유로워드넷 방식에 기반한 한국어와 영어의 명사 상하위어 정렬

김동성*†
한국외국어대학교

Dong-Sung Kim. 2008. Alignment of Hypernym-Hyponym Noun Pairs between Korean and English, Based on the EuroWordNet Approach. *Language and Information 12.1*, 27–65. This paper presents a set of methodologies for aligning hypernym-hyponym noun pairs between Korean and English, based on the EuroWordNet approach. Following the methods conducted in EuroWordNet, our approach makes extensive use of WordNet in four steps of the building process: 1) Monolingual dictionaries have been used to extract proper hypernym-hyponym noun pairs, 2) bilingual dictionary has converted the extracted pairs, 3) WordNet has been used as a backbone of alignment criteria, and 4) WordNet has been used to select the most similar pair among the candidates. The importance of this study lies not only on enriching semantic links between two languages, but also on integrating lexical resources based on a language specific and dependent structure. Our approaches are aimed at building an accurate and detailed lexical resource with proper measures rather than at fast development of generic one using NLP technique. (**Hankuk University of Foreign Studies**)

Key words: 언어간 어휘 정렬 (Interlingual Lexical Alignment), 워드넷 (WordNet), 상하위어 (Hypernym-Hyponym), 의미적 유사성 (Semantic Similarity), 다의어 (Polysemy), 관계적 유사성 (Relational Similarity), 분류적 유사성 (Taxonomic Similarity), MDL (Minimum Description Length), 유로워드넷 (EuroWordNet), 계층적 유사도 (Hierarchical Affinity)

* 이 연구에 참여한 연구자는 '2단계 BK21 사업'의 지원비를 받았음. 논문이 완성되기까지는 여러 분들의 도움이 있어서 감사의 말을 전하고 싶다. 먼저 본 논문은 세 분 심사자들의 엄밀하고 건설적인 조언을 통해서 완성될 수 있었다. 또한 대학원생 방찬성, 이선경의 도움으로 자료를 정리하고 보완하였다. 사전 정의문 자료의 수집 및 가공작업에는 BK21과 연관된 대학원생들이 참여하였다. 그 외에 논문에 사용된 프로그램 작업이나 결과분석, 기타 자료의 수집 및 가공, 논문의 작성은 저자가 직접 수행하였다. 이 연구의 초기 내용은 18th International Congress of Linguists(July 24, Korea University, Seoul)에서 "Alignment of Hypernyms between Two Different Languages, Using Semantic Relations"의 제목으로 Kim, Dong-Sung & Sun Kyoung Lee가 발표할 예정이다. 학회의 발표 선정이 논문 제출보다 더 이르게 이루어졌지만, 발표 날짜가 논문의 출판 일정보다 더 늦게 되었다. 발표할 내용은 논문의 초기 구성에 따른 것으로 논문에서 제시한 연구 내용은 발표할 내용보다 더 확장되었다.

† 경기도 용인시 처인구 모현면 왕산리 산89 한국외국어대학교 언어인지학과. Email: dsk202@hufs.ac.kr

1. 서론

본 연구에서는 유로워드넷(EuroWordNet)의 방식에 기초해서 한국어와 영어의 명사 상하위어들을 정렬하였다. 전반적으로 언어간 정렬을 위한 작업 방식은 다국어 의미 체계를 정렬한 유로워드넷의 방법론에 기초하였다(Vossen, 1999; Farreres, Rigau, and Rodriguez, 1998). 이 연구를 통해서 한국어와 영어의 명사 상하위어 정렬에 있어서의 문제점과 어떠한 세부 연구 방법론이 필요한지를 살펴보았다.

상하위어 관계를 다루는 것은 단순한 어휘관계만을 고려하는 것이 아니고, 개념 관계도 고려하게 된다. Miller (1998)에 따르면 상하위어는 어휘화된 개념으로 의미 체계에서 계층관계를 나타내는 단위이다. 또한 Cruse (2002)에 따르면 상하위어 관계를 설정할 수 있는 것은 단순한 어휘적 관계에 기인하기 보다는 개념 관계의 사상(mapping)에 기인한다고 한다.[1] 따라서 언어간 상하위어 정렬의 문제는 언어간 개념 체계의 정렬과도 연관되어 있다.

여러 언어 자원들에 대한 연구 중에서 유로워드넷은 언어간 정렬을 위해서 영어로 작성된 워드넷(WordNet)을 중계 언어 인덱스(interlingual index)로 활용하였다(김현권, 2000; Vossen, 1999). 유로워드넷에서는 워드넷을 활용하여 다음과 같이 방법론을 구축하여 언어간 정렬을 하였다(Vossen, 1999; Farreres, Rigau, and Rodriguez, 1998). 각 개별 언어의 언어 자원을 개별 언어사전에 근거해서 별도로 정립하고, 이중 언어사전(bilingual dictionary)으로 번역하고, 이를 워드넷의 관계성에 비추어서 정렬하였다. 이에 따라 본 연구도 한국어의 상하위어를 추출하고, 한영사전을 활용해서 번역하고, 워드넷의 관계성에 비추어서 한국어와 영어간에 정렬을 시도하였다.

기존의 여러 연구들이 코퍼스나 기타의 언어자원을 활용한 어휘자원의 자동구축이나(Hearst, 1998; 김민수 · 김태연 · 노봉남, 1995; 신명근, 2006; 김혜경 · 윤애선, 2006; No, 2007; 이경순 외, 2001) 언어간의 정렬에(Yap and Baldwin, 2007; Paik, Bond, and Shirai, 2001; Wang, Ali, and Appaneravanda, 2005; Yeh et al., 2004; Edmonds and Hirst, 2002; 이지선, 2001; 문유진, 1997) 대해서 초점을 맞추었다. 여러 연구들에서 공통으로 목표한 것은 데이터를 분류하고 구축하는 데 많은 노력과 시간이 걸리는 이른바 "지식습득의 병목현상(knowledge acquisition bottleneck)"의 문제를 해결하기 위한 방식으로 더 많은 데이터를 더 정확하게 기계적으로 분석하기 위한 방법론들이다. 그러나 이 연구에서는 이러한 "지식습득의 병목현상"을 해결하기 위한 연구에 집중하지 않고, 언어간에 더 정확하고 더 세밀한 어휘관계를 정렬하기 위한 방법론을 제시하고자 한다.

논문의 구성은 다음과 같다. 2절에서는 기존 연구를 살펴보고 문제점을 지적한다.

---

[1] 언어간 개념간의 정렬이 가능한 인지적 근거는 Turney (2006)에서 논의한 바와 같이 관계적 유사성(relational similarity)을 통한 분류적 유사성(taxonomic similarity)에 있다. 3.3절에서 자세히 논의하게 될 것이다.

3 절에서는 유로워드넷의 방식에 따라 한국어, 영어의 상하위어를 각각 추출하여, 추출된 어휘를 검증하며 정렬하는 절차를 설명한다. 4 절에서는 의미관계에 대한 분석을 통해서 어휘간의 정렬에 대한 논의점을 살펴본다. 5 절은 본 논문의 결론과 문제점 및 향후 목표에 대한 토의이다. 그리고, 논문의 부록부분에서 연구의 결과로 얻어진 정렬된 언어간 상하위어를 제시한다.

2. 기존 연구

본 연구와 연관된 기존 연구는 크게 개별 언어의 상하위어 추출과 관련된 연구, 언어간 정렬과 연관된 연구로 나뉜다. 언어간 정렬은 다시 사전적 어휘 정렬을 위한 연구와 온톨로지 체계를 정렬한 연구로 구분된다. 개별 연구별로 각각 살펴보도록 하자.

2.1 상하위어 추출

수동으로 작업자들의 직관에 의한 상하위어 추출은 많은 노력과 시간이 할애되므로, 기존 연구에서는 코퍼스나 사전의 정의문을 활용한 자동 상하위어 추출이 많이 연구되었다. 이 중에서 Hearst (1998)는 코퍼스를 활용한 어휘 통사적 패턴에 의한 상하위어 추출을 제안하였다. 이 주장에 따르면 명사의 상하위어는 코퍼스에서 일정한 패턴으로 나타나며, 이러한 패턴이 일치하는 문맥을 코퍼스에서 추출하면 명사의 상위하위어가 추출된다고 한다. 예를 들어서, 영어 어구 "NP1, NP2 and/or other NP3" 와 같은 패턴의 경우에 NP3는 NP1, NP2의 상위어가 된다. (1a)의 어구에 동일한 패턴을 활용하면 (1b)와 같은 상하위어 관계가 추출된다.

(1) a. NP {, NP}* {,} and/or other NP
temples, treasuries, and/or other important civic buildings

b. hyponym("temple", "civic building"),
hyponym("treasury", "civic building")

이러한 작업의 장점은 코퍼스를 활용해서 상하위어관계를 추출해내고, 이를 확장하면 어휘사전 자동 구축에 있다. 그러나 (1a)와 같은 패턴이 모두 다 상하위관계를 나타내는 데 활용되지는 않는다. Hearst (1998)의 연구에 따르면 코퍼스에서 (1a)의 패턴으로 등장하는 문맥 중 대략 52%에서만 상하위어 관계성이 발견되었다. 따라서 (1a)와 같은 패턴을 가진 문맥이 모두 상하위어를 추출하는데 활용될 수 없으므로, 실제 패턴을 적용한 방식은 상하위어 추출의 정확도가 낮은 문제점이 있다.[2]

또한, 이러한 상하위어 관계의 자동 추출에 연관된 연구들의 문제점은 추출한 상하위어가 정확하게 상하위어 관계인지에 대한 검증이 결여되어 있거나, 검증의 기준이

[2] 또는 일정한 패턴에 따라서 정확도는 높지만 적용의 범위인 재현율이 낮은 경우가 있을 수도 있다. 즉, 발견되는 패턴이 모두 상하위어 추출에 활용될 수 있으나, 그 수가 너무 적아서 추출의 의미가 없는 경우도 있다. Cederberg and Widdows (2003)에서는 이러한 정확도와 재현율을 높이는 연구를 수행하였다.

명확하지 않다는 것이다. 신명근 (2006)은 워드넷의 상하위어 관계를 활용해서 영어 텍스트에서 발견되는 어휘들 간에 계층적 개념관계를 자동으로 구축하는 방법론을 제시하였다. 김혜경 · 윤애선 (2006)에서도 동사의 어휘의망을 반자동으로 구축하고자, 정확한 규칙을 사용하여서 사전의 정의문에서 중심어를 추출는 방식을 제시하였다. 사전의 정의문을 활용한 이유는 사전의 정의문이 기존의 코퍼스에 비해서 정제된 형태로 쓰여 졌고, 정의문을 활용하는 것이 코퍼스를 활용하는 결과보다 더 정확한 추출 가능성을 제공하기 때문이다. 이 연구들은 Hearst (1998)에서의 패턴에 의한 자동 추출의 문제점을 보완하기 위해서 더 정교한 언어 자원인 워드넷이나 더 정밀한 형태의 규칙을 활용하였다. 그러나 추출된 상하위어가 정확한지에 대한 검증은 결여되어 있다.[3]

반면에 김민수 · 김태연 · 노봉남 (1995)에서는 사전의 정의문에서 발견되는 일정한 패턴을 활용해서 상위어를 추출하고, 추출된 상하위어의 정확도를 화자 직관을 활용하여 검증하였다. 이경순 외 (2001)에서는 기계 가독형 사전과 코퍼스를 활용해서 정확한 명사의 의미 관계를 추출하고, 화자 직관을 동원해서 검증하였다. 이러한 연구들은 화자 직관을 동원하여 어휘 관계를 검증하는 측면에서 앞에서 언급된 연구들과 차별성을 띤다. 그러나, 화자 직관을 동원한 상하위어의 판별이 체계적인 의미 관계를 규정한 하나의 틀 안에서 일관적으로 성립되었는지는 알 수 없다. 화자 직관은 개인별이나 집단별로 차이가 있기 때문에, 체계적으로 성립한 하나의 의미 체계라는 틀에서 검증되었는지는 담보될 수 없다. 그러므로 기존 연구에서 활용한 화자 직관을 통한 판별 작업이 전체 실험을 통해서 살펴볼 때, 일관적인 상하위어 관계성을 의미하는지를 알 수 없다.

반면에 No (2007)에서는 BNC에서 발견되는 등위 명사절의 의미체계를 워드넷을 활용해서 검증하고 있다. 즉 상하위어 패턴들을 활용해서 자동으로 BNC에서 상하위어를 추출하고, 이를 다시 일관적인 의미 체계를 지닌 워드넷과 비교하였다. 또한 No (2007)에서는 영어 화자의 화자직관을 활용해서도 검증하고 있다. 이러한 부분은 기존의 연구와 차별적이다.

본 연구에서는 화자 직관과 워드넷을 활용하였는데 이것은 다음과 같은 의미를 갖는다. 앞에서 살펴본 기존 연구들이 화자 직관을 통한 검증이라는 측면을 갖고 있다. 이러한 측면에서 살펴보면 화자 직관이 상하위어 추출 연구의 기초 요소라는 것을 설명한다. 또한, 본 연구는 유로워드넷의 구축 방식에 따라서 워드넷을 통하여 정렬하였는데 이것은 일관된 의미 체계를 통해서 화자 직관에만 기초한 추출이 실수나 오류를 보완하기 위함이다(Vossen, 1999). 워드넷이 사전의 개념을 참조하여서 화자 직관을

[3] 김혜경 · 윤애선 (2006)의 연구는 사전 정의문에서 중심어를 추출하여 동사의 어휘망을 구축하는데 목적이 있다. 추출된 전체 자료가 논문에 다 기술되어 있지는 않으나, 기술된 자료만을 검토할 때 의미를 분류하기 위한 중심어가 되기에 부족한 어휘도 발견된다. 예를 들어서 "출강하다"의 중심어로 "사신으로 가다"는 어구를 추출하였는데, 이러한 경우에 의미적 분류를 위한 어휘가 아니고, 의미를 설명하기 위한 어휘로 보인다. 일부 어휘들은 개념 분류를 위한 상하위어 관계성이 없고, 오히려 동의어로서의 성격을 지닌다. 신명근 (2006)에서는 제시된 자료가 너무 한정적이어서 정확성을 파악하기에는 부족하다. 이 두 연구들은 이어서 제시될 연구들과 다른 점은 추출 작업을 검증하는 부분에 대한 논의가 없다는 것이다.

적절히 활용하여 구축된 자원이라는 점은 크게 두 가지를 설명한다(Miller, 1998). 첫째는 연구의 틀을 화자 직관에 기초하였다는 것이며, 둘째는 화자 직관에 의해서 구축된 자료를 사전적 개념을 참조해서 일관성 있는 체계로 구축하였다는 것이다. 따라서 이러한 여러 기존 연구의 고찰에 근거하여 본 연구에서도 상하위어 추출에 있어서 화자 직관에 기초하고, 워드넷이라는 일관성 있는 체계를 활용하였다.

2.2 다국어 사전 정렬

다국어 사전을 정렬하는 것은 어휘와 어휘간의 정렬이므로, 대역어 선정의 문제로 귀결된다. 하나의 언어에서 하나의 어휘는 다른 언어로 번역되었을 때, 정확하게 하나의 어휘로 대응되지 않고 여러 개의 어휘로 대응된다. 따라서 어떠한 대역어를 선정할 것인가와 몇 개의 대역어를 적절하게 대응할 것인가가 다국어 사전을 정렬하는데 문제가 된다.

Yap and Baldwin (2007) 에서는 일본어와 영어를 정렬하면서 한 개의 일본어 어휘를 여러 개의 영어 대역어로 바꾸고, 워드넷을 활용해서 정렬하였다. 정렬할 때, 사전에 수록된 어휘들의 빈도에 기반을 둔 TF-IDF(Term Frequency-Inversed Document Frequency) 를 반영하였다.

Paik, Bond, and Shirai (2001) 는 한국어, 중국어, 일본어 어휘들을 정렬하면서, 한자 인덱스로 세 개의 언어를 정렬하였다. 이 연구는 한자가 하나의 주요한 인덱스로 작용할 수 있다는 근거를 제시하였고, 여러 개의 다국어 언어자원을 활용하였다는 측면에서 의의가 있다. 또한 말레이어 및 영어도 일본어를 중심으로 정렬하여 보았다. 이 연구에서 한자를 활용한 정렬의 경우에 각 언어간에 1:1 관계가 성립하지 않고 1:n 관계가 성립함이 발견되었다.

다국어 사전 정렬과 관련된 기존 연구는 동음이의어와 다의어를 모두 포함하여서 언어간 어휘의 관계성이 1:n, 1:1, n:m 과 같이 다양하다는 것을 보여준다. 언어간 정렬은 단의적 (monosemic) 이라기보다는 다의적 (polysemic) 관계에 기초하는데, 정렬에 있어서 다의어 (polysemy) 는 매우 중요하며 조심스럽게 다루어져야 한다. 다의어의 문제는 4 절에서 자세하게 논의될 것이다.

2.3 다국어 온톨로지 체계 정렬

다국어 온톨로지 체계를 정렬하는 것은 여러 방식이 있으나, 크게 두 가지 방식이 많이 활용된다. 하나는 여러 언어자원 중 워드넷을 정렬에 활용하는 것과 다른 하나는 유사도에 의한 측정을 활용한다는 것이다.

워드넷을 활용한 방식 중에서 널리 알려진 것은 유로워드넷으로 개별 언어로 작성된 워드넷을 워드넷에 기초해서 정렬하고 이를 하나의 온톨로지로 연결하는 방식이다. 정렬의 방식은 영어로 된 인덱스를 활용하여 여러 언어와 대응 관계를 만들고, 이 인덱스를 통해서 언어간 정렬을 하였다(Vossen, 1999). 이와 같은 인덱스는 언어간 중계

인덱스(Inter-Lingual Index)로써 워드넷을 기반으로 제작되었다.

워드넷은 다른 언어의 온톨로지 체계를 정렬하거나 구축된 체계의 정확성을 검증하기도 하였다. Khan and Hovy (1997)에서는 아랍어와 영어의 온톨로지 체계를 정렬하면서 영어의 워드넷을 활용하였다. 이 방식에서는 영어 워드넷에서 설정된 관계성을 통해서 아랍어가 영어와 정렬되었다. 또한 Wang, Ali, and Appaneravanda (2005)에서도 워드넷을 활용해서 구축된 상하위어 관계성을 검증하였다.

Yeh et al. (2004)에서는 중국어와 영어의 의학 도메인 온톨로지를 정렬하면서 동의어 집합 단위를 정렬의 단위로 하였다. Yeh et al. (2004)은 중국어와 영어간의 의학 온톨로지 체계를 정렬하기 위해서 두 언어에서 발견되는 동의어 집합들을 유사도에 따라서 정렬하였다. 또한 개념체계의 유사성도 연산하였는데, 워드넷에서 발견되는 유사도 측정에 기반을 두었다.

Edmonds and Hirst (2002)는 영어와 불어의 시소러스의 정렬방식을 설명하면서, 유사도의 측정이 어휘의 정렬이 아닌 개념의 정렬에도 활용될 수 있다고 주장하였다. 이 연구는 영어와 불어의 기계번역에서 정확한 번역을 위해서 정렬이 활용될 수 있음을 보였다.[4]

한국어와 연관되어서 이지선 (2001)은 한국어의 분류체계를 미리 간단히 설정하고, 새로운 개념이 미리 정의된 분류체계에 어떻게 분류되어져야 하는가를 연구하였다. 이 연구에서 새로운 개념은 영어 대역어로 전환하고 워드넷의 분류방식에 따라 분류되는데, 한국어의 분류체계는 워드넷의 분류방식과 연관되어 있어서, 워드넷으로 분류되기만 하면 다시 한국어 체계에 분류되도록 하였다. 이 연구의 차별성은 한국어의 체계와 워드넷의 체계를 미리 정렬하여 언어간 정렬에 분류체계를 활용하였다는 점에 있다. 또한 이전 소절에서 논의된 기존 연구의 문제인 검증의 문제는 분류체계의 정렬을 통해서 해결할 수 있었다. 그러나, 워드넷 분류체계를 활용하여 구축한 한국어의 분류체계가 얼마나 엄밀하고 정확한지에 대한 검증이 없다. 이것은 언어간에 분류체계가 차이점이 있을 수 있다는 점을 인정하지 않는 것으로, 한국어와 영어라는 언어간 분류체계의 차이점을 인정하지 않았다.

문유진 (1997)에서는 한국어 워드넷의 설계 및 구축을 제안하면서 한국어 사전을 통해서 기본 어휘 5,000개를 토대로 영어의 대역어를 추출하고, 대역어를 활용해서 워드넷 의미 체계와 유사한 한국어 워드넷 체계를 만들고자 하였다. 문유진 (1997)은 한국어의 의미체계가 영어의 의미체계와 서로 다를 수 있는데, 이러한 차이점을 고려하지 않았다. 또한 영어의 개념체계를 그대로 활용함으로써 한국어의 의미체계를 반영하지 않은 문제점이 있다. 또한 한국어에서 어의관계 추출에 대한 정확성에 대한

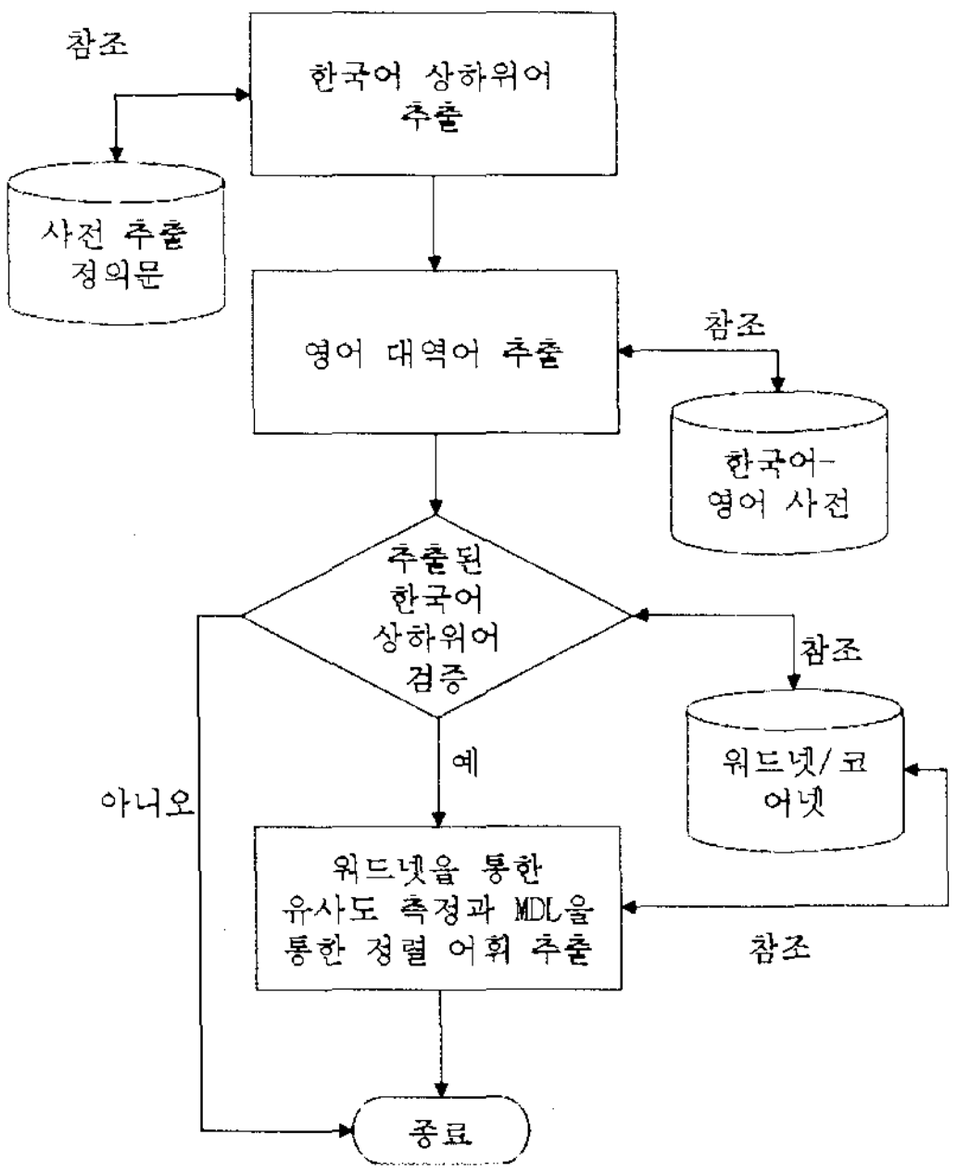
[4] Edmonds and Hirst (2002)는 세 가지 다른 체계의 어휘지식이 연결되어야 정렬이 가능하다고 주장하였는데, 각각은 개념-의미단계 (conceptual-semantic level), 하위개념/스타일-의미단계 (subconceptual/ stylistic-semantic level), 통사-의미단계 (syntactic-semantic level)이다. 언어간 정렬에 있어서 이러한 단계가 필요한지 또는 이러한 단계의 설정에 효용성이 있는지에 대한 검증은 본 논문에서는 다루지 않았지만, 의미 있는 언어학적 단위를 활용하였다는 측면에서 살펴볼만 하다.

검증도 결여되어 있다.

이상의 기존 연구들에서 살펴본 바와 같이 언어자원을 활용하면 정렬을 효과적으로 수행할 수 있는데, 여러 언어자원들 중 워드넷은 널리 활용된다. 그리고 워드넷에서 발견되는 유사도의 측정은 의미체계의 구축에 적절히 활용되는 것을 알 수 있다. 본 연구에서도 워드넷을 활용해서 한국어와 영어의 정렬에 대한 연구를 진행하였고, 유사도에 의한 측정을 통하여 정렬의 정확성을 향상시키려고 시도하였다.

3. 한국어와 영어 상하위어 정렬

정렬을 위해서 본 연구에서 실행한 과정은 크게 사전의 정의문에 기초해서 화자 직관을 활용한 한국어 상하위어 추출,[5] 한국어-영어 사전을 통한 대역어 추출, 워드넷을 활용한 한국어 상하위어 관계성 입증, 유사도 측정 방식과 Minimum Description Length(MDL)를 활용한 유사 단어 클러스터의 추출로 구성된다. 전반적인 과정을 그림으로 나타내면 [그림 1]과 같다.



[그림 1] 전체 작업 공정도

[5] 본 연구는 유로워드넷의 구축 방법에 따라서, 각 개별 언어의 상하위어를 개별 언어사전에 근거해서 별도로 정립하였는데, 영어의 경우에는 워드넷을 그대로 활용하였다.

3.1 한국어 상하위어 추출 및 대역어 검출

한국어에서 상하위어를 추출하기 위해서 기본 어휘를 활용하였다. 임지룡 (1991)과 김광해 (2003)에서 논의한 바와 같이 기본 어휘는 빈도와 연관이 있는데, 이러한 빈도는 한국어의 의미 분포와 연관된 어휘 사용과 관련이 있다.

본 연구는 코퍼스를 통해서 빈도순으로 의미 목록을 작성한 기존 연구를 참조하였다. 서상규 (2000)에서는 '연세말뭉치'를 대상으로 '연세한국어사전'의 의미 기술에 따라서 의미 태깅을 하고, 발견되는 어휘 목록과 통계 자료를 제시하였다. 목록에서는 표제어 단위가 아닌 어깨번호 단위 이하의 세세한 의미 분류 목록까지 제시되었다. 아래의 예는 서상규 (2000)에서 제시한 동사 "가다"의 1번째 표제어 중 I 번째 어깨번호 중에서 ①번째 의미의 통계 자료이다.

(2)

표제어_품사	빈도수	의미기호	의미빈도	의미사용률
가다_VV	4,516	1-Ⅰ-①	292,364	73%

서상규 (2000)에서 제시한 연구는 매우 세세한 의미적 분류에 따른 것으로 기본 어휘를 의미 빈도순으로 고려하기에 매우 적절하다. 그러나 본 연구에서는 연구의 객관성을 높이기 위해서 두 가지의 코퍼스와 두 가지 사전적 의미를 서로 비교하기로 하였다. 사전으로는 '연세한국어사전'과 '표준국어대사전'을 활용하였으며,[6] 코퍼스로는 서상규 (2000)에서 근거한 '연세말뭉치'와 세종계획 2단계 연구 결과 중 한국어 학습용 어휘 선정을 위한 기초 조사에 활용된 '현대 국어 사용 빈도 조사' 코퍼스 150만 어절(1998년에서 2002년, 이하 '세종 150만코퍼스')을 활용하였다. '세종 150만코퍼스'는 '표준국어대사전'에 의거하여 표제어 단위인 동음이의어로만 의미 구분된 태깅코퍼스이다. 코퍼스에 출현하는 동음이의어는 '01, 02, 03,.....'과 같이 '표준국어대사전'의 표제어 단위 분류 숫자에 따라서 구분되었다. (3)은 의미 구분된 '세종 150만코퍼스'의 예이고, [표 1]은 '세종 150만코퍼스'의 기본적인 통계자료이다.

(3) 1/SN + ./SF 아름답다/VA + 은/ETM 그01/MM 시작01/NNG "/SS + 이/NNP 형 01/NNG + ,/SP 잘02/MAG .....

'세종 150만코퍼스'를 통해서 동음이의어가 분류된 어휘 목록을 4,000개를 빈도순을 추출하였다.[7] 이 중에서 명사들 중 (2)와 같이 서상규 (2000)에서 제시한 목록과 일치하는 어휘 1,000여개를 빈도순으로 선별하였다. 선별된 어휘는 '표준국어대사전'의 표제어 단위에 기반을 둔 어휘들이다. 즉, '세종 150만코퍼스'에서 추출된 의미단위는 (2)에서 의미기호 1-Ⅰ-①에서 표제어 단위인 1에만 일치하는 어휘들의 목록이다. 따라서, 서상규 (2000)에서 조사한 세부 의미 빈도 및 의미사용률과 일치하는 목록은

---

[6] 연세한국어사전과 표준국어대사전은 웹사이트를 통해서 활용하였다. < 참고웹사이트> 참조.

[7] 이 빈도순에는 명사 1,860개, 동사 789개, 부사 232개, 형용사 201개로 내용어(content word)가 전체의 77.05% 정도 분포하였다.

전체 어절 수	1,522,137
총 형태소 개수	3,224,866
1 어절 당 형태소 개수	2.11
형태소 타입 (type) 개수	89,005
형태소 타입 (type) 당 토큰 (token) 개수	36.23
동음이의어 형태소 개수	285,277
전체 어절에서 동음이의어 형태소 비율	18.7%
전체 형태소에서 동음이의어 형태소 비율	8.8%
전체 형태소 빈도의 평균	36.23
전체 형태소 빈도의 표준편차	901.97

[표 1] '세종 150 만코퍼스'의 기본 통계자료

'세종 150 만코퍼스'를 활용하여서는 찾아낼 수 없다. 그러므로 표제어 단위로 일치하는 목록들 중에서 '연세한국어사전'과 '표준국어대사전'에 근거한 의미를 찾아내기 위해서는 두 사전의 정의문을 일일이 비교하여 서로 일치하는지를 검증할 필요성이 있다.

본 연구에서는 '연세말뭉치'와 '세종 150 만코퍼스'에서 표제어 단위로 일치되는 어휘를 대상으로 '연세한국어사전'과 '표준국어대사전'의 사전 정의문을 대비하여 서로 일치하는 의미로 기술된 정의문만을 고려하였다. 정의문에서 표제어의 상위어를 추출하기 위해서 사전적 정의를 활용하여 중심어휘를[8] 추출하고 해당 어휘가 제시한 기본 어휘의 상하위어인지를 화자 직관을 활용해서 판별하였다. 예를 들어서 "남자"의 경우에 (4)와 같이 '연세한국어사전'에 정의되어 있다면, 중심어휘를 "사람"으로 판별하고, "남자"의 상하위어로 판단될 경우에만 상위어로 추출하였다.

(4) 남자(男子)【명사】 남성인 사람.

국립국어원에서 발행한 '표준국어대사전'에서는 '연세한국어사전'의 (4)에 해당하는 의미를 아래 (5)와 같이 기술하고 있다.

(5) 남자 男子 「명」「1」남성(男性)으로 태어난 사람.

이 경우에 "남자"의 상위어로 "사람"을 추출하고, (4)와 (5)에 의거하여서 "남자→ 사람"[9]이라는 상하위어쌍을 추출해 냈다. 이러한 작업을 2 명의 연구자가 자료를 번갈아 가면서 확인하고 추출하였다. 작업에서 추출된 어휘는 상위어를 하나의 의미에

[8] 중심어휘는 정의에 있어서 가장 핵심이 되는 말로, 정의문에서 해당 어휘를 더 일반화하는 어휘로 많이 활용된다. 이러한 어휘는 상위어로 쓰일 가능성이 더 많다. 중심어휘에 대한 정의는 김민수 · 김태연 · 노봉남 (1995) 참조.

[9] "X→Y" 표현은 Y는 X의 상위어 관계를 의미한다.

하나씩만이 아니라, 상위어로 의심이 될 만한 모든 어휘가 해당되었다. 다시 말하면, 여러 개의 가능성을 염두에 두고 추출하였다. 많은 어휘가 상위어로 추출이 되면 정확한 어휘 판별에 많은 어려움이 있을 수 있으나, 다음 검증 단계에서 어떤 어휘가 적절한지를 검증할 수 있도록 하였다. 또한 정의문에서 명사로만 추출이 어려운 경우에는 동사나 형용사 같은 어휘도 상위어이라고 판단되면 모두 추출하였다.[10]

'연세한국어사전'과 '표준국어대사전'이 일치하는 목록은 전체 1,000개 명사 목록 중에서 상하위어쌍이 328개로 전체 어휘 수는 656개이다. 연구에서 실제로 고려된 목록은 이 328개의 상하위어들이다. 이 어휘 목록을 조사하면 빈도가 5이상인 어휘가 발견이 되는데, 여러 곳의 정의문에서 상위어로 추출된 경우가 있다. 이러한 어휘 목록은 (6)과 같다.

(6) 사람, 상태, 곳, 기관, 때, 조직

이러한 어휘는 의미체계의 계층구조에서 상위에 존재하는 어휘일 가능성이 높다. Wilks, Slator, and Guthrie (1996)에서는 Longman Dictionary of Contemporary English를 전자사전으로 구축하면서 중심어들의 계층적 구조를 만들었는데, 추출된 중심어들 사이에는 계층구조가 가능하다고 주장한다. 이 계층구조가 가능한 이유는 중심어로 추출된 어휘들 중 중복되는 어휘들을 중심으로 계층구조의 생성이 가능하기 때문이다. Wilks, Slator, and Guthrie (1996)는 중복되는 어휘들을 대표어(genus)라고도 하며, 중복어휘간에 빈도순과 깊은 관련성이 있다고 한다. 본 연구에서 최종적으로 추출된 328개의 어휘쌍을 통하여 중복어휘를 발견하였으나 이들 사이에 개념적 계층관계를 구성하기는 힘들었다. (6)의 어휘들은 대표어로서 추출된 상위어들로 빈도 5이상으로 출현한다.

김민수 · 김태연 · 노봉남 (1995)의 연구에서는 사전 정의문에서 일정한 패턴을 활용하여서 상위어를 자동적으로 추출하였다. 이 연구에서는 패턴을 활용한 자동화 방식을 활용하지 않고, 언어직관을 활용한 수동적인 방식을 활용해서 상위어를 추출하였다. 자동화된 추출이 모든 사전 정의문에 일률적으로 적용되지 않으므로 정교한 작업을 위해서 화자 직관을 활용하였다. Hearst (1998)는 (1)과 같은 어휘 통사규칙의 패턴을 활용하는 방식은 제한적이라고 언급하고 있다. 따라서, 기존의 자동추출 연구들의 방식이 추출 방식의 효율성과 효용성에 치중하였다면, 본 연구는 정확한 어휘의 선정에 초점을 두고 화자 직관을 활용하였다. 특히 기존 연구들이 상위어 추출의 검증을 화자 직관을 활용하는 측면을 고려하면 화자 직관을 활용하는 것이 가장 정확한 추출의 방식이 된다는 것을 의미한다. 그러나, 직관을 동원한 상위어 판별 작업도 일관적이거나 체계적이지 않아서 잘못된 오류를 많이 가지고 있었다. 이러한 직관의 오류는 다음

---

[10] 실제 정렬된 상하위어쌍에는 동사나 형용사는 고려하지 않았지만, 추출 작업에서는 모두를 고려하였다. <부록> 부분에 제시한 어휘들에는 명사-명사 상하위어쌍만이 목록화되었다.

소절에서 논의될 방식으로 영어 워드넷을 활용해서 검증하고 이를 다시 직관을 통해서 제거되었다.

추출 작업에서 두 개의 어휘가 한 어휘의 상위어로 두 사전에서 동일하게 추출된 경우도 33건이나 되었다. 이 경우에 특이한 점은 사람이나 직위를 동시에 가리키는 경우((7a)의 예)와 동의관계로 의심이 되는 경우((7b)의 예)가 발견된다는 것이다. 각각을 살펴보면 다음과 같다.

(7) a. 차장 → 사람
차장 → 직위

b. 대량 → 분량
대량 → 수량

위에서 (7a)에서 상위어로 추출된 "사람" 이나 "직위" 를 통해서 살펴보면 "차장" 이라는 어휘는 사람이나 직위를 동시에 의미한다. '연세한국어사전' 에 "차장" 에 대한 정의는 (8)과 같다.

(8) 차장1(次長)
【명사】보통, 과장보다 높고 부장보다 낮은 직위에 있는 사람, 또는 그 직위

이 경우와 같이 사람과 직위가 하나의 어깨번호 안에 같은 세부의미로 분류되어 있다. 이것은 두 개의 의미가 하나로 분류된 경우인데, 의미적으로 모호성을(vagueness) 띠거나 의미 분화가 하기 힘든 대상으로서 하나의 의미적 대상으로 취급된다(Cruse, 2000).

또한 (7b)에서 "분량" 이나 "수량" 은 동의적 관계에 있는 어휘인 것으로 판단되나, 정확하게 동의적 관계인지 아닌지에 대한 검증이 필요하다. 따라서 워드넷을 활용해서 이 어휘들의 관계가 동의어인지 아닌지에 대한 부분도 검증하였다.

직관을 이용하여 추출한 어휘들이 실제 상위어인지에 대한 검증도 필요하다. 본 작업에서 사전적 정의가 상위어를 체계적으로 포함한 경우에만 가능하나, 사전적 정의문은 정의를 위한 용도로 활용되었을 뿐 상위어를 염두에 두고 작성된 것은 아니다.[11] 따라서 사전 정의문에는 상위어가 아닌 다른 관계의 어휘를 포함되어 있고, 직관이 동원된 작업에서 이러한 어휘들이 상위어로 판별될 수도 있다. (9)의 경우는 직관을 활용한 작업에서 상하위어로 추출된 경우 중 일부이다.

(9) a. 남편 → 남자

[11] 사전 정의문을 활용해서 상위어를 추출하는 까닭은 사전 정의문이 여러 언어 자원 중 의미가 잘 정의된 언어 자원이기 때문이다. 따라서 기존의 여러 연구들도 사전 정의문에 기초해서 상위어 판별 작업을 수행하였다. 2.1 절 참조.

b. 손 → 팔

c. 국내 → 나라

(9a)에서 "남자"는 "남편"의 상위어로서 정확하게 "남편"을 포괄하는 상위 개념인지는 의심스럽다.[12] (9b)에서 보이는 것은 "손"은 "팔"의 부분어이지 상위어는 아닌데, 작업을 통해서 상위어로 잘못 추출된 경우이다. (9c)는 "나라"와 "국내"는 상하위어 관계가 아닌 다른 관계의 어휘들인데, 잘못 추출된 경우이다.

이러한 문제점에 대해서 상하위어 관계를 명확하게 검증할 필요성이 있다. 특히 온톨로지나 시소러스를 구축하면서 상하위어를 설정하는 것은 구축의 기본 골격이 되므로 상하위어의 정확한 검증은 필요하다. 본 논문에서는 이러한 문제점을 극복하기 위해서 언어자원으로 워드넷을 활용하고, 코어넷을 통해서 다시 살펴보았다. 자세한 설명은 다음 소절에서 논의될 것이다.

3.2 추출된 상하위어 검증

추출된 상하위어쌍이 올바른 상하위어인지 아닌지를 검증하기 위해서, 언어 자원을 워드넷과 코어넷을 활용해서 검증하였다. 워드넷을 활용하기 위해서 먼저 상하위어로 추출된 한국어 어휘 목록에 해당하는 어휘들을 영어의 대역어로 바꾸고, 대역어들의 관계가 상하위어 관계인지 아닌지를 워드넷에서 찾아보았다. 또한 코어넷을 활용해서 전반적으로 검증하였다.

대역어를 위한 사전으로는 동아프라임한영사전을 활용하였다.[13] 하나의 어휘에 대해서 영어 사전에 등록된 어휘들은 여러 개로 1:n의 관계가 성립한다. 가령 이전 소절의 작업을 통해서 "여자→사람" 이 상하위어 관계로 추출되었다면, "여자"와 "사람"에 해당하는 모든 어휘들을 (10)과 같이 사전을 활용해서 찾아낸다.

(10) a. 여자

woman, female, lady, girl

b. 사람

man, mankind, person, people, public, character, personality, adult

다음으로 영어 대역어들 사이에 상하위어 관계가 성립하는지를 워드넷을 활용해서 찾아본다. 다시 말하면, (10a)와 (10b)의 어휘들의 각각의 쌍 사이에 상하위어 관계가 성립하는지를 워드넷을 활용해서 살펴본다.

워드넷에서 상하위어는 계층적 구조로 정의되어 있는데, 계층적으로 (11a)와 (11b)에서 "female→person" 과 "woman→people" 의 상하위어의 유사도가 틀리다.

---

12 코어넷(전문용어언어공학연구센터, 2005)의 경우에는 "남편"을 "부부"나 "집단"의 하위어로 분류한다. 따라서, "남편"은 "남자"의 상위어가 되기 어렵다.

13 동아프라임영어사전은 웹 사이트를 활용하였다.

(11) a. **female** → **person** → organism → living thing → object → physical entity → entity

b. **woman** → class → **people** → group → abstraction → abstract entity → entity

"woman" 과 "people" 은 "woman→class→people" 의 관계를 통해서 상하위관계가 성립하지만, "female" 과 "person" 은 "female→person" 의 관계로 상하위어관계가 성립한다. 따라서 "female→person" 의 관계는 "woman→people" 보다 상하위관계성이 더 유사하다. 이러한 상하위어 유사성의 차이점을 수치적으로 측정할 함수로 아래의 방식을 활용하였다. 아래 (12) 에서 w 와 h 는 각각 개별 단어, 상위어를 의미한다.

(12)

$$\alpha(w,h) = \begin{cases} f(distance(w,h)) & \text{if} \quad w \subseteq h \\ 0 & \text{if} \quad w \not\subseteq h \end{cases}$$

(12) 의 방식은 Widdows (2004) 에 의해서 제안되었는데, 워드넷에서 단어간의 상하위어간의 계층적 거리를 비교, 측정하기 위해서 제안되었다. 여기서 상하위어 관계가 아닐 경우에는 상수 0 을 설정하였다. 함수 f 는 (13) 과 같이 정의된다. 수학적으로 두 지점 x, y 의 거리를 측정하는 경우에 거리가 서로 같다면 하나의 공간에서 같은 거리에 존재하는 점들 위에 있을 것이다. 기하학적으로 이러한 점들은 원으로 표현되고, 두 지점은 하나의 원 위에 존재는 점들 사이의 거리로 측정된다. 따라서 x 와 y 의 거리는 비례적으로 동일한 거리로 원 위에서의 거리의 제곱으로 측정될 수 있으며, $\frac{1}{(x-y)^2}$ 와 같이 표현된다. 이를 정리하여 함수 f 에 반영하면 아래와 같다.

(13)

$$f = \frac{1}{distance(w,h)^2}$$

이 공식을 활용하면 "female→person" 의 관계는 $\frac{1}{1^2}$, "woman→people" 의 관계는 $\frac{1}{2^2}$ 으로 측정된다. 이런 방식을 활용해서 많은 대역어들사이에서 발견되는 상하위어쌍의 상하위어 관계성을 수량적으로 측정하였다.

한국어에서 추출된 상하위어쌍 328 개 중 118 개는 워드넷에서 발견되지 않았다. 발견되지 않은 이유는 크게 두 가지로 생각될 수 있다. 첫 번째는 비록 직관을 활용한 작업에서 상하위어로 추출되었다 할지라도 직관에 의한 오류이므로 의미적으로 상하위어 관계가 아니어서 워드넷에서도 이러한 관계를 발견한 경우이다. 이런 경우에

추출된 상하위어쌍들은 상하위어 관계가 아니므로, 본 연구의 대상이 아니다. 두 번째는 상하위어의 목록이 한국어와 영어가 서로 다르므로, 영어에는 이러한 상하위어쌍이 발견되지 않지만 한국어에는 존재할 수 있다. 두 번째 경우는 워드넷를 통한 검증 작업이 무의미한 경우가 될 것이므로, 연구 방식을 처음부터 다시 고려해야 한다. 두 번째 경우를 고려해서 한국어 의미 체계를 구축한 한국어의 다른 언어 자원을 통해서 상하위어쌍을 검증하였다. 연구에서는 코어넷을 활용해서 해당 상하위어쌍이 코어넷에 존재하는지를 살펴보았다. (14)는 본 연구에서 선정된 상하위어쌍이 대역어를 워드넷을 활용해서 검증하였으나 상하위어 관계가 발견되지 않은 경우인데, 코어넷에서도 상하위어 관계성이 나타나지 않은 경우이다. 이러한 경우에 해당 상하위어는 직관을 통해서 잘못 추출된 경우이다.

(14) a. 국내→ 나라 (본 연구); 국내→ 조직 (코어넷)

b. 바람→ 공기 (본 연구); 바람→ 기상 (코어넷)

결과적으로 118개 중 13개의 상하위어만이 코어넷에 상하위어 관계로 등재되어 있었고, 나머지는 등재되어 있지 않았다. 다시 말하면, 13개의 상하위어들만이 워드넷을 통해서 잘못 예측될 수 있는 것들이고, 나머지 105개의 상하위어는 코어넷에서도 발견되지 않은 상하위어로 잘못 추출될 가능성이 높다. 또한 하나의 사전 정의문을 통해서 두 개 이상의 상위어가 추출된 경우도 있었는데, 33쌍이 해당된다. 이 경우에 워드넷을 통해서 살펴보면 둘 중 하나는 상하위어가 아닌 경우가 22쌍이고, 코어넷에도 발견되지 않은 상하위어이었다.

측정 방식 (13)을 활용하면 계층적으로 동일한 단계에서 발견되는 상하위어 관계를 모두 동일하게 측정한다. 예를 들어서, "계급→ 지위"의 경우에 (15)의 모든 대역어들의 유사도가 1인 매우 유사한 상하위어쌍이 발견된다.

(15) a. caste→position

b. caste→status

c. rank→position

d. rank→status

e. grade→rank

f. grade→status

여기서 어떠한 상하위어쌍이 "계급→ 지위"와 가장 적절한 상하위어쌍인지에 대한 판별이 문제가 될 수 있다. 즉, (15a-f)의 상하위어쌍 중 어떤 상하위어쌍이 가장 적합하게 "계급→ 지위" 정렬될 수 있는가는 연구의 대상이 된다.

3.3 한국어-영어 상하위어 정렬

언어간 어휘 관계의 정렬이 가능한 근거는 관계적 유사성에 바탕을 둔 분류적 유사성에 기인한다. Medin, Goldstone, and Gentner (1990)는 유사성을 속성에 의한 유사성 (attributional similarity)과 관계적 유사성으로 구분한다. 전자는 하나의 속성에 의해서 하나의 논항만을 취하는 형태로 정의된다. 예를 들어서 Large(X) 나 Large(Y)와 같이 Large 라는 속성을 X 나 Y 와 같은 논항에 부여하여서 유사성을 발견하는 경우이다. 반면에 후자를 활용하면 두 개의 논항 사이의 유사성을 발견할 수 있다. 예를 들어서 Larger(X, Y) 와 같이 X 와 Y 의 유사성은 "X is larger than Y" 와 같은 관계를 통해서 측정된다.[14]

Turney (2006)은 이러한 관계적 유사성을 통해서 두 개의 단어쌍의 의미 관계의 유사성을 측정하였다. 관계적 유사성은 "A:B::C:D" 로 표현될 수 있으며, "A:B" 의 관계가 "C:D" 의 관계와 유사한 것으로 이해될 수 있다. 예를 들어서 "car::road::boat:river" 와 같이 두 단어쌍의 유사성은 "car:road" 와 "boat:river" 의 사이에서 발견되는 관계의 유사성으로 측정된다. 여기서 두 단어간의 관계의 유사성은 각각 운송수단이라는 관계의 유사성으로 정의되는데, "car" 는 "road" 에서의 운송수단이고, "boat" 는 "river" 에서의 운송수단이 된다.

유사성에 근거한 의미의 분류는 유사한 의미들의 대상을 분류적으로 구분한다. 예를 들어서, {dog, cat, bird} 의 집합에서 {dog, cat} 과 {bird} 라는 두 개의 집합으로 구분할 수 있다. 여기서 "dog:cat" 의 관계는 "dog:bird" 나 "cat:bird" 의 관계들 보다 더 유사하기 때문에 분류가 가능하다. 이러한 분류적 유사성은 유사성에 기반을 두고 대상을 분류할 수 있는 근거가 된다.

이러한 분류적 유사성에 근거해서 어휘 관계를 설정하고 언어간에 정렬하기 위해서는 두 언어간에 발견되는 관계성이 동일하여야 한다. 예를 들어서 "아이→사람::child→person" 의 관계가 성립하기 위해서는, "아이→사람" 과 "child→person" 은 한국어와 영어 사이에서 관계성이 가장 유사한 경우이어야 한다.

하나의 한국어 어휘에 해당하는 영어의 대역어는 여러 개이므로 하나의 한국어 상하위어쌍에 대해서 가능한 영어의 상하위어쌍은 여러 개 일 수 있다. 그러나, 두 언어 사이에 정렬이 가능한 것은 영어와 한국어의 상하위어쌍 사이에 유사성이 존재하고, 그 관계성이 동일하기 때문이다. 예를 들면, 한국어의 "계급→지위" 와 관련된 상하위어쌍이 영어인 경우에 (15)와 같이 여러 개일지라도 가장 유사한 상하위어쌍이 존재하고 그 상하위어쌍이 한국어와 영어에서 발견되는 동일한 관계성을 갖기 때문이다. 따라서, (15)의 여러 어휘쌍 중에서 가장 유사한 쌍을 선택하면 이 어휘쌍은 한국어의

---

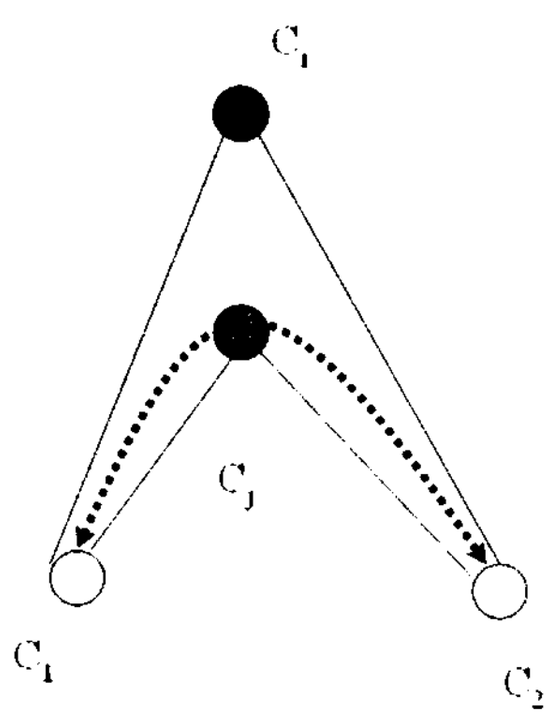
[14] Gentner (1983)에 따르면 속성에 의한 유사성이나 관계에 의한 유사성은 심리적으로 문맥적 판단에 근거하여서 정도의 차이로도 분류된다고 한다. 이러한 심리적 근거는 모든 유사성이 정도의 차이에 따라서 측정되어져야 할 것으로도 해석되며, 관계적 유사성에 근거한 언어간 유사성의 측정도 정도의 차이에 따라서 분류되어 할 것으로 풀이된다.

어휘쌍과 정렬될 수 있을 것이다.

여러 개의 어휘쌍이 관련성이 있을 때, 워드넷의 계층적 관계를 활용해서 여러 가지 형식으로 유사도를 측정하는 방식이 제안되었다. 특히 Pedersen, Patwardhan, and Michelizzi (2004)에서는 기존의 여러 워드넷을 활용한 유사도 측정 방식 중 10가지 방식을 프로그래밍 모듈로 구현하였다.[15] 이러한 측정 방식은 크게 세 가지로 구분될 수 있다. 첫 번째는 하나의 개념이 다른 개념과 얼마만큼의 거리를 갖는가를 측정하는 것이고, 두 번째는 두 개념이 얼마만큼 동질적인가 이다. 마지막 세 번째는 첫 번째와 두 번째를 결합한 것이다.

Pedersen, Patwardhan, and Michelizzi (2004)에서 제시한 모든 방식이 본 연구의 측정방식과 부합되지는 않고, 개념들간의 거리를 측정하는 방식만이 본 연구와 연관성이 있다. 따라서 본 연구에서는 워드넷에서 상하위어간의 거리를 측정하는데 활용 가능한 방식만을 선정하였다. 본 연구를 통해서 적용된 모듈은 wup(Wu and Palmer, 1994), res(Resnik, 1995), hso(Hirst and St-Onge, 1998), path, jcn(Jiang and Conrath, 1997)이다.

첫 번째 부류에 속하는 측정방식은 다음과 같다. 두 개의 개념 C_1과 C_2 사이에 개념적인 거리가 [그림 2]와 같이 계층적이라면, 두 거리 사이에 존재하는 여러 개의 노드 중에서 가장 적절한 노드는 거리가 가장 가까운 노드이다.



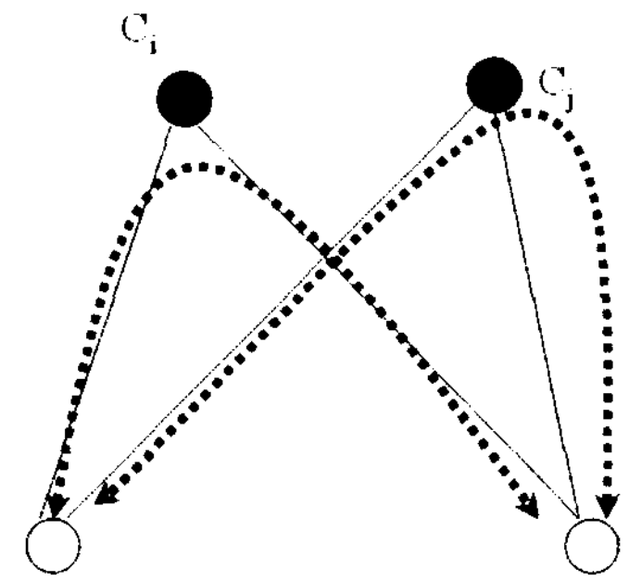
[그림 2] 개념간 거리에 바탕을 둔 측정

[그림 2]에서 C_1에서 C_2까지의 개념적 거리는 C_i나 C_j 간의 거리로 측정되는데, 적절한 거리는 노드간의 거리가 짧은 C_j의 거리를 통과하는 거리일 것이다. 만약 노드가 C_j를 통과하지 않고 C_i를 통과한다면 개념적으로 거리가 더 먼 거리를 통과하게

[15] 관련된 웹 사이트는 http://www.d.umn.edu/~tpederse/similarity.html이다. 프로그래밍 모듈은 객체지향을 활용한 펄(Perl) 프로그래밍 기법으로 작성되어 있으며, 설치하고 구동하면 유사도를 쉽게 측정할 수 있다. 본 연구에서도 해당 모듈을 활용한 펄 프로그램을 작성하여서 측정하였다.

되고, 개념적으로 유사성이 더 멀 것이다. 이러한 점에서 노드간의 거리를 측정하는 것은 각 개념간의 유사성을 측정하는 결과일 것이다. 이런 점을 활용하는 모듈은 wup, hso, path 인데, 이 모듈을 활용해서 두 어휘의 개념간 거리를 측정하였다. 그러나, 이러한 측정 방식은 동일한 개념을 동일하게 취급하므로 (12) 에 의한 측정법과 차이점이 없어서, (15) 에서 제기한 어휘쌍에서 차이점을 발견하지 못한다.

두 번째 부류에 속하는 측정법은 개념 간의 동질성을 개념간의 정보량을 통해서 측정하는 res 방식이다. 이 방식은 두 개념이 얼마나 다른 계층적 개념에 근거하는가를 측정한다. 아래 [그림 3] 에서 C_1 과 C_2 사이의 거리는 상위에 존재하는 노드가 C_i 인지 C_j 인지에 따라서 다르게 측정될 것이다.



[그림 3] 노드에 바탕을 둔 측정

이 방식은 C_1 에서 C_2 로 이어지는 노드들이 어떠한 노드인가를 측정하는데, 각각의 정보량을 측정한다. 따라서 C_i 와 C_j 의 정보량을 측정해서 이를 유사도 측정에 활용한다.[16] 이 방식은 Jiang and Conrath (1997) 에서 지적했듯이 해당 분류체계의 개별 정보량의 밀집도에 따라 값이 다르거나 같게 측정되는 문제점이 있다. 본 연구에서는 res 모듈을 활용해서 측정하였으나, 실제 측정값이 너무 크거나 너무 작은 값일 경우가 산출되어서 신뢰성이 그렇게 높지 않았다.

마지막으로 두 가지 방식을 결합한 jcn 방식을 활용하여서 상하위어간의 유사도를 측정하였다. 두 가지의 장점을 이용해서 기본 골격은 노드간의 거리에 바탕을 두고 노드간의 정보량을 고려한다. 이 방식이 가장 유사성이 높은 상하위어쌍을 고려하기에 적절하였다. 이러한 측정으로 가장 유사성이 높은 어휘를 추출해서 한국어 어휘쌍과 서로 일치시켰다.

여러 언어 간의 정렬의 문제를 다룬 유로워드넷은 하나의 언어에서 다른 언어로

16 이 방식은 Resnik (1995) 에서 제안되었으며, 여기서 정보량은 해당 범주에 얼마만큼의 정보가 있는지에 대한 정보량을 로그방식으로 측정한다.

정렬할 때 1:1 관계성이 아닌 동의어 집합들로 정렬하였는데, 워드넷을 중계 언어 인덱스로 활용하였다. 다시 말하면, 유로워드넷은 어휘간의 관계성이 중계 언어 인덱스를 중심으로 하여 1:n의 관계로 구성되어 있다. 이것은 언어간 정렬의 문제에 시사하는 바가 매우 큰데, 실제 언어간에는 1:1이 아닌 n:n이나 1:n과 같은 관계성으로 정렬할 필요성이 있다. 예를 들어서 "맛→ 감각"의 상하위어 관계가 한국어에서 발견되는데 반해서 워드넷에서 발견되는 어휘쌍은 여러 개이다.

(16) a. aroma → sensation

b. flavor → sensation

c. relish → sensation

d. savor → sensation

e. taste → sensation

f. taste → sense

g. gusto → feeling

h. relish → feeling

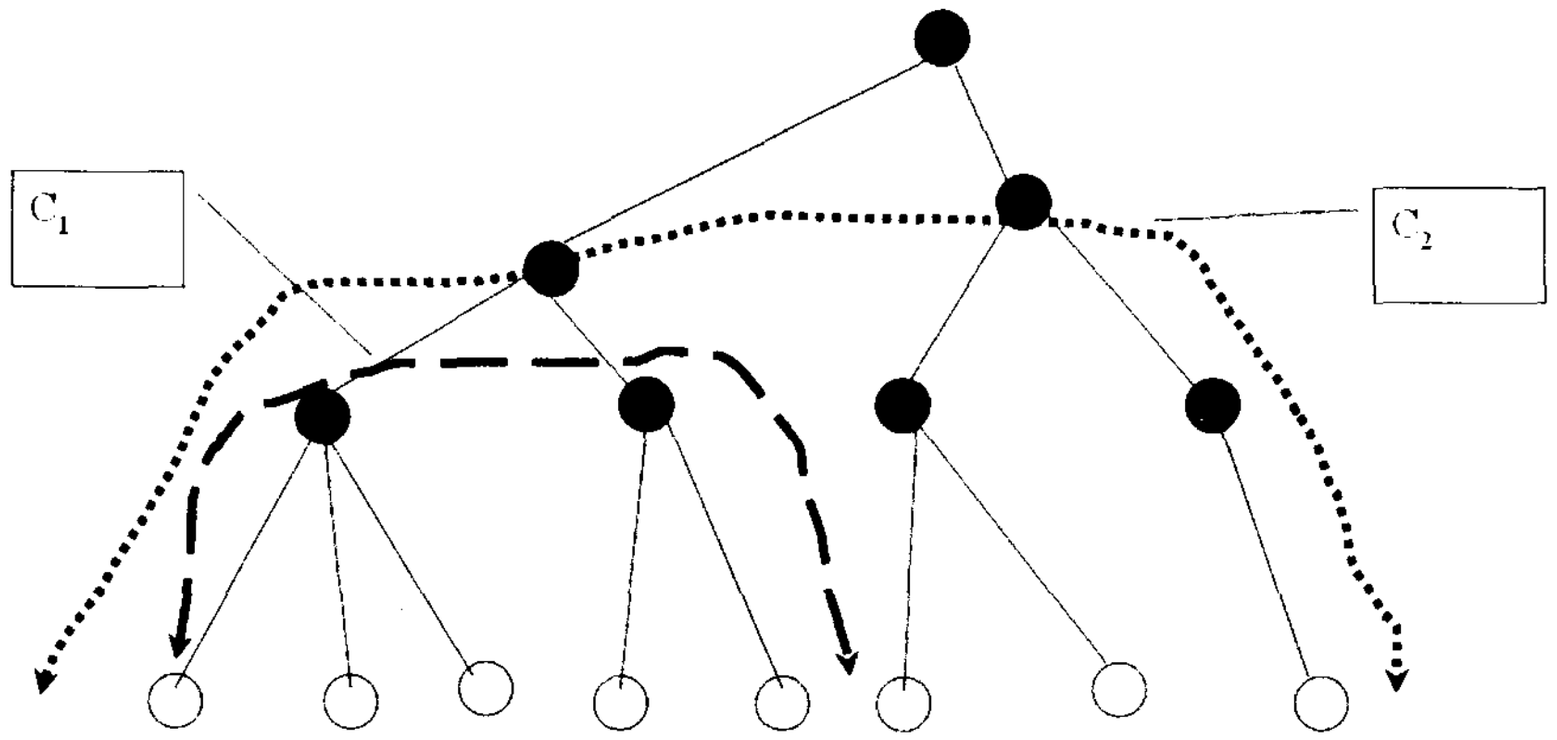
i. zest → feeling

j. taste → feeling

워드넷에서 발견되는 각각의 상위어를 중심으로 볼 때, 상하위어는 {aroma, flavor, relish, savor, taste, **sensation**}, {taste, **sense**}, {gusto, relish, zest, **taste**}의 집합으로도 구분될 수도 있다. 각각의 상위어를 중심으로 구성된 집합들이 한국어 상하위어쌍인 "맛→ 감각"과 정렬될 수도 있을 것이다. 이러한 경우에 위에서 언급한 방식은 두 개의 단어를 중심으로 한 유사도 측정이므로 여러 개의 단어들의 집합인 경우에는 적용이 가능하지 않다. 따라서 다른 측정 방식을 도입하여야 한다.

본 연구에서는 계층적 고려에 바탕을 둔 MDL 방식으로 집합에서 발견되는 유사성을 측정하였다. MDL을 자세히 소개하면 아래와 같다. MDL은 계층적인 차이를 통한 단어의 클러스터를 발견하는 수학적인 모델이며 워드넷을 활용한 의미 분류 연구에서 워드넷의 계층적 차이를 활용한 의미분류 모델로 연구되어 왔다(Hang and Abe, 1998; Tomuro, 2000). 이 모델은 다음 [그림 4]에서 계층적인 모델을 통한 단어 클러스터를 구분한다.

[그림 4]에서 C_1에 의한 클러스터는 C_2에 의한 클러스터보다 더 유사도가 높다. 이러한 계층적 분류는 주어진 모델안에서 어떠한 클러스터가 가장 좋은 데이터인지를



[그림 4] MDL 계층적 분류

검사할 수 있다.[17] 전체 모델이 M이라면 모델 M에 대한 계층적 클러스터 는 인자 (parameter) 를 가질 경우에 다음의 공식으로 표현된다.

(17)

$$M = (\Gamma, \Theta)$$

여기서 하나의 모델 M과 모델이 가지는 최적의 길이인 클러스터는 MDL을 가리키는데, 전체 길이는 L(M,S) 이고 이 모델은 을 만족하게 된다. 이를 수식화 하면 다음과 같다.

(18)

$$L(M, S) = L(\Gamma) + L(\Theta|\Gamma) + L(S|\Gamma, \Theta)$$

여기서 모델이 가지는 길이는 최적의 데이터를 설명하기 위한 클러스터의 인자인 데이터의 길이와 일치하게 된다. 아래 (19) 에서 bird 와 insect 의 계층적 분류에 의해서 나타나는 단어들의 데이터 구조를 고려하자.

(19) a. **bird** ⇒

eagle, swallow, seagull

b. **insect** ⇒

bug, bee, spider

[17] 여기서 모델안의 계층적 클러스터는 트리 구조안에서 나타나는 클러스터이다. 이러한 클러스터를 표현하는 모델을 tree-cut 모델이라고 한다.

위의 (19)를 중심으로 올바른 데이터를 분류하는 것은 올바른 계층적인 분류를 공유하는 데이터들의 클러스터를 추출하는 것이다. MDL을 활용하면 계층적 클러스터를 설명할 수 있으며, 아래 (20)과 같은 데이터들의 클러스터를 MDL로 연산하여서 측정할 수 있다. (19)의 계층적 구조를 잘 반영한 클러스터는 (20a)이며, MDL로 연산된 값이 가장 높다.

(20) a. {bug, bee, **insect**}

b. {swallow, crow, eagle, **bird**}

c. {swallow, eagle, *bug, spider*}

이러한 MDL 연산을 워드넷의 계층적 분류에 적용하면 다음 같다. MDL을 활용하면 계층적 구조의 정보가 양적인 면과 통계적인 면으로 측정되는데, 두 가지 정보가 필요하다. 하나는 모델 자체에서 정보가 어떻게 구성되어 있는가에 대한 측정이고, 다른 하나는 주어진 데이터에서 정보가 어떻게 구성되어 있는가에 대한 측정이다. 첫 번째 정보는 MD(Model Description)이고 두 번째 정보는 DD(Data Description)라고 한다. 이러한 MM과 DD를 측정하기 위한 세부 구성 요소는 다음과 같다.

(21) a. 워드넷에서 하나의 계층인 c의 빈도인 f(c)

b. c의 확률인 p(c)

c. 계층의 크기인 |c|

d. 계층 c에 포함되는 구성요소의 개수인 n

e. 워드넷 계층의 전체 개수인 k

f. 워드넷의 전체 빈도인 |S|

이러한 세부 구성 요소를 모두 구해서 MDL을 연산하는데, MDL은 MD와 DD의 합으로 계산된다. (22)는 MD와 DD를 연산하는 공식이다.

(22) a. $MD = \frac{k}{2} \times \log_2 |S|$

c. $DD = -\sum_n f(c) \times \log_2 p(n) \qquad (p(n) = \frac{p(c)}{n})$

위의 (22)에서 제시한 계산법을 (18)의 전체 모델에 대입하면 전체 모델은 다음과 같이 연산된다.

(23) $L(M,S) = L(\Gamma) + L(\Theta|\Gamma) + L(S|\Gamma,\Theta) = \log|S| + MD + DD$

즉, 전체 모델은 전체 모델의 개수, 전체 모델의 정보량, 그리고 주어진 데이터의 정보량으로 계산된다. 이러한 연산은 (19)에서 단어 데이터의 구조를 바탕으로 해서 (20)의 클러스터들이 이질적인지 동질적인지를 판별하게 된다. 가령, {swallow, eagle, *bug, spider*}, {swallow, crow, eagle, **bird**}와 {bug, bee, **insect**}와 같은 클러스터가 있을 경우에 어떤 집합이 더 이질적이고 어떤 집합이 더 동질적인지를 판가름한다.

이러한 방식을 활용해서 상하위어 목록에 활용된 전체 어휘들 중에서 가장 적절한 어휘 클러스터를 찾아낼 수 있다. MDL을 활용한 계층적 데이터 분류 방식은 관련 어휘가 2개 이상인 경우에도 가능하다. 예를 들어서, 한국어의 상하위어 중 "시대→기간" 인 경우에 "time→period" 와 "age→period" 의 상하위어 목록이 워드넷에서 발견된다. 이런 경우에 jcn 모듈을 활용해서 측정할 수도 있고, MDL 방식을 활용해도 측정이 가능하다. 실제 측정의 경우에 아래 [표 2]와 같다. 이 경우에 "time→period" 가 가장 유사한 상하위어쌍이 된다.

영어 상하위어휘쌍	jcn	MDL
time→period	0.44	13.96
age→period	0.24	9.42

[표 2] MDL과 jcn 비교 값

또는 (17)에서 발견된 상하위어 집합과 같은 어휘군도 같은 방식으로 측정할 수 있다. 연구에서는 jcn과 MDL 방식을 모두 활용해서 측정하였다.

4. 결과분석 및 토의

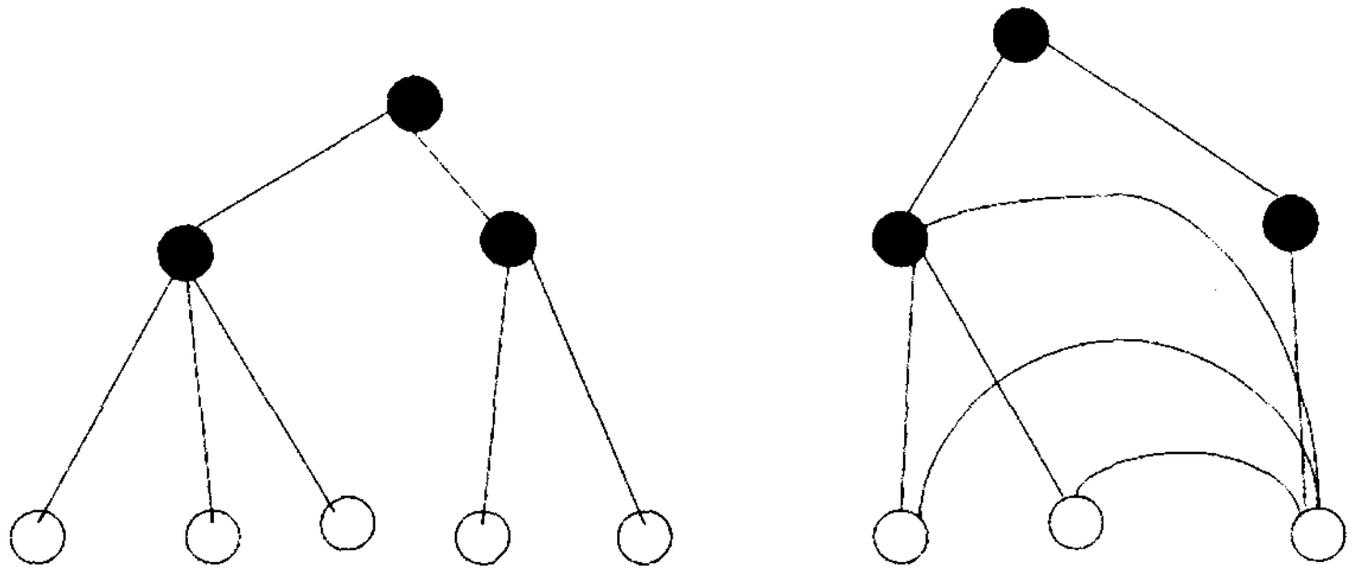
4.1 상하위어 검증 및 계층관계

본 연구에서는 유로워드넷의 작업 방법론에 따라서 한국어에서 상하위어를 추출하는데 있어서 기존의 언어 자원을 그대로 활용하지 않고, 새로이 언어 자원을 구성하였다. 즉, 기존의 언어 자원인 코어넷의 상하위어쌍을 영어의 상하위어쌍과 정렬하는데 직접적으로 사용하지 않고 사전의 정의문에서 직접 추출해서 새로운 상하위어쌍을 구성하였다. 이러한 작업 방식은 Vossen (1999)과 Farreres, Rigau, and Rodriguez (1998)의 제안에 따라서 개별 언어의 언어 자원을 별도로 구성한 것에 따른 것이다.

이러한 연구 방법론을 선택한 연유는 워드넷과 코어넷 체계사이의 차이점 때문이다. 코어넷은 일본 NTT 커뮤니케이션 과학기술연구소의 '어휘대계' 분류체계를 따르고 있고, 워드넷의 분류체계와 다르다. 따라서 워드넷과 코어넷의 상하위어 분류체계는 서로 다르다. 예를 들어서, "역사→ 기록" 의 상하위어쌍이 추출작업의 결과로 발견되었고, 이러한 쌍은 코어넷에서도 발견된다. 또한 워드넷에서도 가장 유사한 상하위어쌍들은 "history→record" 와 "chronicle→record" 가 발견되므로 코어넷에서 발견

된 상하위어쌍을 그대로 연결할 수도 있을 것이다. 그러나, 어휘관계성에서 보면 코어넷은 워드넷에서 발견되는 상하위어쌍과 다른 의미를 포함하고 있다. 코어넷의 경우에 "기록"은 {설화, 희곡, 시가, 소설}의 어휘군과 연관된 의미로 언어로 된 창작물의 분류에 속한다. 반면에 워드넷에서 발견되는 "record"는 {document}와 동의어로 "communication"의 분류에 속한다. 따라서, 코어넷에서 발견되는 상하위어쌍은 영어의 상하위어쌍과 다른 의미를 지니고 있다. 그러므로, 코어넷을 활용한 상하위어의 정렬은 워드넷의 의미 체계와 맞지 않다. Cruse (2002)에 따르면 상하위어 관계성은 개념 관계를 의미하기도 하므로, 단순히 한국어와 영어를 일대일 관계성으로 정렬하는 것은 올바르지 않으므로, 새로운 상하위어쌍을 추출해서 정렬하는 것이 올바른 정렬이 될 것이다.

코어넷과 워드넷의 어휘관계를 구조화는 각각 트리구조와 그래프구조로 차이가 난다. 워드넷은 의미 구조를 그래프구조로 표현하였다. 하나의 의미는 다른 의미에 관련되어서 연결되어 있으며, 어휘간에 관련된 모든 관계성이 표현되었다. 반면에 코어넷은 온톨로지 체계를 따르는 구조로서 트리 형태로 다른 의미들이 중첩되지 않게 연결되어 있다. 아래 [그림 5]는 트리 구조와 그래프 구조의 차이점을 그림으로 표현한 것이다.



[그림 5] 트리 구조와 그래프 구조의 차이점

이러한 트리 구조와 그래프 구조의 차이점을 나타내는 경우는 다음과 같은 경우가 있다. "경영→운영"이나 "경영→관리"는 사전적 정의문을 활용하면 상하위어 목록으로 추출될 수도 있다. 워드넷을 살펴보면 "경영→운영"이나 "경영→관리"의 상하위어 목록에 모두 다음과 같은 상하위어쌍이 발견된다.

(24) a. administration→management

b. management→administration

워드넷에는 "administration"과 "management"는 동의어 관계로 정의되지만,

"administration" 은 "management" 의 하위어로도 관계가 설정되어 있고 "management" 는 "administration" 의 하위어로도 관계가 설정되어 있다. 워드넷은 동의어집합으로 정의되기 때문에 각각의 다른 의미들이 상위어, 하위어로 설정된다. 이 경우에 워드넷을 바탕을 둔 상하위어 관계성에 따르면 "운영" 이나 "관리" 그리고 "경영" 은 동의적 관계로 정의된다. 그러나 코어넷을 조사해 보면, "경영→ 운영" 은 상하위어로 발견되지만, "경영→ 관리" 는 발견되지 않는다. 따라서, 워드넷을 활용한 상하위어 관계성과 코어넷에 정의된 상하위어 관계성은 서로 다르다. (24) 는 워드넷의 상하위어를 코어넷의 상하위어와 정렬하기 어려운 예의 하나이다.

영어 워드넷을 활용해서 한국어에서 추출된 상하위어쌍을 검증하면서, 발견이 가능한 것은 한국어 상하위어 관계에서 계층적 관계가 발견될 수 있다는 점이다. 사전 정의문을 활용해서 추출된 한국어 상하위어쌍 사이에는 계층적 관계성이 없었다. 그러나, 영어 대역어를 통한 워드넷 상에서는 계층적 관계가 발견되었다. 예를 들어서, "줄" 의 상위어로 "선" 과 "물건" 의 두 단어가 추출되었다. 따라서, "줄→ 선" 과 "줄→ 물건" 과 같은 두 개의 상하위어 관계가 성립한다. 이 두 상하위어 간에 이행적 (transitive) 관계성이 성립한다면 "줄→ 선→ 물건" 의 관계가 성립될 것이다. 그러나, "선→ 물건" 의 관계가 발견되지 않았기 때문에 실제적으로 이 세 단어 사이에 이행적 관계가 성립하는지는 알 수가 없었다. 워드넷을 활용해서 어휘 관계성을 살펴보면 이 어휘들 사이에는 이행적 관계성이 성립하므로, "줄→ 선→ 물건" 이 가능한 관계임을 알 수 있었다. "줄→ 선" 의 영어 대역어에 해당하는 상하위어들과 "줄→ 물건" 에 해당하는 대역어 상하위어를 서로 비교하면 다음과 같은 상하위어 관계를 발견할 수 있다.

(25) a. 줄→ 선
string→line

b. 줄→ 물건
line→object

따라서 관계성을 종합하면 "string→line→object" 의 이행적 관계성을 파악할 수 있다. 이것은 "물건" 이 가장 상위 개념에 해당하는 상위어이고, "선" 이 중간 상위어이며 "줄" 이 최하위어인 관계로 도출된다. 이와 같이 한국어와 영어의 상하위어를 비교해보면 한국어에서는 발견되지 않는 계층관계가 도출될 수 있다.

그러나 모든 어휘에서 이행적 관계 설정이 가능한 것은 아니다. 한국어에서 다음과 같은 상하위어쌍이 발견된다.

(26) a. 힘→ 능력

b. 능력→ 힘

c. 이성→ 능력

우선 "이성→능력→힘"과 같은 이행적 관계를 설정하는 것은 옳지 않고, "이성→능력"의 관계 설정은 "능력→힘"이나 "힘→능력"의 관계 설정과 관련성이 없다. 워드넷에 설정된 관계성과 비교해 보면 관련성이 없다는 것을 쉽게 알 수 있다. 워드넷을 조사해 보면, (26)에 대한 영어 상하위어 관계로 (27)과 같은 상하위어 목록이 발견된다.

(27) a. power→strength

b. strength→power

c. reason→power

워드넷에서 동의어 관계를 살펴보면, "power"와 "strength"는 동의어이며 "reason"의 상위어가 되는 "power"는 "strength"와 아무런 연관성이 없다. 워드넷 어휘목록의 "reason"의 상위어인 "power"는 "ability"의 동의어와 관련성이 있지, "strength"의 동의어는 아니다. 따라서 (26)와 같은 한국어의 상하위어쌍은 아무런 관련성이 없다.

워드넷은 영어로 작성된 것이어서 한국어의 체계와는 다르나, 언어 자원을 구축하는데 있어서 매우 유용하다. 특히 워드넷에 정의된 다양한 어휘관계를 적절히 반영하면 언어자원을 구축하는 데에 있어 많은 도움이 될 것이다. 언어 자원의 구축에는 많은 비용과 시간이 소요되므로, 여러 언어 자원의 다양한 측면을 활용하는 전략을 중요하다.[18]

4.2 다의어 관계와 언어간 정렬

언어간 정렬의 문제는 번역의 문제와 일치한다. 하나의 언어 A에서 다른 언어 B로 번역하는 것은 A언어의 표현을 B언어 표현으로 1:1 대응시키는 것이 아니라 B언어에서 가능한 표현(들)으로 대체하는 것을 말한다.[19] 언어간 어휘 정렬의 문제에 있어서도 이와 유사하게 언어간에 동일한 어휘들이 1:1 대응되는 것이 아니라 가능한 어휘(들)로 대응될 수 있다.

언어간에 어휘를 정렬을 한 유로워드넷의 경우에 언어간 정렬은 중계 인덱스를 활용하였는데, 중계 인덱스는 하나의 언어적 요소로 표현되며 언어간 중립적인 것으로 설정되었다(Vossen, 1999). 이러한 언어 중립적 중계 인덱스는 영어로 작성된 워드넷에 기반을 두었는데, 워드넷을 매개로 1:1 교환하는 방식을 취하였다. 언어간에 교환되는 것은 동의어의 집합(들)의 집합으로, 실제로 교환되는 것은 1:1 방식이 아닌 1:n방식이다. 예를 들어서, 영어 워드넷에 정의된 "office"의 여러 의미 중 {authority, bureau, agency, office}의 동의어 집합을 갖는 의미의 "office"는 네덜란드어, 스페인어, 이탈리아어의 다음 어휘들과 정렬된다(Vossen, 1999, 41).

---

[18] 여러 연구 중에서 Soica and Hearst (2004)에서는 워드넷을 활용한 상하위 계층을 구축하면서, 워드넷을 메타데이터 즉 상위계층의 골격으로 활용하였다.

[19] 번역과 관련된 문제는 Bassnett-McGuire (1980)를 참조.

영어 워드넷으로 작성된 중계 인덱스	네덜란드어	스페인어	이탈리아어
{authority, bureau, agency, office}	{dienst}, {kantoor, bureau, bureel, burelen,}, {bureau}, {agentuur}	{agencia, oficina}	{ispettorato}

[표 3] "office"를 중심으로 한 유로워드넷의 정렬

[표 3]에서 하나의 워드넷 동의어 집합에 대해서 네덜란드어는 두 개의 동의어 집합으로 정렬되었으나, 스페인어나 이탈리어는 하나의 동의어 집합으로 정렬되었다. 여기서 네덜란드어는 다의어 관계로 다른 언어와 정렬되어 있다. 언어간에 이러한 다의어 집합을 포착할 수 있는 방식을 활용해서 다의어 정렬이 가능해야 한다.

연구는 다의적 관계성을 포착하는 방식으로 MDL을 활용하였다.[20] 이러한 MDL은 다의적 관계를 구분해 내고, 계층적으로 관련된 어휘집합을 찾아내는 데 역할을 한다. 먼저 다의적 관계를 구분해 내는 MDL이 활용된 경우를 살펴보기 위해서 하나의 일례를 들어보자. 화자 직관을 활용하여 사전 정의문에서 상위어를 추출한 작업에서 "모양→생김새"의 상하위어가 추출되었다. 영어의 경우에 (27)과 같은 두 개의 다른 상하위어쌍이 발견되었다. 두 개의 쌍은 공식 (12)를 활용한 상하위어 유사도 측정에서 똑같이 높은 수치로 측정되므로 상하위어의 정도가 매우 긴밀하게 판단되는 경우이다. 따라서 두 개의 상하위어쌍 모두 정렬의 후보가 될 수 있다.

(28) a. figure→shape

b. figure→form

본 연구에서는 워드넷에서 발견된 모든 상하위어쌍을 jcn과 MDL을 활용해서 측정하였다. 이 경우에 대부분이 jcn과 MDL의 우세도가 일치하였다. 즉, 하나의 상하위어의 jcn 값이 다른 상하위어보다 더 크게 측정되면 MDL도 더 크게 측정되었다. 그런데, (28)의 경우에는 jcn을 측정한 결과 (28a)와 (28b)의 값이 같다. 하지만, MDL을

[20] Fellbaum (1998)에서 소개한 워드넷의 다의적 관계는 일반적으로 예측할 수 있는 다의어 (regular polysemy)와 예측하지 못하는 다의어 (irregular polysemy)로 구분한다. Tomuro (2000)에 따르면 MDL은 다의어 중에서 일반적으로 예측할 수 있는 정규적 다의어 (regular polysemy)를 추출하는 방식이다. 여기에서도 MDL이 일반적으로 예측할 수 없는 정규적 다의어 (regular polysemy)를 구분하는데 활용되었다.

측정한 결과 (28a)의 값이 (28b)의 값보다 더 크게 측정되었다. 이전 절에서 논의한 바와 같이 jcn은 개념간 거리와 정보량에 기반을 둔 유사도 측정방식이고, MDL은 계층적 분류에 기반을 둔 측정방식이다. 따라서 (28a)와 (28b) 상하위어쌍 사이에는 개념간 거리나 정보량은 유사하지만 계층적으로 서로 다른 형태인 다의어 관계를 나타내고 있는 결론에 도달할 수 있다.

더 자세한 고찰을 위해서 워드넷에서 "figure"를 중심으로 살펴본 결과, "form"과 "shape"은 동의어로 "figure"의 서로 다른 두 개의 의미를 정의하는데 활용된다. "shape"과 "form"은 모두 "form"의 두 번째 의미의 동의어 관계로서 사용되고 있으며((29a) 참조), 여섯 번째 의미 첫 번째 상위어로 정의되어 있다((29b) 참조).

(29) a. 동의어 관계

Sense 2

human body, physical body, material body, soma, build, figure, physique, anatomy, **shape**, bod, chassis, frame, **form**, flesh

b. 상위어 관계

Sense 6

⇒ **shape**, **form**

⇒ attribute

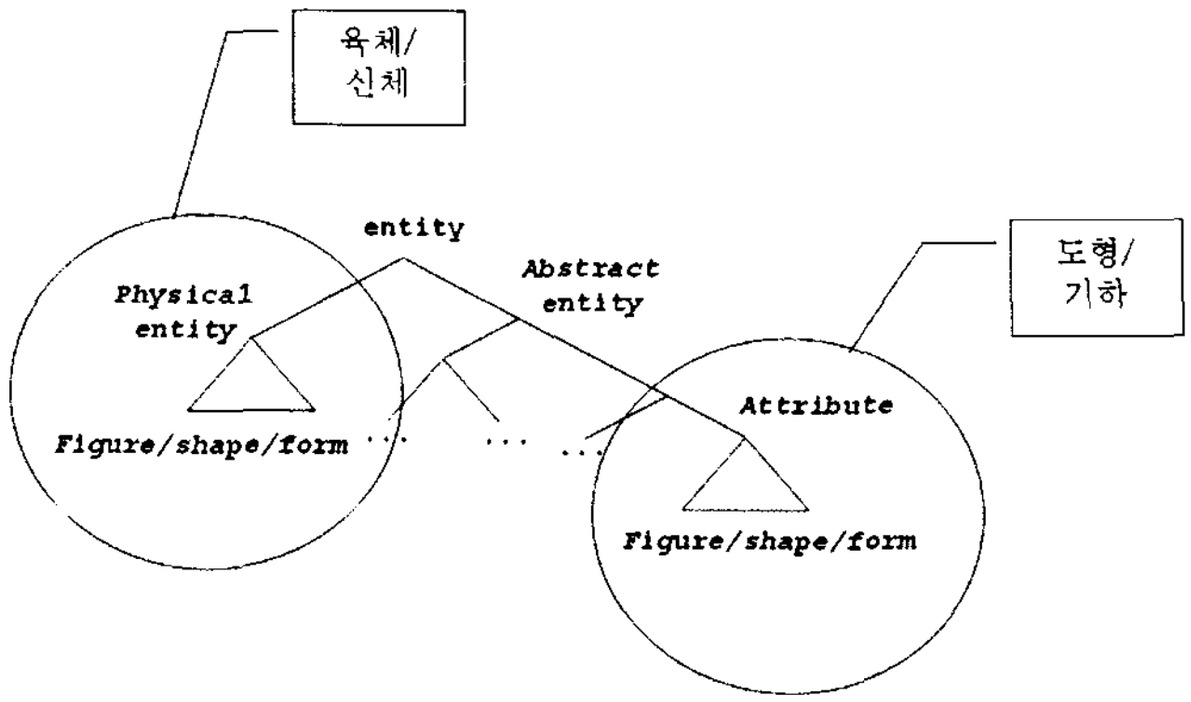
⇒ abstraction

⇒ abstract entity

⇒ entity

아래 [그림 6]을 살펴보면 "figure"의 의미는 서로 다른 "physical entity"와 "abstract entity"로 분류로 되며, 이와 같은 관계성으로 서로 다른 어휘들과 관계를 맺게 된다. 이러한 분류적 기준으로 인해서 워드넷에 "form"과 "shape"은 "figure"의 서로 다른 두 가지 의미의 동의어 집합에 나타난다. 하나는 신체와 연관된 의미이고 다른 하나는 기하학의 도형과 연관된 의미이다. 따라서 의미적으로 연관성이 없지는 않지만, 서로 다른 차원의 두 개의 의미가 얽혀 있다. 의미적으로 두 개의 다른 의미는 유사성이 있는 의미로서 동음이의(homonym)이라기 보다는 다의어(polysemy) 관계이다.

[그림 6]의 "figure"의 다의적 의미는 세 단어 관계성에서 복잡하게 얽혀있다. 워드넷을 자세히 조사해보면 "form"과 "shape"은 육체, 신체 의미의 "figure," 도형, 기하 의미의 "figure"와 각각 다른 관계성이 있다. 워드넷에서 "form"과 "shape"의 상위어를 조사해서 "figure"가 어떤 계층에 포함되어 있는지를 찾아보면 다른 관계성을 발견할 수 있다. "figure"는 "form"의 15번 째 의미의 두 번째 상위어로 정의되지만, "shape"의 경우에는 3번 째 의미의 동의어로 정의된다. 여기서, (30a)의 "form"은 육체, 신체 의미의 상위어 관계로 "figure"와 관계성이 있고, (30b)의 "shape"은



[그림 6] "figure" 의 다의적 의미 구성

기하, 도형 의미로 "figure" 와 관계성이 있다.

(30) a. form 의 15 번 째 의미 상위어
⇒ dummy
⇒ **figure**
⇒ model, simulation
⇒ representation
⇒ creation
⇒ artifact, artefact
⇒ whole, unit
⇒ object, physical object
⇒ physical entity
⇒ entity

b. shape 의 3 번 째 의미의 동의어 집합
human body, physical body, material body, soma, build, **figure**, physique, anatomy, shape, bod, chassis, frame, **form**, flesh

(29) 와 (30) 을 종합하면, "shape" 은 "figure" 의 다른 다의적 관계를 구분하지 않고, 같은 관계성으로 규정하고 있는 반면에 "form" 은 "figure" 의 다른 다의적 관계를 계층적 분류로 구분하고 있다. 유사도 계산에서 jcn 은 개념의 거리와 정보량만을 계산하므로 두 어휘쌍의 다른 관계성을 구분하지 못하나, MDL 은 데이터의 계층적 동질성

여부를 판별하여 연산한다. MDL이 높게 측정되면 같은 계층적 클러스터에 속할 확률이 높은 것이고, 낮게 측정된다면 다른 계층적 클러스터에 포함될 확률이 높다. 여기서 계층적 동질성이 "figure→shape" 보다 "figure→form" 이 더 크므로, "figure→shape"의 MDL이 "figure→form" 보다 더 작게 측정될 것이다. 실제 MDL을 측정해 보면, "figure→shape" 이 "figure→form" 보다 더 작게 측정되며, 계층적으로 동질한지가 수치적으로 나타난다.

MDL은 여러 동의어 집합간의 유사도 측정에도 활용될 수 있다. 예를 들어서 한국어에서 "여자→ 사람" 이라는 상하위어쌍이 추출되었는데, "여자" 와 "사람" 의 영어 대역어는 (31)과 같이 여러 개가 될 수 있다.

(31) a. 여자
woman, female, lady, girl

b. 사람
man, mankind , person, people, public

이 어휘들 사이에 여러 개의 집합이 가능한데, 상위어인 "사람" 에 해당하는 대역어를 중심으로 하위어 "여자" 에 해당하는 대역어(들)을 포함하는 집합은 여러 개일 수 있다.

(32) a. {lady, woman, girl, **person**}

b. {lady, girl, **person**}

c. {lady, girl, **man**}

d. {lady, girl, **people**}

e. .....

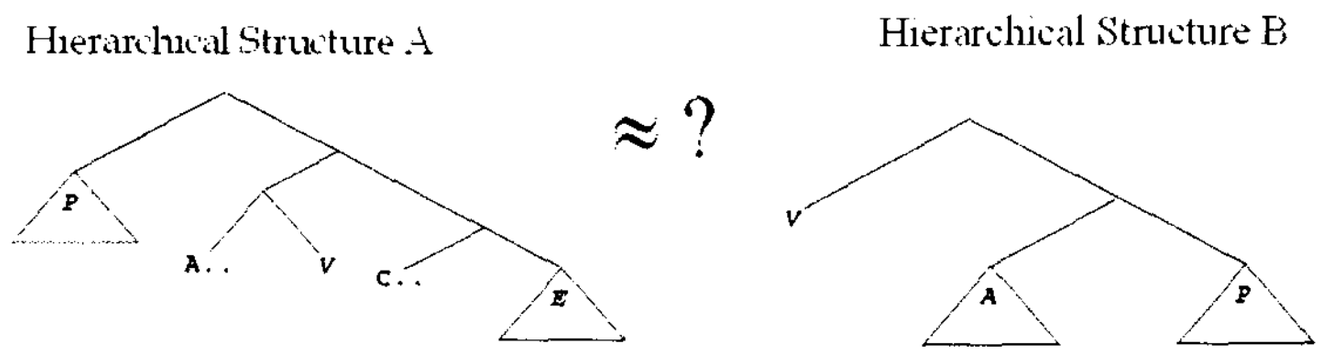
위의 (32)에 포함된 어휘들의 MDL을 측정하면 해당 어휘들이 얼마나 유사한 계층적 구조를 갖는지를 알 수 있다. 만약 계층적으로 유사하다면 동일한 부류에 속하는 집합일 것이다. 또한 두 개 이상의 집합의 선출이 가능하다. 서로 같은 MDL 값이 측정되거나 구분하기에 미미한 MDL 값이 측정되면 동질적인 집합으로도 구분될 것이다. 예를 들어서 (32b)와 (32d)의 MDL값이 서로 같다면 이 경우에 서로 같은 의미적 부류에 속하는 어휘 집합이 되고 (33)와 같이 "여자→ 사람" 으로 정렬될 것이다. 아래 (33)에서 굵은 글씨의 어휘는 해당 어휘 집합의 상위어이다.[21]

[21] <부록>에서는 MDL을 활용되어서 미미한 차이점이 발견될 경우에는 두 개 이상의 동의어집합을 제시하였다.

(33)

한국어	영어
{여자, **사람**}	{lady, girl, **person**}, {lady, girl, **people**}

언어간 정렬은 단순한 어휘쌍과 어휘쌍에 대한 정렬이 아니라 계층적 구조에서 나타나는 어휘쌍들 중 가장 유사한 쌍들로 정렬하는 것이다. Vossen (1999)에서도 논의한 바와 같이 언어간에 1:1 대응관계는 사실상 불가능하므로, 언어간에 가장 유사한 어휘들 간에 대응관계로 대신한다. 아래 [그림 7]과 같이 서로 다른 계층적 구조를 갖는 두 개의 구조를 고려해 보자.



[그림 7] 계층적 관계의 유사성

두 언어는 서로 다른 계층관계를 갖게 되므로, 언어간 정렬은 서로 다른 계층적 구조 간에 이루어진다. [그림 7]에서 A 계층구조와 B 계층구조간에 어휘 V, A, P를 정렬하는 것은 V, P, A가 포함된 계층적 구조를 비교해서 정렬하는 것이다. 따라서 서로 다른 계층구조에서 V, A, P의 구조를 서로 비교해 보고, 이것이 얼마나 유사한지를 비교하는 것이다.

이러한 정렬 작업에는 계층적 구조를 반영하는 유사도 비교가 활용된다. Atserias et al. (1997)에서는 유로워드넷에서 스페인어 워드넷과 영어 워드넷 구조를 정렬하면서 계층적 구조를 연구하였는데, 스페인어와 영어 두 언어간에 성립된 계층구조를 정렬하기 위해서 두 언어의 계층구조가 개념적으로 동질적인지를 비교하였다. 기본적인 연구의 틀은 Agirre and Rigau (1996)에서 차용되었는데, 계층구조의 동질성을 구조적 밀집도(density)에 따른 동질성의 정도를 측정하는데 바탕을 두었다. 구조의 틀의 밀집도는 구조에서 거리, 깊이에 따른 구성을 말하는데, 두 개의 개념이 유사하기 위해서는 동질적인 구조적 밀집성을 갖는다는 것을 의미한다. 두 개의 언어에서 발견되는 의미 계층은 동일한 원칙에 의해서 작성되었다고 할지라도 언어간 차이점에 의해서 동질적일 수 없다 따라서, 구조간의 정렬에 있어서 유사도의 측정에 의한 정렬은 필수적이다.

온톨로지 체계를 비교한 기존의 연구에서도 이와 같은 유사도의 고려가 발견된다. Wang, Ali, and Appaneravanda (2005)나 Yeh et al. (2004)는 온톨로지 체계를 정렬하였는데, 특히 Yeh et al. (2004)에서 선정한 온톨로지 체계는 의학 도메인이므로

어휘관계가 매우 한정적이어서 언어 전반을 다룬 워드넷이나 유로워드넷과 달리 의학 온톨로지는 언어간에 동질적인 면이 많다고도 할 수 있다. 그러나 언어간 계층구조는 차이가 있으므로 이러한 차이점을 고려하는 유사도가 고안되어서 정렬에 활용되었다. 본 연구에서도 유사도의 측정 방식을 활용하였다. 연구에서 활용된 유사도 측정 방식은 크게 두 가지인데, 하나는 워드넷에서 발견되는 계층간 구조와 구조적 정보량을 연산하는 jcn 이고, 다른 하나는 계층적 클러스터를 연산하는 MDL 방식이다.[22]

5. 결론

본 연구에서는 한국어와 영어의 상하위어를 정렬하였다. 먼저 한국어에서 코퍼스와 사전의 정의문을 활용해서 의미적 분포가 많은 상하위어를 추출하고, 이를 영어 대역어로 전환해서 워드넷의 상하위어 관계를 활용 검증하였다, 그리고, 검증된 어휘들에 대해서 유사도 측정을 통해서 한국어와 영어간에 상하위어를 정렬하였다.

연구를 통해서 화자 직관을 이용해서 상하위어를 추출하였는데, 이러한 측면은 정확하고 엄밀한 관계성을 파악하려는 노력의 일환이었다. 또한 유사도의 측정과 통계적 기재를 활용해서 워드넷의 계층 구조에서 적절한 어휘쌍과 어휘집합도 추출하였다. 이를 통해서 한국어와 영어의 정렬을 시도하였다.

본 연구는 유로워드넷의 방법론을 활용해서 언어간 정렬을 하였으며, 의미 계층이 서로 다른 언어간에 정렬을 통해서 적절한 방법론을 찾고자 하였다. 기존의 언어 자원인 코어넷과 워드넷을 정렬하는 것은 상하위어쌍의 정렬에 있어서 분류적 차이에 의한 오류 때문에 채택하지 않았다. 새로이 상하위어쌍을 추출하는 작업을 통해서 언어자원을 소규모로 구축하여, 방법론적 문제점을 연구하고자 하였다.

그러나, 본 연구는 한국어의 전체적인 의미 체계의 구축을 통한 전체 의미 체계와 워드넷 전체 체계와의 정렬이 아니어서 전체적인 언어 자원의 그림을 보기에는 부족한 면이 있다. 이러한 전체적인 그림을 그리기위해서는 많은 비용과 시간 및 인적자원이 소요되는 바, 전체 언어자원의 구축이라는 연구로서의 한계를 느낀다. 이런 점에도 불구하고 본 연구는 앞으로의 세계 언어자원의 통합이라는 커다란 목표에 방법론적 연구라는 면에 기여할 것이다.

< 참고문헌>

김광해. 2003. 국어교육용 어휘와 한국어교육용 어휘. *국어교육연구* 111, 255–293.

[22] 언어간 정렬의 방식에는 여러 가지가 있을 수 있다. 유사도 측정은 데이터 마이닝 기법과 연관된 것으로 수리적으로 유사성의 정도를 측정하는 것이다(Tan, Stenbach, and Kumar, 2006). 본 연구에서는 유사도 측정을 통하여 상하위어의 유사성을 수리적으로 측정하여서 가장 유사성이 높은 상하위어쌍을 선택하여 정렬하였다. Khan and Hovy (1997) 은 아랍어와 영어 워드넷을 정렬하면서 유사도의 측정 방식이 아닌 결정 트리 (decision tree) 방식을 활용하였다. 또한 Maedche and Staab (2001) 에서는 온톨로지 체계를 비교하는 방식을 연구하면서 기호학적 방식을 활용하였는데, 논리적 연산과 수리적 연산을 혼합하여서 활용하였다.

김민수 · 김태연 · 노봉남. 1995. 국어사전을 이용한 한국어 명사에 대한 상위어 자동 추출 및 WordNet의 프로토타입 개발. *한국정보처리학회논문지* 2.6, 847–856.

김현권. 2000. EuroWordNet의 구성원리와 설계. *언어학* 27, 145–177.

김혜경 · 윤애선. 2006. 동사 어휘의미망의 반자동 구축을 위한 사전정의문의 중심어 추출. *언어와정보* 10.1, 47–70.

문유진. 1997. 의미론적 어휘 개념에 기반한 명사 워드넷의 설계와 구축. *정보과학회논문지* 2.4, 437–445.

서상규. 2000. 한국어 교육 기초 어휘 의미 빈도 사전의 개발. 기술문서, 문화관광부. 2000년도 한국어 세계화 추진을 위한 기반 구축 사업 보고서.

신명근. 2006. Concept Hierarchy Creation Using Hypernym Relationship. *한국컴퓨터정보학회 논문지* 11.5, 115–125.

이경순 · 김도완 · 김길창 · 최기선. 2001. 기계가독형 사전과 코퍼스에서 추출한 의미정보를 이용한 명사열의 의미해석. *한국인지과학회 논문지* 12.1-2, 11–24.

이지선. 2001. WordNet을 이용한 한국어 명사 의미지표 자동 구축. 석사학위 논문, 전북대학교 대학원 전산통계학과.

임지룡. 1991. 국어의 기초어휘에 대한 연구. *국어교육연구* 23, 87–131.

전문용어언어공학연구센터. 2005. *어휘의미망 구축론*. KAIST Press.

Agirre, Eneko and German Rigau. 1996. Word Sense Disambiguation using Conceptual Density. In *Proceedings of the 16th Conference on Computational Linguistics*, pp. 16–22, Copenhagen, Denmark.

Atserias, Jordi, Salvador Climent, Xavier Farreres, German Rigau, and Horacio Rodriguez. 1997. Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pp. 143–149, Tzigov Chark, Bulgaria.

Bassnett-McGuire, Susan. 1980. *Translation Studies.* Methuen Young Books. 엄재호 옮김. 1999. 번역학개론. 인간사랑.

Cederberg, Scotts and Dominic Widdows. 2003. Using LSA and Noun Coordination Information to Improve the Precision and Recall of Automatic Hyponym Extraction. In *Proceeding of Association of Computational Linguistics*, pp. 111–118.

Cruse, Alan. 2000. Aspects of the Micro-Structure of Word Meanings. In Yael Ravin and Caludia Leacock (eds.), *Ploysemy: Theoretical and Computational Approaches.* Oxford University Press, pp. 30–51.

Cruse, Alan. 2002. Hyponymy and Its Varieties. In Rebecca Green, Carol Bean, and Myaeng Sung (eds.), *The Semantics of Relationships: An Interdisplinary Persepctives.* Kluwer Academic Publishers, pp. 3–21.

Edmonds, Philip and Graeme Hirst. 2002. Near-Synonymy and Lexical Choice. *Computational Linguistics* 28.2, 105–144.

Farreres, Xavier, German Rigau, and Horacio Rodriguez. 1998. Using WordNet for Building WordNets. In *Proceedings of the Association for Computational Linguistics*, pp. 65–72, Somerset, New Jersey.

Fellbaum, Christiane. 1998. A Semantic Network of English Verbs. In Christiane Fellbaum (ed.), *WordNet: An Electronic Lexical Database.* The MIT Press, pp. 69–194.

Gentner, Dedre. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7.2, 155–170.

Hang, Li and Naoki Abe. 1998. Generalizing Case Frames Using a Thesaurus and the MDL Principle. *Computational Linguistics* 24.2, 217–244.

Hearst, Marti. 1998. Automated Discovery of WordNet Relations. In Christiane Fellbaum (ed.), *WordNet:An Electronic Lexical Database.* The MIT Press, pp. 153–178.

Hirst, Graeme and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum (ed.), *WordNet:An Electronic Lexical Database.* The MIT Press, pp. 305–332.

Jiang, Jay and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the International Conference on Research in Computational Linguistics*, p. 1993, Tapei, Taiwan.

Khan, Latifur and Eduard Hovy. 1997. Improving the precision of lexicon-to-ontology alignment algorithms. In *Proceedings of the AMTA/SIG-IL First Workshop on Interlinguas*, pp. 53–58, San Diego, USA.

Maedche, Alexander and Steffen Staab. 2001. Comparing Ontologies- Similarity Measures and a Comparison Study. Measures and a Comparison Study. Technical report, Institute AIFB, University of Karlsruhe. Internal Report.

Medin, Douglas L., Robert L. Goldstone, and Dedre Gentner. 1990. Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science* 1.1, 64–69.

Miller, George. 1998. Nouns in WordNet. In Christiane Fellbaum (ed.), *WordNet:An Electronic Lexical Database.* The MIT Press, pp. 23–46.

No, Yongkyoon. 2007. Automatic Construction of a Concept Hierarchy from Coordinated Phrases. 언어와정보 11.1, 39–52.

Paik, Kyonghee, Francis Bond, and Satoshi Shirai. 2001. Using Multiple Pivots to align Korean and Japanese Lexical Resources. In *Workshop on Language Resources in Asia*, pp. 63–70, Tokyo, Japan.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity – Measuring the Relatedness of Concepts. In *the Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, pp. 1024–1025, San Jose, USA.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, San Mateo, CA. Morgan Kaufmann.

Soica, Emilia and Marti Hearst. 2004. Nearly-Automated Metadata Hierarchy Creation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 117–120, Boston, Massachusetts, USA.

Tan, Pang-Ning, Michael Stenbach, and Vipin Kumar. 2006. *Introduction to Data Mining.* Pearson Education, Inc.

Tomuro, Noriko. 2000. Automatic extraction of systematic polysemy using tree-cut. In *NAACL-ANLP 2000 Workshop on Syntactic and semantic complexity in natural language processing systems*, pp. 20–27, Seattle, USA.

Turney, Peter D. 2006. Similarity of semantic relations. *Computational Linguistics* 32.3, 379–416.

Vossen, Pick. 1999. *EuroWordNet General Document.* University Amsterdam. http://www.hum.uva.nl/ewn.

Wang, Zijun, Farha Ali, and Rashmy Appaneravanda. 2005. A Web Service for Efficient Ontology Comparison. In *2005 IEEE International Conference on Web Services*, pp. 843–844.

Widdows, Dominic. 2004. *Geometry and Meaning.* CSLI Publications.

Wilks, Yorik, Brian Slator, and Louise Guthrie. 1996. *Electric Words: Dictionaries, Computers, and Meanings.* MIT Press.

Wu, Zhirbia and Matha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138, Las Cruces, New Mexico.

Yap, Willy and Timothy Baldwin. 2007. Dictionary Alignment for Context-sensitive Word Glossing. In *Proceedings of the 2007 Australasian Language Technology Workshop*, pp. 125–133, Melbourne, Australia.

Yeh, Jui-Feng, Chung-Hsien Wu, Ming-Jun Chen, and Liang-Chih Yu. 2004. Automated alignment and extraction of bilingual ontology for cross-language domain-specific applications. In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 1140–1146.

< 참고 웹사이트>

연세한국어사전	http://dic.yonsei.ac.kr
표준국어대사전	http;//www.korea.go.kr/08_new/index.jsp
동아프라임한영사전	http://engdic.empas.com

< 부록>

아래 부록 목록은 다음과 같은 절차에 의해서 만들어진 목록이다.

A. 사전 정의문을 중심으로 상하위어쌍을 추출하였다.

B. 추출된 목록을 코어넷을 참조해서 잘못된 상하위어쌍을 걸러냈다.

C. 한영사전을 중심으로 워드넷에서 상하위어쌍을 살펴보고 이를 토대로 잘못된 상하위어쌍을 걸러냈다.

Hypernym Affinity는 (12)의 공식에 의해서 상하위어의 유사도가 가장 높은 목록을 선출한 것이다. jcn/MDL은 Hypernym Affinity의 여러 두 개의 어휘쌍들을 워드넷을 활용해서 측정 가능한 유사도인 jcn과 MDL로 측정하여 가장 유사도가 높은 어휘쌍 하나만을 나타냈다. MDL은 상위어를 중심으로 발견되는 여러 어휘의 MDL값 중 10이상인 높은 수치를 지닌 어휘 목록을 나타낸 것이다. 연구에 활용된 데이터를 전부 제시하는 것이 연구의 객관성을 담보하기 위해서는 필요하나, 지면의 관계상 연구에서 올바르게 추출되고 정렬된 것만으로, 즉 의미 있는 데이터로만 제한한다.

정의문 작업	Hypernym Affinity	jcn/MDL	MDL
때→ 시기	age→era, occasion→time, time→period	age→era	{age, occasion, time}, {occasion, time, season}
사회→ 집단	world→group, public→group	world→group	{world, society, group}, {community, society, public, group}
집→ 건물	house→building, house→edifice	house→building	{dwelling, house, cross, building}
여자→ 사람	female→person	female→person	{girl, woman, female, person}
정도→ 분량	grade→measure, rate→measure	grade→measure	{grade, rate, measure}
아이→ 사람	baby→person, child→person	child→person	{boy, son, infant, person},{kid, person}
눈→ 기관	eye→organ	eye→organ	{eye, organ}
시대→ 기간	age→period, era→period, time→period	era→period	{era, age, period},{age, period}
친구→ 사람	fellow→man, friend→person	friend→person	{fellow, comrades, company, person}, {pal, company, man}, {company, chum, friend, person}
관계→ 관련	connection→relation, relationship→relation	relationship →relation	{relationship, matter, connection, relation}, {matter, connection, relation}
남자→ 사람	fellow→man, guy→man, man→person	fellow→man	{fellow, guy, man}
물→ 액체	liquid→fluid, water→liquid	water→liquid	{liquid, fluid}, {water, liquid}
역사→ 기록	chronicle→record, history→record	chronicle→record	{annals, chronicle, record}, {history, chronicle, record}
나라→ 세상	country→people, land→people, nation→people, nationality→people, world→people	nation→people	{world, land, people}, {nation, nationality, people}, {nation, country, people}
대학→ 기관	university→institution	university →institution	{university, institution}
밥→ 음식	feed→food	feed→food	{feed, meal, food}

술→ 음료	alcohol→beverage, alcohol→drink, intoxicant→drink, intoxicant→beverage, liquor→beverage, liquor→drink, spirits→beverage, wine→drink	alcohol→drink	{wine, liquor, intoxicant, drink}, {alcohol, wine, spirits, beverage}, {intoxicant, liquor, drink}, {alcohol, wine, drink}
맛→ 감각	aroma → sensation, flavor → sensation, relish → sensation, savor → sensation, taste → sensation, taste → sense, gusto → feeling, relish → feeling, zest → feeling, taste → feeling	taste→sensation	{zest, taste, sensation}, {taste, gusto, flavor, sensation}, {savor, taste, zest, sensation}
아내→ 여자	wife→woman	wife→woman	{wife, female, woman}
작업→ 일	work→business	work→business	{work, business}
도시→ 지역	city→territory	city→territory	{city, territory}
꽃→ 식물	flower→plant	flower→plant	{flower, plant}
교사→ 사람	master→person	master→person	{schoolmaster, instructor, person}, {master, preceptor, person}
산→ 땅	height→place, peak→place	height→place	{height, peak, place}
의원→ 사람	assemblyman→person	assemblyman →person	{assemblyman, person}
선생→ 사람	master→person	master→person	{master, person}, {master, teacher, person}, {teacher, instructor, person}
기능→ 작용	capacity→function	capacity →function	{capacity, function}
기분→ 감정	mood→feeling	mood→feeling	{mood, sensation, feeling}
나무→ 식물	wood→vegetation	wood →vegetation	{wood, tree, shrub, plant}
시절→ 시기	occasion→time, season→period, time→period	occasion→period	{occasion, time, period}
작가→ 사람	author→person, writer→person	writer→person	{writer, author, person}

회장→ 사람	chairman→person	chairman →person	{chairman, president, person}
수사→ 조사	detection→investigation, investigation→inquiry, investigation→research, search→examination, search→investigation	search →investigation	{search, detection, investigation}
기자→ 사람	correspondent→person, reporter→person, writer→person	reporter→person	{reporter, writer, journalist, person}, {reporter, pressman, person}
토지→ 땅	land→area, soil→earth, territory→area	soil→earth	{land, territory, area}
어른→ 사람	adult→person, man→ person	adult→person	{man, adult, person}
감독→ 사람	manager→person, supervisor→person	manager →person	{overseer, foreman, person}, {supervisor, manager, person}
환자→ 사람	sufferer→person, patient→person	patient→person	{convalescent, sufferer, person}, {invalid, patient, person}
대화→ 이야기	chat→conversation, chat→discourse, dialogue→talk	dialogue→talk	{chat, dialogue, talk}
여름→ 계절	summer→season	summer→season	{summer, season}
주인→ 사람	host→person	host→person	{master, landlord, owner, person}
선수→ 사람	athlete→person, player→person	athlete→person	{athlete, player, person}
귀족→ 사람	peer→person	peer→person	{peer, nobleman, person},{aristocracy, nobleman, person}
특성→ 성질	characteristic→attribute, characteristic→property, feature→attribute, feature→property, trait→attribute	feature →property	{feature, property}, {peculiarity, individuality, attribute},{feature, characteristic, attribute}
업무→ 일	duty→work, service→job, service→work	service→work	{duty, business, service, work}
달→ 위성	moon→satellite	moon→satellite	{moon, satellite}
세월→ 시간	times→time	times→time	{times, time}

왕→ 사람	ruler→person	ruler→person	{sovereign, magnate, baron, person}, {monarch, ruler, sovereign, person}
신→ 존재	god→being	god→being	{god, being}
줄→ 물건	line→object	line→object	{cord, line, connection, object}, {string, line, object}
배경→ 경치	background→prospect, background→scene, background→scenery, background→view	background →view	{background, scene},{background, prospect}, {background, scene}, {background, scenery}, {background, view}
논의→ 토의	debate→discussion	debate →discussion	{debate, discussion}
독자→ 사람	reader→person	reader→person	{reader, person}
과제→ 문제	problem→question, question→topic, question→subject	problem →question	{problem, question}, {question, topic, subject}
가을→ 철	autumn→season, fall→season, harvest→ season	fall→season	{fall, autumn, season}
수단→ 방법	design→system, expedient→means, expedient→way, instrument→device, instrument→means, instrument→way, method→manner, method→way	method→way	{step, instrument, measure, method, manner}, {instrument, measure, way}
급→ 등급	grade→magnitude, grade→rank	grade→rank	{grade, magnitude}, {grade, rank}
의사→ 사람	doctor→person	doctor→person	{doctor, physician, person}
영역→ 범위	province→sphere, realm→sphere	province →sphere	{province, realm, sphere}
기사→ 글	article→writing	article→writing	{article, writing}
일행→ 사람	party→person	party→person	{party, company, person}
버릇→ 행동	manners→behavior, manners→conduct	manners →behavior	{manners, habit, behavior}
실체→ 물체	subject→substance, substance→object	substance →object	{subject, substance, object}
학기→ 기간	semester→term, term→period	semester→term	{semester, term, period}

풀→ 식물	herb→plant, weed→plant	herb→plant	{herb, grass, weed, plant}
짐→ 물건	burden→object, cargo→object, charge→stuff, freight→object, load→object, load→object, load→stuff	burden→object	{freight, charge, burden, object}, {luggage, load, baggage, object}
영상→ 모습	image→appearance, reflection→image	image →appearance	{image, reflection, appearance}
의무→ 일	duty→work	duty→work	{duty, work}
탑→ 건축물	tower→structure	tower →structure	{pagoda, tower, structure}
바지→ 옷	pants→clothes, pants→clothing	pants→clothes	{pants, clothes}
겁→ 마음	fear→feeling, fright→feeling	fear→feeling	{fear, dread, feeling}
안경→ 물건	eyeglasses→object, glasses→object, pince-nez→object, specs→object, spectacles→object	glasses→object	{eyeglasses, glasses, spectacles, specs, object}
이상→ 상태	abnormality→condition, disorder→condition, disorder→state, unsoundness→condition	disorder →condition	{disorder, abnormality, derangement, condition}
거실→ 방	parlor→room	parlor→room	{parlor, room}
이성→ 능력	reason→faculty	reason→faculty	{reason, faculty}
향기→ 냄새	aroma→odor, aroma→smell, fragrance→aroma, fragrance→odor, fragrance→scent, fragrance→smell, perfume→odor, perfume→smell, perfume→scent, scent→odor, scent→smell	fragrance→smell	{perfume, scent, fragrance, smell}, {scent, fragrance, aroma}

통→ 그릇	barrel→vessel, box→container, bucket→vessel, can→container, canister→container, bucket→vessel, pail→vessel, tub→vessel	box→container	{can, tub, keg, container}, {tub, keg, bucket. vessel}
떼→ 무리	bevy→flock, drove→crowd, flock→crowd, group→set, herd→throng, herd→crowd, multitude→group, pack→band, pack→lot, pack→set, pack→crowd, swarm→crowd, throng→crowd	herd→crowd	{school, herd, pack, crowd}, {flock, throng, drove, crowd}
들→ 땅	field→earth, field→ground, field→land, plain→earth, plain→ground, plain→land	field→ground	{plain,field, ground}
인민→ 사람	public→people	public→people	{public, people}
골목→ 길	alley→lane, alley→thoroughfare	alley→lane	{alley, lane}
신호→ 부호	gesture→sign	gesture→sign	{gesture, sign}
모양→ 생김새	figure→form, figure→shape, manner→form	figure→form	{shape, figure, form}
줄→ 선	string→line	string→line	{string, line}

접수 일자: 2008 년 4 월 17 일
게재 결정: 2008 년 6 월 9 일