에 사용된 방식에 따라 주어, 목적어에 사용된 명사와 동사를 자연스럽게 클러스터링하는 것이다. 본 논문에서는 [2]의 방법을 IT 분야 온톨로지 구축 시 기타 관계 자동 정의를 위한 용언 클러스터링 방법으로 선택하였다. 3장에서 [2]의 방법에 대하여 자세한 설명을 하였다.

3. EM 알고리즘을 이용한 클러스터링

3.1 EM 클러스터링 방법 개요 [2]

[2]에서는 구문분석으로 주어와 서술어 혹은 목적어와 서술어 기능을 하는 명사와 동사 쌍 들의 빈도수를 수집하고, 이를 EM 알고리즘으로 클러스터링하여 각 동사, 명사들이 속한 latent semantic 클래스를 찾아주는 방법을 제안하였다. 그림 1 에서 [1]에서 수집된 명사 쌍들의 초기 상태를 보여주고 있다.

그림 2에서는 클러스터링된 동사와 명사들의 결과를 보여주고 있다. 그림 2에서는 현재 실험으로 보여주고 있는 동사 24개와 명사 24개가 클러스터링된 결과를 보여주고 있다. 클래스 1 명사들은 asset, bond, interest 등으로 모두 금융, 금리, 돈, 재산 부류의 명사들이다. 해당 클래스에 속해 있는 동사들은 이런 명사들과 함께 자주 등장하여 의존관계를 맺고 있는 acquire, buy, dump 등의 동사들이다.

마찬가지로, 클래스2 와 클래스 3에 속해 있는 동사들과 명사들도 특정 도메인 정보를 공유하고 있는 동사, 명사들로 클러스터링 하게 되어짐을 볼 수 있다.

[2]의 방법은 코퍼스에 출현된 동사 명사 pair 빈도에

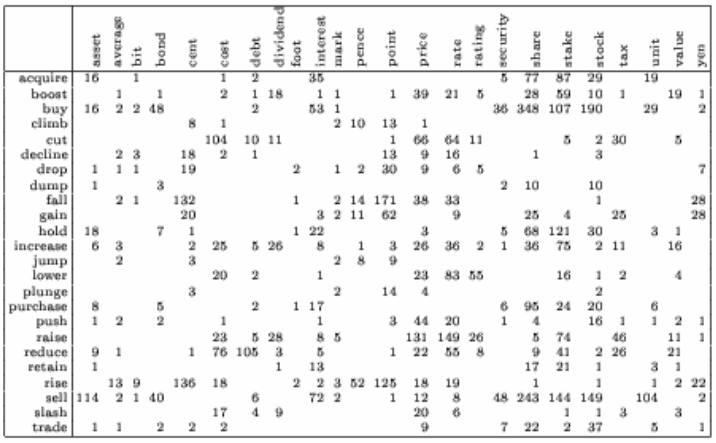

그림 - 1
[1] (명사, 동사) 출현 빈도 정보 (클러스터링 전)

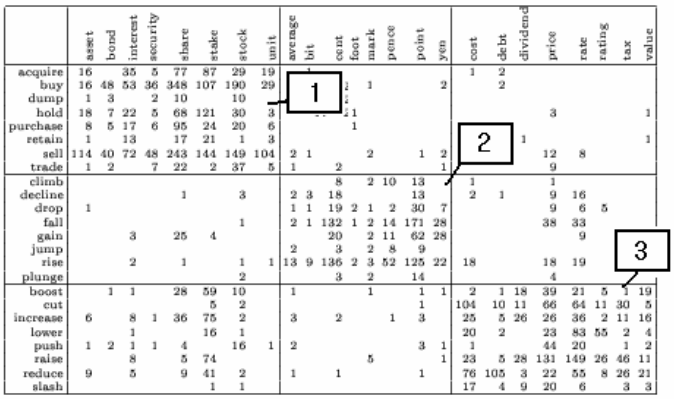

그림- 2 [2] (명사, 동사) 클러스터링 결과

따라 클러스터링 한 것이다. 표-1 에서 [2]의 실험에 cluster 1 에 속한 동사 acquire의 synset을 보여주고 있다. [2]의 실험에 클래스 1에 속해 있는 동사 buy, dump, hold, purchase, retain, sell, trade 등은 wordNet에서 Acquire의 synset 멤버로 한번도 출현하지 않았음을 보여주고 있다. [2]의 클래스1의 기타 동사에 대하여서도 상황은 비슷했다. (부록 표 -11 참조 바람.) wordNet의 결과와 [2]의 실험의 결과가 상의함을 볼 수 있다.

Acquire	{verb: get, acquire}, {verb: assume, acquire, adopt, take_on, take}, {verb: grow, develop, produce, get, acquire}, {verb: acquire, win, gain}, {verb: learn, larn, acquire}, {verb: develop, acquire, evolve}

표 - 1 워드넷 synset

3.2 EM 클러스터링 Algorithm

그림 -1 과 같이 수집된 명사, 동사 쌍에 대하여, 각 명사, 동사들에 특정한 latent class 를 부여하는 것이 [2]에서 클러스터링의 주된 목적이다.

EM 알고리즘을 이용하여 클러스터링 작업을 수행하기 위하여,

(1) 그림-1과 같이 수집된 동사 명사 쌍 및 해당 빈도수 정보를 incomplete data로 설정하고, $Y = V \times N$ 로 표기한다. 여기서 Y는 incomplete data set 을 의미하고 V 는 동사 집합, N은 명사 집합을 각각 의미한다.

(2) $X = C \times V \times N$ 는 complete data 이다. 식 중에서 C는 클러스터링 class 를 나타내고 있다.

(3) $X(y)=\{x \in X | x=(c,y), c \in C\}$ 는 관측된 데이터 y와 관련된 complete data 이다. 즉 data y 가 observe 되었을 때에 해당 y 가 개개의 latent semantic class c 에 속할 확률이 전체 complete data가 되게 된다.

(4) $P_\theta(x) = p(c,v,n)$ 는 각각의 latent semantic class C(클래스 집합) 와 V (동사 집합), 그리고 N (명사 집합) 이 서로 독립적인 관계이기 때문에 $C \times V \times N$ 의 결합 확률이 된다. 이중 θ 는 $\theta = <\theta_c, \theta_{vc}, \theta_{nc} | c \in C, v \in V, n \in N>$ 각 파라미터 (parameter) 들이다.

(5) $P_\theta(y)$는 incomplete data 로서 $P_\theta(y) = \sum_{X(y)} P_\theta(x)$ 식으로 complete data와 연결을 맺고 있다

EM 알고리즘은 $\theta' = \arg\max L(\theta)$ 를 만족하는 θ' 를 구해준다. $L(\theta)$ 는 incomplete data의 log-likelihood 함수이다. 각 parameter, $\theta_c, \theta_{vc}, \theta_{nc}$ 를 구하는 식은 아래와 같다.

$$\theta vc = \frac{\sum_{y \in \{v\} \times N} freq(y)\, p_\theta(x|y)}{\sum_{y} freq(y)\, p_\theta(x|y)} \quad (1)$$

$$\theta nc = \frac{\sum_{y \in \{n\} \times V} freq(y)\, p_\theta(x|y)}{\sum_{y} freq(y)\, p_\theta(x|y)} \quad (2)$$

$$\theta c = \frac{\sum_{y} freq(y)\, p_\theta(x|y)}{|Y|} \quad (3)$$

4 IT 전문 온톨로지에서 기타 관계 자동 정의

본 논문에서는 [2]에서 제안한 방법을 IT 온톨로지 중에서 디지털 TV, 지능형 로봇 분야의 온톨로지 구축에 적용하였다.

먼저 도메인 전문 코퍼스를 수집하고, [2]의 방법과 같이 본 연구실이 보유하고 있는 한국어 의존구문분석기 - KOPA[4]를 이용하여 전문용어와 의존구문관계를 가지는 (주어, 용언)(목적어, 용언) 쌍들을 추출했다. 추출된 (주어/ 목적어, 용언) 쌍은 [2]의 실험과 비교했을 때, 용언들을 클러스터링 하기에는 데이터가 많이 부족할 것으로 판단되었다. 이러한 문제점들을 극복하기 위하여, 아래와 같은 작업을 진행 하였다.

4.1 전문 용어 및 의존구문관계를 지닌 용언 추출

전문용어 및 해당 용어와 의존구문관계를 지닌 용언 쌍을 추출하기 위하여, 본 연구실이 보유하고 있는 의존구문분석기를 아래와 같이 수정하였다.

(1) 전문용어 인식

전문용어들은 단일어인 경우와 복합어인 경우 두 가지가 있다. 복합어일 경우, 전문용어의 인식 정확도를 높이기 위하여, 구문분석 결과를 바탕으로 전문용어를 인식하는 방식을 선택하였다.

(2) 기능용언 처리

서술성 명사는 명사적 특성과 술어적 특성을 동시에 가진 명사를 의미하며 기능동사란 어휘 의미가 미약하거나 없으며 서술성 명사를 뒷받침해 주는 동사를 말한다. 이런 성격의 기능 동사는 자신의 논항을 선택하지 못한다. 기능동사 구문이란 이처럼 서술성 명사가 기능 동사와 함께 자신의 고유한 논항을 선택하는 기본 구문을 의미한다. 예를 들어

"투자를 한 결과 자체 기술로 유럽 수출용 모델을 개발 하였다."

```
| 하였다. 평서형종결 ]
=> 하 CMCN 이 fpd 었 fmbtp 다 fmofd . g
    | 개발 부사어 ] ----> 하였다.
    => 개발 CMCPA
    | 모델을 목적어 ] ----> 하였다.
    => 모델 CMCN 을 fjco
        | 수출용 부사어 ] ----> 모델을
        => 수출용 CMCN
            | 유럽 부사어 ] ----> 수출용
            => 유럽 CMP
```

의 경우, 기능동사 "하였다"가 아닌 서술성 명사 "개발"이 "유럽 수출용 모델"을 논항으로 취하는 실제 지배소가 되어야 한다. 따라서 이러한 문장에서 각각의 쌍을 추출할 때 용언으로 서술성 명사와 기능동사를 함께 기능 동사 구문으로 처리하여 (유럽 수출용 모델, 개발 하였다) 쌍을

추출한다.

4.2 전문 용어와 용언 쌍 희귀성 문제

용언과 의존구문관계유형은 크게 주어, 목적어, 부사어 등으로 나뉠 수 있다. 온톨로지 클래스 설정에 적합한 의존구문관계는 주어와 목적어로 판단하여, 전문 용어와 용언이 주어/서술어인 경우와 목적어/서술어인 경우를 추출하였다. 표 - 2 에서는 디지털 TV 분야, 목적어/서술어 의존구문관계를 지닌 전문용어 및 용언 쌍의 일부분을 보여주고 있다.

표-2에서 각 줄의 마지막 숫자는 같은 줄에 있는 (용언, 전문용어)가 코퍼스 중에 출현한 빈도수이다. 표-2에서 보여준 것과 같이, 이러한 쌍들의 빈도수가 너무 낮아, [1]의 방법을 그대로 적용하여 좋은 결과를 내는 것은 어렵다. 빈도수가 낮은 원인은 코퍼스의 크기가 작은 것이 문제가 될 수 있지만, 전문용어인 만큼 코퍼스의 크기가 커도, 그 출현 빈도는 작을 것으로 판단된다.

```
….
공급하다  | 디지털 셋톱박스     | 목적어   /1
공포하다  | 방송 규정           | 목적어   /1
구성하다  | 인터 리버 깊이      | 목적어   /1
구하다    | 매출 실적           | 목적어   /1
구하다    | 애니메이션 데이터   | 목적어   /1
규정하다  | 제한 수신           | 목적어   /1
규제하다  | 방송 내용           | 목적어   /1
….
```

표 - 2 디지털 TV 분야 전문용어와 용언 쌍 예

전문용어와 구문적 관계를 지닌 용언의 빈도수는 많이 나타나지 않지만, 일반 명사들과 구문적 관계를 지닌 경우는 많다. 용언 정보를 이용하여 온톨로지 관계 정의를 위한 것이라면, 용언과 의존구문관계를 가진 것이 일반 명사이든, 전문용어이든 큰 상관은 없다. 그리하여

(1) 도메인 코퍼스에서 전문용어와 의존구문관계를 지닌 용언 리스트를 추출하였다. 표-2 에서 나타나는 것과 같은 용언들 1179 개를 추출하였다. 본 논문에서는 그중 DTV 코퍼스에 나타난 450개 용언의 사례를 보여주고 있다.

(2) 추출된 용언 리스트를 다시 일반 코퍼스에 넣어, 목적어, 서술어 구문관계를 지닌 명사, 용언 쌍 및 그들의 출현 빈도수를 함께 추출하였다.

먼저 각각의 용언을 아래 그림- 3 과 같이 구글 검색어로 입력하였고, 그 결과로 나온 snippet 데이터 중 최대 3000개까지의 결과에 대해서 수집을 하였다.

그림-3 구글 검색기를 이용한 용언 관련 코퍼스 수집

이러한 방법으로 약 200만개의 snippet 을 추출하였다. 이 코퍼스를 수집 하는 데에 24시간 (Conroe 6300) 소요되었다. 추출된 snippet 들을 문장 단위로 나누어서 의존 구문분석기(KOPA) [4]로 구문 분석 하여 주어/용언, 목적어/용언 쌍을 추출하였다. 모든 문장을 분석 하기까지는 110시간 가량 소요 되었다.

위와 같은 과정을 통해서 DTV와 지능형 로봇에서 사용되는 용언을 클러스터링 하기 위해 필요한 주어/용언, 목적어/용언 쌍을 일반 코퍼스를 통해서 구축하게 된다. 표 – 3은 전문 도메인만을 이용하여 추출한 각 쌍의 개수이다. 표 – 4 는 일반 도메인에서 추출된 쌍의 개수이다. 표 - 5 은 각 클러스터링을 하고자 하는 용언의 개수이며 이를 통해서 표 – 6와 같이 각 용언을 클러스터링하기 위한 데이터의 평균 개수를 구할 수 있다.

	(주어, 용언) 쌍	(목적어, 용언) 쌍
DTV	948	492
ROBOT	3540	2526

표 - 3 전문 도메인을 통해서 추출된 쌍의 개수

	(주어, 용언) 쌍	(목적어, 용언) 쌍
DTV	93694	43591
ROBOT	161869	110671

표 - 4 일반 도메인을 통해서 추출된 쌍의 개수

	(주어, 용언) 쌍	(목적어, 용언) 쌍
DTV	376	139
ROBOT	839	400

표 - 5 각 쌍에서 용언의 개수

	주어, 용언 쌍	목적어, 용언 쌍
DTV	2.5/249.1	3.5/313.6
ROBOT	4.2/380.7	6.3/842.9

표 - 6 용언당 명사의 개수 평균(전문도메인/일반)

표 - 6 에서도 볼 수 있듯이 IT 분야 도메인에서만 추출한 쌍을 가지고서는 클러스터링을 하기에는 데이터가 부족함을 볼 수 있다. 반면 일반 도메인을 통해서 추출한 데이터 쌍은 클러스터링을 하기에 부족하지 않은 데이터를 제공해주게 된다.

Class 10-3	
낮다	0.09
높다	0.08
헐다	0.07
제정되다	0.05
적극 참여	0.04
여전하다	0.04
무관하다	0.04
풍부하다	0.02
지상파 방송 하다	0.02
정의되다	0.02
정도 완성	0.02
내다	0.02
커지다	0.02
위하다	0.02
선명하다	0.01
협의하다	0.01
탑재하다	0.01
탈피하다	0.01
출시하다	0.01
추진 하다	0.01
차다	0.01
정착되다	0.01
전송하다	0.01
전달되다	0.01
적다	0.01
일컫다	0.01
인증되다	0.01
의하다	0.01
비례하다	0.01
분포하다	0.01
병행되다	0.01
미미하다	0.01
다수 확보	0.01
규정되다	0.01
국가체제 개편	0.01
구축되다	0.01
고려하다	0.01
강요하다	0.01
가지다	0.01
가능해지다	0.01

표 - 7 클래스 10-3 (10개 클래스로 클러스터링 됨)

4.3 클러스터링

[2]의 클러스터링 방법에서는 몇 개의 class로 클러스터링을 할 것인지 미리 설정을 해줘야 한다. 본 논문에서는 클래스의 개수를 10개, 20개, 30개로 설정하여 실험을 하였다.

이와 같이 EM 클러스터링을 할 데이터를 구축한 후에 3.2에서 설명한 EM 클러스터링 방법을 용언, 주어 쌍 과 용언, 목적어 쌍에 대해서 적용 하였다. 클러스터링이 끝나는 조건은 각 latent semantic class 에서 용언 및 명사가 가지는 기대값과 이전 iteration에서 가지는 기대 값 사이의 절대값 차이가 거의 변화가 없을 때 종료되도록 하였다. 10개의 클래스, 20개의 클래스, 30개의 클래스로 각각 클러스터링을 수행 하였으며 약 70시간이 소요되었다.

5 실험에 대한 분석

5.1 10개 클래스에서 용언/주어 사이의 의미 관계 추출

표-7는 용언과 주어로 사용되는 용어 사이의 쌍을 바탕으로 하여 10개의 Class로 설정한 상태에서 EM 클러스

터링을 한 뒤 한 개 클래스로 분류된 용언들의 예제이다

오른쪽 숫자는 그 용언이 해당 클래스를 만드는데 기여하는 확률로서 그 용언이 의미적으로 해당 클래스에서 차지하는 비중이다.

결과로 나온 용언을 살펴 보면 사람이 판단할 수 있는 의미를 지닌 한 개의 클래스로 설정할 수 없다는 것을 확인 할 수 있다. 해당 클래스 내의 용언들이 의미 관계에 따라서 각각 다시 묶일 수 있기 때문이다.

번호	용언
1	낮다, 높다, 적다, 커지다
2	제정되다, 헐다, 규정되다, 구축되다 정의되다
3	풍부하다, 적다, 미미하다
4	탑재하다, 가지다
5	선명하다, 미미하다
6	전송하다, 지상파 방송 하다
7	비례하다, 분포하다

표 - 8 subclass 예제

표-8 은 클래스 10-3 내에서 의미적으로 묶일 수 있는 용언들을 나열한 표이다. 보는 바와 같이 10개의 클래스로 나누었을 경우에 각각의 클래스가 하나의 의미 관계를 대표하지 못하고 있고, 그 안에서 다시 여러 개의 의미관계들이 포함되어 있다.

5.2 20개 클래스에서 용언/주어 사이의 의미 관계 추출

다시 20개의 클래스로 클러스터링을 수행하면 표-9과 같은 결과가 나온다.
표 – 10 는 총 20의 클래스로 클러스터링 하였을 경우에 나오는 클래스 중에서 특히 표-7의 10-3 클래스의 용언들이 주로 분포하는 클래스 20-6, 20-7 의 용언들이다.

번호	클래스	용언
1	20-6	낮다, 높다, 적다, 커지다
2	20-6	구축되다
2	20-7	제정되다, 정의되다, 규정되다, 헐다
3	20-6	풍부하다, 적다, 미미하다
4	20-6	가지다
4	20-7	탑재하다
5	20-6	선명하다, 미미하다
6	20-6	전송하다, 지상파 방송 하다
7	20-7	비례하다, 분포하다

표 - 9 클래스 10-3의 용언들이 클러스터링 된 상태

표 - 10 는 표 - 8 의 클래스 10-3 내부에 존재하는 주요 용언들이 클러스터링된 모습이다. 여기서 특이한 점은 비슷한 의미관계로 묶이는 클래스들이 20개의 클러스터에서도 대부분 같은 클래스로 클러스터링 되었다는 점이다. 즉 클래스 개수를 10개에서 20개로 감소 시키자 각 클래스가 가지는 의미 범위가 좀더 구체적으로 줄어들게 되었다.

Class 20-6	
높다	0.18
낮다	0.1
적극 참여	0.08
풍부하다	0.05
지상파 방송	0.05
내다	0.05
선명하다	0.02
탈피하다	0.02
커지다	0.02
출시하다	0.02
정착되다	0.02
정도 완성	0.02
전송하다	0.02
적다	0.02
의하다	0.02
여전하다	0.02
미미하다	0.02
다수 확보	0.02
국가체제 개편	0.02
구축되다	0.02
고려하다	0.02
가지다	0.02
가능해지다	0.02

Class 20-7	
헐다	0.24
제정되다	0.13
정의되다	0.06
낮다	0.06
협의하다	0.03
탑재하다	0.03
추진 하다	0.03
차다	0.03
정도 완성	0.03
전달되다	0.03
일컫다	0.03
인증되다	0.03
위하다	0.03
비례하다	0.03
분포하다	0.03
병행되다	0.03
규정되다	0.03
강요하다	0.03

표 - 10 클래스 20-6, 20-7 (20개 클래스로 클러스터링)

5.3 30개 클래스에서 용언/주어 사이의 의미 관계 추출

번호	클래스	용언
1	30-11,24	낮다
1	30-9,23,22	높다
1	30-10	적다, 커지다
2	30-10	제정되다, 구축되다, 규정되다
2	30-11	정의되다, 헐다
3	30-9	풍부하다, 미미하다
3	30-10	적다
4	30-1,8	가지다
4	30-10	탑재하다
5	30-9	선명하다, 미미하다
6	30-9	전송하다, 지상파 방송 하다
7	30-10	비례하다
7	30-11	분포하다

표 - 11 30개의 클래스로 클러스터링 된 상태

표 - 11은 30개의 클러스터로 클러스터링을 하였을 경우에 표 - 9 의 각 용언들이 클러스터링된 모습이다. 좀더 세분화된 의미로 클래스가 나누어지게 되고, 그에 따라서 표 - 9의 각 클래스가 나누어지기도 하고, 같이 클러스터링되기도 하였다. 대체적으로 같이 클러스터링 되는 경향이 있는 것을 확인할 수 있다. 이와 같이 각 부분 부분의 의미가 클러스터링되는 것을 분석해 보면 클래스가 많아질수록 같은 클래스로 클러스터링 되는 경향이 커지고 있음을 확인할 수 있다. 더 많은 클래스로 설정하여 실험을 할 경우, 더욱 세분화된 의미를 지닌 동사들이 클러스터링이 되면서

동시에 클래스에 속하는 동사가 너무 적어지기도 할 것으로 예상된다. 다만 실험을 돌리는 시간상 관계로 이에 관한 실험은 향후에 계속 진행하고자 한다.

5.4 용언/목적어 사이의 의미 관계 추출

[2] 실험에 사용된 동사/명사 쌍에 제약이 따른다. 즉 출현된 명사의 개수와 동사의 개수가 같거나 비슷해야 한다. 4.2에서 설명한 방법으로 추출된 용언/목적어 쌍에 출현된 명사에 개수는 몇만개 되었다. 그러므로 많은 명사들의 출현 빈도가 낮아, 실험 데이터로 직접 사용하기에는 적절치 못하였다. 이러한 문제를 해결하기 위해서는 실험용으로 추출된 데이터를 추가 가공이 필요한 것으로 생각된다. 이에 대한 연구는 향후에 진행할 계획이다.

6. 온톨로지 클래스 관계 설정

위 실험에서 얻어진 결과를 통해서 전문 용어 코퍼스에서 사용되는 용언들을 클러스터링 하게 된다. 각 용언들은 전문 용어와 함께 쓰이고 있기 때문에 전문 용어도 해당 클러스터링된 용언을 중심으로 클러스터링을 할 수 있게 된다. 이렇게 클러스터링된 결과는 상하위 관계와 같은 클래스 관계 설정을 하는데에 사용될 수 있다.

7. 결론 및 향후 계획

본 논문에서는 전문 도메인 온톨로지 상의 클래스간 관계를 정의하기 위해서 필요한 용언을 EM 알고리즘을 사용하여 클러스터링하였다.

전문 도메인은 많은 양의 코퍼스를 확보하기 어렵다. 본 논문에서 클러스터링 하고자 하는 것은 명사가 아닌 용언이기 때문에 해당 용언이 포함된 일반 도메인의 코퍼스를 사용할 수 있음에 착안하여 일반 도메인에서의 주어와 용언 쌍 및 목적어와 용언 쌍을 추출하여 EM 클러스터링을 수행 하였다. 이때 주어와 용언 쌍에 대한 클러스터링은 의미적으로 묶일 수 있는 관계들로 클러스터링되는 경향이 있었지만 전체적으로 완벽하게 클러스터링되지는 못하였다. 이는 전체 클래스의 수를 의미 관계의 수에 따라서 자동으로 정하지 못한 점에 크게 기인하고 있으며, 같은 클래스 내에서도 다시 의미 관계에 따라서 클러스터링을 제대로 수행하지 못한 점에도 영향을 받았다.

이러한 실험을 통해서 특정 도메인의 용언을 자동 클러스터링할 수 있는 가능성을 볼 수 있었다. 앞으로는 가장 최적화된 클래스 개수를 자동으로 결정하는 문제나 같은 클래스로 구분되더라도 해당 클래스 내의 용언이 모두 같은 하나의 의미로 구분되지 못한 문제점 등을 해결해 나가 볼 것이다. 그래서 전문 도메인에서 사용되는 용언을 효과적으로 클러스터링 할 것이고, 이를 통해서 전문 도메인 분야의 온톨로지 클래스를 자동 분류해 나갈 예정이다.

감사의 글

본 연구는 첨단정보기술 연구센터를 통한 과학재단 및 2007년도 두뇌한국21사업의 지원을 받았고 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고 문헌

[1] 박성배,장병탁,김영택,Self-Organizing Map을 이용한 한국어 동사 클러스터링, 정보과학회 추계 학술발표 논문집, pp. 183-185, 1998

[2] M.Rooth, S.Riezler, D.Prescher, G.Carroll and F.Beil, “Inducing a semantically annotated lexicon via EM-based clustering”, ACL , pp 104-111, 1999

[3] Sabine Schulte im Walde,Clustering Verbs Semantically According to their Alternation Behaviour, Proceedings of the 18th conference on Computational linguistics - Vol.2, pp.747-753, 2000

[4] 김미영, 강신재, 이종혁 "단위(Chunks)분석과 의존문법에 기반한 한국어 구문분석" ,정보과학회 춘계 학술발표 논문집, pp. 327~329, 2000

부록

buy	{verb: buy, purchase}, {verb: bribe, corrupt, buy, grease_one's_palms},
dump	{verb: dump, ditch}, {verb: dump, underprice}, {verb: plunge, dump}, {verb: deck, coldcock, dump, knock_down, floor}]
hold	{verb: keep, maintain, hold}, {verb: hold, take_hold}, {verb: hold, throw, have, make, give}, {verb: have, have_got, hold}, {verb: deem, hold, view_as, take_for}, {verb: harbor, harbour, hold, entertain, nurse}, {verb: restrain, confine, hold}, {verb: retain, hold, keep_back, hold_back}, {verb: bear, hold}, {verb: hold, support, sustain, hold_up}, {verb: hold, bear, carry, contain}, {verb: accommodate, hold, admit}, {verb: hold, carry, bear}, {verb: prevail, hold, obtain}, {verb: contain, take, hold}, {verb: reserve, hold, book}, {verb: defend, guard, hold}, {verb: oblige, bind, hold, obligate}, {verb: defy, withstand, hold, hold_up}, {verb: apply, hold, go_for}, {verb: control, hold_in, hold, contain, check, curb, moderate}, {verb: halt, hold, arrest}, {verb: carry, hold}, {verb: declare, adjudge, hold}, {verb: agree, hold, concur, concord},
purchase	[{verb: buy, purchase}]
Retain	{verb: retain, continue, keep, keep_on}, {verb: retain, hold, keep_back, hold_back},
Sell	{verb: deal, sell, trade}, {verb: betray, sell}
trade	{verb: trade, merchandise}, {verb: trade, trade_in}, {verb: trade, swap, swop, switch}, {verb: deal, sell, trade}

표 - 11 wordNet 결과