공기정보 벡터를 이용한 한국어 명사의 의미구분

Snasam 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, AIITRC, KAIST

요 약

본 논문은 문맥의 공기정보를 사용한 한국어 명사의 의미구분에 관한 연구이다. 대상 명사에 대한 문맥의 저열적인 단어분포는 명사의 의미구분을 위한 의미적 특성을 표현하는데 중분하지 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 정보로 바꾸어 기저벡터에 사용한다. 상의어 정보는 단어의 형태와 사전 경의문의 패턴을 통해 추출한다. 의미 벡터를 통한 의미구분에 실패하였을 경우 혼란데이터에서 가장 많이 나타난 의미로 정답을 제시한다. 실험을 위해 본 논문에서는 SENSEVAL 실험결과를 사용하였으며, 실험 방법으로 공기정보의 가중치 없이 그대로 실험한 방법과 비교하여 최고 42% 정도의 정확률을 나타내었다.

1. 서론


단어가 같은 의미로 쓰일 때 유사한 문맥과 공기정보를 가지고 있는 것은 사실이지만 (Rigau, 1997), 그 정보는 너무 저열적이고 불필요한 정보와 섞어 있어 의미구분에 그대로 적용하기는 부적합하다. 각각의 명사의 백에서 단어의 공기정보는 너무 다양하여서 각 의미별로 정확한 패턴을 추출하기가 어렵기 때문이다. 그러므로, 문맥의 단어들을 그들의 상의어 정보로 일반화 하고, 가중치 비교를 통해 문맥에서 불필요한 정보들을 제거한 후에, 이 정보들을 변별력 있는 형태로 바꿔주면 SVD(Singular Vector Decomposition) 기저벡터로 변환한다. 이 과정은 위에서 선택한 공기정보를 m x n 행렬 형태의 단차원 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
2. 방 법
2.1 공기정황의 추출
본 논문에서 고려하는 문맥의 크기는 대상 영상을 포함하는 문장과 그 문장의 앞뒤 두 문장씩, 5문장으로 한정한다. 문맥은 다양한 문맥특성을 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 충분히 반영하는 범위에 적합하다. 이 문맥 크기에 있는 영상, 동사, 대상영상의 수식어와 대상명사의 지배동사를 구분하여 추출한다.
문맥 안의 명사와 동사는 대상영상의 슈퍼젝터블 보여주고 가정한다. 공기형태는 빈도수를 기반으로 한 벡터로 나타낸다.
명사 수식어는 명사의 의미구분에 적합한 의미 정보를 포함한다. 예를 들어, '밥'은 '시간표현의 밥'과 '과일의 밥'의 두 가지 뜻을 가질 수 있다. 그러나, '있는 밥'에서 '있는'이라는 수식어는 '밥'의 의미가 두 번째 의미의 '밥'을 사용되는 경우에 대부분 나타난다. 반면에, '아무 향'이라는 표현에서 '없는'이라는 수식어는 과일을 뜻하는 '밥'보다는 '밥'이 첫번째 의미로 나타날 때 대부분 등장한다.
문맥 안의 동사를 지배하는 동사는 그 동사에 기 중점이 주어 의미정보에 반영하였다. 대상 명사를 지배하는 동사는 문맥 안의 다른 동사들보다 대상 명사와 의미적으로 더욱 밀접한 관계를 가지기 때문이다.
2.2 사건으로부터의 상의어 추출
추출한 공기정보가 문맥 지배적인 단어정보를 그대로 가지고 있다면, 문맥정보가 분산되고 의미벡터가 너무 많은 측을 가지게 되므로, 정확한 비교가 어려워지게 된다. 그러므로, 본 논문에서는 횡단형 공기정보를 사건에서 추출한 상의어 정보로 일반화하는 방법을 제안한다. 예를 들어, 우리가 '가명국'을 '국가'로, '가시방울'을 '길'로, 또한 '가을바람'을 '바람'이라는 상의어로 일반화할 수 있다. 이런 일반화된 정보를 사용한다면, 추출한 공기정보를 그대로 사용하는 것보다 더 복잡한 의미구분이 가능하다.
본 논문에서는 수동 구축된 의미 세계의 격동 대신 사전 정의어와 표제어의 패턴본성을 통한 상의어를 찾는 방법을 제시한다. 합성 명사의 경우, 그 명사의 상의어 정보를 포함하고 있는 경우가 많다. 예를 들면, '雀나무'처럼 '나무/雀'으로 끝내는 단어들의 상의어는 각각 '나무'와 '雀'이다. '雀'와 '雀'은 또한 사전 정의어의 패턴을 통하여 '석음'이라는 상의어를 유추할 수 있다. 사전 정의어를 통한 상의어 추출의 예를 들어보면, '가정교시'라는 단어의 경우, 이 단어의 사전 정의어인 '가정교시'의 품가르는 주는 사람을 상대방으로, '가정교시'의 상의어인 '사람'이 정의어의 맨 뒤에 위치하는 것을 볼 수 있다. 이처럼, 사전 정의어가 명사로 끝나는 대부분의 경우, 그 끝내는 명사가 대상 표제어의 상의어라는 정의어 패턴을 통하여 상의어를 추출할 수 있다. 앞서 설명한 두 가지 방법으로 사건의 정의어와 표제어의 형태정보를 통하여 비교적 정확하게 상의어 정보를 추출할 수 있었다.
본 논문에서는 약 30여 개의 패턴을 우리나라와 사전(1997)의 모든 표제어에 적용하여 상의어 리스트를 추출하였다.
문맥의 단어가 상의어 리스트 안에 있는 경우, 우리는 해당 단어를 상의어로 변환한 뒤 공기정보를 제공하였다. 공기정보를 상의어로 변환한 결과, 지업적인 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 적어서 의미구분에 적절하게 반영하기 어려운 고유명사 같은 단어들이다. 이런 단어들은 공기정보를 SVD로 변환할 때, 의미벡터의 수를 증가시켜서 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 본토로 나타나기 때문에 의미구분에 중요한 정보가 아니고 오히려 해방해 주는 요인이 된다.
공기정보에서 이 같은 불필요한 단어들을 제거하기 위하여, 본 논문에서는 tf (Term Frequency)와 idf (Inverse Document Frequency)를 사용한다. tf는 훈련 데이터에서 단어의 빈도수를 나타내고, idf는 훈련 데이터에서 그 단어를 포함하는 샘플의 수를 나타낸다.
(Hinrich, 1998). 즉, \( t f \)는 문맥 단어의 빈도수를 나타내고 \( idf \)는 단어의 분포도를 표현하고, 우리는 이 분포 정도와 반도 정도를 하나의 값으로 합쳐서 기준을 삼는다. 그 값의 표현식은 아래와 같다.

\[
T = \sum t f_{ij} \times \log (N/d f_j)
\]

\( t f \)는 변별력의 정도를 표현하는 값이고, \( j \)라는 의미의 \( j \)라는 단어의 출현 경도 \((t f_{ij})\)와 \( i \)의 의미에서 \( j \)라는 단어가 출현하는 샘플의 수 \((d f_j)\)로 곱하여 표현한다. \((N)\)는 훈련 샘플의 수이다. 

본 논문에서는, \( t f \)가 기준 값이 하일 경우 해당 공기정보를 제외하였는데, 기준 값인 6.73은 반도수가 2보다 작고 훈련데이터의 1/3 이상의 훈련 데이터에서 나타났을 경우이다. 이 기준값은 반복 실험을 통해 얻은 결과값이다. 기준값 이상의 저반도와 고반도 단어들은 제외함에 따라서, 다음 과정에서 추출한 의미 벡터에서 노이즈를 제거하는 효과를 얻을 수 있었다.

2.4 의미벡터 추출과 유사도 비교

이전 과정에서 추출한 공기정보를 SVD를 이용하여 기적 벡터화하였다. 이 방법을 통하여, 우리는 대상명사의 각 의미를 기적 벡터로 표현할 수 있고, 벡터의 쪽은 훈련 데이터에서 나타나는 문맥의 단어와 상의하여 나온다. 

본 논문에서는, 추출한 공기정보를 여러 가지 장점을 가지고는 SVD로 변환하였다. SVD를 가지고 선택한 공기정보를 다차원 의미벡터로 합성하고 변환하는 과정에서, 선택한 의미정보를 구조적인 형태로 변환하고, SVD에서도 나타나는 차원 축소 효과는 추출한 공기정보를 더욱 정규화 해서 나타내는 효과를 얻을 수 있다. 본 논문의 이러한 방법으로 의미를 SVD로 단어본의 위치와 모양으로 표현한다. 이 방법은 비슷한 의미와 단어는 유사한 문맥 분포를 가진다는, 공기정보를 이용한 의미 구분의 기존 이론을 포함할 수 있고, SVD가 가지고 있는 특성으로 정점의 구조화와 정규화 과정을 통하여, 공기정보를 의미구분에 적합하고 좀 더 정화된 형태로 가공할 수 있다. 즉, 본 논문에서 훈련데이터의 각 의미의 공기정보로부터 의미벡터를 추출하였고, 성형데이터의 문맥 중 훈련데이터에 나오지 않는 공기정보는 무시하고 훈련데이터가 포함하는 공기정보를 가지고 의미구분을 하였다.

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

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

이는 훈련데이터에서 문맥을 표현하는 벡터의 쪽, \( v \)는 훈련데이터에서 추출한 의미 벡터이고 \( w \)는 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 일치하는 부분이 없다는 점 의미한다. 결국, 훈련 데이터와 평가데이터에서 추출한 의미벡터의 유사도 계산과 비교를 통하여, 대상 명사가 가질 수 있는 의미 중 최대 유사도 값을 가지는 의미를 각 평가데이터의 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 시스템의 의미구분 방법을 적용하는데 실패가 되지 않는 경우에는 정확률과 재현률을 향상시키고, 시스템의 전반적인 성능 향상을 가져올 수 있다.

3. 실험

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 방법들을 추출한 공기정보를 변형하여 의미 구분에 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.
실험 결과의 결과는 다음과 같다. 

### 표 1. 대상 영어의 의미 값

<table>
<thead>
<tr>
<th>대상 영어</th>
<th>의미의 정수</th>
<th>의미번호</th>
</tr>
</thead>
<tbody>
<tr>
<td>짐</td>
<td>k00082</td>
<td></td>
</tr>
<tr>
<td>어떤 특성이나 형태의 부분</td>
<td>k00085</td>
<td></td>
</tr>
<tr>
<td>어떤 사람의 수</td>
<td>k00086</td>
<td></td>
</tr>
<tr>
<td>난고기의 조각</td>
<td>k00087</td>
<td></td>
</tr>
<tr>
<td>바람</td>
<td>k00031</td>
<td></td>
</tr>
<tr>
<td>어떤 일에 대한 기대</td>
<td>k00032</td>
<td></td>
</tr>
<tr>
<td>일반적인 경향</td>
<td>k00033</td>
<td></td>
</tr>
<tr>
<td>난한 행동경향 혹은 유행</td>
<td>k00034</td>
<td></td>
</tr>
<tr>
<td>밤</td>
<td>k00091</td>
<td></td>
</tr>
<tr>
<td>하루 종일 해가 지고 여두운 시간</td>
<td>k00092</td>
<td></td>
</tr>
</tbody>
</table>

### 표 2. 실험 결과의 분포

<table>
<thead>
<tr>
<th>단어</th>
<th>학습값</th>
<th>평가값</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parem</td>
<td>99</td>
<td>48</td>
</tr>
<tr>
<td>jeom</td>
<td>109</td>
<td>40</td>
</tr>
<tr>
<td>baram</td>
<td>143</td>
<td>59</td>
</tr>
</tbody>
</table>

4. 오히려의 분석

결과에 있어 의미에 의한 평가구분에 실패하였거나 잘못된 의미를 정상으로 제시하는 오류를 완전히 분석하여 보았다. 실험 후에 나타나는 오류들을 분석한 결과, 오류는 크게 상의어 추출의 오류와 실험결합의 형태와 분석에 관한 오류, 각가 원인으로 구분되었다.

본 논문에서는 우리말 사전(1997)에 있는 약 270,000 사전의 정의로부터 43,048 개의 상의어를 추출하였고, 대부분의 추출한 정보는 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 나타나지 않으면, 그 단어는 일반화 과정이 생략된 채 지연된 데이터 그대로 남아있게 되고 이는 공기정보에서 평가 데이터를 이용할 때 변별력을 떨어뜨리는 요소로 여겨져 적절하게 정의한다. 비록 상의어 추출에 성공하였으나, 상의어 정보에는 부적절한 패턴의 적용으로 인한 2%의 오류를 포함하고 있다. ‘거만’이라는 명사의 경우, 이 단어는 ‘전방점’과 ‘백만장자’라는 두 가지 뜻을 가지고 있지만, 추출한 상의어는 ‘식물’이었다. 이런 경우는 잘못된 상의어 추출으로 인해 내포하고 있던 공기정보의 의미를 왜곡하는 경우이다. 또한 다른 문제는 문맥의 단어가 의미적인 예제성을 내포하고 있을 때 발생한다.

'강화'는 단어는 그 문맥에 따라 ‘불’, 행동이나
상태, 섬' 같은 다양한 상의어를 가질 수 있다. 결국, 문맥 단어의 의미 매개성의 문제는 대상 단어의 의미구분에서 흔히 인식되는 가장 큰 문제이다. 문맥의 형식과 본질 대립에 대한 실험 결과에 관계 있는 어휘로도 많이 나타난다. 적어도, 문맥이 많은 고유명사와 대명사를 포함하되 혹은 지나치게 많은 문장일 경우, 문맥의 어떤 단어들은 매우 낮거나 혹은 높은 반도로 인하여 불필요한 정보로 인식하여 제거하기가 쉽다. 때문에 의미구분에 필요한 의미 정보의 부족으로 인하여 정확한 의미를 찾는데 실패하는 경우가 종종 발생한다. 표4의 경우, 문맥은 많은 고유명사를 포함하고 있다. 그러나 고유명사: 분위수가 낮기 때문에 불필요한 정보의 제거 과정에서 대부분의 고유명사 정보가 없어져 버린다. 결국, 이 문장은 적어도 반지 정보라는 형태로 인하여 정확한 의미구분에 실패하게 된다. 심지어 대명절 <head>/예제 <head>/예제에 있는 명사가 대상 명사이고 굳은 깊은 단어들은 문맥을 포함하고 있는 고유 명사이다.

<instance id = "bam.13">  
<answer instance = "bam.13" sensid = ""></answer></instance>

이것이 토종 번나무는 신의주와 합동을 하는 선악자로서에서 특히 식이 잘 된다. 또한 우리 농부 에도부터 맛고 곡물로 유명하게, 삼국 지 3국시 중 마한 백악 원에 의하면, 마한에는 곡기가 매우 많은 밭이 많다고 한다. 또한 연달라 향이 풍선나는 소비 조리에 백색에 큰 <head> låf</head>가 이나 기름을 풍부하다. 그러나 실제로 토종받은 많이 찬한 절을 빚긴 밭의 연산 슬픔을 마침 마치 빚나다.</answer></instance></context>

<instance id = "bam.136">  
<answer instance = "bam.136" sensid = ""></answer></instance>

직접상관에게 하듯 대대정의 보고는 확실히했다. 사단장은 속으로만 힘을 썼다. 최 보좌관은 그린 것을 많이 차지하는 듯 대담이 없었다. 잠재 적책의 질서가 호랑이 뒤에 있어 그 사단장의 결리가 달랐다.

"수고를 많이하셨습니다, 선배님." "난 하지 않아나가 그러네..." '뭐, 기분 나쁘시 않으신히나, 선배님' 단순히 일반적으로 놀아 오는 <head> 바람</head>에 사단장은 적이 당황하지 않을 수 없었다.

"아니라, 그렇 일이 뭐 있었겠어 아무것도 아닌가? 최보좌관은 호랑이중에 잠재적 결을 찾을 수 없었다. 기분은 안정되었습니다. 하지만 이제를 해주실시오. 아무래도 저희들이 많은 일이 있느냐 만큼 어질 수가 없습니다. 선배님." 내일 깨에 얘기 를 해야한다. 사단장은 생각했다." 이해는 하지만 말해야... 여간 내 예배대야. 내가
</instance>
연구이다. 문맥에서 훈련데이터의 명사, 동사, 대상명
사의 수식어와 대상명사의 지배 동사들을 추출하여 공
기정보로 사용하고 대상명사의 수식어와 지배 동사들은
대상명사와의 의미적 연관성 정도가 더욱 높은 사실을
고려하여 가중치를 주어 반영하였다. 본 논문에서는 또
한 좀 더 변화에 있는 정보를 사용하기 위하여 공기 정보
에서 블록한 정보들을 제거하고 지연적인 문맥정보
를 상위여로 바꾸는 작업을 추가하였다.

SVD는 차원 축소 과정과 정규화 과정을 통하여 관
련 있는 의미의 정보들을 모아주기 때문에, 공기 정보를
SVD를 사용하여 의미벡터로 바꾸는 과정은 의미
구분에 더욱 적절한 형태로 변환한다. 정답은 유사도
비교를 통하여 결정된다. 또한, 각 단어마다 훈련데이
터에서 가장 많이 나타나는 의미를 그 단어의 기본 의
미로 결정하고, 의미벡터를 이용한 유사도 비교 방법으
로 정답 결정에 실험하는 경우, 해당 단어의 기본 의미
를 자동적으로 정답으로 계산하는 방법을 통하여, 본
시스템의 적절한 성능을 확장시킬 수 있었다.

6. 향후 계획

본 논문에서 적용해 본 방법은 향상시키기 위한 요
인 중 하나로, 좀 더 효율적이고 정확한 문맥 정보의
인기가지는 의미의 의미체계 구축이 중요하다. 본
논문에서는 상의 어추출을 위해 표제어의 형태적 패턴
과 의미 정의문에 관한 20여 개의 허리스
틱을 사용하였다. 이러한 상의 정보를 가지고 문맥의
단어를 추출한 정보가 포함한 경우에만 적용하였다. 그
러나 좀 더 정확하고 광범위한 상의 정보를 사용할
수 있다면, 더욱 효과적인 상의에 공기정보를 일반화하
여 사용할 수 있을 것이다. 그러므로, 상의 어추출을
위한 패턴에 관련 있는 허리스틱을 보완하고, 반 자동
적 수작업을 첨가하여 상의에 정보를 확장하고, 또한
공극적으로 의미구분에 적합한 시스템을 구축하여 적
용하여 보고 그 결과를 비교해 볼 필요가 있다.

또한 맵모드에 많이 등장하고 있는 고유명사가 내포
한 의미의 정보의 추출 방법도 개선하여야 한다. 본 논
문에서는, 대부분의 고유 명사들이 낮은 농도수를 가지
기 때문에 기준치 이상의 테이블을 가지지 제거하지 않고
의미벡터가 포함하기 쉽지 않다.
비록 비교적 고빈도의 고유명사를 의미벡터가 포함함
라도, 그것들이 가지는 의미적 특성을 제대로 일반화
할 수 없기 때문에, 고유명사가 포함하고 있는 의미적
인 정보들을 정확하게 반영할 수 없다. 그러므로, 개체
명 인식과 관련한 방법론들을 사용하여 고유 명사들의
의미적 패턴들을 일반화하여 의미 벡터에 반영할 필요
가 있다.

그리고, 문맥에 등장하는 고유 명사들이 내포한 원
래 의미를 추적하여 이용할 수 있다면, 같은 문맥에서
다양한 의미적 특성을 추출할 수 있을 것이다. 마지막으
로, 공기정보 의미 벡터를 가지고 의미구분이 힘들
있거나 문맥 독립적 의미의 의미구분 정확도 향상
을 위하여, 적용할 공기정보를 확장하거나, 훈련 데이터
의 문맥에 나타나는 문장의 등장 패턴을 통한 학습
방법도 가능할 것이다.

참고 문헌


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


Michael W.Berry, Susan T.Dumais and Todd A.Letsche (1995) “Computational Methods for...
Intelligent Information Access " SuperComputing
95,}