

20170221 Jaehyeon Myung

### 1. 문제 정의

다양한 경우에 대해 Restricted 된 intensifier를 자동으로 찾기 위해 먼저 두가지 설정을 하였다. Intensifier로 사용되는 단어 중 예시를 비롯한 대부분이 adjective를 수식하는 adverb형태이거나 noun을 수식하는 adjective 두가지에 대한 자동화를 하였다. 또한 "Restricted"에 대해서 피수식자 집단의 유사도가 클수록 intensifier가 더 restricted 되었다고 정의하였고, 이에 따라 피수식자 집단의 similarity와 단어의 반복성으로 이를 계산하였다. 해당 과제의 경우 Brown Corpus의 전체 corpora를 활용하였다.

### 2. Degree of Restricted

앞서 말했듯 Restricted된 정도를 파악하기 위해 수식된 형용사와 명사의 유사도를 검사하였다. 먼저 전체에 대한 intensifier-adj/noun 쌍을 조사하여 모두 저장한 다음 각 intensifier에 대해 피수식자들끼리 hand-shaking 방식으로 중복에 상관없이 이들의 유사도를 검사하여 평균값을 구하였다(이 과정에서 시간복잡도가 증가하여 execution time이 오래 걸림). 또한 wordnet에 내장된 wup\_similarity()의 경우 verb와 noun에 한정되어 결과값이 나오기 때문에 이를 해결하기 위해 adjective를 noun의 꼴로 변형하는 "adjective\_to\_noun()"을 정의하였다.

### 3. Intensifier

많은 부사와 형용사들 중에서 Intensifier를 정의하기 위해 manual로 HW1에서 사용한 대표적인 intensifier 11가지를 제공하였다. 이 단어들에 대한 유사도가 WUP 유사도 기중 0.3이상일Eo intensifier 라고 정의하였다.

### 4. 추가 작업들

정확도를 높이기 위해 몇가지 설정들을 추가하였다.

- ➔ 출력 값에서 restricted된 정도는 매우 좋으나 appearance가 1회 이하로 극히 적은 경우 제외하였다.
- ➔ Restricted된 정도를 파악하기 위해 Scoring을 하는 방법을 2가지로 구현했는데, 단순히 빈도수/전체 빈도수 로 점수를 측정한 경우와, 각 adj, adv에 대한 다른 scoring 방식이다. 전자의 경우 정확도가 조금 더 떨어진다고 판단이 되었으며, 앞서 정의한 restricted와 거리가 있어 후자의 결과로 제출을 하였다.
- ➔ 미리 제시한 Manual에 있는 intensifier의 경우 대중적으로 사용되는 intensifier로 간주하여 해당 intensifier에 의해 수식되는 경우는 제외하였다.

### 5. Quality of Output

Csv 형태로 출력된 output 값이 완전히 예상한 값처럼 나오지는 않았다. 전체적으로 봤을때 앞서 정의한 "restricted"를 측정하는 방법에 있어 빈도수가 적은 경우 피수식 집단 유사도에서 유리한 점수를 가져가게 되는 결과가 나왔다. 따라서 등장 횟수가 적은 intensifier가 우선순위로 나타났으며, brown corpus의 특성상 비슷한 주제(뉴스 거리) 들 안에서 4회 이하로 언급된 단어들이 우선순위를 가져 정확성이 떨어졌다고 생각한다.

### 6. Improvement

- ⇒ Sentiment 분석을 활용하여 adj와 adv를 noun으로 변환하기 않고 유사도를 측정해야 좋은 정확도가 나올 것으로 예상된다. 현재는 명사로 전환시 부정확도가 높음
- ⇒ 앞서 정의한 adverb의 경우 adjective외에도 adverb자체에 대한 수식도 하므로 이에 대한 구현은 더 많은 모집단과 정확도를 높일 것이다.
- ⇒ 강한 intensity를 가진 intensifier가 restricted된 정도가 더 강하다는 점을 고려하여 manual에 추가해주면 개선될 것이다.
- ⇒ Scoring에 있어 더 많은 실험으로 최적의 모델을 찾아야 더 정확도가 올라갈 것이다. Similarity와 단어의 출현 빈도 사이에 수식적으로 여러가지 모델을 세워(ML등을 사용) 학습시키면 점점 더 좋은 결과가 나올 것으로 예상된다.