Journal of Korean Institute of Intelligent Systems
Vol. 25, No. 2, April 2015, pp. 111-118
http://dx.doi.org/10.5391/JKIIS.2015.25.2.111

ISSN(Print) 1976-9172
ISSN(Online) 2288-2324

JKIIS

개체중의성해소에서 의미관련도 활용 효과 분석: 한국어 위키피디아를 사용하여

An Effect of Semantic Relatedness on Entity Disambiguation: Using Korean Wikipedia

강인수†
In-Su Kang†

경성대학교 공과대학 컴퓨터공학부
School of Computer Science & Engineering, College of Engineering, Kyungsung University

요 약

개체 링킹은 텍스트에 출현하는 개체 표현을 위키피디아 등의 지식베이스 항목으로 연결하는 작업이다. 동일한 개체 표현을 공유하는 서로 다른 개체들의 존재로 인해 개체 링킹에서는 개체 표현의 중의성을 해소할 필요가 있다. 개체 중의성 해소를 위한 최근 연구에서는 공기 개체 의미관련도를 중심으로 개체 출현 선험 확률와 공기 용어 정보 등을 결합하는 시도들이 주류를 형성하고 있다. 그러나 의미관련도의 왕성한 활용에도 불구하고 의미관련도 기반 방법이 개체중의성해소에 미치는 순수 효과를 분석 제시한 연구는 찾기 힘들다. 이 연구는 NGD, PMI, Jaccard, Dice, Simpson 등 서로 다른 의미관련도 지표의 차이, 공기개체집합 내 중의성 정도의 차이, 개별적/집단적 중의성해소 방식의 차이의 세 가지 관점에서 의미관련도 기반 개체중의성해소 방법들을 한국어 위키피디아 데이터를 사용하여 실험적으로 평가한 결과를 제시한다.

키워드 : 개체링킹, 개체중의성해소, 의미관련도, 위키피디아

Abstract

Entity linking is to link entity's name mentions occurring in text to corresponding entities within knowledge bases. Since the same entity mention may refer to different entities according to their context, entity linking needs to deal with entity disambiguation. Most recent works on entity disambiguation focus on semantic relatedness between entities and attempt to integrate semantic relatedness with entity prior probabilities and term co-occurrence. To the best of my knowledge, however, it is hard to find studies that analyze and present the pure effects of semantic relatedness on entity disambiguation. From the experimentation on Korean Wikipedia data set, this article empirically evaluates entity disambiguation approaches using semantic relatedness in terms of the following aspects: (1) the difference among semantic relatedness measures such as NGD, PMI, Jaccard, Dice, Simpson, (2) the influence of ambiguities in co-occurring entity mentions' set, and (3) the difference between individual and collective disambiguation approaches.

Key Words : Entity Linking, Entity Disambiguation, Semantic Relatedness, Wikipedia

Received: Nov. 5, 2014
Revised : Feb. 16, 2015
Accepted: Mar. 10, 2015
†Corresponding author(dbaisk@ks.ac.kr)

1. 서 론

개체 링킹(entity linking)은 텍스트에 출현하는 개체 표현(entity mention)을 지식베이스(예: 위키피디아) 내의 해당 개체 항목에 대응시키는 작업이다[1]. 예를 들어 다음 예문 텍스트에 출현한 개체 표현 '시카고'는, 미국 일리노이 주에 위치한 도시 시카고를 의미하므로 개체 링킹을 위키피디아 지식베이스로 연결하는 경우, 웹 URL 개체 "http://ko.wikipedia.org/wiki/시카고"에 대응되어야 한다.

"그 비행기는 미국 **시카고**에서 출발한 다음 ..."

그러나 아래 예문의 개체 표현 '시카고'는 뮤지컬 제목 시카고를 의미하므로 위키피디아 지식베이스의 경우 웹 URL 개체 "http://ko.wikipedia.org/wiki/시카고 (뮤지컬)"에 대응되어야 한다.

"브로드웨이 뮤지컬 **시카고**가 올해 한국에서 다시 ..."

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2012R1A1A1011668).

이처럼 개체 링킹에서는 위 예의 '시카고'와 같이 동일한 개체 표현을 공유하는 서로 다른 개체가 다수 존재할 수 있어 개체 표현의 중의성을 해소하는 절차가 필수적으로 요구된다. 개체 표현과 개체의 관계는 어의중의성해소(word sense disambiguation, WSD)에서 단어와 의미의 관계에 대응되며, 위키피디아 지식베이스의 경우 개체는 특정한 위키피디아 페이지에 해당된다.

개체 표현 중의성 해소(entity disambiguation, 이후 개체중의성해소)를 위한 기존 연구에서는 개체 출현의 선험 확률(prior probability), 공기(co-occurrence)하는 용어 정보, 공기 개체의 의미관련도(semantic relatedness) 등이 활용되고 있다[1,2,3,4,5,6,7]. 개체 출현 선험 확률은, 위키피디아와 같은 대용량 개체 링크 부착 코퍼스로부터 획득된 동형이의 개체들의 사전 출현 확률을 의미하며 다른 조건이 동일할 경우 이 확률값이 최대인 개체를 선호하는 방식으로 활용될 수 있다. 공기 용어 정보는, 특정 개체는 공기하는 고유의 용어 집단을 갖는다는 가정 하에 각 개체의 고유 용어 집단(예: 특정 개체의 위키피디아 페이지 텍스트 등)을 미리 구축해 둔 다음, 중의성 해소 대상 개체 표현의 인접 용어 집단과 가장 유사한 기구축 인접 용어 집단을 갖는 개체를 선호하는 방식으로 활용된다. 마지막으로 공기 개체 의미관련도는, 공기하는 개체들은 각 개체의 중의성 해소에 상호 기여한다는 가정 하에, 공기 개체들과 가장 높은 의미관련도를 갖는 개체를 선호하는 방식으로 활용된다.

최근 연구에서는 공기 개체들의 의미관련도를 활용한 시도들이 개체 링킹 방법론의 주류를 형성하고 있다. Medelyan[2], Milne[3] 등은 문서 내 비중의성 개체 표현들만을 공기 개체 집합으로 고려하여 각 개체 표현의 중의성 해소를 개별적으로 수행하였다. Ferragina[4] 등은 개별적 개체중의성해소를 위해 문서 내 모든 개체 표현으로부터의 모든 공기 개체 후보들을 활용하였다. Kulkarni[5], Han[1] 등은 문서 내 모든 개체 표현들의 모든 가능한 공기 개체 후보들을 사용하여 각 개체 표현의 중의성 해소를 개별적이 아닌 집단적 방식으로 수행하였다.

그러나 이들 의미관련도 활용 연구들에서는 개체 선험 확률 및 공기 용어 정보와 의미관련도를 결합 사용한 경우의 개체중의성해소 효과를 제시하고 있어 공기 개체 의미관련도 기반 방법의 순수 효과 수준을 가늠하기 어렵다. 예를 들어 Medelyan[2], Milne[3] 등은 개체 선험 확률과 의미관련도를 결합 사용한 개체중의성해소 방법을 개체선험확률만을 사용한 경우의 성능과 비교하여 제시하였다. Ferragina[4] 등은 개체선험확률과 의미관련도를 결합 사용한 성능을 Milne 등의 방법과 비교하였다. Kulkarni[5], Han[1], Hoffart[6], Ratinov[7] 등에서도 개체선험확률, 공기 용어 정보, 의미관련도를 결합 사용한 다양한 성능을 비교 제시하거나 기존 방법들과 비교하였으나 의미관련도의 순수 개체중의성해소 성능은 확인하기 힘들다.

이 연구에서는 개체중의성해소와 관련하여 의미관련도 활용 기법의 순수 효과를 비교 제시하고 분석한다. 이를 위해 거의 대부분의 기존 연구에서 사용된 의미관련도 지표인 NGD(Normalized Google Distance)를 포함하여 PMI(Pointwise Mutual Information), Jaccard, Simpson, Dice와 같은 다양한 의미관련도 지표를 사용하여 개별적/집단적 의미관련도 기반 방법들이 개체중의성해소에 미치는 영향을 살핀다. 또한 의미관련도 계산에 사용되는 공기 개체 집합의 중의성 정도의 차이가 개체중의성해소에 미치는 영향을 분석한다.

논문의 구성은 다음과 같다. 2장에서 개체중의성 해소의 관련 연구를 기술한다. 3장과 4장에서는 논문에 사용된 의미관련도 지표 및 의미관련도 기반 개체중의성해소 기법에 대해 상술한다. 5장에서는 실험과 결과를 기술하고 6장에서 결론을 맺는다.

2. 관련 연구

개체 링킹은 개체표현 인식, 개체중의성해소, 링크 개체 선택의 세 가지 부분 문제를 안고 있다. 개체표현 인식은 입력 문서 내 모든 가능한 개체표현 후보들을 추출하는 것으로, 영어에서는 문서 내 모든 가능한 n-gram 용어들 중 통제용어집합에 해당하는 것들만 추출하는 방식[8]이 시도되었다. 통제용어집합은 일반적으로 지식베이스 내 모든 개체들의 정규화된 명칭들로 구성된다. 링크 개체 선택은 추출된 개체 표현 후보들 중 지식베이스로의 링크를 부여할 개체 표현들을 선발하는 것으로, keyphraseness 강도를 사용하거나[8], 링크가 수작업 부여된 학습 문서로부터 링크 여부를 결정하는 분류기를 학습하는 방법[3]이 시도되었다.

이 연구에서 다루고 있는 개체중의성해소와 관련된 기존 연구에서는 개체 출현 선험 확률, 공기 용어 정보, 공기 개체 의미관련도 등이 활용되고 있다. Medelyan 등은 목표 개체 표현의 각 후보 의미(개체)에 대해 비중의성 공기 개체들과의 평균 의미관련도와 개체선험확률의 곱이 최대가 되는 의미를 취하는 방식으로 개체중의성해소를 시도하였다[2]. Milne와 Witten은 비중의성 공기 개체들만을 사용하여 평균 의미관련도와 개체선험확률 각각이 중의성해소에 미치는 중요도들을 위키피디아 링크부착말뭉치로부터 학습한 분류기를 통해 개체중의성해소를 수행하였다[3]. Ferragina와 Scaiella는 중의성 개체를 포함한 전체 공기 개체를 사용하여, 목표 개체 표현의 각 후보 의미에 대해 공기하는 각 개체 표현의 의미기여도의 총합에 기반하여 의미를 결정하는 방식을 제안하였다. 이 때 의미기여도는 개체선험확률을 가중치로 사용한 의미관련도의 가중 평균으로 계산된다[4]. Ratinov 등은 개체선험확률, 지역문맥유사도, 개체쌍 의미관련도 자질의 중요도를 링크부착말뭉치로부터 학습한 분류기를 통해 개별적 개체중의성해소를 시도하였다[7].

Kulkarni 등은 개체중의성해소문제를 선형계획법(linear programming) 문제나 Hill-climbing 문제의 해를 구하는 방식으로 고려하여 집단적 개체중의성해소를 시도하였다[5]. 이를 위해 개체선험확률, 공기 용어정보에 기반한 지역문맥유사도, 공기 개체들의 의미관련도들을 활용하였다. Han 등은 Pagerank 알고리즘을 통해 집단적 개체중의성해소를 시도하였다[1]. 그들은 Pagerank 적용을 위한 그래프의 노드 집합으로, 문서 내 각 개체 표현에 대응하는 개체 표현 노드들과 개체 표현의 모든 가능한 의미(개체)에 대응하는 개체 노드들을 생성하였고, 노드 간 edge 가중치로는 개체표현 노드와 개체 노드 사이에 지역문맥유사도를 설정하고, 임의의 두 개체 노드 간에 개체쌍의 의미관련도를 설정하였다. Hoffart 등은 Han의 것과 구조적으로 동일한 그래프로부터, 모든 개체 표현 노드들을 포함하면서 각 개체 표현이 하나의 개체 노드로만 연결된 부분그래프를 탐색하는 방식으로 집단적 개체중의성해소를 시도하였다[6].

한편 Bollegala 등은 웹 검색 엔진을 통해 얻어지는 웹페이지출현횟수(hit count)와 텍스트 snippet에 기반하여 두 단어의 의미유사도를 계산하는 방법을 제안하고, 이를 웹 문서 내 개체 표현의 중의성해소에 활용하였으며, Jaccard, Simpson, Dice, PMI 등의 지표들과 비교하였다[9]. Islam 등은 코퍼스 기반의 단어 유사도 계산을 위한 NGD, PMI, Jaccard, Dice, Simpson 등 지표의 성능을 비교하였다[10]. 또한 이 연구의 주제와 관련된 연구로, Li 등은 의미관련

도와 개체선험확률을 결합한 개별적 개체중의성해소 기법의 성능을 NGD, Jaccard, Dice 지표에 대해 비교하였다[11].

한국어의 경우 교차언어 개체링킹 환경에서의 개체중의성해소 방법을 다룬 시도들이 있었다. 대표적으로, Kang 등은 한-영 교차언어 개체링킹에서의 한국어 개체중의성해소를 위해 개체선험확률과 '(의미부착말뭉치에서의) 두 의미의 동일 문장 내 출현 횟수'에 기초한 문맥유사도의 중요도들을 기계학습하는 방법을 시도하였다[12]. 또한 Kang은 영-한 개체링킹에서 위키피디아 영-한 링크목록, 영-한 대역어사전, 개체표현의 어구길이 등을 절차적으로 적용하는 방법에 기반하여 영어 개체 표현에 대한 한국어 대역 개체를 결정하였다[13].

3. 의미관련도 지표

의미관련도(혹은 개체관련도)는 두 의미(혹은 개체)의 상호 관련된 정도를 수치화한 것이다. 이는 비교되는 각 단어의 의미들을 특정하지 않고 계산되는 단어(혹은 용어) 간 관련도와 차이가 있다. 의미관련도 계산을 위한 정보원으로 워드넷(WordNet) 등의 의미망을 사용하거나[14], 의미부착말뭉치를 활용할 수 있으나, 위키피디아 개체 링킹에서는 위키피디아 전체 문서 집합을 의미부착말뭉치로 고려하여 의미관련도를 계산하는 것이 일반적이다.

"일본 J리그의 교토 퍼플 상가를 시작으로 잉글랜드 프리미어리그의 맨체스터 유나이티드 등 여러 팀에서 활동하였다."

위의 축구선수 '박지성'에 대한 위키피디아 발췌글에서 알 수 있듯이 위키피디아 텍스트는 다른 위키피디아 페이지로의 링크(위 예의 밑줄 표시 부분)를 포함하고 있어 링크가 부착된 개체 표현의 의미를 특정하고 있다. 예를 들어 실제 한국어 위키피디아에는 개체 표현 '프리미어리그'에 해당하는 개체가 30개 이상 존재하나, 위 예의 '프리미어리그'에 부착된 링크는 "http://ko.wikipedia.org/wiki/프리미어리그"로 이는 잉글랜드 축구리그라는 의미를 특정하고 있다.

대부분 의미관련도 수식들은 두 의미가 공기하는 정도가 높을수록 높은 유사도를 갖는다는 가정을 사용하므로 이 연구에서도 의미 공기의 기초 증거로 기존 연구에서 사용된 방식인 '두 의미의 동일 문서 내 출현 여부'를 활용한다. 예를 들어 위의 위키피디아 발췌글을 의미부착말뭉치로 고려하면, 공기하는 아래 두 의미들이 상호 관련성을 갖는다는 기초 증거를 획득할 수 있다. 이후 두 의미가 공기하는 문서의 개수가 의미관련도 계산의 주요 인자가 된다.

- "http://ko.wikipedia.org/wiki/프리미어리그"
- "http://ko.wikipedia.org/wiki/맨체스터 유나이티드 FC"

이 연구에서는 서로 다른 의미관련도 지표의 차이를 함께 고려하기 위해, Bollegala 등이 단어 관련도 지표 연구에 사용한[9], NGD, PMI, Jaccard, Simpson, Dice 지표들을 선택하였다.

3.1 NGD

NGD 지표는 용어의 hit count를 이용하여 두 용어의 의미적 거리를 계산하기 위해 최초 제안되었고[15], 이후 개체중의성해소에 적용하기 위해 용어의 hit count 대신 의미(특정 위키피디아 페이지)의 링크 집합을 사용하는 방식으로 활용되었다[3]. NGD는 위키피디아 개체중의성해소를 위한 거의 대부분의 방법론에서 활용되고 있다.

두 의미 s_1, s_2 사이의 NGD는 다음 식으로 정의된다[3]. L_1, L_2는 각각 s_1, s_2로의 링크를 포함하는 위키피디아 개체(페이지)들의 집합들이며 N은 전체 위키피디아 개체의 개수이다.

$$\mathrm{NGD}(s_1,s_2)=\frac{\max\left(\log(|L_1|),\log(|L_2|)\right)-\log(|L_1\cap L_2|)}{\log(N)-\min\left(\log(|L_1|),\log(|L_2|)\right)} \quad (1)$$

그러나 위 수식은 두 의미 간의 거리를 표현하고 있으므로 의미관련도로 사용하기 위해 기존 연구에서 1-NGD의 형태로 변경되어 활용되었다[1,6]. 이 연구에서는 이를 다음 식과 같이 NGD1으로 명명한다.

$$\mathrm{NGD1}(s_1,s_2)=1-\mathrm{NGD}(s_1,s_2) \quad (2)$$

NGD 지표는 이론적으로 0에서 +∞ 사이의 한 값으로 두 의미 간 거리를 수치화한 것이다. Gracia 등은 NGD를 0에서 1 사이의 의미관련도로 정규화하는 수식 NGDn을 아래 식과 같이 제안하였다[16]. 이 수식은 개체중의성해소의 활발한 연구가 시작되기 전에 제안되었음에도 불구하고 개체중의성해소의 주요 연구들에서 사용된 사례를 찾기 힘들다.

$$\mathrm{NGDn}(s_1,s_2)=e^{-2\times NGD(s_1,s_2)} \quad (3)$$

3.3 PMI

PMI 지표[17]는 두 의미의 출현이 상호 종속일 확률과 상호 독립일 확률의 비를 정량화한 것으로, Latinov 등[7]이 아래 수식과 같이 log를 취하지 않은 형식으로 개체중의성해소에 사용하였다.

$$\mathrm{PMI}(s_1,s_2)=\log\frac{p(s_1,s_2)}{p(s_1)p(s_2)}\approx\frac{\frac{|L_1\cap L_2|}{N}}{\frac{|L_1|}{N}\times\frac{|L_2|}{N}} \quad (4)$$

3.4 Jaccard

Jaccard 지표는 두 의미의 공기 출현 정도를 공기 출현 및 개별 출현을 포함한 전체 출현 크기와 비교한 것으로 다음 식과 같이 정의된다[18].

$$\mathrm{Jaccard}(s_1,s_2)=\frac{|L_1\cap L_2|}{|L_1|+|L_2|-|L_1\cap L_2|} \quad (5)$$

3.5 Simpson

Jaccard 방식의 공기 의미 출현 비율은, 개별 의미 출현 집합들의 크기 차이가 심한 경우, 출현 집합이 큰 의미에 의해 공기 출현 비율이 왜곡되는 단점을 가질 수 있다[19]. Simpson 지표는 이러한

문제를 극복하기 위해 다음 식과 같이 두 의미 출현 집합 크기 중 작은 값을 기준으로 공기 의미 출현 비율을 계산하는 방식을 사용한다.

$$\mathrm{Overlap}(s_1,s_2)=\frac{|L_1\cap L_2|}{\min(|L_1|,|L_2|)} \quad (6)$$

3.6 Dice

Dice 지표는 다음 식과 같이 공기 의미 출현 비율을 두 의미 출현 집합 크기의 평균을 기준으로 계산하는 방식을 사용한다[19,20].

$$\mathrm{Dice}(s_1,s_2)=\frac{2\times|L_1\cap L_2|}{|L_1|+|L_2|}=\frac{|L_1\cap L_2|}{(|L_1|+|L_2|)/2} \quad (7)$$

4. 의미관련도 기반 개체중의성해소 기법

4.1 개별적 개체중의성해소 기법

입력 문서 D에 출현한 개체 표현들의 집합 T={t_1, t_2, … , t_n}에 대해 개체중의성해소를 수행한다고 가정하고 개체 표현 t_i의 가능한 후보 의미(위키피디아의 경우 위키피디아 페이지에 대한 웹 URL 링크가 의미에 해당함)들의 집합은 링크(link)들의 집합(set)이라는 의미로 lset(t_i)로 표현하기로 한다. 현재 개체중의성해소 대상인 목표 개체 표현을 t(∈T)라고 가정하면 t의 개체중의성해소를 위한 개별적 개체중의성해소 절차를 다음 식으로 정의한다.

$$Sense(t)=\underset{l\in lset(t)}{\mathrm{argmax}}\,CSR(t,l)$$
$$CSR(t,l)=f_m\left(\bigcup_{t'\in C(T,t)} f_s\left(\bigcup_{l'\in lset(t')} SR(l,l')\right)\right) \quad (8)$$

위 식은 특정 목표 개체와 공기하는 각 개체들의 의미관련도 값들을 결합 활용하는 기존 방법들을, 의미관련도 순수 효과 평가라는 본 연구의 목적을 위해, 개체선험확률을 배제하는 방식으로 단순화하고 일부 변형을 가한 것이다.

위 식에서 CSR(t,l)은 공기개체집합(혹은 공기개체문맥) 의미관련도(Contextual Semantic Relatedness)이며 l은 목표 개체 표현 t의 후보 의미 중 하나이다. 따라서 위 식은 t의 모든 후보 의미 중 CSR(t,l)이 최대가 되는 의미를 결정하도록 동작한다. C(T,t)는 t의 개체중의성해소 문맥으로 이는 T의 부분집합 중 하나로 정의된다. SR(l,l')은 두 의미 l, l'에 대해 계산된 의미관련도 값으로 3장의 의미관련도 지표 수식에 해당한다.

위 식에서 f_s는 t의 후보 의미 l에 대해 공기하는 개체 표현 t'의 (아직 결정되지 않은) 의미들이 l과 관련된 정도를 계산하는 함수이다. f_s가 max인 경우, t의 후보 의미 l에 대해 t'의 모든 가능한 의미 l'과의 의미관련도들을 계산하여 그 최대값을 취한다. f_m은 t의 후보 의미 l에 대해, t의 전체 공기 개체 표현 집합 C(T,t)와의 의미관련도를 계산하는 함수로, f_m이 sum인 경우 각 공기 개체 표현에 대해 f_s를 통해 얻어진 의미관련도 값들의 총합을 취한다.

위 식에서 SR(l,l')은 수학적으로 {SR(l,l')}로 표현되어야 하나 이를 간략 표기한 것이다. 또한 위 식에 사용된 합집합은 중복 원소를 서로 다른 원소로 고려하는 합집합 연산으로 정의한다. 즉 {0.1}, {0.5}, {0.5}의 합집합을 {0.1, 0.5, 0.5}로 계산한다.

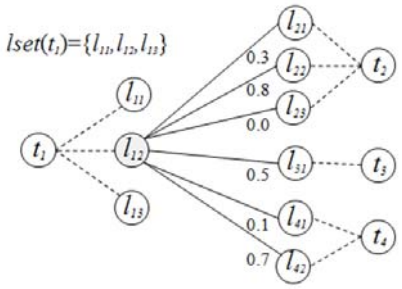

그림 1. 개별적 개체중의성해소 중간 과정 예
Fig. 1. Example of individual entity disambiguation

Fig. 1의 예에서 f_s가 max이고 f_m이 sum인 경우, 목표 개체 표현 t_1의 의미 l_{12}에 대한 CSR(t_1,l_{12}) 값은 다음과 같이 계산된다.

CSR(t_1,l_{12})=sum({max({0.3,0.8,0.0}),max({0.5}),max({0.1,0.7})})=sum({0.8,0.5,0.7})=2.0

4.2 집단적 개체중의성해소: Han's 방법

이 연구에서는 집단적 개체중의성해소 방법론으로 Han 등[1]의 방법 일부를 사용하며 그 절차는 다음과 같다.

(1) 입력 텍스트에 출현한 각 개체 표현의 모든 가능한 의미들을 노드로 설정하고 노드 간 edge의 가중치를 의미관련도로 설정하여 그래프 G를 생성한다.

(2) G에 Pagerank 알고리즘[21]을 적용한다. 이후 G의 각 노드는 Pagerank 점수가 부여되어 있다.

(3) 입력 텍스트의 각 개체 표현 t에 대해 t의 모든 가능한 의미들에 해당하는 (G의) 노드들 중 Pagerank 점수가 최대인 노드에 해당하는 의미를 t의 의미로 결정한다.

5. 실 험

5.1 평가 시스템

실험에서는 순수 의미관련도 기반 개체중의성해소 기법의 성능을 한국어 위키피디아를 대상으로 평가한다. 이를 위해 중의성/비중의성 공기 개체 집합 활용의 차이, 서로 다른 의미관련도 지표 사용의 차이, 그리고 개별적, 집단적 중의성해소기법의 차이의 세 가지 측면을 고려한다. 전술한 세 가지 각 측면의 모든 가능성의 조합으로부터 실험에 사용될 각 개체중의성해소기법의 명칭을 명명한다(Table 1 참조).

표 1. 실험용 개체중의성해소기법 명명 기준
Table 1. Criteria for naming experimental entity disambiguation methods

Type of co-occurring terms	Semantic relatedness	Disambiguation method
Unambiguous a**M**biguous **A**ll	**NGD1** **NGDn** **PMI** **Jacc**ard **Simp**son **Dice**	**M**ax_**S**um
		Page**R**ank

다음은 위의 방식을 따라 명명된 몇 가지 개체중의성해소기법 명칭의 예들이다.

- U-NGD1-M_S: 비중의성(**U**nambiguous) 공기 개체 집합 사용, 의미관련도 지표 NGD1 사용, 개별적 개체중의성해소기법의 f_s, f_m에 각각 Max와 Sum 사용
- M-NGDn-M_S: 중의성(a**M**biguous) 공기 개체 집합 사용, 의미관련도함수 NGDn 사용, f_s, f_m에 각각 Max와 Sum 사용
- A-PMI-PR: 모든(**A**ll) 공기 개체 집합 사용, 의미관련도함수 PMI 사용, 집단적 개체중의성해소기법 Pagerank 사용

개체중의성해소기법의 성능을 비교하기 위한 두 가지 베이스라인 방법은 다음과 같다.

- Baseline I (unambiguous only): 이 방법은 입력 문서 내 개체 표현 중 비중의성 개체 표현에 대해서만 개체중의성해소를 수행한다.
- Baseline II (MFS): 이 방법은 입력 문서 내 각 개체 표현 t에 대해 t의 최빈 의미(the Most Frequent Sense)로 개체중의성해소를 수행한다. 이는 1장에서 소개한 개체 출현 선험 확률을 활용한 개체중의성해소 기법과 동일하다.

5.2 평가 데이터 및 평가 지표

평가용 테스트셋 구축을 위해 Milne 등[3]의 연구를 따라 한국어 위키피디아 dump(2014년 3월)로부터 50개 이상의 위키피디아 링크를 갖는 위키피디아 페이지 500개를 무작위 추출하였다. 테스트셋에 포함된 500개 문서를 제외한 나머지 위키피디아 문서 집합은 의미관련도 계산을 위한 의미부착말뭉치로 고려하였다. 테스트셋의 각 문서는 최소 50개 이상의 개체 표현에 대해 이미 정답 의미(즉, 위키피디아 URL 링크)가 부착되어 있다. 이 연구에서는 이들 의미 부착된 개체 표현에 대해 5.1절의 다양한 개체중의성해소기법들을 적용한 성능을 평가한다. 따라서 성능 평가 지표로 다음의 두 가지 정확률(accuracy)를 사용한다. iAcc (micro **Acc**uracy)는 개별 개체 표현 단위의 정확률이며 aAcc (macro **Acc**uracy)는 문서 단위로 계산된 iAcc들의 평균 정확률이다.

$$\text{iAcc} = \frac{\text{정답 의미로 결정된 개체표현 개수}}{\text{테스트셋 내 개체표현 전체 개수}}$$

$$\text{aAcc} = \text{테스트셋 내 문서별 iAcc들의 평균}$$

5.3 실험 결과

Table 2는 비중의성 공기개체집합을 사용한 중의성해소기법들의 성능을 베이스라인 성능과 함께 비교한 것이다. 베이스라인 성능들로부터 테스트셋 내 위키피디아 문서들은 중의성 개체 표현들이 전체의 약 20%(≈1-0.8088)를 차지하며 이들 중 약 65% (≈ ((1-0.8088)-(1-0.9316)) / (1-0.8088))는 MFS 의미 선택 방식을 통해 중의성이 해소됨을 알 수 있다.

표 2. 비중의성 공기 개체 기반 중의성해소기법 성능
Table 2. Performance of entity disambiguation using unambiguous co-occurring entity mentions

System	Disambiguation scheme	iAcc	aAcc
Baseline I	unambiguous only	0.8088	0.8162
Baseline II	MFS	0.9316	0.9319
Individual disambiguation	U-NGDn-M_S	0.9441	0.9437
	U-NGD1-M_S	0.9439	0.9434
	U-PMI-M_S	0.9436	0.9429
	U-Dice-M_S	0.9432	0.9431
	U-Jacc-M_S	0.9430	0.9430
	U-Simp-M_S	0.9379	0.9380
Collective disambiguation	U-NGDn-PR	0.9436	0.9436
	U-NGD1-PR	0.9434	0.9433
	U-Jacc-PR	0.9430	0.9433
	U-Dice-PR	0.9429	0.9431
	U-PMI-PR	0.9427	0.9428
	U-Simp-PR	0.9379	0.9381

Table 2에서 비중의성 공기 개체들을 사용한 의미관련도 기반 개체중의성해소 기법들은 개별적, 집단적 방식 모두 대부분의 의미관련도 지표들에서 MFS 베이스라인보다 1% 이상 높은 성능을 보였다. 수치적으로 1%의 차이는 미미하게 해석될 수 있으나 이 결과는 다음 두 가지 측면에서 큰 의미를 갖는다. 첫째, MFS 베이스라인은 개체중의성해소와 밀접히 관련된 기존 WSD 연구에서조차 그 성능을 능가하는 기법의 고안이 쉽지 않을 만큼[22] 강력한 베이스라인이라는 점이다. 둘째, Table 2의 의미관련도 기반 기법들은 MFS 정보를 활용하지 않고 MFS의 성능을 능가하였다는 점이다. 거의 대부분의 기존 개체중의성해소 기법들이 제안되는 기법들과 결합할 기본 정보로 MFS를 활용하고 있는 점을 감안할 때, Table 2의 성능은 순수 의미관련도 기반 기법의 개체중의성 해소능력을 실험적으로 보인 하나의 증거이다.

비중의성 개체 집합에 전적으로 의존한 개체중의성해소 기법[3]은 주어진 하나의 문서 내에서 비중의성 개체 집합의 크기가 중의성 해소 측면에서 항상 충분하지 않을 수 있다는 제약을 안고 있다[1,4,7,11]. 이와 관련하여 Fig. 2는 한 문서 내 비중의성 개체 집합의 크기에 따른 비중의성 공기 개체 기반 중의성해소기법의 성능을 제시한 것이다. 그러나 전체 비중의성 개체 집합의 부분만을 사용하는 경우 어떠한 비중의성 개체들이 선택되느냐에 따라 목표 개체의 중의성해소 성능의 차이가 발생할 수 있다. 이러한 점을 감안하여 Fig. 2에서는 x축의 각 크기에 대해 해당 크기만큼의 임의 선택된 비중의성 개체

들을 사용하여 중의성해소를 수행하는 절차를 10회 반복한 성능들의 평균을 표시하였다.

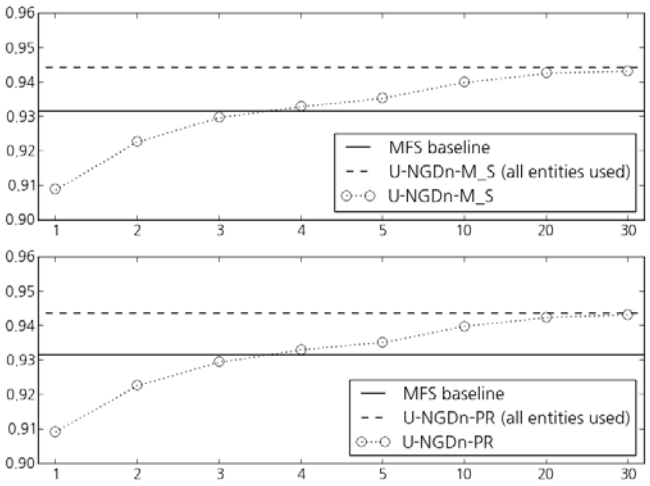

그림 2. 비중의성 개체 집합 크기에 따른 중의성해소 성능 변화 (x축: 비중의성 개체 집합 크기, y축 iAcc)

Fig. 2. Performance of entity disambiguation according to varying sizes of unambiguous co-occurring entity mentions (x-axis: count of unambiguous co-occurring entity mentions, y-axis: iAcc)

Fig. 2에서 단일 비중의성 개체의 사용은, U-NGDn-M_S, U-NGDn-PR 각 방법에 대해 이미 90% 이상의 중의성 해소 성능을 보였으며, 비중의성 개체의 전체 집합이 적용된 성능 대비 96% 성능 수준을 보였다. 전체적으로, 보다 많은 비중의성 개체들의 사용은 중의성 해소 성능을 점증적으로 향상시켰으며, 대량의 비중의성 개체들이 적용된 구간에 비해 소량의 비중의성 개체 적용 구간에서 더 큰 폭의 성능 향상을 가져왔다. 특히 Fig. 2의 결과로부터 NGDn 지표와 결합된 의미관련도 기반 방법들은 5개 미만의 비중의성 개체만으로도 MFS 베이스라인의 성능 수준을 확보할 수 있음을 알 수 있다.

Fig. 3은 개별적 중의성해소기법 **M**ax_**S**um과 집단적 중의성해소기법 **P**age**R**ank에 대해, 비중의성/중의성/전체 공기 개체 집합을 적용한 각 경우의 중의성해소 성능을 서로 다른 의미관련도 지표들에 대해 비교한 것이다. 그 결과는 대체로 비중의성, 전체, 중의성 공기 개체 집합 적용 순으로 성능이 감소하였다.

또한 Fig. 3에서 중의성이 있는 공기 개체만을 사용한 경우에도 성능의 저하는 크지 않았으며 Max_Sum 방식과 Simpson 함수를 결합한 경우를 제외한 모든 경우에서 MFS 베이스라인 이상의 성능을 보였다. 특히 개별적 개체중의성해소기법에 비해 집단적 중의성해소기법은 중의성 공기 개체의 추가로 인한 성능 저하가 크지 않았다. 이는 집단적 중의성해소기법의 경우 공기 개체 집합의 중의성에 보다 강건하게 동작하는 것으로 해석된다. 이와 관련하여 집단적 중의성해소기법과 결합된 NGD1, NGDn 및 PMI 지표의 경우 비중의성/중의성/전체 공기개체집합들의 차이에 무관하게 거의 일정한 성능 수준을 유지하고 있어 의미관련도 기반 중의성해소 기법의 적절한 결합 방식으로 판단된다.

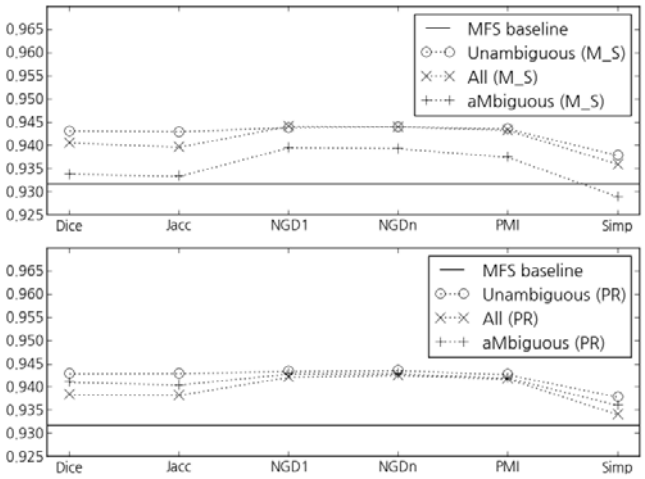

그림 3. 비중의성/중의성/전체 공기 개체 집합 적용에 따른 중의성해소 성능 변화 (x축: 의미관련도 함수, y축: iAcc, M_S: Max_Sum, PR: PageRank)

Fig. 3. Performance of entity disambiguation according to unambiguous/ambiguous/all co-occurring entity mentions (x-axis: semantic relatedness metric, y-axis: iAcc, M_S: Max_Sum, PR: PageRank)

Fig. 4는 서로 다른 의미관련도 지표들의 개체중의성해소 성능을 공기개체집합(U,A,M)의 차이와 개별/집단적 중의성해소기법(M_S,PR)의 차이들을 구별하여 비교한 것이다. 예를 들어 x축의 U-M_S는 비중의성 공기개체집합과 개별적 중의성해소기법 Max_Sum을 결합한 것을 의미한다. 전체적으로 의미관련도 지표들은 NGDn, NGD1, PMI, Dice, Jaccard, Simpson 순으로 중의성해소 성능의 차이를 보여 기존 연구에서의 NGD 지표의 왕성한 활용을 지지하는 결과를 보였다. 한편 상대적 성능 저하가 두드러진 Simpson 지표를 제외하면 NGDn, NGD1, PMI 지표 그룹들과 Dice, Jaccard 지표 그룹들은 각각 그룹 내에서 비슷한 성능을 보였다. 특히 비중의성 공기 개체들만 사용된 경우(Fig. 4에서 U-M_S, U-PR), 중의성 개체가 포함된 경우들과 달리, Simpson을 제외한 나머지 의미관련도 지표들의 성능은 현저한 차이가 발견되지 않았다. 이는 역으로 공기 개체들의 중의성 정도가 증가하는 상황에서는 의미관련도 지표의 적절한 선택이 필요함을 의미한다.

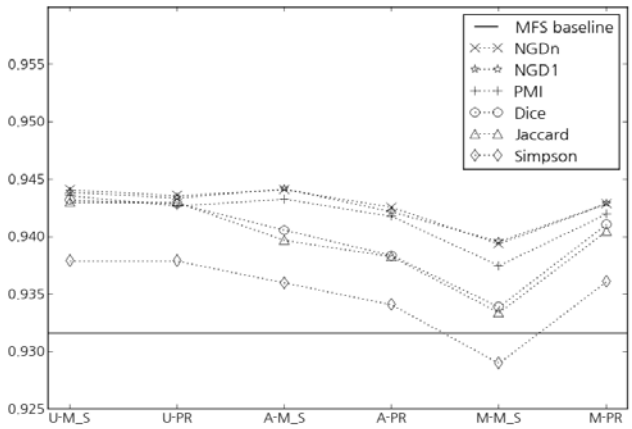

그림 4. 서로 다른 의미관련도 지표들의 개체중의성해소 성능 (x축: 공기개체집합(U,A,M)과 개별/집단적(M_S,PR) 중의성해소기법

구분, y축: iAcc)

Fig. 4. Performance of entity disambiguation according to different semantic relatedness metrics (x-axis: label for co-occurring entity mention sets (U,A,M) and individual/collective disambiguation methods (M_S,PR), y-axis: iAcc)

6. 결 론

이 연구에서는 공기 개체들의 의미관련도에 기반한 개체중의성해소 기법의 평가를 공기개체집합의 중의성 정도의 차이, 의미관련도 지표의 차이, 개별적/집단적 중의성해소 방식의 차이 관점에서 한국어 위키피디아 데이터를 대상으로 시도하였다. 그 결과 실험에 사용된 각 의미관련도 기반 중의성해소 방법들은, 개별적 중의성해소방식과 Simpson 지표의 결합을 제외하면, 그 자체로 MFS 베이스라인의 성능을 능가하여 의미관련도 기반 중의성해소의 유용성을 보였다. 또한 NGD1, NGDn 혹은 PMI 의미관련도 지표와 결합된 집단적 중의성해소 기법은 공기 개체 집합의 중의성 정도에 무관하게 거의 일정한 중의성해소성능을 보여 의미관련도에 기반한 중의성해소의 적절한 방안으로 판단되었다. 특히 비중의성 공기 개체들에 전적으로 의존하는 경우에도 NGDn 지표는 개별적/집단적 중의성해소방식에 무관하게 소량의 비중의성 개체만으로도 MFS 베이스라인의 성능 수준을 확보할 수 있음을 보였다.

이러한 결과들은 개체중의성해소의 최근 주요 연구들이 NGD, PMI 지표에 기반한 집단적 기법에 집중되는 현상을 설명하는 경험적 분석 평가 자료가 될 뿐 아니라, 향후 새로운 의미관련도 기반 개체중의성해소 방법론 개발을 위한 참조 자료가 된다는 점에서 그 의의가 있다.

References

[1] X. Han, L. Sun, J. Zhao, "Collective entity linking in web text: a graph-based method," *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011

[2] O. Medelyan, I. H. Witten, D. Milne, "Topic indexing with Wikipedia," *Proceedings of the Wikipedia and AI workshop at AAAI-08*, 2008.

[3] D. N. Milne, I. H. Witten, "Learning to link with Wikipedia," *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008.

[4] P. Ferragina, U. Scaiella, "TAGME: on-the-fly annotation of short text fragments (by Wikipedia entities)," *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, 2010.

[5] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, "Collective annotation of Wikipedia entities in web text," *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[6] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum, "Robust disambiguation of named entities in text," *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.

[7] L. Ratinov, D. Roth, D. Downey, M. Anderson, "Local and global algorithms for disambiguation to Wikipedia," *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[8] R. Mihalcea, A. Csomai, "Wikify!: linking documents to encyclopedic knowledge," *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 2007.

[9] D. Bollegala, Y. Matsuo, M. Ishizuka, "Measuring semantic similarity between words using web search engines," *Proceedings of the 16th International Conference on World Wide Web*, 2007.

[10] A. Islam, E. E. Milios, V. Keselj, "Comparing word relatedness measures based on Google n-grams," *Proceedings of COLING 2012: Posters*, 2012.

[11] C. Li, A. Sun, A. Datta, "A generalized method for word sense disambiguation based on Wikipedia," *Proceedings of the 33rd European Conference on IR Research*, 2011.

[12] I. Kang, S. Kang, "A single-step machine learning approach to link detection in Wikipedia: NTCIR Crosslink-2 Experiments at KSLP," *Proceedings of the 10th NTCIR Conference*, 2013.

[13] S. Kang, "English-Korean cross-lingual link discovery using link probability and named entity recognition", *Journal of The Korean Institute of Intelligent Systems*, vol. 23, no. 3, pp. 191-195, 2013.

[14] S. Hassan, R. Mihalcea, "Semantic relatedness using salient semantic analysis," *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011.

[15] R. Cilibrasi, P. M. B. Vitányi, "The Google similarity distance", Available: http://arxiv.org/pdf/cs/0412098.pdf, 2004, [Accessed: October 29, 2014]

[16] J. Gracia, R. Trillo, M. Espinoza, E. Mena, "Querying the web: a multiontology disambiguation method," *Proceedings of the 6th International Conference on Web Engineering*, 2006.

[17] K. W. Church, P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22-29, 1990.

[18] P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bull. Soc. Vaud. Sci. Nat.*, vol. 44, pp. 223-270, 1908.

[19] G. G. Simpson, "Notes on the measurement of faunal resemblance," *American Journal of Science*, vol. 258a, pp. 300-311, 1960.

[20] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297-302, 1945.

[21] S. Brin, L. Page, "The anatomy of a large-scale hypertextual Web search engine," *Computer Networks*, vol. 30,

pp. 107-117, 1998.
[22] R. Navigli, "Word sense disambiguation: a survey," *ACM Computing Surveys*, vol. 41, no. 2, 2009.

저 자 소 개

강인수(In-Su Kang)
1995년 : 경북대학교 컴퓨터공학 공학사
1999년 : POSTECH 컴퓨터공학 공학석사
2006년 : POSTECH 컴퓨터공학 공학박사
1995년~1997년 : 포스데이타
2006년~2008년 : 한국과학기술정보연구원
2008년~현재 : 경성대학교 컴퓨터공학부

관심분야 : 자연어처리, 정보검색, 시맨틱 웹
Phone : +82-51-663-5147
E-mail : dbaisk@ks.ac.kr