인지과학, 제27권 제1호
Korean Journal of Cognitive Science
2016, Vol. 27, No. 1, 43～60.

Word2Vec을 이용한 반복적 접근 방식의 그래프 기반 단어 중의성 해소*

오 동 석　　강 상 우†　　서 정 연
서강대학교 컴퓨터공학과

지식기반을 이용한 비지도 방법의 단어 중의성 해소 연구는 그래프 기반 단어 중의성 해소 방법에 중점을 두고 있다. 그래프 기반 방법은 중의성 단어와 문맥이나 문장에서 같이 등장한 단어들과 의미그래프를 구축하여 연결 관계를 보고 중의성을 해소한다. 하지만, 모든 중의성 단어를 가지고 의미 그래프를 구축하게 되면 불필요한 간선과 노드 정보가 추가되어 오류를 증가시킨다는 단점이 있다. 본 연구에서는 이러한 문제를 해결하고자 반복적 접근 방식의 그래프 기반 단어 중의성 해소 방식을 사용한다. 이 방식은 모든 중의성 단어들을 특정 기준에 의해서 단어를 매칭 하고 매칭 된 단어들을 반복적으로 그래프를 재구축하여 단어중의성을 해소한다. 본 연구에서는 Word2Vec을 이용하여 문맥이나 문장 내에 중의성 단어와 의미적으로 가장 유사한 단어끼리 매칭하고, 매칭 된 단어들을 순서대로 그래프를 재구축하여 중의성 단어의 의미를 결정하였다. 결과적으로 Word2Vec의 단어 벡터정보를 이용하여 이전에 연구 되었던 그래프 기반 방법과 반복적 접근 방식의 그래프 기반 방법보다 더 높은 성능을 보여준다.

주제어 : 단어중의성해소, 그래프 기반, Word2Vec, 비지도 학습, 자연어처리

---

* 본 연구는 미래창조과학부 및 정보통신기술진흥센터의 대학ICT연구센터육성 지원사업의 연구결과로 수행되었음(IITP-2016-R0992-15-1011) 그리고 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.NRF-2013R1A1A2010190)

† 교신저자: 강상우, 서강대학교 컴퓨터공학과, 서울특별시 마포구 백범로 35 R904
연구분야: 자연어처리
Tel: 02-706-8954, E-mail: gahng.sw@gmail.com

서 론

자연어에서는 하나의 단어가 둘 이상의 의미를 가지는 경우가 빈번하다. '다리'라는 단어는 사람이나 동물의 몸통 아래 붙어있는 신체의 부분을 의미하기도 하고, 물을 건너거나 또는 한편의 높은 곳에서 다른 편의 높은 곳으로 건너다닐 수 있도록 만든 시설물을 의미하기도 한다. 인간은 이러한 단어가 문장에 사용되었을 때 어떤 의미로 사용되고 있는지 쉽게 판단할 수 있지만 컴퓨터는 그렇지 않다. 따라서 둘 이상의 의미로 사용되는 중의성 단어가 문맥에서 어떤 의미로 사용되는지를 정확하게 파악 하는 작업이 필요하다. 이 작업을 단어 중의성 해소라고 한다.

단어 중의성 해소 방법은 지도 방식 및 비지도 방식 방법이 있다. 지도 방식 단어 중의성 해소 방법은 높은 성능의 결과를 내기위해서 많은 양의 정제된 학습 데이터가 필요하고, 대부분 테스트 셋의 단어 의미 범위는 일반적으로 제한되어 있다. 그러나 모든 언어의 단어와 의미를 포함하는 학습 데이터를 생성하는 것은 매우 많은 비용이 든다. 이러한 문제점들로 인해서 비지도 방식 단어 중의성 해소 방법이 제시되었다. 하지만, 비지도 방식을 이용한 단어 중의성 해소 방식은 높은 성능을 기대하기 힘들다. 본 연구에서는 대량의 의미 태그 된 말뭉치를 필요로 하지 않고 높은 성능을 내기위해 지식 기반을 이용한 비지도 방식의 단어 중의성 해소 방법을 제시한다.

지식 기반을 이용한 비지도 방식 단어 중의성 해소 방법은 그래프 기반(Agirrem & Soroa, 2009; Hessami, 2011; Mihalcea, 2005; Navigli & Lapata, 2010; Sinha & Mihalcea, 2007)과 유사도 기반(Banerjeem & Pedersen, 2002; Florentina, 2004; Lesk, 1986)이 가장 많이 연구되어 왔다. 그래프 기반 방법은 중의성 단어와 함께 문맥이나 문장에서 나타난 단어들과의 의미 그래프를 구축하여 연결 관계를 보고 중의성을 해소한다. 유사도 기반 방법은 중의성 단어와 함께 문맥이나 문장에서 나타난 단어들과의 사전에 나온 정보를 이용하여 의미 유사성을 계산하고, 높은 유사성을 가지는 단어의 의미를 선택한다. 하지만, 유사도 기반 방법은 사전에 나온 단어들이 모두 일치되어야 하기 때문에 사전으로부터 나오는 정보만으로는 높은 성능을 기대하기 힘들다. 이러한 이유로 그래프 기반 방법보다 낮은 성능을 보인다(Navigli & Lapata, 2010).

하지만 그래프 기반 방법은 모든 중의성 단어를 가지고 의미 그래프를 구축하기 때문에 불필요한 간선과 노드 정보가 추가되어 오류를 증가시킨다는 단점이 있다. 따라서 본 연구에서는 이러한 문제를 해결하고자 반복적 접근 방식의 그래프 기반 단어 중의성 해소 방식을 사용했다. 이 방식은 모든 중의성 단어들을 특정 기준에 의해서 단어를 매칭하며 매칭 된 단어들을 반복적으로 그래프를 재구축 하여 단어 중의성을 해소한다. 하지만, 같은 문맥과 문장에 나타난 단어라도 조건 없이 단어 매칭을 하게 되면 올바른 간선과 노드를 가지지 못한 그래프가 생성되게 된다. 본 연구에서는 단어 매칭을 위한 기준으로 Word2Vec 모델을 이용하였으며 Word2Vec은 인공 신경망을 통해 문장들을 학습하고, 문장에 속한 단어들을 파악하여 의미적으로 비슷한 단어끼리 가까운 벡터공간에 표현해주는 모델이다(Mikolov et al., 2013). 이러한Word2Vec 특성을 이용하여 문장이나 문맥에서 나온 단어들 중 의미적으로 유사한 단어끼리 매칭하고 보다 의미있는 간선과 노드정보를 가진 그래프를 구축하여 기존에 연구되었던 그래프 기반 단어 중의성 해소 방법들 보다 더 높은 성능을 보여주었다.

관련 연구

그래프 기반 단어 중의성 해소 방식은 크게 세 가지 과정으로 진행된다. 첫 번째로 문맥이나 문장에 나온 단어들을 지식 사전에 정의 되어있는 단어로 매칭하기 위해서 단어를 기본형으로 바꾸는 작업이다. 예를 들면 “characterized” 라는 과거 동사가 들어왔을 때, 이 단어의 본동사인 “characterize“로 변경해준다. 마찬가지로 과거동사 뿐만 아니라 이 외에 복수형태의 단어 등도 기본형으로 변경한다. 기본형으로 변경하는 것은 식 (1)과 같이 정의 할 수 있다. 입력으로 들어오는 단어의 집합 W의 원소를 기본형으로 바꾸어 기본형 단어 집합 L에 추가한다.

$$W(w_1, \ldots, w_n) \rightarrow L(l_1, \ldots, l_n) \tag{1}$$

두 번째로는 문장이나 문맥에 나타난 중의성 단어들이 가지는 의미 그래프 구

축 단계이다. 기존의 연구들은 종속트리 경로나 최단거리 경로를 이용하여 의미간의 경로가 정의된 대규모 어휘 의미망 사전으로부터 탐색 알고리즘을 적용하여 단어 간의 의미 그래프를 구축했다. 식 (2)에서 S_L 은 기본형 단어들이 가지는 의미들의 집합을 의미하며 E_L은 의미들 간의 연결된 간선 정보의 집합이다.

$$S_L := \cup_{i=1}^{n} R(l_i),\ E_L := \phi \tag{2}$$

식 (2)와 같이 l_i가 가지는 의미들을 집합 S_L에 추가하고, 추가된 모든 의미들 간에 어휘 의미망에 표현된 간선을 탐색하여 탐색된 간선을 E_L에 추가하고, 그래프를 구축한다. 기존에 연구되었던 탐색방법은 위에서 언급한 두 가지 방법이 있다. 종속트리 경로 방법은 어휘의미망 사전에서 의미 집합 S_L에 속한 의미 $s_a \in S_L$와 새로운 의미인 $s_b \in S_L$을 지정된 길이 L만큼 경로를 모두 추출한다. 추출된 모든 경로 정보를 통해 각 의미 s_a와 s_b에 연결된 모든 간선들을 DFS (Depth-First Search) 알고리즘을 사용하여 추출한다(Navigli & Lapata, 2010). 최단거리 경로 방법은 BFS(Breath First Search)알고리즘을 사용하여 추출한다(Gutirrez et al., 2013). 하지만, Manion과 Sainudiin의 실험 결과를 볼 때, BFS알고리즘보다 DFS알고리즘이 보다 올바른 간선과 노드 정보를 가지는 그래프를 구축하는 것을 증명하였다(Manion & Sainudiin, 2014). 각 의미 s_a와 s_b에 연결된 모든 간선들 정보는 식 (3)과 같이 정의할 수 있다.

$$P_{a \to b} = \{\{s_a,\ s\},\ \ldots,\ \{s',\ s_b\}\},\ |P_{a \to b}| \leq L \tag{3}$$

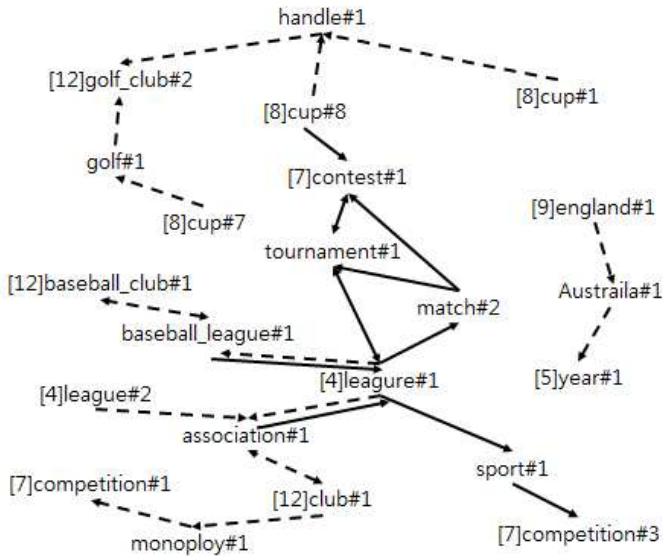
세 번째로는 구축된 그래프의 연결성을 측정하여 중의성 단어의 의미를 결정하는 단계이다. 그래프의 연결성 측정은 $\phi(s_{i,j})$로 정의한다. 연결성 측정을 통해 중의성 단어의 가장 높은 값을 가지는 의미를 선택하는 것은 식 (4)와 같이 정의할 수 있다.

$$\hat{s}_{i,*} = \mathrm{argmax}_{s_{i,j} \in R(l_i)} \phi(s_{i,j}) \quad (4)$$

그래프 연결성 측정은 다양한 방식으로 연구되어왔다. 하나의 노드에서 들어오는 간선의 수와 나가는 간선의 수를 통해 각 노드의 스코어를 측정하는 Degree Centrality 방법(Navigli & Lapata, 2007; Navigli & Ponzetto, 2012)과 시멘틱웹에서 많이 사용되고 있는 연결 구조를 이용해서 "Hub"와 "Authorities"를 찾는 알고리즘인 HITS(Hypertext Induced Topic Selection)(Kleinberg, 1999)와 노드의 상대적 중요성을 측정하기 위해 그래프 기반으로 노드의 우선순위를 정하는 PageRank 방법(Brin & Page, 1998)이 있다. Social Network Analysis방법은 사이 중앙성(Betweenness Centrality)을 사용하는데 이것은 전체 네트워크에서 그 노드의 중요성을 결정하는 측정치로 네트워크에 존재하는 모든 노드 쌍의 최단 경로와 그 최단 경로가 특정 노드를 경유하는 비율을 측정한다(Freeman, 1979). Word Sense Disambiguation Focused 방법에는 그래프에서 다른 의미들에 연결된 모든 경로들의 길이 역을 합산하여 각각의 의미에 점수를 매기는 방식인 Sum Inverse Path Length 방식(Navigli & Ponzetto, 2012)이 있다. 하지만, 기존에 많이 사용되는 그래프 연결성 측정 방식들 중 Degree Centrality 방식이 가장 높은 성능을 보이는 것을 증명하였다(Manion & Sainudiin, 2014; Navigli & Lapata, 2010).

기존의 그래프 기반 단어 중의성 해소는 문맥이나 문장 내에서 나타나는 모든 중의성 단어에 대해서 그래프를 구축하여 연결 관계를 분석한 후 중의성 단어의 의미를 결정한다. 하지만, 대규모 어휘 의미망 사전은 단어의 각 의미들 간의 수많은 의미 경로 정보가 포함되어 있기 때문에 의미경로필터를 하지 않으면, 그래프 기반 단어 중의성 해소 방법으로 중의성 단어의 의미를 결정하는데 상당한 오류를 발생 시킨다. 이와 같은 문제를 해결하고자 반복적 접근 방식의 그래프 기반 단어 중의성 해소 방법이 연구되었다. 이 방법은 모든 중의성 단어들을 특정 기준에 의해서 단어 매칭을 하고 매칭 된 단어들을 반복적으로 그래프를 재구축하면서 연결 관계를 보고 중의성 단어의 의미를 결정한다(Manion & Sainudiin, 2014). Manion과 Sainudiin의 연구에서 반복적 그래프와 비 반복적 그래프를 비교하는 예제를 보여주었다. 이 예제에 포함된 예문(*SemEval-2013 d011.s007)*인 "Spanish football players

playing in the All-Star League and in powerful clubs of the Premier League of England are during the year very active in league and local cup competitions and there are high-level shocks in the European Cups and European Champions League."에 나온 중의성 단어들을 반복적으로 그래프를 구축한 것과 비 반복적으로 그래프를 구축한 것을 비교하였다. 하지만 모든 중의성 단어들의 의미 경로를 불러와 비교하지 않고, 중의성 단어 "cup"에 대한 의미 경로만 추출하여 비교하였다.



(그림 1) 반복적 접근 방식의 그래프 구축과 일반적인 방식의 그래프 구축 비교

(그림 1)은 중의성 단어 "cup"에 대한 모든 의미 경로를 추출한 것이다. 점선 경로는 일반적으로 그래프를 구축한 방법을 표현한 것이고, 비점선 경로는 반복적 접근 방식으로 그래프를 구축한 방법을 표현한 것이다. 예문(SemEval-2013 d011.s007)에서 "cup"의 의미는 cup#8의 의미로 "경기에서 우승하면 주는 트로피"가 올바른 의미이다. 하지만, 점선 경로만을 보고 Out Degree 방식으로 cup의 알맞은 의미를 선택하면, cup#1, cup#7, cup#8의 나가는 간선의 수가 모두 같기 때문에 가장 먼저 추출된 의미인 cup#1이 선택되어 올바르지 못한 결과를 내게 된다.

같은 방법으로 비점선 경로를 통해 cup의 의미를 선택하면 cup#8이 선택되어 올바른 의미로 선택되게 된다. 이와 같이 반복적 접근을 이용하여 의미 그래프를 구축하면 대규모 어휘 의미망 사전으로부터 의미경로필터가 적용된 정확한 그래프를 추출할 수 있다.

반복적 접근 방식의 그래프 구축 방법은 특정 기준에 의해서 반복적으로 그래프를 재구축 한다. 하지만, 같은 문맥과 문장에 나타난 단어라도 조건 없이 단어를 매칭 하게 되면 올바른 간선과 노드를 가지지 못한 그래프를 추출된다. 따라서, 연결성 측정을 먼저 진행하고 그 다음 순서로 의미를 결정할 때 중의성 단어의 올바른 의미를 결정하지 못하게 된다.

본 연구에서는 반복적 접근 방식의 그래프 기반 방법으로 시스템을 구축하였고, 다음 장에서는 올바른 간선과 그래프를 구축하기 위해 Word2Vec을 이용한 단어 매칭에 중점을 두어 설명한다.

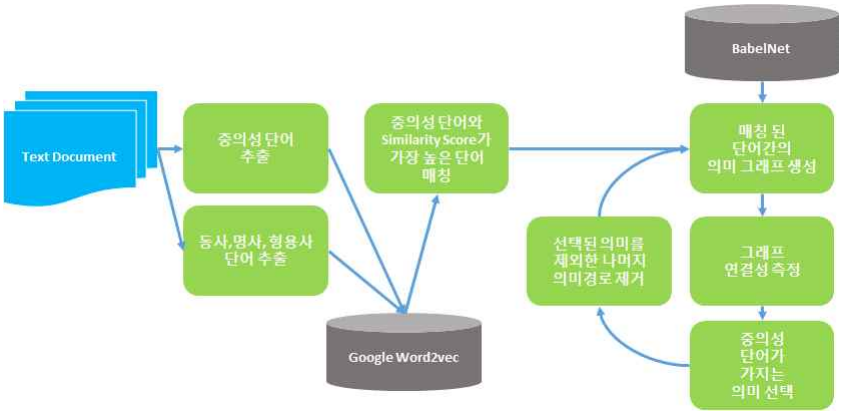
Word2Vec을 이용한 반복적 접근 방식의 그래프 기반 단어 중의성 해소 방법

본 연구에서는 반복적 접근 방식의 그래프 기반 시스템으로 문장이나 문맥에서 나타난 단어들 중 의미적으로 유사한 단어끼리 매칭하여 올바른 간선과 노드를 가진 그래프를 재구축하는 시스템을 제안한다((그림 2)).

제안 시스템은 2단계로 이루어져 있다. 1단계에서는 반복적으로 그래프를 재구축하기 위한 중의성 단어들을 매칭하는 과정을 진행하며, 2단계에서는 매칭 된 단어들 간의 의미그래프를 구축하고 연결 관계를 통해 중의성 단어의 의미 결정하는 과정을 진행한다.

단어 매칭

의미 그래프는 두 개 이상의 단어가 의미적으로 관계가 있을 때 더 정확한 그래프가 구축되며 그래프 연결성 측정 시에 더 정확한 의미를 선택할 수 있게 도와



(그림 2) 제안 시스템

준다. 그래서 의미적으로 관계가 있는 단어 매칭을 위해서 Word2Vec 결과를 이용하였다. Word2Vec은 입력 문장을 인공신경망을 통해 학습하고 유사한 주변단어를 갖는 단어는 의미적으로 유사한 단어들로 가까운 벡터공간에 표현되게 된다. 이러한 Word2Vec의 특성을 이용하여 (그림 3)과 같이 문장이나 문맥 내에 나온 단어들 중에 중의성 단어와 가장 벡터공간에 가까운 단어를 매칭한다.

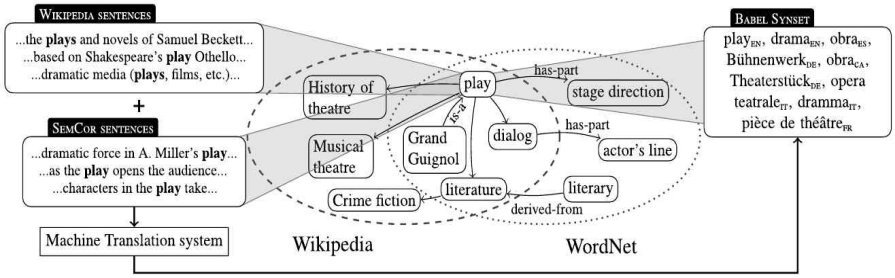
(그림 3) 문장 내 중의성 단어와 유사한 단어 매칭

Word2Vec 결과는 Google에서 제공하는 데이터를 이용하였다. 이 데이터는 1000억개의 단어들로 이루어진 뉴스데이터를 학습하여 3백만 개의 단어와 구로 이루어진 300차원의 벡터결과를 제공한다.

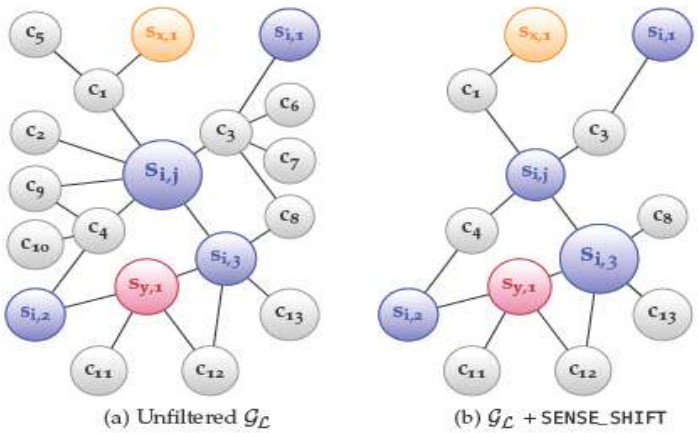
BabelNet 어휘 의미망 사전

본 연구에서는 BabelNet 어휘 의미망 사전을 지식기반으로 사용하여 다중언어에서의 단어 중의성 해소도 해결할 수 있도록 하였다(Navigli & Ponzetto, 2012).

BabelNet은 다국어 어휘 의미망 사전으로 그림 4와 같이 영어 WordNet에 방대한 다국어 정보를 가진 Wikipedia 백과사전을 연결하여 만들어졌다. BabelNet은 또한 의미 경로 필터를 자체적으로 지원한다. 그래프를 구축 할 때 어휘의미망 사전의 그래프가 정점이나 간선에 의한 경로들이 정제 되어 질 경우 그 그래프 구축은 연결 관계를 보고 중의성 단어의 의미를 결정하는데 효과적일 것이다. 그래프를 구축하는 과정에서 같은 기본형 단어로부터 얻은 다의어 단어의 의미들은 서로가 매우 밀접하게 있으면, 이 의미들은 그래프 연결성 측정을 할 때, 각각 서로의 스코어에 많은 영향을 미치게 된다. BabelNet은 SENSE_SHIFT라 불리는 필터로부터 같은 기본형 단어로부터 얻은 의미들 공유하는 그래프에 추가된 경로를 제거시켰다.



(그림 4) BabelNet의 형태



(그림 5) 그래프 경로 필터링이 되지 않은 그래프와 된 그래프

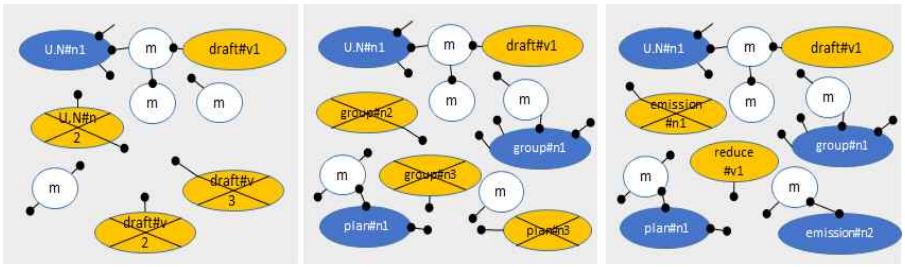
이 필터의 효과는 (그림 5)에서 확인할 수 있다. 예를 들어, s의 의미들을 가지는 단어가 문장에서 쓰이는 의미로는 $s_{i,3}$인데, In-Degree방식을 이용해 s의 의미들 중 하나를 선택하려고 한다면 (그림 5).(a)의 경우 $s_{i,j}$가 선택될 것이다. 반대로, 필터를 통해서 (그림 5).(b)의 그림처럼 그래프를 구축하게 되면, $s_{i,j}$가 아닌 원하는 결과의 $s_{i,3}$가 선택될 것이다.

반복적 접근의 그래프 구축

반복적 접근 방법을 이용하여 그래프를 구축하기 위해서는 반복적으로 그래프를 구축하기 위한 특정 기준이 필요하다. 본 연구에서는 입력문장으로부터 포함되어 있는 중의성 단어들을 나온 순서대로 반복적으로 그래프를 재구축하여 단어 중의성을 해소하였다.

그래프를 구축은 종속트리 경로를 이용하여 구축하였는데, 중의성 단어와 매칭된 단어가 가지는 모든 의미들을 지정된 길이만큼 어휘 의미망 사전에 정의된 의미 경로 정보를 추출하고, 지정된 길이만큼 추출된 모든 의미들의 경로정보를 이용해 DFS로 검색하여 서로의 의미 간에 탐색되는 간선과 의미노드를 추출하여 그래프를 생성하였다.

반복적으로 그래프를 재구축하는 과정은 (그림 6)과 같다. (그림 6)은 (그림 3)의 예제 문장에서 중의성 단어들이 나오는 순서대로 반복적으로 그래프를 재구축 해 나가면서 의미를 결정하는 과정을 보여준다. 이 문장에서 중의성 단어는 “U.N.”, “group”, “plan”, “emission”이다. 먼저 첫 번째로 나오는 중의성 단어 “U.N.”의 의미를 결정하기 위해서 “Word2Vec”에 의해 매칭 된 단어 “draft”와 그래프를 구축한다. 그래프 연결성 측정은 In Degree 방식을 사용하였다. In Degree 방식은 해당 노드에 종료된 간선의 수가 많은 노드의 스코어가 가장 높게 나오는 방식이다. 이 측정 방식을 이용하여 각 중의성 단어가 가지는 의미 노드들 중에서 종료된 간선의 수가 많은 의미 노드를 해당 중의성 단어의 의미로 결정한다. In Degree 방식을 통해 가장 종료된 간선의 수가 많은 U.N.#n1과 draft#v1의 의미를 제외한 나머지 노드들은 의미 그래프에서 삭제하고, 삭제된 나머지 그래프 정보를 반환한다.



(그림 6) Word2Vec결과를 이용해 반복적 그래프 기반 단어 중의성 해소 과정

다음으로 나오는 중의성 단어 “group”의 의미를 결정하기 위해서 “group”의 중의성 단어와 Word2Vec을 통해 문장 내 가장 유사도가 높은 중의성 단어로 매칭 된 “plan”의 의미 경로를 앞서 U.N#n1과 draft#v1으로부터 반환된 그래프 정보에 추가하고, 새롭게 그래프를 재구축한다. In Degree 그래프 연결성 측정을 통해 group#n1과 plan#n1의 의미를 결정하고, 결정된 U.N#n1, draft#v1, group#n1, plan#n1의 의미들을 제외한 나머지 노드들은 삭제하고 나머지 그래프 정보는 반환한다. 위와 같은 방식으로 문장 내에 마지막 중의성 단어인 “emission”의 의미를 결정하기 위해 “emission”의 의미경로와 Word2Vec으로 매칭된 단어 “reduce”의 의미경로를 반환된 그래프에 추가하고, 새롭게 그래프를 재구축한다. 그래프 연결성 측정을 통해 emission#n2와 reduce#v1 의미를 결정하면, 문장 내에 해결하고자 하는 중의성 단어인 U.N, group, plan, emission의 모든 의미를 결정되고, 반복적으로 그래프를 재구축하는 것을 종료한다.

실험 및 평가

본 연구에서 실험을 위해 사용한 데이터는 *SemEval-2013-English* 데이터를 이용해 테스트를 했다. 이 데이터는 2013년도 에 *개최한 SemEval Task 12 Multilingual Word Sense Disambiguation* 파트에서 제공한 데이터로 2010년에서 2012년까지 스포츠에서 금융까지 다양한 도메인이 포함된 13개의 뉴스기사로 이루어져 있다(Navigli, 2013).

중의성 단어의 의미는 문맥에 나타난 단어의 정보에 의해서 의미가 결정될 수 있다. 그래서 문맥의 정보를 사용하느냐 안하느냐에 따라 성능이 크게 바뀐다. 본 연구의 연구 과정에서 행한 실험은 문맥정보를 이용하지 않고 오로지 문장 정보만을 이용한 실험과 문맥정보를 이용한 실험을 하여 각각 성능을 비교하였다.

문장 단위 내에서 단어 중의성 해소 시스템 성능 비교

실험은 반복적 접근 방식을 이용한 그래프 구축 방식인 Iter 방법과 전체 중의성 단어를 불러와 탐색 알고리즘만 적용하여 그래프를 구축한 No Iter방식을 각각 적용하여 실험 하였다. 그래프를 구축하기 위한 탐색 알고리즘은 *UMCC-DLSI-Run 1 Sen*은 BFS방식을 적용하였고 나머지 시스템은 모두 DFS를 적용하였다. 그래프 연결성 측정은 *UMCC-DLSI-Run 1 Sen*을 제외하고 모두 In-Degree 측정법을 적용하여 스코어를 계산하였다. *UMCC-DLSI-Run 1 Sen*의 측정 방식은 기존의 측정법과 다르게 PPR+Freq 측정법을 적용하여 스코어를 계산하였다(Gutirrez et al., 2013).

본 연구에서 제안하는 모델은 반복적 접근 방식을 적용하여 그래프를 구축한 *Word2Vec Iter Sen* 시스템과 그렇지 않은 *Word2Vec No Iter Sen* 시스템이다. *No Iter Sen* 시스템은 반복적 접근 방식을 적용하지 않고 전체 중의성 단어를 불러와 탐색 알고리즘만 적용하여 그래프를 구축한 시스템이고, *Sudoku Iter Sen* 방식은 중의성 단어의 의미의 개수가 가장 작은 기준으로 그래프를 반복적으로 재구축한 시스템이다(Manion & Sainudiin, 2014). *UMCC-DLSI-Run 1 Sen* 시스템은 BabelNet 이외에 ISR-WN이라는 어휘 의미망 사전을 추가하여 새롭게 지식 기반을 구축하였고, 구축된 어휘 의미망 사전으로부터 *No Iter* 방식으로 탐색 알고리즘만 적용하여 그래프를 구축한 시스템이다.

<표 1>의 결과에서 본 연구에서 제안하는 반복적 시스템*(Word2Vec Iter Sen)*과 비 반복적 시스템*(Word2Vec No Iter Sen)*이 다른 그래프 방식의 시스템들보다 더 좋은 성능을 내었다. *UMCC-DLSI-Run 1 Sen* 시스템은 *SemEval 2013 Task 12 Multilingual Word Sense Disambiguation* 대회에서 가장 높은 성능을 보였으나 많은 비용이 증가되는 확장된 지식 기반과 문맥정보를 사용하였다. 제안 시스템은 비록 *UMCC-DLSI-Run 1 Sen* 시스템보다는 다소 낮은 결과를 보여주었지만 추가적인 비용

〈표 1〉 문장 단위 내에서 구축한 단어 중의성 해소 시스템 성능 비교

Method	Precision	Recall	F_1-Measure
No Iter Sen	60.83	50.70	55.30
Sudoku Iter Sen	61.80	56.23	58.88
UMCC-DLSI-Run1 Sen	67.70	67.70	67.70
Word2Vec No Iter Sen	**57.11**	**56.55**	**56.83**
Word2Vec Iter Sen	**62.70**	**62.09**	**62.39**

증가 없이 유사한 성능을 보여주었음을 증명하였다.

문맥 단위 내에서 단어 중의성 해소 시스템 성능 비교

본 연구에서 제안하는 시스템은 한 문서 전체의 문맥을 사용하여 실험한 결과와 지정된 문맥 정보만 이용하여 그래프를 구축하여 성능 비교를 하였다. 본 연구에서 제안하는 시스템 중 단어 *Doc*이 포함된 방식(*Word2Vec No Iter Doc, Word2Vec Iter Doc*)은 한 문서 전체에서 중의성 단어와 가장 유사한 단어를 매칭하여 그래프를 구축한 시스템이고, *Word2Vec Iter Context(3)*은 중의성 단어가 포함된 문장 외에 위에 3문장 내에 포함된 단어와 가장 유사한 단어를 매칭하여 그래프를 구축한 시스템

〈표 2〉 문맥 단위 내에서 구축한 단어 중의성 해소 시스템 성능 비교

Method	Precision	Recall	F_1-Measure
No Iter Doc	61.70	55.51	58.44
Sudoku Iter Doc	65.39	63.74	64.55
UMCC-DLSI-Run2 Doc	68.50	68.50	68.50
Word2Vec No Iter Doc	**70.44**	**69.75**	**70.10**
Word2Vec Iter Doc	**75.20**	**74.46**	**74.83**
Word2Vec Iter Context(3)	**65.53**	**64.88**	**65.20**
Word2Vec Iter Context(5)	**68.77**	**68.09**	**68.43**

이다. *Word2Vec Iter Context(5)*는 중의성 단어가 포함된 문장 외에 위에 5문장 내에 포함된 단어와 가장 유사한 단어를 매칭하여 반복적으로 그래프를 구축한 시스템이다.

<표 2>는 본 연구에서 제안하는 한 문서 전체의 문맥 정보를 고려하여 구축한 본 시스템이 가장 높은 성능을 보인 것을 보여준다. *UMCC-DLSI-Run 1 Doc*시스템은 문장 정보만을 이용하여 그래프를 구축한 것보다 성능은 증가하였지만, 반복적으로 그래프를 구축하지 않았기 때문에 성능이 높은 증가율을 보이지 못하였다. 또한 지정된 문맥정보를 이용하여 반복적으로 그래프를 구축하여도 한 문서 전체의 정보를 이용한 반복적 방식보다 더 좋은 성능을 보였고, *UMCC-DLSI-Run 1 Doc*시스템과 비교해도 성능 차이가 크지 않았다.

결과적으로, Word2Vec을 이용한 단어 매칭이 문맥 정보를 이용할 시에 단어 간의 의미적 유사도가 더 높은 단어끼리 매칭되어 그래프를 구축할 때, 문장 정보만을 이용하여 그래프를 구축한 것보다 더 올바른 간선과 노드가 추가된 그래프를 구축되어 더 높은 성능을 낼 수 있었다. 그리고 한 문서 전체의 문맥 정보를 이용하지 않아도, 정확한 단어 매칭만 이루어진다면, 반복적으로 그래프를 구축할 시에 더 올바른 간선과 노드 정보가 추가된 그래프를 구축되어 높은 성능을 기대할 수 있었다.

결 론

본 연구에서 제안하는 시스템을 분석해보면, 문맥 정보를 사용하여 반복적으로 그래프를 구축하는 시스템이 높은 성능을 보여 주었다. 이러한 이유는 문장 정보만 이용하여 중의성 단어와 문장에서 나온 단어들과 매칭하게 되면, Word2Vec 거리 값이 가깝지 않은 단어 간의 매칭이 되는 경우가 많다. 그래서 매칭 된 두 단어 사이에 그래프를 구축할 때, 불필요한 간선과 노드가 많이 추가되어 오히려 성능을 저하시킨다. 반대로 문맥 정보를 이용하면, 문맥에 나온 모든 단어들 중 Word2Vec 거리 값이 가까운 단어 간의 매칭이 되는 경우가 많아 그래프를 구축할 때, 성능을 낮추는 불필요한 간선과 노드를 줄일 수 있게 된다.

끝으로 본 연구의 결과를 바탕으로 하여 더 좋은 성능을 발휘하는 단어 중의성 해소 시스템을 개발하기 위해 생각해 볼 수 있는 향후 연구 과제로 다음과 같은 문제를 제안한다.

그래프 연결성 측정 방식을 개선하는 방법이다. 본 연구에서는 단순하게 해당 노드에 종료되는 간선의 수가 많으면 많을수록 노드의 스코어가 높아지는 In-Degree 방식을 제안하였지만, 이러한 방법이 아닌 *UMCC-DLSI-Run* 시스템에 적용한 PPR+Freq 측정방법을 적용하면 더 나은 성능이 예상된다. 이 측정방식은 Personalized Page Rank의 개선 방법으로 각 단어의 의미 패턴 빈도를 Personalized Page Rank 식에 추가하여 더 좋은 성능을 내는 방식이다. 이 방식은 단어가 가지는 의미 패턴 빈도를 그래프 연결성 측정 방식에 추가하였기 때문에 많은 문맥 정보를 사용하지 않아도 해당 중의성 단어의 의미를 결정하는데 상당히 플러스 요인이 될 것이다.

참고문헌

Agirre, E. & Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. *Proceedings of EACL*, 33-41.

Banerjee, S. & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Proceeding*, 136-145.

Brin, S. & Page, L. (1998) The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems*, 107-117.

Florentina, V., Langlais, P. & Lapalme, G. (2004). Evaluating variants of the Lesk approach for disambiguation words. *Proceedings of the Conference on Language Resources and Evaluation*, 633-636.

Freeman, L. C. (1979). Centrality in Social Networks Conceptual Clarification. *Social Networks*, *1*(3), 215-239.

Gutirrez, Y., Orqun, F., Camara, F., Castaeda, Y., Gonzlez, A., Montoyo, A., Muoz, R.,

Estrada, R., Piug, D., Abreu, I. & Prez, R. (2013). UMCC DLSI: Reinforcing a Ranking Algorithm with Sense Frequencies and Multidimensional Semantic Resources to solve Multilingual Word Sense Disambiguation. *Proceedings of the 7th International Workshop on Semantic Evaluation*, 241-249.

Hessami, E., Mahmoudi, F. & Jadidinejad, A. (2011). Unsupervised Graph-based Word Sense Disambiguation Using lexical relation of WordNet. *International Journal of Computer Issues*, *8*(6), 225-230.

Kleinberg, M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, *46*(5), 604-632.

Lesk, M. (1986). Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone. *Proceedings of the SIGDOC*, 24-26.

Manion, L. & Sainudiin, R. (2014). An Iterative Sudoku Style Approach to Subgraph-based Word Sense Disambiguation. *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics*, 40-50.

Mihalcea, R. (2005). Unsupervised large-vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. *Proceedings of HLT/EMNLP*, 411-418.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality, *Proceedings of The 27th Annual Conference on Neural Information Processing Systems.*

Navigli, R., Jurgens, D. & Vannella, D. (2013). SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantcis (*SEM 2013).*

Navigli, R. & Lapata, M. (2007). Graph Connectivity Measures for Unsupervised Word Sense Disambiguation. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 1683-1688.

Navigli, R. & Lapata, M. (2010). An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, *32*(4), 678-692.

Navigli, R. & Ponzetto, P. (2012). *BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence*, 217-250.

Navigli, R. & Ponzetto, P. (2012). Joining Forces Pays Off: Multilingual Joint Word Sense Disambiguation. *Proceedings of EMNLP/CoNLL*, 1399-1410.

Navigli, R. & Velardi, P. (2005). Structural Semantic Interconnections: A Knowledge-based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(7), 1075-1086.

Sinha, R. & Mihalcea, R. (2007). Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity. *Proceedings of the IEEE International Conference on Semantic Computing*, 363-369.

1차원고접수 : 2016. 01. 04
1차심사완료 : 2016. 03. 07
2차원고접수 : 2016. 03. 16
최종게재승인 : 2016. 03. 17

(*Abstract*)

An Iterative Approach to Graph-based Word Sense Disambiguation Using Word2Vec

Dongsuk O Sangwoo Kang Jungyun Seo

Computer Science and Engineering Sogang University

Recently, Unsupervised Word Sense Disambiguation research has focused on Graph based disambiguation. Graph-based disambiguation has built a semantic graph based on words collocated in context or sentence. However, building such a graph over all ambiguous word lead to unnecessary addition of edges and nodes (and hence increasing the error). In contrast, our work uses Word2Vec to consider the most similar words to an ambiguous word in the context or sentences, to rebuild a graph of the matched words. As a result, we show a higher F1-Measure value than the previous methods by using Word2Vec.

Key words : Word Sense Disambiguation, Graph-based, Word2Vec, Unsuperviesed Learning, NLP.