공기정보 벡터를 이용한 한국어 명사의 의미구분

신사임 이주호 최용석 최기선
전문용어언어공학연구센터, 첨단정보기술연구센터, 한국과학기술원
{miror, mywork, angelove, kschoi}@world.kaist.ac.kr

Word Sense Disambiguation Using of Co-occurrence Information Vectors

Sa-Im Shin, Juho Lee, Yong-Seok Choi, Key-Sun Choi
KORTERM, AITRC, KAIST

요 약

본 논문은 문맥의 공기정보를 사용한 한국어 명사의 의미구분에 관한 연구이다. 대상 명사에 대한 문맥의 지엽적인 단어분포는 명사의 의미구분을 위한 의미적 특성을 표현하는데 충분하지 못하다. 본 논문은 의미별로 수집한 문맥 정보를 기저 벡터화 하는 방법을 제안한다. 정보의 중요도 측정을 통하여 의미구분에 불필요한 문맥정보는 제거하고, 남아있는 문맥의 단어들은 변별력 강화를 위하여 상의어[1] 정보로 바꾸어 기저벡터에 사용한다. 상의어 정보는 단어의 형태와 사전 정의문의 패턴을 통해 추출한다. 의미 벡터를 통한 의미구분에 실패하였을 경우엔 훈련데이터에서 가장 많이 나타난 의미로 정답을 제시한다. 실험을 위해 본 논문에서는 SENSEVAL 실험집합을 사용하였으며, 제시한 방법으로 공기정보의 가공 없이 그대로 실험한 방법과 비교하여 최고 42% 정도의 정확률 향상을 나타내었다.

1. 서론

훈련데이터의 문맥정보를 사용하는 것은 단어의 의미구분에서 잘 알려진 방법 중에 하나이다(Agirre 1996, Esscudero 2000, Gruber 1991). 한 명사가 제한된 거리 안에 공기 하고 있는 단어들이 유사한 경우에는 애매한 여러 의미 중 같은 의미를 나타내는 경향이 있다. 변별력 있는 문맥정보를 척도화 하여 정확한 의미구분을 하기 위한 중요한 문제는 의미구분에 가장 영향력 있는 패턴을 찾아내는 일이다.

단어가 같은 의미로 쓰일 때 유사한 문맥과 공기정보를 가지는 것은 사실이지만 (Rigau, 1997), 그 정보는 너무 지엽적이고 불필요한 정보와 섞여 있어 의미구분에 그대로 적용하기는 부적합하다. 각각의 명사의 맥에서 단어의 공기정보는 너무 다양하여서 각 의미별로 정확한 패턴을 추출하기가 어렵기 때문이다. 그러므로, 문맥의 단어들을 그들의 상의어 정보로 일반화하고, 가중치 비교를 통해 문맥에서 불필요한 정보들을 제거한 후에, 이 정보들을 변별력 있는 형태로 바꿔주는 SVD (Singular Vector Decomposition) 기저벡터로 변환한다. 이 과정은 위에서 선택한 공기정보를 m x n 행렬 형태의 다차원 벡터로 합성하고 변환하게 되는데, 여기서 m은 의미구분에서 가능한 의미의 수이고 n은 추출한 문맥정보의 수를 의미한다(Michael, 1995). 이 과정은 추출한 공기정보를 구조적인 형태로 변환하고 의미 있는 정보를 중심으로 벡터를

[1] 의미적으로 다른 단어들을 포함하는 단어

형상화하여 차원축소의 효과를 얻을 수 있다(Jen, 1998). 각각의 단어는 사전 정의문과 단어 형태에서 추출한 상의어를 가지고 일반화한다. 이 과정은 기저벡터를 좀 더 일반적이면서도 작은 차원으로 정보를 표현할 수 있도록 해준다. 그리고 tf/idf값을 기반으로 한 기준 값을 사용하여 의미구분에 영향을 주지 않는 불필요한 정보를 제거한다.

본 논문의 순서는 다음과 같다 : 2장은 본 논문에 적용한 방법들을 보여준다. 3장은 실험을, 4장은 오류분석을 통한 토론을 다룬다. 5장에서는 결론을, 6장에서는 향후 계획을 제시한다.

2. 방 법

2.1 공기정보의 추출

본 논문에서 고려하는 문맥의 크기는 대상 명사를 포함하는 문장과 그 문장의 앞뒤 두 문장씩, 5문장으로 한정한다. 문맥은 다양한 문맥특징을 충분히 반영할 수 있어야 한다. 포함하는 문맥의 크기가 너무 넓다면, 문맥은 관련 있는 정보만을 일관성 있게 포함할 수 없다. 그러므로, 문맥의 크기는 다섯 문장이 관련 있는 정보를 충분하게 반영하는데 적합하다. 이 문맥 크기 안에 있는 명사, 동사, 대상명사의 수식어와 대상명사의 지배동사를 구분하여 추출한다.

문맥 안의 명사와 동사는 대상명사와의 공기형태를 보여준다고 가정한다. 공기형태는 빈도수를 기반으로 한 벡터로 나타난다.

명사 수식어는 명사의 의미구분에 적합한 의미 정보를 포함한다. 예를 들어, '밤' 은 '시간 표현의 밤' 과 '과일의 밤' 의 두 가지 뜻을 가질 수 있다. 그러나 '맛있는 밤' 에서 '맛있는' 이라는 수식어는 '밤' 의 의미가 두 번째 의미의 '밤' 으로 사용되는 경우에 대부분 나타난다. 반면에, '어두운 밤' 이라는 표현에서 '어두운' 이라는 수식어는 과일을 뜻하는 ' 밤' 보다는 '밤' 이 첫번째 의미로 나타날 때 대부분 등장한다.

또한, 대상 명사를 지배하는 동사는 그 동사에 가중치를 주어 의미정보에 반영하였다. 대상 명사를 지배하는 동사는 문맥 안의 다른 동사들보다 대상 명사와 의미적으로 더욱 밀접한 관계를 가지기 때문이다.

2.2 사전으로부터의 상의어 추출

추출한 공기정보가 문맥 지엽적인 단어정보를 그대로 가지고 있다면, 문맥정보가 분산되고 의미벡터가 너무 많은 축을 가지게 되므로, 정확한 비교가 어려워지게 된다. 그러므로, 본 논문에서는 지엽적인 공기정보를 사전에서 추출한 상의어 정보로 일반화하는 방법을 제안한다. 예를 들어, 우리가 '가맹국' 을 '국가' 로, '가시밭길' 을 '길' 로, 또한 '가을바람' 을 '바람' 이라는 상의어로 일반화할 수 있다. 이런 일반화된 정보를 사용한다면, 추출한 공기정보를 그대로 사용하는 것보다 더 변별력 있는 의미구분이 가능하다.

본 논문에서는 수동 구축된 의미 체계의 적용 대신 사전 정의문과 표제어의 패턴분석을 통하여 상의어를 찾는 방법을 제시한다. 합성 명사의 경우, 그 명사의 상의어 정보를 포함하고 있는 경우가 많다. 예를 들어, '벚나무' 처럼 '~나무/~꽃' 으로 끝나는 단어들의 상의어는 각각 '나무' 와 '꽃' 이다. '나무' 와 '꽃' 은 또한 사전 정의문의 패턴을 통하여 '식물' 이라는 상의어를 유추할 수 있다. 사전 정의문을 통한 상의어 추출의 예를 들어보면, '가정교사' 라는 단어의 경우, 이 단어의 사전 정의문인 '가정에서 공부를 가르쳐 주는 사람' 을 살펴보면, '가정교사' 의 상의어인 '사람' 이 정의문의 맨 뒤에 위치하는 것을 볼 수 있다. 이처럼, 사전 정의문이 명사로 끝나는 대부분의 경우, 그 끝나는 명사가 대상 표제어의 상의어라는 정의문 패턴을 통하여 상의어를 추출할 수 있다. 앞서 설명한 두 가지 방법으로 사전의 정의문과 표제어의 형태정보를 통하여 비교적 정확하게 상의어 정보를 추출할 수 있었다.

본 논문에서는 약 20여 개의 패턴을 우리말 큰 사전(1997)의 모든 표제어에 적용하여 상의어 리스트를 추출하였다.

문맥의 단어가 상의어 리스트 안에 있는 경우, 우리는 해당 단어를 상의어로 변환한 뒤 공기정보를 재계산 하였다. 공기정보를 상의어로 변환한 결과, 지엽적인 공기정보보다 65% 정도로 줄어든 단어로 공기정보를 표현할 수 있었다.

2.3 불필요한 단어의 제거

문맥에서 추출한 단어 중에는 의미구분에 영향을 주지 않는 단어들을 포함하는데, 이런 단어들은 크게 두 가지로 나눌 수 있다. 첫째, 문맥에서 단어의 빈도수가 너무 작아서 의미구분에 적절하게 반영하기 어려운 고유명사 같은 단어들이다. 이런 단어들은 공기정보를 SVD로 변환할 때, 의미벡터 축의 수를 증가하게 해서 불필요하게 기저벡터를 복잡하게 만든다. 두 번째 문제는 대명사와 같은 고 빈도 단어들이다. 이런 단어들은 대상 명사의 의미와 관계없이 어디서나 너무 많이 나타난다. 그러므로, 이 같은 고 빈도 단어들은 문맥에서 높은 빈도수를 가지고 있음에도 모든 의미에서 유사한 분포로 나타나기 때문에 의미구분에 중요한 정보가 아니고 오히려 방해가 되는 요인이 된다.

공기정보에서 이 같은 불필요한 단어들을 제거하기 위하여, 본 논문에서는 tf (Term Frequency)와 idf (Inversed Document Frequency)를 사용한다. tf는 훈련 데이터에서 단어의 빈도수를 나타내고, idf는 훈련 데이터에서 그 단어를 포함하는 샘플의 수를 나타낸다

(Hinrich, 1998). 즉, tf는 문맥 단어의 빈도수를 나타내고 idf는 단어의 분포정도를 표현하고, 우리는 이 분포 정도와 빈도 정도를 하나의 값으로 합쳐서 기준을 삼는다. 그 값의 표현식은 아래와 같다.

$$T = \sum tf_{ij} \times log(N/df_j)$$

T는 변별력의 정도를 표현하는 값이고, 이는 i라는 의미일 때 j라는 단어의 출현 정도 (tfij) 와 i의 의미에서 j라는 단어가 출현하는 샘플의 수 (dfj)를 곱하여 표현한다. (N는 훈련 샘플의 수이다.)

본 논문에서는, T값이 기준 값 이하일 경우 해당 공기정보를 제거하였는데, 기준 값인 6.73은 빈도수가 2보다 작고 훈련데이터의 1/3 이상의 훈련 데이터에 나타났을 경우이다. 이 기준값은 반복 실험을 통해 얻은 결과값이다. 기준치 이상의 저빈도와 고빈도 단어들을 제거함에 따라서, 다음 과정에서 추출한 의미 벡터에서 노이즈를 제거하는 효과를 얻을 수 있었다.

2.4 의미벡터 추출과 유사도 비교

이전 과정에서 추출한 공기정보를 SVD를 이용하여 기저 벡터화 하였다. 이 방법을 통하여, 우리는 대상명사의 각 의미를 기저 벡터로 표현할 수 있고, 벡터의 축은 훈련 데이터에서 나타나는 문맥의 단어와 상의어로 나타난다.

본 논문에서는, 추출한 공기정보를 여러 가지 장점을 가지는 SVD로 변환하였다. SVD를 가지고 선택한 공기정보를 다차원 의미벡터로 합성하고 변환하는 과정에서, 추출한 의미정보를 구조적인 형태로 변화하고, SVD에서 나타나는 차원 축소 효과는 추출한 공기정보를 더욱 정규화 하여 나타내는 효과를 얻을 수 있다. 본 논문의 이러한 방법으로 의미를 SVD로 단어분포의 위치와 모양으로 표현한다. 이 방법은 비슷한 의미와 단어는 유사한 문맥 분포를 가진다는, 공기정보를 이용한 의미구분의 기본 의도를 포함할 수 있고, SVD가 가지고 있는 특성으로 정보의 구조화와 정규화 과정을 통하여, 공기정보를 의미구분에 적합하고 좀 더 정화된 형태로 가공할 수 있다. 즉, 본 논문에서는 훈련데이터의 각 의미별 공기정보로부터 의미벡터를 추출하였고, 실험데이터의 문맥 중 훈련데이터에 나오지 않는 공기정보는 무시하고 훈련데이터가 포함하는 공기정보를 가지고 의미구분을 하였다.

SVD의 결과를 가지고 훈련데이터와 평가데이터에서 추출한 의미 벡터의 유사도 비교를 통하여 정답을 선택하였다. 유사도는 두 의미 벡터 사이의 코싸인 값으로 측정하였고, 그 식은 다음과 같다.

$$sim(v,w) = \frac{\sum_{i=1} v_i w_i}{\sqrt{\sum_{i=1} v_i^2 \sum_{i=1} w_i^2}}$$

i는 훈련데이터에서 문맥을 표현하는 벡터의 축, v는 훈련데이터에서 추출한 의미 벡터이고 w는 평가데이터의 의미 벡터이다. 유사도를 두 벡터 사이의 코싸인 값으로 결정하기 때문에, 유사도 값은 0과 1 사이에서 존재한다. 유사도 값이 1인 경우에는 두 벡터의 완벽한 일치를, 반대로 유사도 값이 0인 경우에는 전혀 일치되는 부분이 없다는 걸 의미한다. 결국, 훈련데이터와 평가데이터에서 추출한 의미벡터의 유사도 계산과 비교를 통하여, 대상 명사가 가질 수 있는 의미 중 최대 유사도 값을 가지는 의미를 각 평가 데이터의 정답으로 결정하게 된다.

2.5 기본 의미의 설정

본 논문은 대상 명사마다 기본 의미를 제안하여, 의미 벡터를 이용한 의미구분이 실패할 경우 이 기본 의미를 정답으로 선택한다.

기본 의미는 훈련 데이터에서 가장 많이 나타난 의미를 선택하고 의미구분이 실패할 경우, 기본 의미를 자동적으로 정답으로 결정한다.

이 방법은 실제 자연 언어 처리 시스템에 의미구분 방법을 적용하는데 걸림돌이 되고 있는 정확률과 재현률을 향상시키고, 시스템의 전반적인 성능 향상을 가져올 수 있다.

3. 실 험

성능의 향상 정도를 평가하기 위하여 본 논문에서는 두 가지 시스템을 비교하였다. 하나는 본 논문에서 제안한 변형과정 없이 직접적인 공기정보를 가지고 그대로 훈련한 시스템이고, 다른 시스템은 본 논문에서 제시한 방법들로 추출한 공기정보를 변형하여 의미구분에 적용한 경우이다.

사용한 평가집합은 2001년 SENSEVAL에서 공개된 한국어 평가집합을 사용하였다. 이 평가집합은 10개 명사에 대한 훈련과 평가 데이터를 제공하는데, 본 논문에서는 그들 중 세 개의 명사에 적용하여 성능을 비교하였다. 표1과 표2는 실험한 세 단어의 데이터 분포와 의미 사전을 보여준다.

대상 명사	의미 정의	의미번호
점	작고 둥근 선	k00082
	어떤 특성이나 형태의 부분	k00085
	어떤 사물의 수	k00086
	날고기의 조각	k00087
바람	날씨에서 공기의 흐름	k00031
	어떤 일에 대한 기대	k00032
	일반적인 경향	k00033
	나쁜 행동경향 혹은 유행	k00034
밤	과일, 밤나무의 열매	k00091
	하루 중 해가 지고 어두운 시간	k00092

<표1> 대상 명사의 의미 사전

단어	학습샘플	평가샘플
Baram	99	48
jeom	109	40
bam	143	59

<표2> 실험 집합의 분포

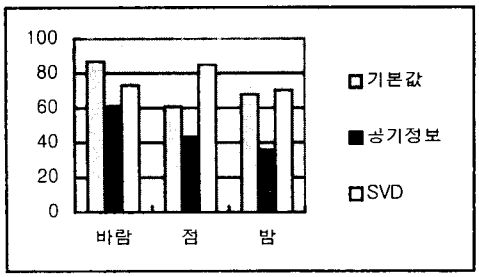
실험 집합에서 샘플은 대상 명사를 포함하는 하나의 문장과 그 문장의 앞뒤 두 문장씩을 포함하는 5문장의 묶음을 의미한다.

비교 시스템에는 훈련집합에서 추출한 같은 공기정보를 본 논문에서 제시한 여과과정 없이 베이시안(Baysian) 훈련 모델로 훈련하여 정답을 결정하는 방법을 사용하였다.

표3과 그림1은 각 시스템의 정확률을 보여준다. 기준선의 결과는 평가 집합의 모든 샘플의 정답을 기본 의미로 선택하였을 때의 정확도이다. 표3과 그림1에서 보여주듯이, 본 논문에서 제안한 방법들을 사용한 결과가 공기정보를 변형과정 없이 사용한 시스템의 결과보다 더 높은 정확률을 나타내고 있다.

(%)	바람	점	밤
기본값	87	60.78	67.8
공기정보	61	43.14	35.59
SVD	72.92	85	70.21

<표3> 결과의 비교



<그림1> 결과의 비교

이 결과는 같은 정보를 사용하더라도 정보의 가공 형태와 훈련 방법에 따라 다른 정확률을 가질 수 있다는 사실을 보여주고 있다. 이 실험에서, 우리는 또한 본 논문에서 제안한 공기정보의 가공 방법이 의미구분 시스템에 적용하고 성능을 향상시키는데 적합하다는 사실도 알 수 있다. 이 실험의 결과로써, 의미 없는 정보를 제거하고 공기정보를 상의어 정보로 일반화시키고, 이러한 정보를 SVD를 이용하여 의미벡터로 변형하는 본 논문의 방법이 의미구분에서 공기정보를 의미구분에 더욱 적합하고 변별력 있는 형태로 가공하고 있다는 것을 증명할 수 있었다.

4. 오류의 분석

결과 중에 의미 벡터로 의미구분에 실패하였거나 잘못된 의미를 정답으로 제시하는 오류를 원인별로 분류하여 보았다. 실험 후에 나타나는 오류들을 분석한 결과, 오류는 크게 상의어 추출의 오류와 실험집합의 형태와 분석에 관한 오류, 두 가지 원인으로 구분되었다.

본 논문에서는 우리말 큰사전(1997)에 있는 약 270,000 사전 정의로부터 43,048 개의 상의어를 추출하였고, 대부분의 추출한 정보를 명사의 형태적인 패턴과 의미 정의에서 추출하였다. 중복된 정보를 제거한 후 37,088개의 상의어 정보를 문맥 공기 정보를 변환하는데 사용하였다. 그러나, 상의어 추출에 실패한 단어가 데이터에 나타난다면, 그 단어는 일반화 과정이 생략된 채 지엽적인 데이터 그대로 남아있게 되고 이는 공기정보에 평가 데이터를 적용할 때 변별력을 떨어뜨리는 요소로 여전히 작용하게 된다. 비록 상의어 추출에 성공하였다 하여도, 상의어 정보에는 부적절한 패턴의 적용으로 인한 2%의 오류율을 포함하고 있다. '거만' 이라는 명사의 경우, 이 단어는 '건방짐' 과 '백만장자' 라는 두 가지 뜻을 가지고 있지만, 추출한 상의어는 '식물' 이었다. 이런 경우는 잘못된 상의어 추출로 인하여 내포하고 있던 공기정보의 의미를 왜곡하는 경우이다. 또 다른 문제는 문맥의 단어가 의미적인 애매성을 내포하고 있을 때 발생한다. '강화' 라는 단어는 그 문맥에 따라 '불, 행동이나

상태, 섬' 같은 다양한 상의어를 가질 수 있다. 결국, 문맥 단어의 의미 애매성의 문제는 대상 단어의 의미구분에서 치명적인 걸림돌이 되고 있다.문맥의 형태와 분석 방법에 대한 실험 집합에 관계 있는 어려움도 많이 나타났다. 첫째로, 문맥이 많은 고유명사와 대명사를 포함하거나 혹은 지나치게 짧은 문장일 경우, 문맥의 이런 단어들은 매우 낮거나 혹은 높은 빈도수로 인하여 불필요한 정보로 인식하여 제거하기가 쉽다. 때문에 의미구분에 필요한 의미적 정보의 부족으로 인하여 정확한 의미를 찾는데 실패하는 경우가 종종 발생한다. 표4의 경우, 문맥은 많은 고유명사를 포함하고 있다. 그러나 고유명사는 빈도수가 낮기 때문에 불필요한 정보의 제거 과정에서 대부분의 고유명사 정보가 없어져 버린다. 결국, 이 샘플은 적은 양의 공기 정보로 인하여 정확한 의미구분에 실패하게 된다. 샘플에서 태그 집합 <head></head> 아에 있는 명사가 대상 명사이고 굵은 글씨의 단어들은 문맥이 포함하고 있는 고유 명사이다.

<instance id ="bam.13">
<answer instance ="bam.13" senseid =""/><context>
이것이 토종 밤나무는 **신의주**와 **함흥**을 잇는 선 아래지역에서 특히 식생이 잘 된다 . 또한 우리 밤은 예로부터 알이 굵기로도 유명한데, **삼국지 三國志** 중 **마한 馬韓** 편에 의하면, **마한**에는 굵기가 배 첼 만한 밤이 난다고 했다 . 또한 **당나라** 때 편찬된 **수서 隋書**라는 책에도 **백제**에 큰 <head>*밤*</head>이 난다고 기록되어 있다 . 그러나 실제로 토종밤은 알이 잘고 껍질을 벗긴 밤알이 노란 빛을 띠며 맛이 뛰어나다 .</context></instance>

<표4> 지나치게 많은 고유명사를 포함하는 경우

또한 문맥의 문장들이 인접하여 나타난다 하더라도 관계없는 문장들로 샘플을 구성한 경우에는 의미벡터를 이용하여 정확한 대답을 찾아낼 수 없었다. 표5의 샘플은 소설의 차례에서 추출한 문맥이기 때문에, 관계없는 문맥으로 구성되었다.

표6과 같이 샘플을 짧은 대화나 소설 같은 문학 작품 중에서 추출했을 때, 이러한 샘플들의 문맥도 또한 기사나 설명문 보다는 문맥의 의미적 관련성이 떨어지기 때문에, 문맥 정보만을 가지고 정답을 결정하는 의미벡터 방법으로 정확한 답을 찾아내기 어렵거나 정답 찾기에 실패하기가 쉽다.

단어 의미의 의미적 특징이 문맥 독립적일 경우, 그 의미에 대한 의미구분의 정확도는 다른 문맥 의존적인 의미의 의미구분 정확도와 비교하여 현저하게 떨어지는 현상을 볼 수 있었다. '점' 의 의미구분의 경우, 정답이 '어떤 특성이나 형태의 부분' 이라는 의미인 경우에는, '~ 라는 점' 같은 문장 형태의 패턴과

<instance id ="bam.128">
<answer instance="bam.128"senseid =""/><context>
서울의 만가 (상) 김성종 저. 본 데이터의 무단 전재 및 복제를 금합니다. 원하는 목차로 커서를 옮겨 누르십시오. 작가 소개 1. 사라진 소녀 2. 무서운 밤 3. 기다리는 <head>*밤*</head> 4 .돌아오지 않는 소녀 5. 어두운 거리 6. 미행 7. 거래 8. 살인 9. 수사 10. 김 교수 11. 교수와 소녀 12. 소녀의 눈물 13. 애꾸눈 14. 몽타주의 여인 15. 빨간 티셔츠 16. 유인 17. 살인자의 손 </context></instance>

<표5> 관계없는 문맥의 샘플의 경우

<instance id ="baram.136">
<answer instance = "baram.136" senseid = ""/><context>
직속상관에게 하듯 대대장의 보고는 확실했다. 사단장은 속으로만 혀를 찼다. 최 보좌관은 그래도 뭔가 성이 차지 않는 듯 대답이 없었다. 잠시 어색한 침묵이 흐른 뒤에야 그는 사단장의 곁으로 다가왔다. **" 수고가 많으셨습니다 , 선배님 ." " 난 하는 게 없다니까 그러네...." 뭐, 기분 나쁘신 일 있으십니까, 선배님"** 단도직입적으로 물어 오는 <head>*바람*</head>에 사단장은 적이 당황하지 않을 수 없었다. **" 아니야. 그럴 일이 뭐 있겠어 아무것도 아냐. "** 최 보좌관은 호남형으로 잘생긴 얼굴에 웃음을 담으며 고개를 끄덕였다. **" 기분을 알겠습니다. 하지만 이해를 해주십시오. 아무래도 저희들이 맡은 일이 있느니만큼 어쩔 수가 없습니다, 선배님. "** 내친 김에 얘기를 해야겠다고 사단장은 생각했다. **" 이해는 하지만 말이야..... 여긴 내 예하부대야. 내가**
</context></instance>

<표6> 소설에서 추출한 대화체의 샘플의 경우

함께 자주 나타나는 것을 볼 수 있기는 하지만, 문맥의 의미적 특성에 상관없이 나타나고 있으므로 의미벡터를 가지고 정답을 판별하는 일은 쉽지 않다. 따라서 본 논문의 시스템의 실험 결과에서 평균 정확도보다 더 많은 오답을 선택하는 경향이 나타났다.

마지막으로, 본 논문의 실험 데이터는 그 크기와 규모가 크지 않고 다양한 문서로부터 추출하였다. 그러므로, 대규모 말모둠에서 일반적으로 나타나는 데이터 편중 현상이 그대로 나타난다면, 훈련데이터 부족으로 인하여 충분한 훈련 데이터를 가지는 다른 의미보다 그 의미구분의 정확도가 현저하게 떨어지는 것도 볼 수 있었다.

5. 결 론

본 논문은 한국어 명사의 의미구분에 관한

연구이다. 문맥에서 훈련데이터의 명사, 동사, 대상명사의 수식어와 대상명사의 지배 동사들을 추출하여 공기정보로 사용하고 대상명사의 수식어와 지배 동사들은 대상명사와의 의미적 연관성 정도가 더욱 높은 사실을 고려하여 가중치를 주어 반영하였다. 본 논문에서는 또한 좀 더 변별력 있는 정보를 사용하기 위하여 공기 정보에서 불필요한 정보들을 제거하고 지엽적인 문맥정보를 상의어로 바꾸는 작업을 추가하였다.

SVD는 차원 축소 과정과 정규화 과정을 통하여 관련 있는 의미적 정보들을 모아주기 때문에, 공기 정보를 SVD를 사용하여 의미벡터로 바꾸는 과정을 통하여 의미구분에 더욱 적절한 형태로 변환한다. 정답은 유사도 비교를 통하여 결정된다. 또한, 각 단어마다 훈련데이터에서 가장 많이 나타나는 의미를 그 단어의 기본 의미로 결정하고,의미벡터를 이용한 유사도 비교 방법으로 정답 결정에 실패하는 경우, 해당 단어의 기본 의미를 자동적으로 정답으로 제시하는 방법을 통하여, 본 시스템의 전반적인 성능을 향상시킬 수 있었다.

6. 향후 계획

본 논문에서 적용해 본 방법을 향상시키기 위한 요인 중의 하나로, 좀 더 효율적이고 정확한 문맥 정보의 일반화를 위한 양질의 의미체계 구축이 중요하다. 본 논문에서는 상의어 추출을 위해 표제어의 형태적 패턴과 의미 정의문의 패턴에 관련한 약 20여 개의 휴리스틱을 사용하였다. 이러한 상의어 정보를 가지고 문맥의 단어를 추출한 정보가 포함할 경우에만 적용하였다. 그러나 좀 더 정확하고 광범위한 상의어 정보를 사용할 수 있다면, 더욱 효과적인 상태로 공기정보를 일반화하여 사용할 수 있을 것이다. 그러므로, 상의어 추출을 위한 패턴에 관련 있는 휴리스틱을 보완하고, 반 자동적인 수작업을 첨가하여 상의어 정보를 확장하고, 또한 궁극적으로 의미구분에 적합한 시소러스를 구축하여 적용하여 보고 그 결과를 비교해 볼 필요가 있다.

또한 말모둠에 많이 등장하고 있는 고유명사가 내포한 의미적 정보의 추출 방법도 개선하여야 한다. 본 논문에서는, 대부분의 고유 명사들이 낮은 빈도수를 가지기 때문에 기준치 이상의 T값을 가져서 제거하지 않고 의미벡터가 포함하기 쉽지 않다.
비록 비교적 고빈도의 고유명사를 의미벡터가 포함하더라도, 그것들이 가지는 의미적 특성들을 제대로 일반화할 수 없기 때문에, 고유명사가 포함하고 있는 의미적인 정보들을 정확하게 반영할 수 없다. 그러므로, 개체명 인식과 관련한 방법론들을 사용하여 고유 명사들의 의미적 패턴들을 일반화하여 의미 벡터에 반영할 필요가 있다.

그리고, 문맥에 등장하는 고유 명사들이 내포한 원래 의미를 추적하여 이용할 수 있다면, 같은 문맥에서 더 많은 의미적 특성들을 추출할 수 있을 것이다.마지막으로, 공기정보 의미 벡터를 가지고 의미구분이 힘들었거나 문맥 독립적 의미들의 의미구분 정확도 향상을 위하여, 적용할 공기정보를 확장하거나, 훈련 데이터의 문맥에 나타나는 문장의 등장 패턴을 통한 학습 방법도 가능할 것이다.

참고 문헌

Horacio Rodriguez Hontoria (1998) "Automatic Acquisition of Lexical Knowledge from MRDs", PhD Thesis, Departament de Llenguatges i Sistemes Inform' atics.-- Universitat Polit' ecnica de Catalunya.

Rigau G., Atserias J. and Agirre E. (1997) "*Combining Unsupervised Lexical Knowledge Methods for Word Sense Disambiguation*", Proceedings of joint 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics ACL/EACL'97. Madrid, Spain.

Agirre E. and Rigau G. (1996) "*Word Sense Disambiguation using Conceptual Density*" Proceedings of 15th International Conference on Computational Linguistics, COLING'96. Copenhagen, Denmark.

Daudé J., Padró L. and Rigau G. (1999) "*Mapping Multilingual Hierarchies using Relaxation Labeling* " Proceedings of Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC'99). Maryland, United States.

Escudero G., Màrquez L. and Rigau G. (2000) "*Boosting Applied to Word Sense Disambiguation* " Proceedings of the 11th European Conference on Machine Learning, ECML 2000. Barcelona, Spain.

Springer Verlag. Lecture Notes in Artificial Intelligence 1810. R. L. de Màntaras and E. Plaza (Eds.).

Grefenstette G. (1994) "*Explorations in automatic Thesaurus Discovery*" Kluwer Academic Publishers.

Gruber T. R. (1991) "Subject-Dependent Co-occurrence and Word Sense Disambiguation" ACL.

Jen Nam Chen and Jason S.Chang (1998) "Topical Clustering of MRD Sense Based on Information Retrieval Techniques" Computational Linguistics.

Hinrich Schutze (1998) "*Automatic Word Sense Disambiguation,* Computational Linguistics".

Michael W.Berry, Susan T.Dumais and Todd A.Letsche (1995) "*Computational Methods for*

Intelligent Information Access " SuperComputing 95'.

한글학회 (1997) " 우리말 큰 사전 ", 어문각.