술어를 활용한 명사 논항간의 유사도 계산[1)]

조병철○, 석미란, 김유섭
한림대학교 유비쿼터스 컴퓨팅학과
max91128@nate.com, smr4880@hanmail.net, yskim01@hallym.ac.kr

Similarity Estimation of Argument Between Noun using Predicate

Byeong-Cheol Jo○, Mi-Ran Seok, Yu-Seop Kim
Dept. of Ubiquitous Computing, Hallym University

요 약

본 논문에서는 명사간의 유사도 추정을 위하여 명사 어휘와 술어-논항 관계에 있는 동사들의 유사도를 측정하여 이를 활용하는 연구를 제안한다. 어휘 유사도 추정은 정보 통합과 정보 검색 분야에서 중요한 역할을 한다. 본 연구에서는 유사한 명사 어휘들은 유사한 문맥을 가지고 있으며 동시에 명사 어휘의 문맥에 있어 가장 중요한 문맥 정보는 명사 어휘와 직접적인 구문 관계를 가지고 있는 술어 정보임을 가정하였다. 실험을 위하여 본 연구에서 제시된 유사도와 명사 계층 클래스간의 유사도간의 상관관계를 계산하였다.

주제어 : 유사도계산. 상관계수 . Propbank.

1. 서론

어휘 의미 유사도 측정은 자연어 처리와 정보 검색 분야에서 많은 기회를 제공한다[1][2]. 때때로 우리는 애매모호한 단어의 의도된 의미를 파악하는데 어려움을 느낀다.

본 연구에서는 명사 논항간의 유사도 추정을 위하여 술어-논항 관계를 활용하는데, 술어들의 성질을 반영하기 위하여 수동으로 구축된 의미역 표지 부착 말뭉치를 사용한다. 여기서는 의미역 표지 부착 말뭉치로 Proposition Bank(이하 PropBank)[3] 체계를 따르는 한국어 PropBank[4]를 자체적으로 구축하여 활용하여 명사와 술어-논항 관계에 있는 술어들의 유사도를 계산함으로서 추정하였다.

개념간의 의미 유사도를 측정하는 방법에는 간선기반 측정방법[6]-[8], 정보량(IC) 측정 방법[9]-[11], 두 개의 방법을 결합한 하이브리드 측정 방법[12]-[15] 등이 있다. 간선 기반 유사도 측정 방법은 개념간의 최단 경로 수를 계산하거나, 계층 구조상에서의 개념의 깊이를 계산하거나, 관계 종류를 기준으로 유사도를 측정한다. 정보량 유사도 측정 방법은 노드의 확률을 기반으로 정보량을 측정하는 접근 방법이다.

이들 방법론들은 기본적으로 명사 어휘가 문맥에서 독립적이라고 가정한다. 일반적으로 어휘들은 주변 어휘들과 밀접한 의미 관계를 가지고 있는데, 이러한 관계 중에서 술어-논항 관계는 문장 전체 의미 분석에 있어 가장 기본적인 관계이다. 때문에 본 연구에서는 명사 어휘와 술어-논항 관계에 있는 술어들의 유사도를 분석함으로써 명사 어휘의 유사도를 추정하고자 하는 것이다.

2장에서는 술어-논항 관계를 파악할 수 있는 PropBank 말뭉치의 개략적인 내용을 기술하고, 3장에서는 계층 구조에 기반 한 유사도 계산 방법을 기술한다. 그리고 4장에서는 실험 결과를, 5장에서는 결론을 논한다.

2. PropBank 표지 부착 말뭉치 구조

그림[1]은 문장 “가게에서 사 가지고 간 사탕이나 초콜릿에 대해서는 선생님이 별 말이 없는 것을 알아차린 영재는 생일날이 다가오자 과자를 만들어 달라고 조르기 시작했다.” 의 구문표지 부착 결과를 보여준다.

1) 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No.20105-0010612)

```
1       2                                       가게[용언불가능보통명사]+에서[부사격조사]
2       4                                  사[일반동사]+이_가지[기타보조용언]+고_가[기타보조용언]+ㄴ[관형사형전성어미] (목적격)
3       4                                  사랑[용언불가능보통명사]+이니[접속조사]
4       8                                초콜릿[용언불가능보통명사]+에_대하_어서는[부사격조사]
5       8                                선생님[용언불가능보통명사]+이[주격조사]
6       7                                      를[성상관형사]
7       8                                  말[용언가능보통명사]+이[주격조사]
8       9                              없[성상형용사]+는[관형사형전성어미] (동격)
9       10                                 것[기타의존명사]+를[목적격조사]
10      11                             알아차리[일반동사]+ㄴ[관형사형전성어미] (주격)
11      16                        영재[인명고유명사]+는[일반보조사] (주격)
12      13                             생일날[용언불가능보통명사]+이[주격조사]
13      16                        다가오[일반동사]+자[종속연결어미]
14      15                                  과자[용언불가능보통명사]+를[목적격조사]
15      16                        만들[일반동사]+어_달[기타보조용언]+라고[종속연결어미]
16      0                   조르[일반동사]+기_시작하[기타보조용언]+었[과거시제선어말어미]+다[평서형종결어미]+.[문미기호]
```

[그림 1] 입력 문장의 구문 표지 부착 결과

[그림 1]의 구문 표지 부착 말뭉치와 연결되는 한국어 PropBank 말뭉치는 아래 [그림 2]와 같다.

```
etri#doc_11.txt 2 사다 ----- 1:1-ARGM-ADV(SRC) 2:0-rel
etri#doc_11.txt 10 알아차리다 ----- 9:4-ARG1(THM) 10:0-rel
etri#doc_11.txt 13 다가오다 ----- 12:1-ARG0(THM) 13:0-rel
etri#doc_11.txt 15 만들다 ----- 14:1-ARG1(THM) 15:0-rel
etri#doc_11.txt 16 조르다 ----- 11:6-ARG0(AGT) 13:2-ARGM-DIS()
15:2-ARGM-DIS() 16:0-rel
```

[그림 2] 의미역 표지 부착 결과

[그림 2]의 의미역 표지 부착 결과 중 네 번째 술어 '만들다' 와 관련한 의미역 표지에 대한 설명은 다음과 같다. 먼저, 술어 '만들다' 는 1개의 논항 '과자를' 을 갖는다. '과자를' 은 PropBank에서 행위의 수동자에 해당하는 'ARG1' 의미역으로 표지 부착 된다. 본 연구에서는 수많은 술어-논항 관계 중에서 술어와 문법적 연관성이 높은 논항 ARG1 관계만을 그 대상으로 한다. 일반적으로 ARG1 논항을 가진 술어는 목적어를 가진 타동사이고, 주어나 그밖에 부사격 수식어 보다는 목적어가 타동사와 문법적 연관성이 더 높다. 그러므로 본 논문에서는 명사 어휘의 유사도를 추정하는데 있어 해당 명사 어휘가 ARG1 의미역을 가지는 술어들의 유사도를 계산하고 이를 통합한다. PropBank에서 일반적으로 사용되는 의미역은 다음 [표 1]과 같다.

[표 1] PropBank 의미역

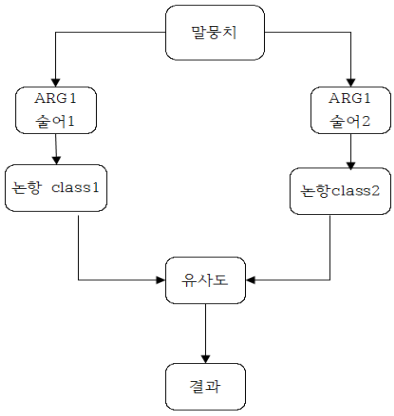
ProBank 의미역	
ARG0(주체자)	ARGM-EXT(크기)
ARG1(수동자)	ARGM-INS(도구)
ARG2(시작점)	ARGM-LOC(장소)
ARG3(끝점)	ARGM-MNR(방법)
ARGM-ADV(부사적 어구)	ARGM-NEG(부정)
ARGM-CAU(원인)	ARGM-PRD(술어의 자격)
ARGM-CND(조건)	ARGM-PRP(목적)
ARGM-DIR(방향)	ARGM-TMP(시간)
ARGM-DIS(문장의 연결)	

3. 계층적 개념 체계

단어 간의 유사도를 계산하기 위해 BOLA(Bank of language resources)의 한국어 개념 기반 어휘의미망인 코어넷을 사용하였다. 코어넷은 총 2,987개의 계층적 개념과 총 92,448개 어휘의미가 연결되어있다. 본 논문에서는 코어넷 중 CBL1(한국어 동사편)을 활용한다.

[그림 3]은 계층적 개념 체계 알고리즘을 보여준다. 수동으로 태깅한 약 10,000개의 의미역 말뭉치에서 ARG1으로 태깅한 술어1에 해당하는 논항 class1 Number 과 술어 2에 해당하는 논항 class2 Number 의 유사도를 구하면 ARG1으로 태깅한 술어1과 술어2의 유사도를 구할 수 있다.

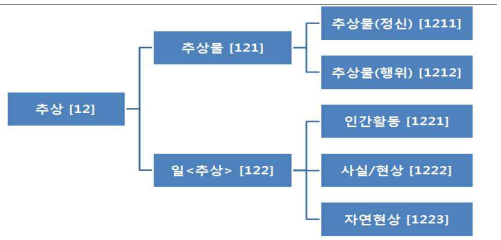
본 논문에서는 각 클래스 간의 경로의 길이를 측정함으로써 유사도를 계산하였다. 예를 들어 [그림 4]은 '꼽다' 동사와 '꽂다' 동사의 유사도를 계산한다고 했을 때 [12212151],[122117531]와[12222632]의 사이의 경로의 길이를 구함으로써 두 단어 사이의 유사도를 계산 할 수 있다.



[그림 3] 계층적 개념 체계 알고리즘

꼽다	꽂다
손동작[12212151, 308]	찔러넣기[12222632, 2191]
선택[122117531, 1562]	

[그림 4] 동사 class 번호



[그림 5] 한국어 어휘 의미망

[그림 5]은 CBL1의 계층 구조로 각 숫자는 개념체계 내에서의 위치에 관한 정보, 즉, 상위개념과 단계정보를 제공한다. 개념번호의 자릿수는 단계를 알려주며, 마지막 자릿수를 제거한 번호가 상위개념에 해당한다.[그림 5]에서 자연현상을 나타내는 1223는 3단계이며 상위개념으로는 일<추상>을 나타내는 122가 있고, 더 상위인 12는 추상을 나타낸다. 자연현상과 추상물(정신) 사이의 유사도를 계산하기 위해 경로의 길이를 측정 해 보면 1223과 122사이의 에지, 122와 12사이의 에지, 12와 121사이의 에지,121와 1211사이의 에지로 총 경로의 길이가 4라는 것을 알 수 있다[5].

[그림 4]에서 '꼽다' 단어와 '꽂다' 단어의 유사도를 계산한다고 했을 때, '꼽다'에 해당하는 '손동작'과 '선택'은 인간활동에 포함되는 단어이고 '꽂다'에 해당하는 '찔러넣기'는 사실/현상에 포함되는 단어이므로 '손동작'과 '찔러넣기', '선택'과 '찔러넣기'는 1221과 1222 사이의 경로와 1221과 1222의 길이를 구함으로써 두 단어 사이의 유사도를 계산할 수 있다. 경로의 길이가 짧을수록 유사도가 높다.

4. 실험

본 논문에서는 10,160개의 수동으로 만든 의미역 말뭉치에서 추출한 술어-논항 데이터 중, ARG1에 해당하는 명사 데이터를 이용한다. [표 2]는 명사 어휘 '금액'과 ARG1 관계에 있는 술어 '바라다'와 '늘어나다'의 계층 체계에서의 정보를 보여주고 있다. 또한 같은 방식으로 '급료'와 관련한 정보를 보여준다. 이를 통하여 '금액'과 관련 있는 술어의 개념 4개와 '급료'에 해당하는 클래스번호 19개의 거리를 1:1로 비교하여 총 76개의 클래스간의 거리를 얻을 수 있다. 계산 결과 '금액'과 '급료'의 유사도는 평균 8.8947이다.

[표 2] 코어 넷 동사 클래스 번호

ARG1 금액			ARG1 급료		
바라다	늘어나다		받다		
바람12211541	늘임12222A211		적합123361	쥐기12212156	
	증가12222A111		적합123361	수령122127B2	
	진보12222A41		수여122127B13	접촉1222281A12	
			수령122127B2	상처1222323	
			사기122127A2	수령122127B2	
			상속122123121	수령122127B2	
			연결123372	채취122128243232	
			개방12222661	수령122127B2	
			충돌1222281A2	조응123373	
			받치기1221282672		

본 논문에서 제시한 유사도와 명사 계층 구조를 이용하여 직접 거리를 측정하여 추정한 유사도 사이에는 0.311 상관 계수를 보여주었는데 [16]기준으로는 보통의 양의 상관관계가 있다고 볼 수 있다. 한편, 코어넷 상의 거리만을 고려한 유사도 추정 방법은 어의 중의성 문제가 해결되지 않았기 때문에 실용적으로는 사용 할 수 없다. 따라서 본 논문에서 제시한 방법은 이 문제에 자유롭기 때문에 실용적으로 다양한 분야에서 사용 될 수 있다.

[표 3] 상관관계 수준[16]

0.0~0.1	거의 관계 없음
0.1~0.2	약한 양의 상관관계
0.2~0.4	보통의 양의 상관관계
0.4~0.6	비교적 강한 양의 상관관계
0.6~0.8	강한 양의 상관관계
0.8~1.0	매우 강한 양의 상관관계

5. 결론

본 논문에서는 지금까지 PropBank의 ARG1 의미역으로 태깅한 데이터를 이용한다. 이 논문에서 제시한 방법으로 이용한 유사도와 명사 유사도 계층의 유사도를 계산하여 상관계수를 구한다.

향후에는 보다 다양한 방법으로 명사 단어 간 유사도를 계산할 수 있는 방법들을 적용시켜 본 문제에 가장 적합한 방법을 찾고자 한다.

참고문헌

[1] Q. Zhao, "Time-dependent semantic similarity measure of queries using historical click-through data", Proceedings of the 15th international conference on WWW, May 2006.

[2] A. Sebti and A. Barfroush, "A New World Sense Similarity Measure in WordNet", Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 369-373, March 2008.

[3] M. Palmer, D. Gildea, and Paul Kingsbury, "The Proposition Bank: An Annotated Corpus of Semantic Rules", Computational linguistics, 31(1), 71-106, 2005

[4] Young-Bum Kim, Heemoon Chae, Benjamin Snyder and Yu-Seop Kim*,"Training a Korean SRL System with Rich Morphological Features",The 52nd Annual Meeting of the Association for Computational Linguistics, 637-642, 2014

[5] 석미란, 윤영신, 김유섭, "개념 계층구조 상의 유사도를 이용한 이종 의미역의 자동 변환", 한국정보과학회 2014 한국컴퓨터 종합 학술대회 논문집,1773-1775, 2014

[6] Sussna, "WordSense Disambiguation for Free-text Indexing Using a Massive Semantic Network", in Proceedings of the Second International Conference on Information and Knowledge Management, Nov. 1993.

[7] C. Leacock and M. Chodorow, "Combining local context and WordNet similarity for word sense identification", in:C. Fellbaum(Ed.), WordNet: An eletronic lexical databases, MIT Press, pp. 265-283, May 1998.
[8] G. Hirst and D. Onge, "Lexical Chains as Representations of Context for the Detection and Correction of Malapropisms WordNet", C. Fellbaum, Cambrige, MA. The Mit Press, 1995.
[9] Philip Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", in Proceedings of the 14th IJCAI, April 1995.
[10] J. Jiang and D. Conrath, "Semantic Similarity based on corpus statistics and lexical taxonomy", In Proceedings on International Conference on Research in Computational Linguistics, Tiawan, pp. 19-33, Sep. 1997.
[11] D. Lin, "Using syntactic dependency as a local context to resolve word sense ambiguity", In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, pp. 64-71, May 1997.
[12] M. Rodriguez and M. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, Issue 2, pp. 442-456, May 2003.
[13] G. M. Euripides and G. Petrakis, "Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies", In 4th Workshop on Multimedia Semantics(WMS'06), pp. 44-52, April 2006.
[14] F. Lin and K. Sandkuhl, "A Survey of Exploiting WordNet in Ontology Matching", In Proceedings of the IFIP, Vol. 276, pp. 341-350, May 2008.
[15] V. Cross and X. Hu, "Using Semantic Similarity in Ontology Alignment", CEUR Workshop Proceedings, Vol. 814, Oct. 2011.
[16]Rea,L.M.,&Parker,R.A.(2005).Designing & Conducting Survey Research A Comprehensive Guide (3rd Edition).San Francisco, CA: Jossey-Bass