형태소 사이의 유사도를 이용한 용례의 의미별 분류*

백 대 호°, 이 호, 임 해 창
고려대학교 전산과학과 자연어 처리 연구실

Conceptual Clustering of Korean Concordances using Similarities between Morphemes

Dae-Ho Baek°, Ho Lee, Hae-Chang Rim
Natural Language Processing Lab.,
Department of Computer Science, Korea University

요 약

본 논문에서는 정보 검색에서 사용하는 계층적 클러스터링 기법을 이용하여 용례들을 중심어의 의미에 따라 분류하고자 한다. 분류에 필요한 용례 사이의 유사도는 형태소 사이의 유사도를 이용하여 계산한다. 형태소 사이의 유사도 계산에는 상호 정보, 상호 정보의 유사도, 벡터 유사도 등을 사용한다. 품사 태깅된 17만 코퍼스에서 명사 4개와 동사 4개를 중심어로 사용하여 추출된 용례에 대해서 각 방법의 정확도를 실험한 결과 상호 정보와 상호 정보 유사도를 더한 값을 형태소 사이의 유사도로 사용한 방법이 90.16%의 정확도를 보였다. 제안된 방법에서 사용하는 정보들은 의미 태깅되지 않은 코퍼스에서 추출할 수 있기 때문에, 정보의 획득이 쉬운 장점이 있다.

1. 서 론

최근의 자연어 처리 연구는 대용량 코퍼스에서 획득한 다양한 언어 정보를 이용하는 추세이다. 이들 언어 정보들은 용례들로부터 획득될 수 있다. 그러나 대용량 코퍼스에서 용례를 추출할 경우, 추출되는 용례가 매우 많고 형태소들이 의미 중의성을 가지고 있기 때문에 원하는 언어 정보를 얻기가 매우 힘들다.

예를 들어 '눈을 감다', '실을 감다', '머리를 감다' 의 경우 '감다:동사'의 형태는 같지만 그 의미는 각각 '위아래 눈시울을 맞닿게 붙이다', '실이나 끈 따위를 다른 물체에 빙 두르다', '몸이나 머리를 물에 잠가서 씻다' 로 모두 다르다.

본 논문에서는 이러한 문제를 해결하기 위하여, 용례들을 용례 중심어의 의미별로 분류하는 방법을 제안한다. 용례 중심어란 '감다:동사' 용례의 경우 형태소 '감'이 용례의 중심어이다. 용례의 분류 방법은 정보 검색에서 사용하는 계층적 클러스터링 기법을 사용하고, 분류에 필요한 용례들 사이의 유사도는 용례에 나타나는 형태소 사이의 유사도를 이용하여 계산한다. 형태소 사이의 유사도는 상호 정보와 상호 정보 유사도, 벡터 유사도 등을 사용하여 계산한다.

2. 기존 연구

통계 기반의 자연어 처리에서는 학습 코퍼스에서 학습되지 않은 정보의 획득을 위해 단어의 클래스를 이용하거나 유사 단어를 이용한다. 이를 위해 자동으로 단어의 클래스를 만들거나, 유사 단어를 추출하는 방법에 대한 많은 연구들이 있었다.

Brown은 코퍼스에서 좌우 50단어 이내에 인접해 나타나는 단어들간의 상호 정보를 이용하여 단

* 본 연구는 시스템공학연구소 위탁과제 "한국어 정보 처리를 위한 언어 정보 획득 도구의 개발"의 지원을 받은 것입니다.

어들을 분류하였고[Brown 92], Yarowsky는 단어 자체가 아니라 단어의 클래스를 이용하여 의미 중의성 있는 단어의 클래스를 찾았는데, 여기서 단어 클래스는 Roget's Category를 이용하였다[Yarowsky 92]. Dagan은 상호 정보의 유사도를 이용하여 유사 단어를 찾는 방법과[Dagan 93], 상대 엔트로피를 사용해서 유사 단어를 찾는 방법을 사용하였다[Dagan 94]. 그 밖에도 목적어와 동사의 분포 유사도를 상대 엔트로피를 이용하여 계산함으로써 단어간의 유사도를 계산하는 연구와[Pereira 93, 조정미 95], 단어의 바이그램 빈도를 이용하여 단어의 유사도를 계산하고 품사가 비슷한 단어 클래스를 만드는 연구도 있었다[이공주 96]. 대부분의 연구들이 단어 사이의 유사도를 단어의 분류, 학습 코퍼스에서 학습되지 않은 통계 정보를 대체하기 위한 수단, 기계 번역에서 단어의 의미 중의성 해결 등에 사용하였다. 본 연구에서는 형태소 사이의 유사도를 용례를 중심어의 의미별로 분류하는데 적용한다.

3. 용례 사이의 유사도

용례를 분류하려면 각 용례 사이의 유사도를 계산하여야 한다. 본 논문에서는 각 용례 사이의 유사도를 계산하기 위해 용례에 같은 형태소가 나타나는 빈도를 이용하는 방법과, 형태소 사이의 유사도를 이용하는 방법을 사용한다.

3.1 같은 형태소가 나타나는 빈도를 이용하는 방법

이 방법은 "두 용례에 같은 형태소가 많이 나타날 수록 두 용례의 의미가 더 유사하다"는 가정을 바탕으로 한다. 그러나 문장에 형태소가 많을수록 같은 형태소가 나타날 가능성이 많으므로, 용례사이의 유사도가 문장의 길이에 영향을 받지 않게 하기 위해서, 두 용례에 나타나는 같은 형태소 수를 용례의 형태소 수의 평균으로 나누어준다. 같은 형태소가 나타나는 빈도를 이용하여 용례들 사이의 유사도를 계산하는 식은 다음과 같다.

$$S_E(E_1, E_2) = \frac{2\ n(E_1 \cap E_2)}{n(E_1) + n(E_2)}$$

E_1, E_2는 동일한 형태소가 중심어인 용례이며, $S_E(E_1, E_2)$는 두 용례 E_1, E_2 사이의 유사도이다. $n(E_1)$은 용례 E_1에 있는 형태소의 수이고, $n(E_1 \cap E_2)$는 두 용례 E_1, E_2에 동시에 나타나는 형태소의 수이다.

이 방법을 사용할 경우에는 실제 용례에 사용되는 형태소가 매우 다양하여 두 용례에 같은 형태소가 한번도 나타나지 않는 경우가 많고, 두 용례에 같은 형태소가 나타나지 않으면 용례 사이의 유사도를 계산할 수 없으므로 용례를 분류하는데 어려움이 있다.

3.2 형태소 사이의 유사도를 이용하는 방법

이 방법에서는 두 용례 사이의 유사도를 계산하기 위해 두 용례에 나타나는 형태소 사이의 유사도를 사용한다. 이를 식으로 나타내면 다음과 같다.

$$S_E(E_1, E_2) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} S_M(E_{1i}, E_{2j})}{mn}$$

E_{1i}는 용례 E_1의 i번째 형태소이고, E_{2j}는 용례 E_2의 j번째 형태소이며, $S_M(E_{1i}, E_{2j})$는 두 형태소 E_{1i}, E_{2j} 사이의 유사도이다. m, n은 각각 용례 E_1, E_2의 형태소 개수이다.

형태소 사이의 유사도는 상호 정보, 상호 정보의 유사도, 벡터 유사도 등을 이용하여 계산한다.

3.2.1 상호 정보를 이용하는 방법

상호 정보를 이용한 방법은 "문장내의 형태소들은 그 문장의 의미를 결정하는데 기여를 하므로, 같은 문장에 나타나는 형태소들은 어느 정도의 의미 유사도를 가지고 있다"는 가정에 바탕을 두고 있다. 상호 정보를 사용하면 형태소가 두 용례에는 같이 나타나지 않더라도, 학습 코퍼스에서 같은 문장에 나타난 적이 있다면 형태소 사이의 유사도를 계산할 수 있고 이를 이용하여 용례 사이의 유사도를 계산할 수 있다. 상호 정보를 계산하는 식은 다음과 같다.

$$S_M(M_1, M_2) = MI(M_1, M_2) = \log_2 \frac{P(M_1, M_2)}{P(M_1)P(M_2)} = \log_2 \frac{f(M_1, M_2)N}{f(M_1)f(M_2)}$$

$P(M_1, M_2)$는 형태소 M_1, M_2가 학습 코퍼스에서 같은 문장 내에서 나타나는 확률이고, $f(M_1)$는 학습 코퍼스에서 형태소 M_1의 출현 빈도이다. M_1, M_2가 같은 형태소인 경우에는 $f(M, M) = f(M)$으로 계산한다.

$$S_M(M, M) = \log_2 \frac{N}{f(M)}$$

유사도 값이 0보다 작은 경우는 유사도를 0으로 하여, 유사도는 항상 0보다 크거나 같은 값을 가진다.

이 방법을 사용할 경우에는 학습 코퍼스에서 두 형태소가 한번도 같이 나타난 적이 없으면 상호 정보가 0이 되어서 형태소 사이의 유사도를 계산할 수가 없다. 실제로 본 실험에서 사용된 코퍼스에서 95.76%의 형태소 상호 정보가 0이었다.

3.2.2 상호 정보의 유사도를 이용하는 방법

Dagan은 "유사한 두 단어는 다른 단어와의 상호 정보가 비슷하다"는 가정을 가지고 상호 정보가 유사한 단어들을 유사한 단어들로 간주하여, 단어의 유사도 계산에 왼쪽 상호 정보와 오른쪽 상호 정보를 이용한 다음 식을 사용하였다[Dagan 93].

$$Sim(w_1, w_2) = \frac{\sum_{w \in lexicon} \min(I(w, w_1), I(w, w_2)) + \min(I(w_1, w), I(w_2, w))}{\sum_{w \in lexicon} \max(I(w, w_1), I(w, w_2)) + \max(I(w_1, w), I(w_2, w))}$$

본 논문에서는 상호 정보를 계산할 때 방향성을 고려하지 않기 때문에 아래의 식으로 3.2.1에서 구해진 상호 정보들의 유사도를 계산하고, 이 유사도를 형태소 사이의 유사도로 사용한다.

$$S_M(M_1, M_2) = \frac{\sum_{M \in lexicon} \min(I(M, M_1), I(M, M_2))}{\sum_{M \in lexicon} \max(I(M, M_1), I(M, M_2))}$$

3.2.3 벡터 유사도를 이용하는 방법

이 방법에 사용된 속성 벡터는 모든 형태소이고 그 형태소와 같이 나타나는 빈도가 그 속성 요소의 값이 된다. 벡터 유사도를 계산하는 방법에는 Dice Coefficient, Cosine Coefficient, Jaccard Coefficient 등이 있다[Salton 89, Frakes 92].

각각의 계산 방법은 아래와 같다.

$$Dice\ Coefficient = \frac{2\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2}$$

$$Cosine\ Coefficient = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}}$$

$$Jaccard\ Coefficient = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N} x_i y_i}$$

x_i, y_i는 각각 형태소 M_1, M_2의 속성 벡터의 i번째 속성 요소이다.

4. 용례의 계층적 분류

3장에서 설명된 방법으로 모든 용례들 사이의 유사도를 계산할 수 있으면, 이를 이용하여 용례들을 분류할 수 있다. 용례의 분류에는 정보 검색에서 많이 사용하는 계층적 분류 기법을 사용하였다. 계층적 분류에는 클러스터 사이의 유사도를 계산하는 방법에 따라 Single Link Clustering, Complete Link Clustering, Group Average Link Clustering 등이 있다. Single Link Clustering은 두 클러스터 안의 가장 유사한 두 요소간의 유사도를 클러스터 간의 유사도로 사용하는 방법이고,

용례 중심어	출현 빈도	용례에서 사용된 의미
눈_명사	225	사람이나 동물의 보는 감각을 일으키는 기관
	31	공중에 떠다니는 김이 찬 기운을 만나 얼어서 땅위로 떨어지는, 희고 여섯모가 난 결정체
	1	싹이 막 터져 돋아나오는 자리
밤_명사	58	저녁 어두운 뒤로부터 새벽 밝기까지의 동안
	13	밤나무의 열매
배_명사	10	배나무의 열매
	8	사람이나 동물의 몸뚱이에 가슴과 엉덩이 사이의 위장, 창자, 콩팥 등이 들어 있는 부분
	5	사람이나 물건을 싣고 물위에 떠다니도록 나무나 쇠로 만든 기구
이상_명사	114	정상적인 상태와 다름
	18	일정한 표준보다 많거나 나은 것
	3	생각할 수 있는 가장 완전한 상태
감_동사	37	위아래 눈시울을 맞닿게 붙이다
	8	실이나 끈 따위를 다른 물체에 빙 두르다
	2	몸이나 머리를 물에 잠가서 씻다
걷_동사	73	다리를 번갈아 떼어 옮기다
	17	널려 있거나 흩어져 있는 것을 한데 모아 들이다
	3	가리거나 늘어뜨린 것을 말아 올리거나 추켜올리다
묻_동사	103	남에게 대답이나 설명을 구하다
	14	가루, 액체, 끈끈한 것 따위가 그보다 큰 물체에 들러붙다
	7	물건을 흙이나 다른 물건 속에 넣어 감추다
쓰_동사	155	어떤 일에 재료, 돈 따위를 들이다
	44	글을 짓다
	10	모자 따위를 머리 위에 얹어 덮다
	2	맛이 소태나 씀바귀의 맛과 같다

[표 1] 실험에 사용된 용례 중심어의 코퍼스 내 의미와 출현 빈도

Complete Link Clustering은 두 클러스터 안에서 가장 유사하지 않은 요소간의 유사도를 클러스터 간의 유사도로 사용하는 방법이며, Group Average Link Clustering은 두 클러스터에 있는 모든 요소 간의 유사도의 평균을 클러스터 간의 유사도로 사용하는 방법이다[Salton 89, Frakes 92].

5. 실험 및 평가

앞에서 제안된 모델을 가지고 [김진동 96] 품사 태거에 의해서 품사 태깅된 약 175,000단어, 16,000 문장으로 이루어진 코퍼스로부터 고려대학교 자연어 처리 연구실에서 개발된 용례 추출기를 이용하여 추출된 용례를 분류하였다. 용례 추출에 사용된 중심어는 명사 4개 동사 4개이고, 각 형태소의 용례 내에서의 의미와 빈도는 [표 1]과 같다[한글학회 95]. 형태소 사이의 유사도 계산에 사용된 형태소는 용례 중심어 어절을 제외하고 품사 태그가 체언, 용언, 관형사, 부사, 감탄사인 형태소만을 대상으로 하였다. 조사, 어미 등을 포함할 경우 정확도가 낮아지고, 컴퓨팅 타임과 메모리 등 자원이 더 많이 필요하다.

정확도 평가 기준은 중심어의 의미가 같은 용례가 같은 클러스터에 있을 때 정확하게 분류된 것으로 간주하고, 다음 식에 의해서 평가하였다.

$$\text{정확도} = \frac{\text{정확하게 분류된 용례 수}}{\text{전체 용례 수}} \times 100$$

실험은 3장에서 설명된 여러 가지 유사도 계산 방법과 4장에서 설명된 계층적 클러스터링 기법을 사용하였다. 실험 1은 3.1절에서 설명된 같은 형태소가 나타나는 빈도를 사용한 방법의 결과이고, 실

용례 중심어	클러스터링 방법	실험 1 클러스터 수	실험 1 정확도 (%)	클러스터 수	실험 2 정확도 (%)	실험 3 정확도 (%)	실험 4 정확도 (%)	실험 5 정확도 (%)	실험 6 정확도 (%)
눈_명사	Single	18	88.71	18	87.93	87.54	89.88	87.93	87.93
	Complete	36	89.49	18	88.71	89.10	90.27	91.82	92.60
	Average	18	88.71	18	91.43	89.49	91.43	90.27	88.71
밤_명사	Single	8	83.09	8	84.50	85.91	81.69	84.57	84.50
	Complete	18	85.91	8	88.73	84.50	81.69	88.73	88.73
	Average	8	83.09	8	91.54	84.50	81.69	88.73	88.73
배_명사	Single	5	60.86	5	65.21	73.91	82.60	65.21	65.21
	Complete	11	82.60	5	82.60	65.21	71.91	82.60	82.60
	Average	5	65.21	5	78.26	78.26	65.21	78.26	78.26
이상_명사	Single	10	89.62	10	88.14	89.62	88.14	88.14	88.14
	Complete	23	92.50	10	94.07	97.03	93.33	95.55	96.29
	Average	10	94.81	10	97.03	95.55	94.07	97.03	97.03
감_동사	Single	3	89.36	3	82.97	89.36	89.36	82.97	82.97
	Complete	5	95.74	3	100.00	95.74	91.48	100.00	100.00
	Average	3	95.74	3	100.00	95.74	95.74	100.00	100.00
걷_동사	Single	15	91.39	15	83.87	86.02	82.79	83.87	83.87
	Complete	29	93.54	15	90.32	92.47	92.47	90.32	90.32
	Average	15	89.24	15	86.02	86.02	90.32	87.09	87.09
묻_동사	Single	15	83.87	15	87.90	86.29	85.48	87.90	88.70
	Complete	27	85.48	15	87.90	87.90	84.67	87.09	85.48
	Average	15	84.67	15	85.48	83.87	83.87	85.48	86.29
쓰_동사	Single	20	76.30	20	77.25	77.72	76.77	77.72	78.19
	Complete	39	85.30	20	81.99	84.83	86.25	83.41	85.30
	Average	20	80.56	20	81.99	84.83	84.83	85.30	82.46

[표 2] 용례의 의미별 분류에 대한 실험 결과

클러스터링 방법	실험 1	실험 2	실험 3	실험 4	실험 5	실험 6
Single Link	82.90	82.21	84.54	84.58	82.28	82.43
Complete Link	88.82	89.29	87.09	86.50	89.94	90.16
Group Average Link	85.25	88.96	87.28	85.89	89.02	88.57

[표 3] 각 실험의 정확도 (%)

험 2는 3.2.1절의 상호 정보를 이용한 방법의 결과이다. 그리고 실험 3은 3.2.2절에서 설명된 상호 정보의 유사도를 사용한 방법을 적용시킨 결과이며, 실험 4는 벡터 유사도를 Cosine Coefficient를 사용하여 계산하여 실험한 결과이다. 실험 5는 ["상호 정보" + α × "벡터 유사도"]를 형태소 사이의 유사도로 사용하여 실험한 결과이며, 실험 6은 ["상호 정보" + α × "상호 정보 유사도"]를 형태소 사이의 유사도로 사용하여 실험한 결과이다. 실험 5에서는 α가 1일 때, 실험 6에서는 α가 2일 때 정확도가 가장 좋았다.

Dice Coefficient와 Jaccard Coefficient를 사용하여 벡터 유사도를 계산하여 실험한 결과는 Cosine Coefficient를 사용한 경우 보다 클러스터 수가 많고 정확도도 좋지 않았다.

클러스터 수는 실험 1의 Complete Link Clustering 방법의 클러스터 수의 반 정도로 하여 실험하였다. 그 이유는 클러스터 수가 다르면 정확도를 비교할 수가 없으므로 클러스터 수를 같게 주었고, 클러스터 수가 너무 많으면 어떤 방법이 좋은지 비교할 수 없기 때문이다. 실험 1의 Complete Link Clustering의 결과가 클러스터 수가 많은 것은 두 용례에 같은 형태소가 나타나지 않아서 용례 사이의 유사도를 계산할 수 없는 경우가 많기 때문이다.

클러스터링 방법은 Complete Link Clustering이 정확도가 가장 좋았고, 형태소 사이의 유사도는 상호 정보와 상호 정보 유사도를 합한 값을 사용한 실험 6이 가장 좋은 정확도를 보였다.

6. 결 론

본 논문에서는 형태소 사이의 유사도를 이용하여 용례들을 중심어의 의미별로 분류하였다. 두 용례에 같은 형태소가 나타나면 의미적 유사성이 있다는 가정이 80%이상 정확하다는 것을 실험 1을 통해서 알 수 있었다. 같은 형태소가 나타나지 않는 용례들도 형태소 사이의 유사도를 사용하여 의미적 유사도를 계산할 수 있었고, 상호 정보와 상호 정보 유사도의 합을 형태소 사이의 유사도로 계산하고, Complete Link Clustering 기법을 사용하여 용례를 분류했을 때 90.16%의 정확도를 보였다. 본 연구에서 사용된 상호 정보와 상호 정보 유사도, 벡터 유사도는 의미 태깅되지 않은 코퍼스에서 추출할 수 있기 때문에 정보의 획득이 쉬운 장점이 있다.

참고 문헌

[Brown 92] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai and Robert L. Mercer. 1992. "Class-Based n-gram Models of Natural Language", *Computational Linguistics*, Vol. 18, No. 4, pp. 467-479

[Dagan 93] Ido Dagan, Shaul Marcus and Shaul Markovitch. 1993. "Contextual Word Similarity And Estimation From Sparse Data", *In Proceedings of the 31th Meeting of the ACL*, pp. 164-171

[Dagan 94] Ido Dagan, Fernando Pereira and Lillian Lee. 1994. "Similarity-Based Estimation of Word Cooccurrence Probabilities", *In Proceedings of the 32th Meeting of the ACL*, pp. 272-278

[Frakes 92] William B. Frakes and Ricardo Baeza-Yates. 1992. "Information Retrieval", *Prentice Hall*

[Luk 95] Alpha K. Luk. 1995. "Statistical Sense Disambiguation with Relatively Small Corpora Using Dictionary Definitions", *In Proceedings of the 33th Meeting of the ACL*, pp.181-188

[Pereira 93] Fernando Pereira, Naftali Tishby and Lillian Lee. 1993. "Distributional Clustering Of English Words", *In Proceedings of the 31th Meeting of the ACL*, pp. 183-190

[Salton 89] Gerard Salton. 1989. "Automatic Text Processing", *Addison-Wesley Publishing Company*

[Yarowsky 92] David Yarowsky. 1992. "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", *Proc. of COLING-92*, pp. 454-460

[김진동 96] 김진동. 1996. "어절 문맥을 고려하는 형태소 단위의 한국어 품사 태깅 모델", *고려대학교 전산과학과 석사학위논문*

[이공주 96] 이공주, 김재훈, 김길창. 1996. "한국어에서의 단어 자동 분류와 품사 분류 체계", *1996년도 한국정보과학회 봄 학술발표논문집 Vol. 23, No. 1*, pp. 961-964

[이호 94] 이호. 1994. "언어 정보 획득을 위한 한국어 코퍼스 분석 도구", *고려대학교 전산과학과 석사학위논문*

[조정미 95] 조정미, 김길창. 1995. "분포 정보를 이용한 의미 중의성을 지닌 한국어 동사의 의미 분별", *95 한글 및 한국어 정보처리 학술발표논문집*, pp. 56-61

[한글학회 95] 한글학회. 1995. "흔글 우리말 큰사전 1.0", *한글과 컴퓨터*