

한국어 단어 공간 모델을 이용한 단어 의미 중의성 해소

Word Sense Disambiguation using Korean Word Space Model

박용민, 이재성

충북대학교 디지털정보융합학과/컴퓨터정보통신연구소

Yong Min Park(yongmin@cbnu.ac.kr), Jae Sung Lee(jasonlee@cbnu.ac.kr)

요약

한국어 단어의 의미 중의성 해소 방법들은 주로 소규모의 의미 태그 부착 말뭉치나 사전 정보 등을 이용하여 엔트로피 정보, 조건부 확률, 상호정보 등을 각각 계산하고 이를 중의성 해소에 이용하는 방법 등으로 다양하게 제안되었다. 본 논문에서는 대규모로 구축된 의미 태그 부착 말뭉치를 이용하여 한국어 단어 벡터를 추출하고 이 벡터들 사이의 유사도를 계산하여 단어 의미 중의성을 해소하는 단어 공간 모델 방법을 제안한다. 세종 형태의미분석 말뭉치를 사용하여 학습하고 임의의 200문장(583 단어 종류)에 대해 평가한 결과, 정확도가 94%로 기존의 방법에 비해 매우 우수했다.

■ 중심어 : | 단어 의미 중의성 해소 | 단어 공간 모델 | 단어 벡터 | 동형의어 |

Abstract

Various Korean word sense disambiguation methods have been proposed using small scale of sense-tagged corpra and dictionary definitions to calculate entropy information, conditional probability, mutual information and etc. for each method. This paper proposes a method using Korean Word Space model which builds word vectors from a large scale of sense-tagged corpus and disambiguates word senses with the similarity calculation between the word vectors. Experiment with Sejong morph sense-tagged corpus showed 94% precision for 200 sentences(583 word types), which is much superior to the other known methods.

■ keyword : | Word Sense Disambiguation | Word Space Model | Word Vector | Homograph |

1. 서론

자연언어 단어 분석은 SNS 분석이나 정보검색, 기계 번역 등 여러 분야에서 많이 사용되고 있다. 그러나 자연언어의 단어는 중의성이 있을 수 있으며, 이에 대한 잘못된 해석은 텍스트 의미 처리의 신뢰도를 떨어뜨릴 수 있다. 예를 들어 '배다'의 경우, '배'는 명사나 동사 등으로 해석될 수 있으며, 명사일 경우는 먹는 배, 타는 배, 신체의 일부인 배 등으로 또 다시 구분되고, 동사일

경우는 '새끼를 배다'와 같이 수태한 경우를 나타내는 뜻으로 구분된다. 만약 '먹는 배'를 검색할 경우, 단어 중의성을 고려하지 않고 결과를 출력한다면 여러 다른 뜻의 '배'가 포함되어 그 정확도를 떨어뜨릴 것이다.

단어의 중의성은 크게 문법적 중의성과 의미적 중의성으로 나눌 수 있다. 앞에서와 같이 '배다'의 경우, 동사에는 하나의 의미만 있다면 품사 구분만으로도 그 의미를 결정할 수 있으나, 명사의 경우 문맥을 파악하여 가능한 여러 의미중 하나를 선택해야 한다. 자연언어

* 이 논문은 2010년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었음.

접수번호 : #120604-006

접수일자 : 2012년 06월 04일

심사완료일 : 2012년 06월 25일

교신저자 : 이재성, e-mail : jasonlee@cbnu.ac.kr

처리에서 단어 의미 중의성 해소(WSD: Word Sense Disambiguation) 방법은 주로 품사 태깅된 결과를 입력으로 받아, 철자와 품사는 같지만 의미 중의성(모호성)이 있는 단어를 대상으로 그 문맥에서 나타내는 의미를 결정한다. 즉, 품사적 중의성은 품사 태거가 해결하고, 의미 중의성은 WSD 프로그램으로 해결한다[1][그림 1].

한국어 단어의 의미 중의성 해소는 주로 소규모의 의미 태그 부착 말뭉치를 이용하여 이루어지거나, 자료 부족 문제를 해결하기 위해 사전 정보 등을 이용하여 처리되어 왔다. 이 자료는 공기 정보, 상호정보, 엔트로피 정보 등으로 계산하여 이용하였다[2-5].

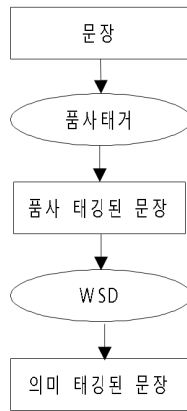


그림 1. 의미 중의성 해소 단계

단어 공간 모델(Word Space Model)은 단어의 좌우 문맥 정보를 벡터로 표현하고 각 단어들에 표현된 벡터 공간내에서 서로의 단어를 벡터 거리 등으로 비교하여 그 유사도를 측정하는 모델이다[6]. 이는 각 단어의 의미는 그 단어와 함께 쓰인 다른 단어들(단어 또는 문맥 벡터)로 표현될 수 있다는 가정하에 만들어진 모델이며 주로 자율 학습(unsupervised learning) 방법으로 단어의 의미 차이만을 비교하기 위해 사용되었다. 본 논문에서는 교사 학습(supervised learning)으로 단어 공간 모델을 구축하고, 이를 이용하여 의미 중의성을 해소하는 모델을 제시한다. 이 모델은 비교적 큰 규모로 구축된 세종 형태의미분석 말뭉치를 이용하여 학습하며, 학습에 사용하지 않은 자료를 이용하여 그 성능을 평가한다.

II. 관련 연구

WSD는 크게 교사 학습 방법, 사전 이용 방법, 자율 학습 방법 등으로 연구되어 왔다[7]. 교사 학습은 의미 태그가 붙어 있어 의미 구분이 가능한 단어들을 학습데이터로 사용하고, 베이지안 분류방법[8], 상호정보를 이용한 방법[9] 등으로 연구되었다. 이 방법은 동형의어어가 쓰인 문맥에서 함께 나타나는 단어들을 단서로 이용하여 처리한다. 하지만, 의미 태그가 부착된 학습 자료는 쉽게 충분한 양을 구할 수 없으므로 충분한 단서 단어들을 추출할 수 없어 그동안 실용적인 수준의 연구에 한계가 있었다.

사전 기반 방법은, 충분한 의미 부착 말뭉치가 없는 경우에 어휘 부족 문제를 해결하기 주로 사용한다. 즉, 이미 많이 구축되어 있는 언어 자원인 사전, 시소러스, 번역물을 이용하여 더 많은 단서 단어들이나 보다 포괄적인 단어들을 추출하는 방식이다. 이러한 예로서, 동형의어를 판별하기 위한 단서 어휘로 사전의 정의에 사용된 용어를 추출하는 방법이나[10], 단어들이 전체 문맥의 주제와 같은 의미로 사용되는 성질을 이용하여 문맥에 사용된 단어들을 단서 어휘로 추출하는 시소러스 사용 방법[11], 번역된 단어의 어휘 차이를 이용하여 의미 차이를 처리하는 방법[12][13] 등이 연구되었다.

자율 학습은 비용이 많이 드는 의미 태그 부착 말뭉치나, 사전, 시소러스 등과 같은 언어 자원을 사용하지 않고, 순수한 원시 말뭉치만을 사용하여 단어의 의미 구분을 시도한 것이다[6]. 이 방법은 같은 문장 혹은 일정한 거리내에 공기되어 나타나는 단어들을 벡터로 표현하고, 이 단어 벡터들의 집합을 단어 공간으로 설정한다. 또한, 이 단어 공간 벡터를 이용하여 2차적으로 다시 각 문맥에 나타난 단어들을 각 단어 벡터의 요소들을 표현하고 그들의 중앙점(centroid)을 취하여 이를 문맥 벡터로 사용한다. 이 방법에서는 특정한 의미 태그를 미리 정해두지 않고, 각 문맥 벡터들을 클러스터링하여 의미(혹은 의미 벡터)로 사용한다[6].

한국어에 대한 단어 의미 중의성 해소 연구도 여러가지 방법으로 시도되었으며, 주로 교사 학습 방법과 사전 기반 방법을 혼합한 것들이다. 대표적인 몇 가지들

보면 다음과 같다.

[2]에서는 1,000만 어절 원시 말뭉치에서 대상 용례(미리 선정된 중의성이 있는 단어의 용례)를 추출한 후, 의미 태그를 수작업으로 부착하고 이를 이용하여 학습하였다. 여기에서는 어떤 속성의 엔트로피 값이 적을수록 의미 구별 능력이 있는 것으로 판단하여 이를 이용하여 단어 의미를 분류하였다. 4개의 단어에 대해 실험한 결과, 학습데이터는 96.7%, 비학습데이터는 84.6%의 성능을 보였다.

[3]은 사전 뜻풀이 말에서 동형의어어가 사용된 경우에 그 문맥 정보(명사와 용언)와 간단한 계층 정보를 추출하고, 문맥 정보와 동형의어어간의 조건부 확률을 계산하였다. 이를 이용하여 9개의 동형의어어를 평가해본 결과 학습데이터는 96.11%, 비학습데이터는 80.73%의 정확률을 보였다.

[4]도 사전 뜻풀이 말을 이용하였지만, 이 뜻풀이 말에 의미 태그를 수작업으로 부착하고, 이를 학습데이터로 사용하여 각 단어 의미에 공기하는 단어 비율을 계산하였다. 또 이들의 성능을 높이기 위해 단어에 해당 의미가 나타나는 비율인 선형 확률(prior probability)을 단어 종류로 계산하여 성능을 높였고, 또한 공기하는 단어의 위치 정보를 활용하였다. 46개의 중의성 높은 단어를 대상으로 실험한 결과 74.09%의 성능을 나타냈다.

[5]은 말뭉치 및 사전의 뜻풀이 말에서 공기하여 나타나는 단어들간의 상호정보량을 계산하여 이를 이용하였다. 또한, 한국어 복합명사가 실제 문장에서 자주 공기하는 단어라는 성질을 이용하여 수작업으로 복합명사 사전을 구축하고 의미 태깅한 후, 이를 의미 중의성 해소에 추가로 사용하였다. 이 방법들을 종합하여 200여개의 단어에 대해 실험한 결과 88.82% 정확률을 보였다.

III. 한국어 단어 공간 모델

본 논문에서는 [6]의 모델을 변형하여 한국어 단어 공간 모델을 작성한다. 이 모델은 기본적으로 정보검색

의 벡터 공간 모델을 이용한 것이다. 즉, 원시 말뭉치에서 각 단어(w)가 쓰인 문맥내에 나타난 단어(v)들을 추출하여 각 단어에 대한 벡터를 추출한다. 이 단어 벡터들의 모임을 단어 공간이라 하고, 새로운 단어(q)가 단어 벡터 형태로 주어지면 이와 가장 유사한 단어 벡터를 계산하여 선택함으로써 그 의미를 결정한다. 논문 [6]에서는, 이 단어 벡터를 한 단계 더 처리하여, 벡터 요소에 해당되는 각 단어의 단어 벡터를 단어 공간에서 추출하여 그 벡터들의 중심을 새로운 문맥 벡터로 사용하였다. 또한, 이 문맥 벡터들을 클러스터링하여 의미 벡터를 생성하고, 이 의미 벡터를 비교하여 단어의 의미 구분에 이용하였다.

본 논문에서는 이러한 단어 공간 개념을 이용하되 의미 태그된 말뭉치를 이용하여 교사 학습을 한다. 이 방법을 사용할 경우, 단어 벡터만 추출한 후, 클러스터링을 하지 않고 바로 의미 태그가 부착된 단어를 의미 벡터(각 의미가 구분된 단어 벡터)로 사용할 수 있다.

[그림 2]는 단어 공간 모델을 이용한 의미 중의성 해소 방법에 대한 개념도이다. 단어공간 구축 프로그램이 형태의미분석 말뭉치를 학습데이터로 사용하여 한국어 단어 공간을 미리 만들어 두면(그림의 우측 부분), WSD 프로그램이 단어 공간을 이용하여 입력된 문장내의 동형의어어에 대한 의미 태깅을 한다. 이 그림은 “배를 먹다” 라는 문장이 품사 태그되어 입력되면, 이를 벡터 형태로 바꾸고 WSD 프로그램으로 의미 태깅하는 과정을 보여준다.

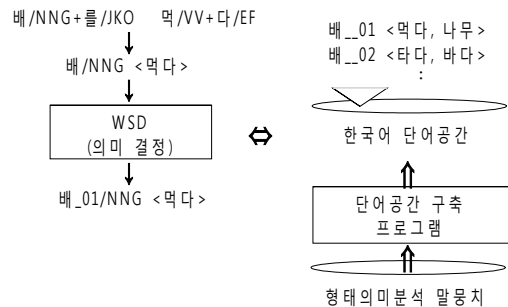


그림 2. 단어 공간 모델 기반 중의성 해소 개요

IV. 단어 벡터 공간 구축

말뭉치에서 각 단어의 단어 벡터를 추출하기 위해, 각 단어가 쓰인 문장의 좌우 n개의 단어를 벡터 요소로 추출한다. 본 논문에서 단어 벡터 요소를 명사, 동사, 형용사의 어휘로만 한정하고, 추출 범위도 대상 단어의 좌우 5개로 하였다. 이는 기존 연구에서 실질 형태소들이 중의성 해소에 어느 정도 효과적인 것으로 나타났기 때문이다[7].

문맥에서 추출한 단어 벡터 요소는 각 단어 벡터의 차원(dimension)에 해당된다. 즉, 실제 추출한 단어 벡터 요소의 종류수가 차원의 크기가 된다. 본 실험의 경우, 차원은 학습데이터에 나타난 명사, 동사, 형용사의 어휘의 종류의 총 개수이므로 크기가 크며, 정보검색 모델에서와 같이 각 단어에 대한 벡터는 포스팅 파일 형식으로 저장한다.

- 빈도 가중치 단어 벡터: w_i^f

각 벡터 요소의 값은 일반적으로 출현 빈도로 나타내며, 각 단어 벡터별로 정규화하여 사용한다. 이는 식(1)과 같이 나타낸다.

$$w_i^f = \langle v_{i1}, v_{i2} \dots, v_{in} \rangle$$

$$\text{where } v_{ij} = \frac{f_{ij}}{\sum_{k=1}^n f_{ik}} \quad (1)$$

(f_{ij} 는 빈도)

- 위치값을 고려한 빈도 가중치 단어 벡터: w_i^d

일반적으로 문맥 단어는 대상 단어에 가까울수록 영향력이 있다[4]. 본 논문에서는 이를 수치화하여 대상 단어로부터 1단어 떨어지면 5, 2단어 떨어지면 4로 하며, 같은 방법으로 계속하여 5단어 떨어진 곳은 1의 값을 갖도록 한다(대상 단어의 좌우에 관계없이 떨어진 거리에 반비례하여 부여한다.) 이 값을 d로 나타내며 이를 빈도 가중치에 곱하여 사용하고 식(2)로 표현한다.

$$w_i^d = \langle v_{i1}, v_{i2} \dots, v_{in} \rangle$$

$$\text{where } v_{ij} = \frac{f_{ij} \times d_{ij}}{\sum_{k=1}^n (f_{ik} \times d_{ik})} \quad (2)$$

(f_{ij} 는 빈도, d_{ij} 는 위치값)

각 단어 벡터 w_i 는 의미 태그를 포함하여 처리하며, 벡터 요소로 사용된 단어의 의미 태그는 무시한다. 예를 들어 '배_01'과 '배_02'가 벡터 요소로 문맥에 나올 경우, 모두 같은 단어인 '배'로 취급하여 처리하며, 단어 벡터를 구할 경우는 서로 다른 단어로 처리한다.(세종 말뭉치에서는 어휘 뒤에 숫자를 붙여 사전에 나타난 각각의 의미를 나타낸다.)

V. 단어 의미 결정

의미 중의성을 해결하고자 하는 단어 벡터 w_q 가 주어지면 이 단어 벡터를 다음 식에 의해 단어 공간 모델의 각 단어 벡터들과 비교하고 식(3)과 같이 최대의 확률을 갖는 (의미가 구분된) 단어 벡터 w_i 를 추출한다.

식(4)는 w_i 와 동형의어 관계가 있는 모든 단어의 빈도수에 대한 비율을 계산한 것이며, $h(w_i)$ 함수는 w_i 의 품사가 같은 동형의어 단어를 뜻한다. 이는 각 의미의 선형 확률(prior probability)로 볼 수 있다. 식(5)는 전형적인 코사인(cosine) 유사도이며, w_i 대신 w_i^f 나 w_i^d 로 대체하여 사용한다.

$$\text{argmax}_i P_{prio}(w_i) \times \text{Sim}(w_i, w_q) \quad (3)$$

$$P_{prio}(w_i) = \frac{\text{freq}(w_i)}{\sum_{h(w_k)=h(w_i)} \text{freq}(w_k)} \quad (4)$$

$$\text{Sim}(w_i, w_q) = \frac{\sum_k (v_{ik} \times v_{qk})}{\sqrt{\sum_k (v_{ik} \times v_{ik})} \times \sqrt{\sum_k (v_{qk} \times v_{qk})}} \quad (5)$$

VI. 실험 및 결과

실험에는 세종계획에 의해 구축된, 총 9,524,183어절의 형태미분석 말뭉치를 이용하였다[14]. 실험을 위해 말뭉치 중 90%를 임의 추출하여 학습데이터로 사용하였고, 나머지 10%중에서 임의로 100, 150, 200문장을 각각 추출하여 식(6)의 평가식으로 정확률을 계산하였다.

$$\text{정확률} = \frac{\text{올바르게 구분한 동형이의어 수}}{\text{전체 동형이의어 수}} \quad (6)$$

실험 결과에서 말뭉치 오류는 평가 대상에서 제외하였다. 이런 말뭉치 오류 단어는 일부는 동형이의어로 의미 태그(번호)가 부착되어 있으나, 일부는 부착되어 있지 않은 것으로 올바른 평가가 어려운 것들이다. 또, 기존의 다른 연구 방법과 최대한 같은 조건에서 평가하기 위해, 자료 희귀성(data sparseness) 문제로 인해 학습 데이터에 나타나지 아니한 동형이의어는 제외하였다.

평가 결과는 [표 1]과 같으며 앞에서 제시한 방법을 조합하여 5가지 모델로 표현하였다. 여기에서 prio는 식(4)의 선형 확률 P_{prio} , cos는 식(5)에 w_i^f 를 사용한 코사인 계산식, cos+dist는 식(5)에 w_i^d 를 사용한 코사인 계산식을 각각 나타낸다. 모델1은 가장 많이 쓰인 의미(선형 확률이 가장 높은 것)를 무조건 선택하는 최하한(low bound) 모델이며, 비교적 높은 성능을 보인다. 이 결과로 볼 때, 동형이의어가 대부분은 특정 의미로 편중하여 사용되고 있음을 보여준다. 따라서, 선형 확률이 고려되는 경우의 모델(모델 4, 5)이 선형 확률을 고려하지 않은 다른 모델(모델 2, 3)에 비해 더 우수한 결과를 보였다. 하지만, 코사인 유사도와 위치 정보가 모두 성능향상에 도움을 주어 이를 모두 사용한 모델5가 최상의 결과를 나타냈다.

[그림 3]은 평가 데이터 수에 따른 각 모델의 성능 변화이다. 평가 문장수를 늘림에 따라 [표 2]와 같이 동형이의어 단어수와 종류수도 증가한다. [그림 3]에서 보듯이 선형 확률을 이용한 모델들(모델 1, 4, 5)는 모두 평

가 데이터 수가 증가함에 따라 어느 정도 증가했다. 이는 전체적으로 많이 쓰인 의미를 갖는 동형이의어가 의미가 평가 데이터에 지속적으로 증가하여 포함되기 때문으로 해석된다.

[표 3]은 본 논문의 방법과 기존의 방법들을 비교하여 정리한 것이다. 평가 데이터가 서로 다르므로 정확한 비교는 어려우나, 기존의 방법들과 비교해 볼 때, 대체적으로 매우 우수함을 알 수 있다.

표 1. 각 모델의 성능 평가 결과(200문장)

모델	정확률
Model1 (prio)	92.09%
Model2 (cos)	89.97%
Model3 (cos+dist)	90.16%
Model4 (prio+cos)	93.82%
Model5 (prio+cos+dist)	94.02%

표 2. 평가 데이터내의 동형이의어 단어 및 종류수

구분	100문장	150문장	200문장
단어 수	589	808	1037
종류(type) 수	395	488	583

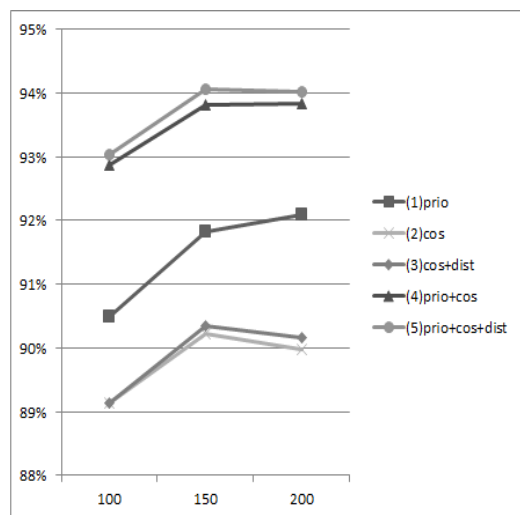


그림 3. 평가 데이터 크기에 따른 성능 변화

표 3. 기존 연구와 비교

저자	사용된 자료 및 정보	평가 단어 수	정확률
이호(1997)[1]	의미 태그 말뭉치, 엔트로피	4개	84.6%
허정(2001)[2]	사전 뜻풀이 말 문맥, 공기관계정보	9개	80.73%
김준수(2003)[3]	사전 뜻풀이 말 문맥, 공기관계, 단어 위치 정보	46개	74.09%
허정(2006)[4]	사전 뜻풀이말, 원시 말뭉치, 의미 태깅된 복합명사 사전, 상호정보량	200여개	88.82%
본 논문	의미 태그 말뭉치, 단어 공간 모델, 단어 위치 정보	583개	94.02%

VII. 결론

동형이의어의 의미 구분은 보다 정확한 언어 분석의 기초가 된다. 한국어 단어 의미 중의성 해소는 그동안 적은 양의 의미부착 말뭉치나 사전 정의를 이용하여, 엔트로피, 조건부 확률, 상호정보 등을 계산하고 이를 이용하여 의미 중의성을 해소하였다. 본 논문에서는 비교적 많은 양의 의미부착 말뭉치를 이용하여 단어 백터를 구축하고 이를 비교하는 단어 공간 모델 방법을 제안하였다.

세종 형태의미분석 말뭉치를 이용하여 실험한 결과 유사도 계산에 선형 확률, 위치값, 코사인 유사도를 모두 결합한 방법이 가장 우수했다. 또, 기존의 다른 방법들과 비교해 볼 때, 본 결과가 우수하다고 할 수 있으며, 본 논문에서 제안한 단어 공간을 이용한 방법이 유용하다고 할 수 있다.

본 논문에서 제안한 방법은 기본적인 빈도, 위치 정보를 단어 백터에 사용하고 있다. 따라서, 앞으로 더 다양한 정보(각 단어의 중요도 등)를 단어 백터에 포함하거나 단어 백터를 문맥 백터로 확장하고, 이를 정보검색의 다양한 유사도 계산식에 적용하여 성능을 평가해 볼 필요가 있다.

참고 문헌

- [1] 안광모, 한규열, 서영훈, “어휘별 중의성 제거 규칙과 통계 정보를 이용한 한국어 품사 태깅”, 한국콘텐츠학회논문지, 제9권, 제2호, pp.18-26, 2009.
- [2] 이호, 백대호, 임해창, “분류 정보를 이용한 단어 의미 중의성 해결”, 정보과학회논문지(B), 제24권, 제7호, pp.779-789, 1997.
- [3] 허정, 옥철영, “사전의 뜻풀이말에서 추출한 의미 정보에 기반한 동형이의어 중의성 해결 시스템”, 정보과학회논문지 소프트웨어 및 응용, 제28권, 제9호, pp.688-698, 2001.
- [4] 김준수, 최호섭, 옥철영, “가중치를 이용한 통계 기반 한국어 동형이의어 분별 모델”, 정보과학회 논문지 소프트웨어 및 응용, 제30권, 제11·12호, pp.1112-1123, 2003.
- [5] 허정, 서희철, 장명길, “상호정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소”, 정보과학회논문지 소프트웨어 및 응용, 제33권, 제12호, pp.1073-1089, 2006.
- [6] H. Schutze, “Automatic Word Sense Discrimination,” Computational Linguistics, Vol.24, No.1, 1998.
- [7] Manning, D. Christopher and Schutze, Hinrich, *Foundations of Statistical Natural Language Processing*, MIT Press, pp.229-261, 1999.
- [8] W. A. Gale, W. C. Kenneth, and D. Yarowsky, “A method for disambiguating word senses in a large corpus,” Computers and the Humanities, Vol.26, pp.415-439, 1992.
- [9] P. F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer, “Word-sense disambiguation using statistical methods,” In Proceedings, 29th Annual Meeting of the Association for Computational Linguistics, pp.264-270, 1991.
- [10] M. Lesk, “Automatic sense disambiguation:

How to tell a pine cone from an ice cream cone," In Proceedings of the 1986 SIGDOC Conference, pp.24-26, 1986.

- [11] D. E. Walker, "Knowledge resource tools for accessing large text files," In Sergei Nirenburg(ed.) Machine Translation: Theoretical and methodological issues, Cambridge: Cambridge University Press, pp.247-261, 1987.
- [12] Dagan Ido, and Alon Itai., "Word sense disambiguation using a second language monolingual corpus," Computational Linguistics, Vol.20, pp.563-596, 1994.
- [13] Dagan Ido, Alon Itai, and Ulrike Schwall, "Two languages are more informative than one," In Proceedings, 29th Annual Meeting of the Association for Computational Linguistics, pp.130-137, 1991.
- [14] 국립국어원, 21세기 세종계획 최종 성과물(2011년 12월 수정판), 2011.

이재성(Jae Sung Lee)

정회원



- 1983년 2월 : 서울대 컴퓨터공학과(학사)
 - 1985년 2월 : KAIST 전산학과(석사)
 - 1999년 2월 : KAIST 전산학과(박사)
 - 1985년 ~ 1988년 : 큐닉스컴퓨터(주) 과장
 - 1988년 ~ 1993년 : 마이크로소프트 차장
 - 1999년 ~ 2000년 : 전자통신연구원 팀장
 - 2005년 ~ 2006년 : 미국 아리조나 대학 방문교수
 - 2000년 9월 ~ 현재 : 충북대 컴퓨터교육과 / 디지털 정보융합학과 교수
- <관심분야> : 정보검색, 자연언어 처리, 컴퓨터교육

저자 소개

박용민(Yong Min Park)

준회원



- 2011년 8월 : 충북대학교 컴퓨터교육과(학사)
- 2012년 3월 ~ 현재 : 충북대학교 디지털정보융합학과(석사 과정)

<관심분야> : 정보검색, 인공지능, 자연언어 처리