온톨로지 구축을 위한 다의어 의미 구분
- 카파 통계를 활용한 의미 구분 작업 일치도 검사 -

김동성

한국외국어대학교 언어인지학과 BK21 사업팀

1. 서론

본 연구는 온톨로지 구축을 위한 다의어들의 의미 구분과 관련된 문제를 논의한다. 자연언어에서 사용되는 어휘들은 중의적이어서, 개념 체계를 구성하는 온톨로지에 활용하기 위해서는 중의성을 제거하여야 한다. 이러한 작업을 위해서 본 연구에서는 카파통계를 제시한다.

온톨로지 체계는 자연언어가 가지는 복잡한 의미 체계를 일관적인 계층구조로 분석할 수 있다. 이러한 온톨로지는 개념의 체계로 구성되며, 이러한 개념의 체계를 구성하기 위해서는 자연언어에서 사용하는 어휘 의미를 분석하여야 한다. 문제가 되는 것은 자연언어의 중의적 의미를 어떻게 분석하는 가이다. 어휘의미는 동음이의어와 같은 전혀 다른 의미가 하나의 어휘에 연결되어 있는 경우와 다의어와 같은 비슷한 의미가 하나의 어휘에 관련되어 있는 경우이다. 정확한 온톨로지 체계를 구성하기 위해서는 다의성을 해결하여서 연결하여야 한다.

이러한 어휘 의미를 분석하기 위해서 본 연구에서는 의미 구분이 필요한 단어를 코퍼스에서 추출하고, 사전을 통해서 분류할 의미를 정의하고, 코퍼스에서 구분할 어휘가 포함된 문장을 추출해서 6~7명의 작업자를 통해서 의미를 구분하였다. 이러한 작업의 결과로 얻어진 결과물이 신뢰성이 있는지 또 어떠한 의미를 선택해야 하는지에 대한 결정이 필요하였다. 이러한 연유로 의미 구분작업에 카파통계를 활용하여서 작업의 신뢰도 및 의미를 결정하였다.

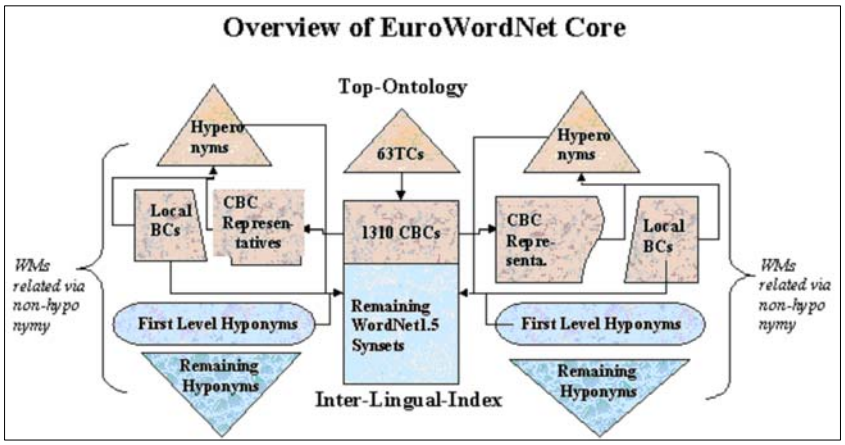
본 논문의 구성은 다음과 같다. 2절은 다의성과 온톨로지 체계를 설명하며, 다국어체계를 연결하고 온톨로지를 상위에 연결한 유로워드넷을 설명한다. 3절에서는 다의성 구분 작업을 설명한다. 4절은 다의성 구분 작업을 분석하는 통계 기재인 카파통계를 설명한다. 5절은 이 논문의 결론이다.

2. 다의성과 온톨로지

다의성은 온톨로지 체계에서 매우 중요한 문제이다. 자연언어의 다의성 구조는 매우 복잡하여서 일관성있는 온톨로지 체계를 구성하는 것은 어렵다. 반대로 일관적인 온톨로지 체계를 구성하여서 어휘 구조와 연결하면 어휘 구조의 일관적인 의미 체계를 구성할 수 있게 된

다.여러 유럽 언어들의 워드넷을 정렬하고 통합한 유로워드넷의 경우에 가장 상위 개념으로 온톨로지를 제시하였다(Vossen 1999).

유로워드넷의 구성방식은 다음과 같다. 개별 언어들이 워드넷 방식으로 구성되어 있고, 언어간은 영어 워드넷을 중심으로 한 중계 인덱스(ILI: Inter Lingual Index)로 연결하고, 이 중계 인덱스는 상위 온톨로지 체계와 연결되어 있다. 전체 구성을 그림으로 살펴보면 다음과 같다.



[그림 1] 유로워드넷의 전체 구성도(Vossen 1999; 58)

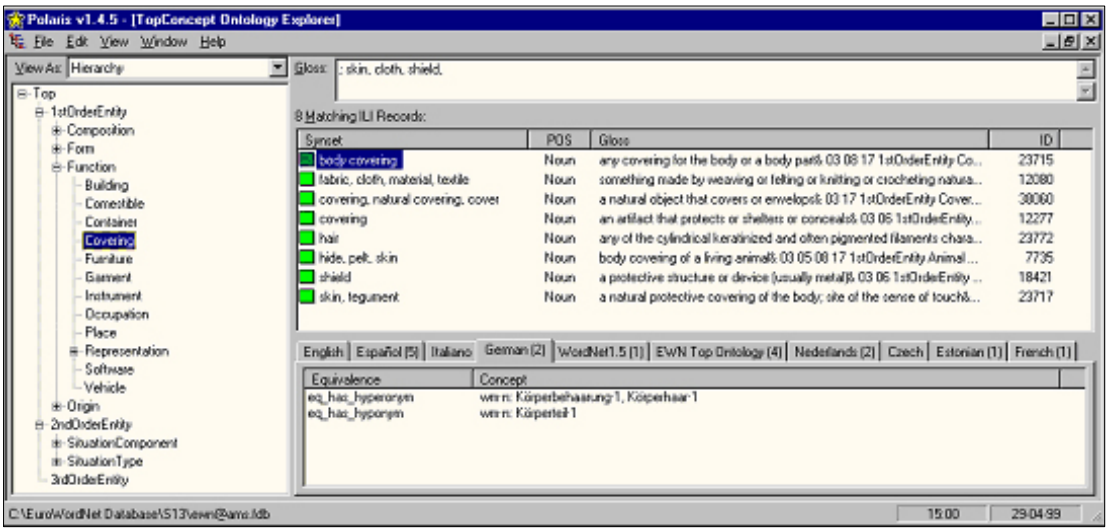
유로워드넷은 개별 언어를 워드넷 형식으로 만들었으므로, 개별 언어는 동의어 집합으로 구성된다. 이러한 동의어 집합을 개별 언어들 간에 정렬되어 있다. 아래 표는 Vossen(1999; 41)에서 제시한 표로 office라는 의미로 쓰인 동의어 집합들을 언어간에 정렬한 것이다.

ILI record	Dutch Synsets	Spanish Synsets	Italian Synsets
{**office**}-1980921 where professional or clerical duties are performed; "he rented an office in the new building"	{kantoor; werkkamer; werkruimte}	{oficina}	{ufficio; studio}
{role; part; **office**; function}-399406 the actions and activities assigned to or required or expected of a person or group: "the function of a teacher";"the government must do its part" or "play its role" or "do its duty"	{functie; rol} {emplooi}	{función; papel; oficio}	{ufficio; mansione; carica}
{situation; place; spot; **office**; slot; berth; post; position}-344376 a job in an organization or hierarchy; "he ocupied a post in the treasury"	{ambt; ambtsbediening; bediening; officie; officium} {betrekking; baan; dienstbetrekking; dienstverband; functie; job; positie; werk; werkkring} {arbeidsplaats; plaats}	{cargo; puesto}	{lavoro; impiego; occupazione}
{authority; **office**; bureau; agency}-5301461 an administrative unit of government; "the Central Intelligence Agency"; "the Census Bureau"; "Office of Management and Budget"; "Tennessee Valley Authority"	{dienst} {kantoor; bureau; bureel; burelen} {bureau} {agentuur}	{agencia; oficina}	{ispettorato}
{office staff; **office**}-5303509 professional or clerical workers in an office; "the whole office was late the morning of the blizzard"	{kantoorpersoneel}		

[표 1] 유로워드넷의 office 어휘의 정렬

위의 [표 1]를 보면 office가 다의적으로 사용되었으며, 언어간에 많은 차이를 보임을 알 수 있다. 이미 설명한 바와 같이 여러 언어들의 동의어 집합(들)이 워드넷을 중심으로 된 중계 인덱스(ILI)를 중심으로 정렬되어 있다. 이 표에서 주의 깊게 살펴볼 점은 네덜란드어에서는 office가 다른 언어와 달리 여러 개의 의미로 구성되며, 이러한 의미들은 다의성을 지닌다. 예를 들어서 {situation; place; spot; office; slot; berth; post; position}에 해당하는 의미를 나타내는 네덜란드어 집합은 세 개의 동의어 집합으로 구성된다.

온톨로지를 활용하면 이러한 의미적 관계를 계층적으로 구분할 수 있다. 유로워드넷에서는 온톨로지를 상위어 배치하여서 의미를 구분하고 있다. 유로워드넷에서 활용하고 있는 소프트웨어인 Periscope을 통한 연결을 살펴보면 아래 그림과 같다.



[그림 2] 상위 온톨로지와 연결된 유로워드넷(Vossen 1999; 74)

유로워드넷의 이러한 체계는 자연언어가 가지는 다의성으로 인하여 복잡하게 구성되었다. 다의성은 온톨로지 체계에서 매우 중요한 문제이다. 따라서, 온톨로지 체계를 구성하면서 다의적 의미를 구분하는 것이 매우 필요하다.

3. 다의적 의미 분류

본 연구에서는 한국어에서 발견되는 다의적 의미를 구분하기 위해서 다음과 같은 공정을 거쳤다.

- 의미 구분이 필요한 단어를 코퍼스에서 추출한다.
- 사전을 통해서 분류할 의미를 정의한다.
- 코퍼스에서 구분할 어휘가 포함된 문장을 추출한다.
- 6~7명의 작업자를 통해서 의미를 구분한다.
- 의미 구분작업에 카파통계를 활용하여서 작업의 신뢰도 및 의미를 결정한다.

먼저 온톨로지를 구성하기 위해서 필요한 단어를 코퍼스에서 추출하였다. 추출에 활용한 코퍼스는 의미 구분이 된 자료로 세종코퍼스를 활용하였다. 코퍼스로는 서상규(2000)에서 근거한 '연세말뭉치'와[1] 세종계획 2단계 연구 결과 중 한국어 학습용 어휘 선정을 위한 기초조사에 활용된 '현대 국어 사용 빈도 조사' 코퍼스 150만 어절(1998년에서 2002년, 이하 '세종150만코퍼스')을 활용하였다. '세종150만코퍼스'는 '표준국어대사전'에 의거하여 표제어 단

1) 서상규(2000)에서 제시한 자료는 어휘 목록만이 제시된 자료이다. 서상규(2000)에서 밝힌 바와 같이 어휘 목록 자료는 '연세말뭉치'에 근거한 것으로 코퍼스에 바탕을 둔 자료이다.

위인 동음이의어로만 의미 구분된 태깅코퍼스이다.

두 자료를 검토해서 서로 일치하는 어휘 목록을 추출하여서 빈도순으로 대략 600여개의 어휘를 산출하였다. 이를 토대로 사전을 사용해서 구분할 어의를 뽑아냈다. 작업에 사용된 사전은 '연세한국어사전'과 '표준국어대사전'이다. 다음으로는 구분해야 할 어휘가 포함된 문장 100개를 세종 1000만 코퍼스를 통해서 추출하였다. 그리고, 마지막으로 7명의 작업자들이 구분할 어의를 대상으로 구분작업을 하였다.

예를 들어서 살펴보면 다음과 같다. 우주라는 단어는 '연세한국어사전'에 다음과 같이 정의되어 있다.

(1) 우주(宇宙)[우 : 주]【명사】
1. 온 세계를 둘러싸고 있는 공간.
[예문] 무한한 우주 속에서 우리 모두는 길 잃은 아이인지도 모른다.
2. (철학에서) 질서 있는 통일체로서의 세계.
[예문] 유교는 가족, 사회, 우주를 일관하는 위계 질서의 원리를 근간으로 한다

'표준국어대사전'에는 우주가 다음과 같이 정의되어 있다.

(2) 우주02 (宇宙) [우 : -] 「명」
「1」 무한한 시간과 만물을 포함하고 있는 끝없는 공간의 총체.
¶우주 만물/우주에 가득 차다/대의를 우주에 밝혀서 천하의 충신 의사로 하여금 나의 고충을 알리려 함이었던 것이다.≪박종화, 임진왜란≫§
「2」 『물』 물질과 복사가 존재하는 모든 공간.
「3」 『천』 모든 천체(天體)를 포함하는 공간.
¶광할한 우주/우주를 왕복하다/우주에 관하여 연구하다. §
「4」 『철』 만물을 포용하고 있는 공간. 수학적 비례에 의하여 질서가 지워져 전체적으로 조화를 이루고 있는 상태를 강조할 때에 사용되는 피타고라스학파의 용어이다.

'연세한국어사전'과 '표준국어대사전'의 정의는 서로 다르다. '연세한국어사전'의 정의만을 활용하면 의미 구분이 너무 적어서 '표준국어대사전'의 정의를 참조하고 이를 다시 목록화하였다.

(3) 우주
1. 온 세계를 둘러싸고 있는 공간. 무한한 시간과 만물을 포함하고 있는 끝없는 공간의 총체.
2. 『물』 물질과 복사가 존재하는 모든 공간.
3. 『천』 모든 천체(天體)를 포함하는 공간.

4. (철학에서) 질서 있는 통일체로서의 세계. 『철』 만물을 포용하고 있는 공간. 수학적 비례에 의하여 질서가 지워져 전체적으로 조화를 이루고 있는 상태를 강조할 때에 사용되는 피타고라스학파의 용어이다.

위의 사전 정의를 토대로 하여서 의미 구분이 되어 있지 않은 코퍼스에서 (4)와 같이 문장을 추출해낸다.

(4)

a. 아직 탄생 초기여서 별을 만들고 남은 [우주] 가스들이 그 주위에 남아서 빛을 발하고 있기 때문에 뿌옇게 보이는 것이다.
b. 올라가는 현상이, 우주가 수축할 때에 [우주] 내 물질의 온도가 올라가는 현상과 거의 흡사하다는 점이다.
c. 있는 수천만 개의 별, 더 나아가서 이 [우주] 안에 있는 수천억 개의 은하들 안에 어디에도 생명체가 존재할 수 있다는 가능성을 점칠 수 있게 된다.
d. 자기들 행동 범위 내에서는 매우 좋은 [우주] 모델이 되지만 그것이 올바른 것이라고는 할 수 없다.
e. 전문가들이 소련이 판매하기로 결정한 [우주] 프로그램 관련 상품을 검사하기 위해 소련으로 비밀 여행을 떠나고 있다고 말했다.

이를 토대로 작업자에게 화자 직관을 활용해서 의미 구분을 하게 하였다. 실제 작업의 결과를 보면 아래 표와 같이 직관이 서로 상이하게 다른 경우가 발생한다.

	작업자1	작업자2	작업자3	작업자4	작업자5	작업자6
(4a)	3	3	3	3	3	3
(4b)	3	3	3	3	3	2
(4c)	3	3	3	1	1	3
(4d)	2	3	0[2)]	1	3	1
(4e)	3	3	3	1	3	3

[표 2] 의미 구분 작업

표를 자세히 살펴보면 (4a)의 경우에는 의미 구분하기에는 문제가 없으나, (4b)의 경우에는 작업자6이 다른 의미를 표기하고 있다. 또한 (4d)의 경우에는 어떤 의미로 우주가 사용되었는지 이해하기 힘들다. 따라서 별도의 기재를 활용해서 이를 처리해야 한다. 본 연구에서는 카파통계를 활용해서 작업의 신뢰도 및 의미를 구분해낸다.

2) 개별 숫자는 (3)에서 표기된 숫자를 의미하고 0은 (3)에서 표기된 의미로 설명되지 않는 경우나, 의미 구분이 불가능한 경우, 기타의 경우에 활용된다.

4. 카파통계[3)]

카파통계는 화자 직관을 활용한 여러 실험에서 그 실험의 통계적 의미와 더불어 집단의 결정을 정하는 경우에 활용된다(Carletta 1995, Passonneau 2004). 카파통계는 Cohen(1960)에 의해서 제시되었으며, 집단의 일치도를 통계적으로 입증하기 위해서 고안되었다. 집단의 일치는 완전한 일치인 1과 완전한 불일치인 0 사이에 통계적으로 분포한다. 실험에 참가한 실험자의 일치는 우연한 일치와 필연적 일치로 구성되며, 일치도의 연산은 전체 일치한 것에서 신뢰성이 떨어지는 일치를 제외한 일치로 구성된다. Cohen(1960)에서 제시한 수식은 다음과 같다.

(5) $\kappa = \frac{P_o - P_e}{1 - P_e}$

위에서 P_o는 관찰되어진 일치이고 P_e는 우연한 일치로 신뢰성이 떨어지는 일치를 말한다. Rietveld(1993)에서 제시한 P_o와 P_e의 측정법은 다음과 같다.

(6)

N: 전체 분류할 항목들

$\sum n_{ii} = \sum P_i$

P_o: $\frac{\sum n_{ii}}{N}$

P_e: $\sum \left(\frac{n_{i.}}{N} \times \frac{n_{.i}}{N} \right) = \sum P_i \cdot P_{.i}$

$P_{i.}$: $\left(\frac{n_{i.}}{N} \right)$, i번째 행의 나머지

$P_{.i}$: $\left(\frac{n_{.i}}{N} \right)$, i번째 열의 나머지

이 방식을 적용하면 카파의 분포범위가 결정되며, 이 결정을 토대로 해서 분포의 범위를 토대로 신뢰할 만한 데이터인지를 결정한다. Landis and Koch(1977)에서 제시한 결정 수준은 다음과 같다.

(7) 0.00 – 0.20: 'slight'
0.21 – 0.40: 'fair'
0.41 – 0.60: 'moderate'
0.61 – 0.80: 'substantial'

3) 카파통계는 통계자료가 명명척도(nominal scale)로 측정이 되는 경우에만 가능하다.

0.81 – 1.00: ‘almost perfect’

본 연구에서는 3절에서 제시된 자료의 일치도를 검증하고 일치의 수준이 0.61이상일 경우에는 작업이 신뢰성이 있는 것으로 판단하였다. 이를 토대로 분류할 의미들을 결정하였다. 예를 들어서 [표 2]를 통해서 도출된 자료가 카파통계값 0.61이상일 경우에는 이 자료의 일치도가 신뢰성이 있는 것으로 판단된다. 그리고, 이 신뢰성을 근거로 3번 우주의 의미의 분포가 신뢰성이 있으므로 온톨로지에서 분류할 의미를 ‘모든 천체(天體)를 포함하는 공간’으로 결정하게 된다.

5. 결론

본 연구는 온톨로지 체계를 구성하기 위한 의미를 추출하는 작업을 설명하였다. 의미 관계에서 다의적 관계를 해결하기 위해서 직관을 활용해서 이를 구분하였다. 화자 직관은 개별 화자에 따라서 달리 분포하므로, 실제 작업을 통해서 도출된 자료를 처리하기 위한 기재가 필요하다. 이에 화자 직관을 분석할 기재로 카파통계를 제안하였으며, 이를 활용해서 의미를 구분하였다.

카파통계는 언어학적 연구에서 많이 활용되는 통계로서, 화자 직관이나 기타 일치를 요하는 실험에서 이용된다. 화자 직관이 개인과 집단 및 실험 방식에 따라서 변화하는 것을 감안하면, 화자 직관을 활용한 실험이 신뢰성이 있는지가 검토되어야 한다. 또한 신뢰성이 있는 자료들의 일치도를 연산해서 어떠한 일치를 선택해야 할지를 결정한다.

<참고문헌>

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

Carletta J. (1995) Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics, Vol 22, pp. 249-254.

Cohen, J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement, Vol. 20, pp. 37-46.

Landis JR & G. Koch. (1977) The measurement of observer agreement for categorical data. Biometrics 33(1). 159-174.

Passonneau R. (2004) Computing reliability for coreference annotation. In Proceedings of LREC, Lisbon.

Rietveld T. (1993) Statistical Techniques for the Study of Language and Language Behaviour. Mouton de Gruyter.

Vossen P. (1999) EuroWordNet General Document. University Amsterdam. http://www.hum.uva.nl/ewn.