U-WIN을 이용한 의미 유사도 측정과 활용1)

임지희[O1], 배영준[1], 최호섭[2], 옥철영[1]
[1]울산대학교 컴퓨터정보통신공학과
{arisu80, young4862, okcy}@ulsan.ac.kr
[2]한국과학기술정보연구원 정보기술개발단 정보시스템개발팀
hschoe@kisti.re.kr

A Measure of Semantic Similarity and its Application in User-Word Intelligent Network

Ji-Hui Im[O1], Young-Jun Bae[1], Ho-Seop Choe[2], Cheol-Young Ock[1]
[1]Dept. of Computer Engineering and Information Technology,
University of Ulsan
{arisu80[O], young4862, okcy}@ulsan.ac.kr
[2]Information System Development Team,
Korean Institute of Science and Technology Information
hschoe@kisti.re.kr

요 약

개념 간의 유사도 측정 방법은 의미망에서의 두 개념의 최단 경로의 수·노드의 깊이·관계의 종류 등의 정보를 이용하는 링크(Link) 기반 방법, 대용량의 말뭉치에서의 개념의 발생빈도를 확률로 계산한 정보량(Information Content) 기반 방법, 관련 단어들의 공기정보를 활용한 의미(Gloss) 기반 방법이 있으며, 이미 국외에서는 WordNet과 같은 의미적 언어자원을 활용하여 많은 연구가 진행되고 있다. 그러나 국내에서는 아직 한국어 의미망을 바탕으로 한 개념간의 유사성 측정 방법이나 이를 활용하는 방법에 대한 연구가 미흡하다.

본 논문에서는 이를 바탕으로 링크 타입·노드의 깊이·최단경로·정보량 등의 요소를 이용한 의미 유사도 측정방법을 제안하고 이를 활용하여 명사-용언간의 연계 정보를 확보함으로써, 효율적으로 명사-용언간의 네트워크를 구축하도록 한다.

1. 서 론

국외에서는 개념간의 유사도 측정 방법이 단어 중의성 해소, 오용어 인식, 정보검색, 자연언어 학습과 같은 다양한 분야에서 광범위하게 연구되고 있다. 그러나 국내에서는 의미망·의미태깅된 말뭉치 등 의미적 언어자원의 부족으로 인해, 현재까지는 이러한 의미적 언어자원 구축에 관한 연구가 진행되고 있거나, 영어 어휘망인 WordNet에서의 개념간 유사도 측정 방법을 활용하는 연구[2] 등이 진행될 뿐, 한국어 의미망을 바탕으로 한 개념간의 유사성 측정 방법이나 이를 활용하는 방법에 대한 연구가 미흡하다.

개념간의 유사도 측정 방법을 크게 세 가지로 분류하면, 1) 의미망에서의 두 개념의 최단 경로의 수·노드의 깊이·관계의 종류 등의 정보를 이용하는 링크(Link) 기반 방법, 2) 대용량의 말뭉치에서의 개념의 발생빈도를 확률로 계산한 정보량(Information Content) 기반 방법, 3) 관련 단어들의 공기정보를 활용한 주석(Gloss) 기반 방법이 있다.

각 방법들은 유사도 측정 기준을 어디에 두느냐에 따라 분류한다. 그러나 개념 간의 유사도는 이러한 요소들을 모두 고려해야 하므로, 본 논문에서는 기존의 유사도 측정 방법들을 알아보고[2], 이를 바탕으로 링크 타입·노드의 깊이·최단경로·정보량 등의 요소를 이용한 의미

1) 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다.

유사도 측정방법을 제안한다. 그리고 제안한 방법을 활용하여 명사-용언간의 연계 정보를 확보함으로써, 효율적으로 명사-용언간의 네트워크를 구축하도록 한다.

2. 의미망 (U-WIN : User-Word Intelligent Network)

자연 언어의 어휘적 의미, 구문적 의미, 담화적 의미를 바탕으로 행위나 현상, 상태 등에 담긴 의미론적·개념론적 특성을 포함하고 있는 의미적 언어 자원 구축에 대한 연구는 다양하게 이루어지고 있다. 국외에서는 Word Net, EuroWordNet, Cyc, HowNet, Lexical FreeNet, EDR 등이 대표적이며, 국내에서는 카이스트의 CoreNet, ETRI의 어휘개념망, 부산대의 KorLex 등이 대표적이라 할 수 있다. [13]

한국어 사용자 어휘지능망 (User-Word Intelligent Network, 이하 U-WIN)은 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 이를 어휘의 의미적·개념적 네트워크로 형성한 온톨로지적 어휘망이라 할 수 있다.

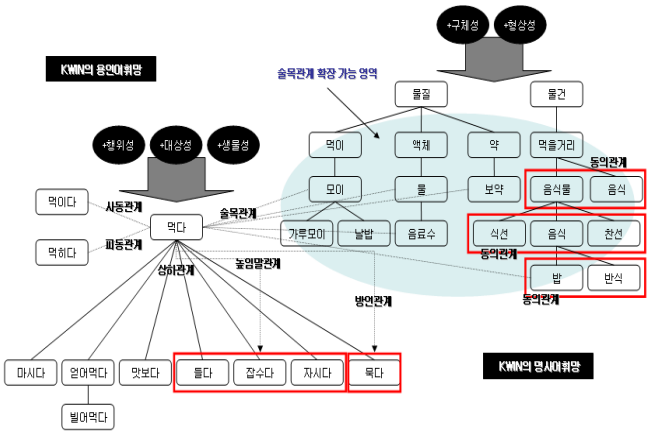

그림 1. U-WIN의 구축 사례 일부

U-WIN은 현재 30만 여 어휘가 구축된 상태이다. U-WIN의 구축 대상은 한국어 어휘 전체(모든 품사 및 언어 단위)로서, 핵심적 대상은 명사, 동사, 형용사이며, 부수적 대상은 부사, 관형사, 대명사, 감탄사, 조사, 수사, 의존명사 등이며, 북한어, 방언, 옛말, 전문용어, 고유명사, 어근, 어미 등 한국어 어휘 전체를 대상으로 연구 개발 중이다. [그림 1]과 [그림2]는 각각 구축 사례의 일부 모습과 분포도이다.

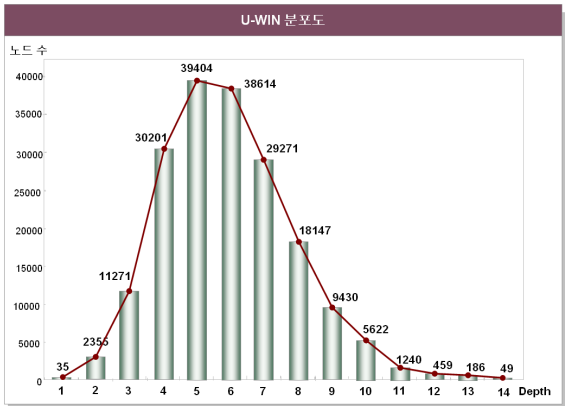

그림 2. U-WIN의 분포도

U-WIN은 한국어정보처리를 비롯하여 정보검색, 기계번역, 시맨틱 웹 등 다양한 분야에 이용될 수 있는 대규모 어휘 지식 베이스를 목표로 하고 있다. 현재 온톨로지 기반 의미적 주석(ontology-based semantic annotation)과 유사한 단어 중의성 해소(word sense disambiguation)와 의미 태깅(semantic tagging) 기술에 활용되고 있으며, 이 외에도 복합명사 자동 생성, 전문분야별 개념체계 자동 생성, 정보검색에서의 질의 확장, 어휘 학습 시스템 등 다양한 기술에서 활용되고 있다.

3. 개념간 의미 유사도 측정방법

기존의 개념간 유사도 측정방법에 대해 살펴보고, 본 논문에서 제안한 유사도 측정방법을 소개한다.

3.1 링크(Link) 기반 측정방법

링크를 기반으로 한 개념의 의미 유사도 측정방법은 개념간의 최단 경로 수를 계산하거나, 계층 구조상에서의 개념의 깊이, 관계 종류 등을 고려할 수 있다. 이러한 방법은 이론상으로 어떤 분류체계·시소러스·온톨로지에도 적용이 용이하여 쉽게 접근할 수 있으나, 사용하는 분류체계·시소러스·온톨로지의 종류에 따라 결과가 다를 수 있다.

대표적인 링크 기반 측정 방법에는 Rada, et. al[6]., Leacock and Chodorow[7], Wu and Palmer[11], Hirst and St.Onge[12] 등이 있다.

Rada, et. al[6]는 가장 기본적인 방법으로, 의학용어 시소러스(MeSH)를 대상으로 두 개념 간 최단 경로의 링크(Link) 개수를 기반으로 유사도를 측정하였다.

Leacock and Chodorow[7]는 Rada, et. al.의 방법에 계층 구조의 깊이를 함께 고려한다.

2) 본 논문에서는 링크 기반, 정보량 기반 측정방법에 대해서만 언급한다.

$$Sim_{lch}(c_1, c_2) = max[-log(length(c_1, c_2)/(2 \cdot D))] \quad \cdots\cdots (1)$$

length(C_1,C_2) 는 계층 구조에서 두 개념을 연결하는 최단 경로의 링크(Link) 개수이고, *D*는 분류체계·시소러스·온톨로지 등의 계층구조의 최대 깊이이다.

그리고 **Wu and Palmer[11]**와 **Hirst and St.Onge [12]**은 각각 계층구조의 깊이, 관계종류를 기준으로 유사도를 측정하였다.

3.2 정보량(Information Content:IC) 기반 측정방법

정보량(Information Content)은 대용량 말뭉치 내 개념의 발생 빈도를 기반으로 MLE(Maximum Likelihood Estimate)방법으로 얻는다. 많은 정보량이 할당된 개념은 특정 주제에 매우 세부적인 개념이고, 적은 정보량이 할당된 개념은 더 일반적인 개념으로 판단할 수 있다.

$$IC(concept) = -log(P(concept)) \quad \cdots\cdots (2)$$

또한, 상위어의 개념 빈도는 하위어의 모든 개념 빈도를 포함하므로, 계층 구조 내에서 상위에 위치한 개념일수록 더 높은 빈도를, 하위에 위치한 개념일수록 더 낮은 빈도를 가짐으로써, 최상위어(root node)는 수식(2)에 의해 가장 낮은 정보량을 가진다.

개념 빈도는 의미태깅된 말뭉치가 있을 경우 쉽게 측정할 수 있지만, 그렇지 않다면 다른 방법을 강구해야 한다. 다시 말해, 단어별 빈도를 해당 단어의 동형이의어/다의어 개수로 나누어 할당하거나, 단어별 빈도를 해당 단어의 동형이의어/다의어에 그대로 할당하는 것이다. Resnik에서는 첫 번째 방법을 사용하였으며, 본 논문에서는 두 번째 방법을 사용하였다.

대표적인 정보량 기반 측정 방법에는 Resnik[10], Jiang and Conrath[8], Lin[9] 등이 있다. **Resnik[10]**은 정보량을 사용하여 수식(3)에 의해 유사도를 측정한다.

$$Sim_{res}(c_1, c_2) = IC(lcs(c_1, c_2)) \quad \cdots\cdots (3)$$

*lcs(c1,c2)*는 개념 c1과 c2의 공통 상위어 중에서 가장 하위에 위치한 개념[3](LCS : lowest common subsumer)을 의미한다. 수식(3)에 의해, 부모 노드가 같은 개념들의 유사도는 최소 공통 상위어가 같아서 항상 같은 값을 가진다. 그러나 주로 계층이 큰 덩어리 형태로 이루어진(coarse-grained) 동사 어휘망은 동일한 최소 공통 상위어를 가지는 개념들이 많으므로, Resnik[]은 가장 좋은 coarse-grained measure로 알려져 있다.

Jiang and Conrath[8]은 Resnik 기반의 명사들의 유사도 측정방법이다. $dist_{jcn}(c_1, c_2)$이 작을수록, 즉 각 개념의 정보량의 합과 최소 공통 상위어의 정보량의 차이가 작다면, 두 개념은 유사도가 높다.

$$dist_{jcn}(c_1, c_2)$$

$$= IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \quad \cdots\cdots (4)$$

$$sim_{jcn}(c_1, c_2) = \frac{1}{dist_{jcn}(c_1, c_2)} \quad \cdots\cdots (5)$$

Lin[9]은 Jiang and Conrath과 비슷한 방법으로, 문서간의 유사도 측정방법 중에 하나인 Dice Coefficient를 이용한 방법이다.

$$sim(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad \cdots\cdots (6)$$

3.3 제안한 의미 유사도 측정방법

앞서 살펴본 기존의 방법들은 각기 다른 특징을 가지고 있다. 링크 기반 측정방법은 의미망의 계층구조에 의존적인 경향이 있으며, 정보량 기반 측정방법은 개념별 정보량 측정 과정에서 엄밀한 의미의 개념 발생빈도가 아닌 단어 발생빈도를 측정함으로써, 오류 발생 가능성을 내포하고 있다.

본 논문에서는 기존의 두 가지 측정방법을 결합하여 새로운 의미 유사도 측정방법을 제안한다. 앞서 살펴보았듯이, 계층구조의 깊이·최단경로정보·정보량 등의 요소가 개념간 유사도를 측정하는 데 적절한 요소로 판단이 되므로, 본 논문에서는 이러한 요소를 결합한 식(7)을 제안하였다. 개념 정보량은 의미태깅된 사전 뜻풀이, 용례에서 추출한 동형이의어 단어 빈도를 다의어에 그대로 할당하여 측정하였다. 이것은 의미망 계층구조에 대

3) 이후 개념 c_1과 c_2의 공통 상위어 중에서 가장 하위에 위치한 개념을 '최소 공통 상위어'라고 한다.

한 의존성을 줄이기 위해 정보량을 사용하였으며, 개념 발생빈도가 아닌 단어 발생빈도로 인한 오류 가능성은 발생 빈도를 의미태깅된 말뭉치에서 측정함을써 해결하도록 하였다.

$$sim(c_1, c_2) = \alpha \times \frac{link(c_1, c_2) \times depth(c_1, c_2)}{IC(los(c_1, c_2))} \quad \cdots\cdots (7)$$

$$depth(c_1, c_2) = \frac{d_{los}}{d_1 + d_2}$$

$$link(c_1, c_2) = -log\left(\frac{length(c_1, c_2)}{2 \times D}\right)$$

4. 명사-용언의 연계 정보 추출

의미망을 한국어정보처리를 비롯하여 정보검색, 기계번역, 시맨틱 웹 등 다양한 분야에 이용하려면, 명사 의미망뿐만 아니라 문장에서 중요한 역할을 담당하는 용언 의미망 구축과 동시에 명사·용언 의미망을 아우르는 명사-용언 네트워크가 구축되어야 한다. 그리고 명사-용언 네트워크는 명사·용언 의미망 사이의 상호 연계정보를 토대로 구축해야 한다.

명사-용언의 상호 연계성은 실제 말뭉치에서 용언과 명사들 간의 관계정보(술목·술주 관계)를 통해 확인할 수 있다. 즉, 용언과 공기하는 명사들의 클러스터링 과정을 통해 용언의 논항정보를 파악할 수 있고, 반대로 클러스터링된 명사들과 공기하는 용언을 그룹핑하여, 관련 용언들의 분포상태를 파악할 수 있으며, 이러한 과정을 통해 명사 및 용언의 구문적·의미적 속성정보를 자동으로 추출할 수 있다.

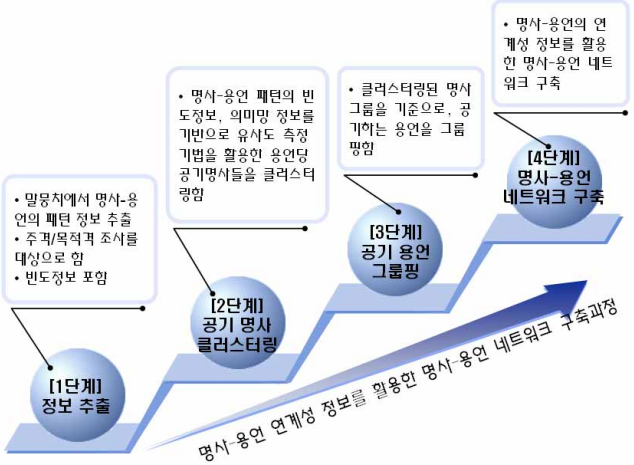

그림 3. 명사-용언의 연계정보 추출과정

[그림 3]은 명사-용언 네트워크 구축을 위한 연계 정보를 추출하는 간략한 과정이다. 우선 주격조사, 목적격조사를 대상으로 말뭉치 상에서 명사-용언의 패턴정보를 추출한 다음, 명사-용언 패턴(빈도 포함)·의미망·정보량을 이용하여, 본 논문에서 제안한 유사도 측정기법을 통해 용언의 공기명사들을 클러스터링한다. 이때, 명사-용언을 연결하는 조사의 종류(주격/목적격)를 구분한다. 그리고 클러스터링된 명사들과 공기하는 용언을 그룹핑하여, 용언들의 분포상태를 바탕으로 명사-용언 네트워크 구축과정의 효율성을 기대할 수 있다.

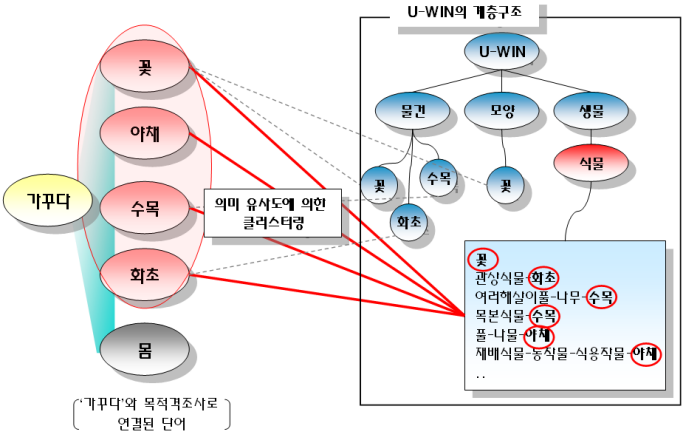

그림 4. 의미 유사도를 이용한 클러스터링 모습

[그림 4]는 말뭉치에서 추출한 패턴 중에서 동사 '가꾸다'와 목적격 조사로 연결된 '꽃, 야채, 수목, 화초, 몸'이 의미 유사도를 통해 클러스터링된 모습이다. U-WIN의 개념 노드는 다의어 수준으로 구축되어 있으므로, 명사 '꽃, 야채, 수목, 화초, 몸'은 다의어 수준의 총 34개 의미로 세분화한 다음, 의미 유사도를 측정하여 식물의 하위 개념으로 연결됨을 알 수 있다.

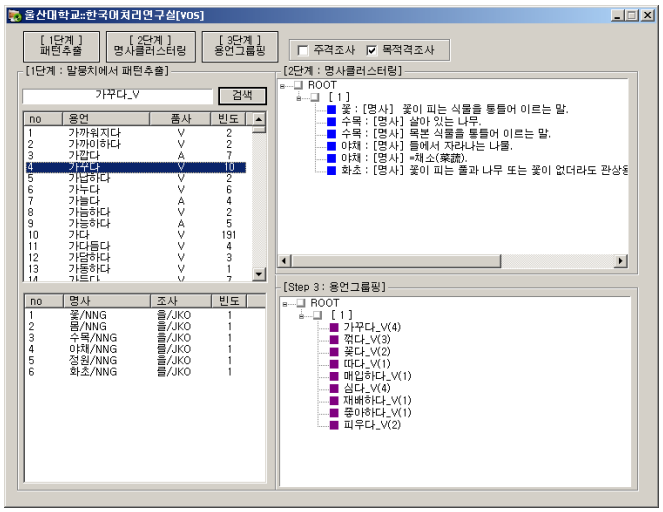

그림 5. 명사-용언 연계성 검증

그리고 [그림 5]에서 '가꾸다(4), 꺾다(3), 꽂다(2), 따다(1), 매입하다(1), 심다(4),[4] 재배하다(1), 좋아하다(1), 피우다(2)'의 용언들이 클러스터링된 명사들과 공기함을 알 수 있다.

5. 결론 및 향후 연구방향

본 논문에서는 기존의 의미 유사도 측정방법을 바탕으로 U-WIN 및 정보량을 활용한 새로운 의미 유사성 측정방법을 제안하고, 이것을 명사-용언 네트워크 구축을 위해 명사-용언의 연계성을 확보하는 데 활용하였다. 제안한 의미 유사도 측정방법은 링크 기반 방법과 정보량 기반 방법의 단점을 보완하였으며, 한국어 의미망을 활용한 개념적인 클러스터링을 가능하게 하였다.

향후 한국어 의미망과 의미 유사도을 활용하여 WSD·격틀사전 구축·정보검색·클러스터링 등의 다양한 분야에서 의미적인 요소를 추가하는 방법으로 활용 가능할 것이다.

참고문헌

[1] Dongqiang Yang and David Powers, "Measuring Semantic Similarity in the Taxonomy of WordNet", ACSC'05 Australasian Computer Science Conference, pp315-322, 2005.

[2] 조미영, 최준호, 김판구, "개념 기반 이미지 검색 시스템을 위한 WordNet 적용 방안", 정보과학회 가을 학술발표논문집, 2002.

[3] Pedersen, Ted and Banerjee, Satanjeev, and Patwardhan, Siddharth "Maximizing Semantic Relatedness to Perform Word Sense Disambiguation" In: University of Minnesota Supercomputing Institute Research Report UMSI 2005/25 March, 2005.

[4] Corley, Courtney and Mihalcea, Rada "Measuring the Semantic Similarity of Texts" In: Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, June, 2005.

[5] Mustapha Baziz , Mohand Boughanem , Nathalie Aussenac-Gilles , Claude Chrisment, Semantic cores for representing documents in IR, Proceedings of the 2005 ACM symposium on Applied computing, 2005.

[6] R. Rada, H. Mili, E. Bicknell, M. Blettner, Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man and Cybernetics 19 (1) 17-30, 1989.

[7] C. Leacock, M. Chodorow, Combining local context and WordNet similarity for word sense identification, in: C. Fellbaum (Ed.), WordNet: An electronic lexical database, MIT Press, pp. 265-283, 1998.

[8] J. Jiang, D. Conrath, Semantic similarity based on corpus statistics and lexical taxonomy, in: Proceedings on International Conference on Research in Computational Linguistics, Taiwan, pp. 19-33, 1997.

[9] 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, 1997.

[10] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: Proceedings of the 14th International Joint Conference onArtificial Intelligence, Montreal, pp. 448-453, 1995.

[11] Wu, Z., Palmer, "Verb semantics and lexical selection", 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico State University, LasCruces, New Mexico, 1994.

[12] Hirst, G. and D. St.Onge. "Lexical Chains as Represenations of Context for the Detection and Correction of Malapropisms. WordNet". C. Fellbaum. Cambridge, MA, The Mit Press, 1995.

[13] 최호섭, 임지희, 배영준, 최수일, 옥철영, "온톨로지 구축 방법과 사례", 정보과학회지, 제24권, 제4호, pp. 31~44, 2006.

[14] 최호섭, 임지희, 옥철영, "대규모 지능형 한국어 어휘망 구축-우리말 어휘지능망(U-WIN)을 중심으로-", 609돌 세종날 기념 한글 학회 전국 국어학 학술대회 발표자료집, 2006.

4) 괄호 안의 숫자는 클러스터링된 명사들과의 공기빈도이다.