말모둠에서 동사 분포 연구

최용석, 이운재, 최기선
한국과학기술원 전산학과/전문용어언어공학연구센터/첨단정보기술연구센터
{angelove, wjlee, kschoi}@world.kaist.ac.kr

A Study on Verbs Statistics in Corpus

Yong-Seok Choi, Woon-Jae Lee, Key-Sun Choi
Department of Computer Science KAIST/KORTERM/AITrc

요 약

말모둠은 특성에 따라서 여러 성격을 나타내게 된다. 하지만 말모둠의 특성을 자동적으로 알아내는 방법은 간단하지 않다. 중요 단어를 가지고 있으면 말모둠에서 통계적으로 많은 부분에 적용시켜 말모둠의 특성을 파악할 수 있다.

본 논문에서는 한국어 말모둠에서 나타나는 동사류 단어들의 빈도를 분석한다. 또한, 사람이 직접 중요도를 평가한 사전의 단어들과 말모둠에 나타나는 단어들을 비교해서 통계적 차이점을 알아보고, 그 차이점을 통해 앞으로 연구할 일에 대해서 토론한다. 간단한 실험을 통해 사람의 평가한 중요도 점수의 효용성도 알아본다.

1. 머릿글

말모둠에서 출현하는 단어들에 대한 빈도 분석은 있어왔다. 기계번역 작업 같은 경우 많은 일들을 해야한다. 중요 단어를 선정하는 것은 일의 순서를 결정하는 데 중요한 역할을 한다. 본 논문에서는 말모둠에 출현하는 고빈도 동사류 단어와 사람이 중요하다고 생각하는 단어 사이에 어떤 연관이 있는지 다룬다. 또한 사람이 뽑은 중요어휘를 통한 실험으로 중요어휘 가치의 가능성에 대해서 살펴본다.

본 논문[1]의 구성은 다음과 같다. 2장에서는 기존 연구에 대해서 살펴보고 본 논문에서 접근하려는 방법에 대해 비교한다. 3장에서는 단어 분포에 관한 통계적 사실에 관해 다루고, 4장에서는 기본 동사를 선정하기 위해 사람이 평가하는 방법에 관해 다룬다. 5장에서는 실제로 말모둠에서 뽑은 분포 통계와 사람의 채점한 중요도와 비교해 본다. 6장에서는 실험을 통해서 중요도 정보의 이용가능성에 대해서 알아보고, 마지막으로 7장에서는 결론과 향후 과제에 대해서 살펴본다.

2. 기존 연구

자연언어의 통계적 정보를 구축하는데 필요한 말모둠을 구축하는데 많은 노력이 있어왔다. 초기에는 통계정보를 위한 말모둠의 크기를 늘리는데 주요 목적이 있었다. 하지만, 말모둠의 양의 점점 늘어나면서 말모둠의 특성을 알아내고 평가하려는 노력들이 이루어져 왔다.

최용석(1999)은 말모둠을 통해 격틀을 구축하면서 격틀의 정보들을 평가하려 했다. 이 방법은 기본적으로 격틀 평가를 위해 시도한 방법이지만 말모둠을 통해 이루어진 방법으로 거꾸로 말

1) 본 연구는 전문용어언어공학연구센터에서 수행한 과학기술부와 KISTEP의 핵심소프트웨어사업 중 "대용량 국어정보 심층처리 및 품질관리 기술개발"과제의 일환으로 수행되었으며, 첨단정보기술연구센터를 통하여 과학재단의 지원도 받았습니다. 본 연구를 가능하게 해 준 지원에 감사드린다.

모둠을 격틀을 통해서 평가할 수 있는 방법이 될 수도 있다.

신중호(1999)는 말모둠에서 격틀로 이용할 수 있는 동사정보를 추출했고, 그 동사정보를 클러스터링 기법을 이용해서 분류해 보고 있다.

기존 연구는 말모둠에서 정보를 추출해서 사용하려 했다. 본 논문에서는 구축해 놓은 정보를 통해서 말모둠을 살펴보려 한다. 격틀의 주 정보인 동사/형용사 정보를 이용하여 말모둠의 특징을 표현하고 이용한다.

3. 단어 분포

25만 8791단어를 가진 말모둠을 분석했다. 단어들을 빈도순으로 나열하고 순서대로 어느 정도 분포하는지 알아봤다. 제일 많이 나온 단어가 전체 말모둠의 4.64%를 차지하고 있음을 알 수 있다. 50개의 단어로 말모둠의 49.08%를 채울 수 있었고, 말모둠의 90.00%는 4890개의 단어로 이루어져 있었다. 다음은 고빈도 순의 단어수와 그 단어의 적용 범위를 나타낸 것이다.

1 = 4.64%
2 = 7.39%
3 = 9.83%
4 = 12.23%
5 = 14.44%
6 = 16.17%
7 = 17.84%
8 = 19.49%
9 = 21.10%
10 = 22.67%
11 = 24.22%
12 = 25.67%
13 = 26.92%
14 = 28.18%
15 = 29.38%
16 = 30.53%
17 = 31.65%
18 = 32.73%
19 = 33.71%
20 = 34.59%
21 = 35.44%
22 = 36.25%
24 = 37.68%
25 = 38.36%
26 = 39.03%
28 = 40.30%
30 = 41.38%
32 = 42.41%
34 = 43.36%
36 = 44.27%
38 = 45.10%
41 = 46.25%
44 = 47.31%
47 = 48.25%
50 = 49.08%
54 = 50.13%
58 = 51.06%
63 = 52.12%
68 = 53.09%
74 = 54.18%
79 = 55.03%
86 = 56.11%
93 = 57.08%
101 = 58.01%
112 = 59.08%
123 = 60.04%
136 = 61.02%
151 = 62.04%
167 = 63.00%
185 = 64.01%
205 = 65.01%
228 = 66.03%
253 = 67.03%
280 = 68.02%
311 = 69.02%
346 = 70.02%
385 = 71.02%
429 = 72.02%
479 = 73.00%
538 = 74.01%
604 = 75.00%
682 = 76.00%
771 = 77.00%
874 = 78.01%
991 = 79.00%
1127 = 80.00%
1282 = 81.00%
1461 = 82.00%
1669 = 83.00%
1916 = 84.00%
2208 = 85.00%

#	2554	= 86.00%
#	2977	= 87.00%
#	3491	= 88.00%
#	4118	= 89.00%
#	4890	= 90.00%
#	5862	= 91.00%
#	7107	= 92.00%
#	8732	= 93.00%
#	10911	= 94.00%
#	13983	= 95.00%
#	18605	= 96.00%
#	26276	= 97.00%
#	41571	= 98.00%
#	87265	= 99.00%
#	258791	= 100.00%

단어 분포로 보아 중요 단어부터 자연언어 처리 작업을 한다면 초기의 성능을 갖추는 데, 역할을 할 수 있다. 가령 4890개의 단어에 대한 작업을 한다면 말모둠의 90%에 대한 작업이 이루어지고, 기본적인 계산으로 90%의 성능이 나온다고 생각할 수 있다. 물론, 마지막 최종 성능을 내기 위해서는 점점 더 많은 일을 해 줘야 하지만, 초기의 성능을 좌우하는 것은 기본 단어 선정이라고 할 수 있다.

4. 기본 동사 선정

기본적인 동사를 선정해서 자연언어 처리의 기본 단어로 다루려고 한다. 사전에 나오는 모든 동사를 대상으로 하지 않고 꼭 필요하다고 생각하는 동사를 대상으로 하는 방법이 효율적이다. 기본 어휘는 단순히 말모둠에 나오는 단어의 빈도에만 의지할 수 없다. 예를 들어 "발톱"이라는 단어가 빈도적으로 한 번도 말모둠에 나타나지 않은 단어라고 해도 이 단어를 기본어휘에서 제외할 수 없다.

기본 동사를 선정하기 위해서 형태소 분석기[이운재 1999]의 동사, 형용사 사전을 이용했다. 이 사전은 동사 3468개 형용사 1626개로 5097 용언류 단어가 들어있다. 형태소 분석기 사전에 들어있는 용언류가 기본적으로 필요한 단어라 할 수 있다. "하다"류의 동사를 제외하기 때문에 사전에 나오는 단어수와 많은 차이가 난다. 이 단어들에 수동으로 중요도 점수를 부여했다. 한 단어에 3명이 교차 평가할 수 있도록 했다. 형태소 분석기에 쓰일 가치가 있는 단어인가를 평가하는 것으로 점수를 부여하는 기준은 아래와 같았다.

O. 우리말 큰 사전에 나오는 단어일 경우

1 점 : 우리말 큰 사전에는 나오지만 형태소 분석기 사전에 들어가면 오류발생시킬 가능성이 큰 단어.

써다 : 썰물.

2 점 : 우리말 큰 사전에는 나오고 평소에 안 쓰는 단어(뜻을 모름)

3 점 : 우리말 큰 사전을 안 보고도 뜻을 짐작할 수 있는 단어

4 점 : 그냥 보면 아는 기본 어휘

O. 우리말 큰 사전에 나오지 않는 단어

0 점 : 틀린 단어

1 점 : 오류 발생시킬 가능성이 큰 하찮은 단어

2 점 : 뜻을 잘 모르겠는 단어

3 점 : 많이 쓰이는 단어

4 점 : 사전에서 빠진 것이 믿겨지지 않을 정도의 중요단어

한 단어에 대해서 3명이 평가한 점수의 합을 구해 총점을 냈다. 따라서, 한 단어는 0~12점 사이의 점수를 가진다. 각 점수에 대한 빈도 그래프는 아래 그림과 같다.

12점을 받은 용언의 단어는 366개였다. 이 366개를 [채영숙 1999] 분류법에 따라서 다시 형태분류를 했다. 이 중 327(89.34%)가 단순 동사/형용사 형태로 출현했다. 13개(3.55%)는 [채영숙 1999] 사전에 출현하지 않았다. 동사 "밝다" 같은 경우 "바르다"의 뜻을 가진 준말로 사전에 없었으며, 동사 "키다"는 "키우다"의 뜻을 가진 준 말로 역시 등재되지 않았다. 빈도는 그림 1, 2와 같았다.

0	50
1	25
2	105
3	147
4	80
5	118
6	672
7	297
8	188
9	354
10	1158
11	1537
12	366
합	5097

그림 1 단어의 점수 빈도

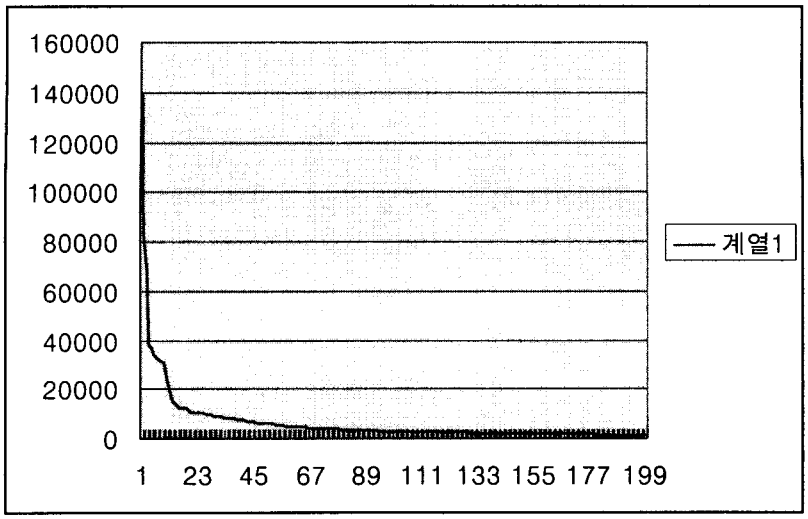

그림 3 단어 빈도

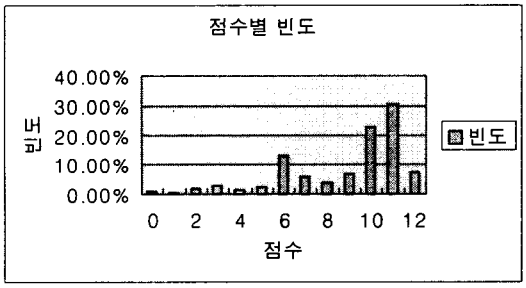

그림 2 단어의 점수변 빈도

그 외에 단순 형태가 아니었던 것은 "떠들다"는 단순형태와 동시에 합성 동사로서의 뜻이 있다. 이 동사는 단순 형태의 뜻으로 위의 기본 327개 단어에 하나 더 추가할 수 있다. "까닥거리다, 깜박이다, 꿀꿀거리다, 너울거리다, 들썩이다, 떵떵거리다" 등은 '동사파생접미사와 결합' 형태로 구분되어 있고, "갸웃하다, 떠들썩하다"는 '부사 + 하다' 형태이다. "가로놓이다, 가로채다, 발가벗다"는 '합성동사 중 비용언 형태와 결합' 형태이며, "내밀다, 덧나다"는 '접두사가 결합'한 형태이다. "도드라지다, 두드러지다, 뭉그러지다"는 '-지다 중 어근이 사전에 없는 것'이고, 동사로 기록한 "무릅쓰다"는 형용사로 관용표현에 해당한다. "기어오르다, 나오다, 내려치다, 떠받들다"는 '합성동사' 형태이다.

5. 비교 평가

사람이 수동으로 평가한 동사 중요도와 말모둠에서 나타나는 동사들의 빈도를 비교해 보았다. 동사들의 빈도를 구하기 위해 문화체육부와 과학기술처의 연구과제 국어정보처리기반구축과

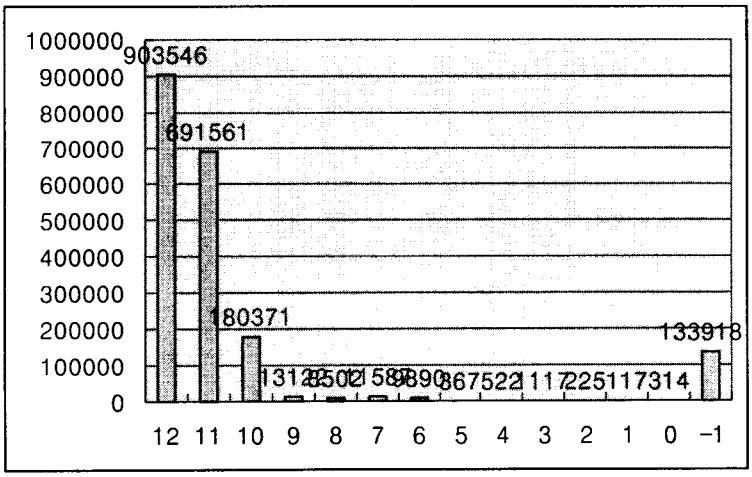

그림 4 점수별 단어 빈도

%	-1	0	1	2	3	4	5
수필	5.797	0.015	0.009	0	0.021	0.009	0.006
역사소설	7.159	0.008	0.004	0.014	0.055	0.028	0.038
장편소설	6.411	0.019	0.005	0.008	0.061	0.026	0.018
중단편 소설	6.426	0.017	0.005	0.01	0.056	0.035	0.013
인문/사회/학술	5.652	0	0.022	0.006	0.022	0.017	0

%	6	7	8	9	10	11	12
수필	0.37	0.643	0.603	0.491	7.167	35.637	49.231
역사소설	0.43	0.521	0.513	0.706	8.778	33.839	47.909
장편소설	0.521	0.577	0.494	0.673	9.387	35.306	46.495
중단편 소설	0.493	0.595	0.357	0.622	9.585	35.631	46.156
인문/사회/학술	0.202	0.601	0.163	0.416	6.855	37.813	48.23

그림 5 분야별 단어 분포표

STEP2000에서 구축된 과기원 말모둠 중에서 1997년에 만든 품사부착 말모둠을 사용했다. 이 품사부착 말모둠은 1160만 어절 수준이다.

총 2만 1361개의 서로 다른 동사류가 존재했다. 전체 동사들의 출현 횟수는 195만 5159번이다. 가장 많이 나온 동사는 일반 동사 "하다"로 13만 9019번(7.11%) 출현했다. 그림 3은 상위 빈도 200개의 출현횟수를 그래프로 나타낸 것이다.

말모둠에서 추출한 동사에 사람이 부여한 점수를 부여했다. 2만 1361개의 단어 중 4045개만 형태소 분석기 사전에 출현했고, 이 중에서 빼버려야 할 단어 12개도 들어가 있다. 일반 동사 "되다"(1.95%)가 형태소 분석기 사전에 들어가 있지 않았다. 13만 3918개(6.85%)의 동사류가 사전에 들어가 있지 않았다. 가장 좋은 점수를 받은 12점짜리 단어들은 90만 3546번(46.21%) 출현한다. 일반 동사 "득하다"는 빈도가 한 번인데도 12점으로 평가된 단어이다. 이런 형태로 서로 상호 조사하면서 오류를 수정해 나갈 수 있다. 11점짜리 단어는 69만 1561번(35.37%) 출현으로 12점과 11점 단어들을 합치면 80% 정도의 적용범위를 얻을 수 있다. 10점짜리 단어는 18만 371번(9.23%) 발생한다. 그림 4는 점수별 단어 빈도를 보여준다.

6. 중요도 이용가능성 실험

사람이 평가한 점수가 유용한 가에 대해 알아보기 위해서 단순한 실험을 해 본다. 문화체육부와 과학기술처의 연구과제 국어정보처리기반구축과 STEP2000에서 구축된 과기원 말모둠 중에서 1997년에 만든 품사부착 말모둠을 이용한다. 이 말모둠은 기본적으로 수필, 역사 소설, 장편 소설, 중단편 소설, 인문/사회/학술 5개 분야로 분류되어 있다. 중요도 점수에 따른 동사/형용사의 분야별 분포표는 그림 5와 같고, 그에 따른 분포도는 그림 6과 같다. 이 분포표로 말모둠의 특성을 파악할 수 있는지 이들 분야를 자동으로 묶어주는 실험을 해 본다. 실험에는 확률 벡터 모델을 사용한다.

6.1. 확률 벡터 모델

각 분야는 14차원의 확률 벡터모델[Wong 1987]에서 벡터로서 표현된다. 이 벡터간의 비교를 위해서 교차 엔트로피(Cross Entropy)를 사용한다. 확률 벡터(Probability Vector)는 다음과 같이 정의된다.

n차 벡터 $\vec{P}=(p_1,p_2,\cdots,p_n)$ 가 다음 식을 만족하면 n차

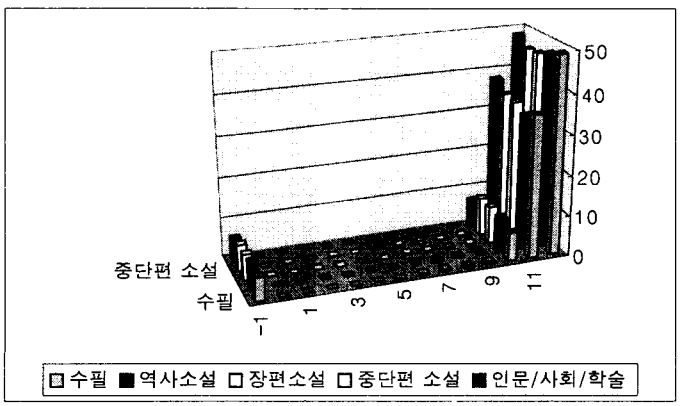

그림 6 분야별 단어 분포도

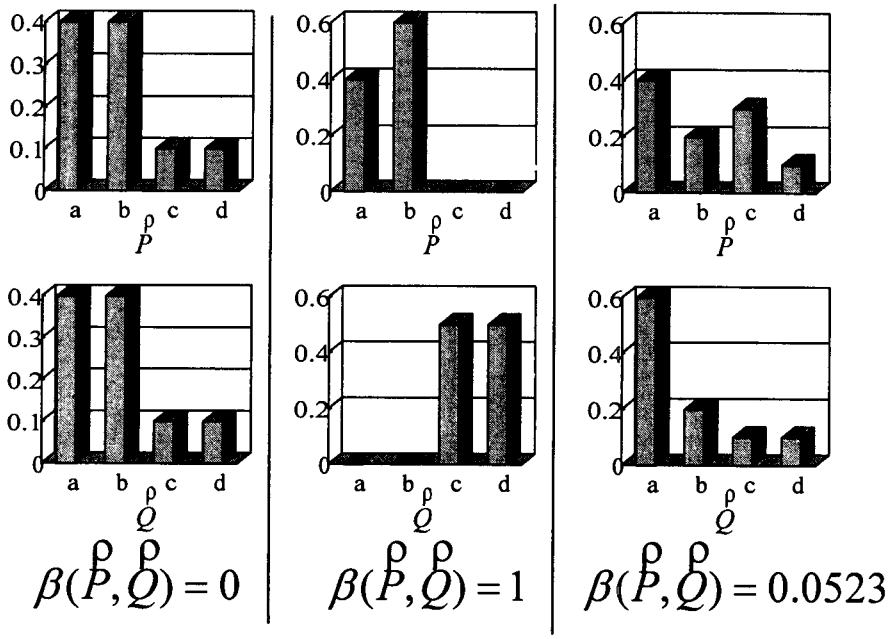

그림 7 확률벡터간의 의미거리 예

확률벡터이다.

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \geq 0, \quad i = 1,2,\cdots,n$$

분야 하나를 확률벡터 $\vec{P}$로 본다면, 각 p_i는 i번째 점수 단어의 분야에서 나타난 빈도를 확률로 나타낸다.

확률벡터 $\vec{P}$에 대한 엔트로피를 다음과 같이 정의한다.

$$H(\vec{P}) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

확률벡터 $\vec{P}$의 각 요소(element) p_i들의 불확실성(uncertainty)을 $-\log_2 p_i$ 의 값으로 측정할 수 있다. 그러므로, 엔트로피는 확률벡터 $\vec{P}$의 정보 불확실성(information uncertainty)에 대한 기대값이다. $H(\vec{P})$는 모든 요소들의 확률이 같을 때 최대값을 가지며, 한 요소만이 1이고 나머지는 0일 때 $H(\vec{P})$는 최소값 0을 가진다.

확률벡터 $\vec{P_1}$와 $\vec{P_2}$가 같은 차원의 벡터일 때, $\lambda \in [0,1]$ 에 대해서 벡터 $\vec{P} = \lambda\vec{P_1} + (1-\lambda)\vec{P_2}$ 도 역시 확률벡터이다. 이 때, 확률벡터 $\vec{P}$를 $\vec{P_1}$와 $\vec{P_2}$의 복합 확률벡터(Composite Probability Vector)라고 한다. 그리고, $\vec{P_1}$와 $\vec{P_2}$를 $\vec{P}$의 구성 확률벡터(Components Probability Vector)라고 한다.

확률벡터 $\vec{P} = (p_1, p_2, \cdots, p_n)$ 와 $\vec{Q} = (q_1, q_2, \cdots, q_n)$ 가 주어졌을 때, 복합 확률벡터 $\frac{1}{2}\vec{P} + \frac{1}{2}\vec{Q}$ 와 이 벡터의 구성 확률벡터들 사이의 엔트로피의 차이를 교차 엔트로피(Cross Entropy)라고 하며 다음과 같이 정의한다.

$$\beta(\vec{P},\vec{Q}) = H(\frac{1}{2}\vec{P} + \frac{1}{2}\vec{Q}) - \frac{1}{2}[H(\vec{P}) + H(\vec{Q})]$$

이 때, β는 다음의 부등식을 항상 만족한다.

$$0 \leq \beta(\vec{P},\vec{Q}) \leq 1$$

β의 값은 두 개의 확률벡터가 복합될 때, 불확실성의 증가 정도를 나타내고 있다. 만약 두 확률벡터가 관련되어 있으면 각각의 확률벡터 요소들의 확률 분포(probability distribution)가 유사하며 두 확률벡터가 관련이 많을수록 β의 값은 작아진다. 즉, 불확실성의 증가 정도가 적어진다. 이러한 β의 값은 두 확률벡터의 관계 차이를 나타낸다. 그러므로, β를 비관련도 계수(dissimilarity coefficient)로 해석할 수 있다. β를 분야간의 의미거리(concept distance) 값으로 사용한다.

그림 7은 확률벡터 모델에서 두 확률벡터간의 의미거리에 관한 예를 보인 것이다. 첫 번째 경우는 완전히 벡터가 일치할 의미 거리는 0이 된다. 두 번째 경우는 완전히 상이한 벡터인 경우로 이 때 의미거리는 1이 나온다. 세 번째는 비슷한 벡터를 비교하면 의미거리 값이 0에 가까이 나옴을 알 수 있다.

6.2. 분야 묶어주기

클러스터링 기법을 사용해서 분야별로 묶어 줬다. 확률벡터를 이용해 자동으로 묶인 순서는 처음에 장편 소설과 중단편 소설을 묶어 줬다. 그리고, 거기에 다시 역사 소설을 묶고 마지막으로 인문/사회/학술을 수필과 묶어서 2개로 분류했다. 간단히 그림으로 표현하면 그림 8과 같다.

이 실험은 단순한 실험이지만 묶어주는 순서가 사람의 직관과 같다는 것을 알 수 있었다. 이 실험을 통해서 직관적으로 중요도 점수가 효용성을 가짐을 알 수 있다. 그러나, 좀 더 명확한 평가를 위해서는 실험을 다시 설계해서 전체적으로 학습하고 학습에 따라서 분야를 자동으로 찾을 수 있는가에 대한 실험이 필요하다

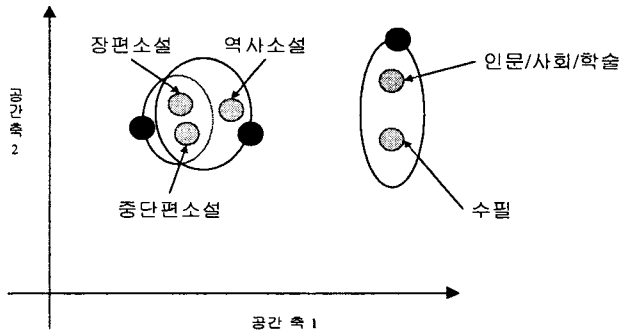

그림 8 분야 묶어주기

7. 맺음글

본 논문에서는 동사류 단어 빈도에 대해서 살펴보았다. 특히 사람이 수동으로 평가한 단어 중요도와 실제 발생하는 단어 빈도 사이의 연관성을 살펴보았다. 사람 평가가 중요 단어를 정한다는 것을 알 수 있었다. 하지만, 발생 빈도 순위 4위인 "되다" 일반 동사가 형태소 분석기 사전에 없어서 애초에 평가할 기회가 없기도 했다.

평가한 점수가 유용한 가를 알기 위해서 분야를 묶어주는 실험을 했다. 직관적으로 가까이 있는 분야가 묶이는 결과를 통해서 중요도 점수의 효과를 알 수 있었다. 이를 통해 문서 분야는 물론 사람의 문체 특징까지 표현할 수 있는 방법을 개발할 수 있다.

사람의 평가한 점수를 가지고, 형태소 분석기 사전의 수준을 결정할 수 있고, 번역기의 기본 어휘를 선택할 수 있다. 사람이 점수를 부여할 때 나누는 작업을 단계별로 하되 명확한 기준으로 통일해야 한다. 높임말, 맞춤법에 그른 단어들에 대한 처리도 필요하다. 사람이 평가할 때, 사전을 보면서 작업하면 모르던 단어도 아는 단어가 될 수 있으므로 사전을 보고 할 것인지 여부에 대한 정책이 필요하다.

우리는 본 논문에서 사람의 평가와 기계적으로 뽑은 빈도에 의한 평가가 상호 보완적으로 작용해서 완벽한 사전을 구성하는데 보탬이 된다는 것을 알았다. 앞으로는 수동 평가의 보완 작업을 통해 좀 더 좋은 형태소 분석기 사전을 구성할 수 있도록 노력해야 하며, 좀 더 명확한 실험을 통해 단어 중요도 점수의 유용성을 입증하는 것도 필요하다.

감사의 글

형태소 분석기 사전에 점수를 부착하는데 도와준 이주호, 오종훈, 김길연, 서충원에게 감사의 마음을 표한다.

참고문헌

[신중호 1999] 신중호, 박혁로, 한선화, "클러스터링 기법을 이용한 동사분류", 한국인지과학회 춘계 학술대회, 232-238쪽, 서울, 1999년 5월 29일

[이운재 1999] 이운재, 김선배, 김길연, 최기선, "모듈화된 형태소 분석기의 구현", 한글 및 한국어 정보처리 학술대회-형태소 분석기 및 품사태거 평가 워크숍, 123-136쪽, 전주, 1999년 10월 8-9일

[채영숙 1999] 채영숙, "구문분석을 전제로 한 전자사전 구축", 한국과학기술원 전문용어언어공학연구센터, 센터내부 메모, 1999년

[최용석 1999] 최용석, 이주호, 최기선, "격틀 자동구축과 격틀 평가 방법에 관한 연구", 한글 및 한국어 정보처리 학술대회, 272-279쪽, 전주, 1999년 10월 8-9일

[Wong 1987] S.K.M. Wong and Y.Y. Yao, "A Statistical Similarity Measure", Proceeding of the Tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 3-12, 1987.