용언의 구문관계를 이용한 명사 분류

김현진*, 박세영**, 장명길*, 박재득*, 박동인*

* 시스템공학연구소 자연어정보처리연구부

** 한국전자통신연구원 소프트웨어연구부

* {jini,mgjang,jdpark,dipark}@seri.re.kr

** sypark@com.etri.re.kr

Clustering Noun Using Syntactic Relations

Hyun-Jin Kim*, Se-Young Park**, Myung-Gil Jang*, Jay-Duke Park*, Dong-In Park*

* Natural Language Processing Department, SERI

** Natural Language Processing Section, ETRI

요 약

자연언어를 처리하는 응용시스템에서는 의미적으로 유사한 집합으로 분류된 단어들을 이용하는 것이 필요하다. 특히 한국어에서는 명사마다 함께 쓰이는 용언들이 제한되어 있다. 이 논문에서는 문장에서 용언과 명사의 구문 관계로 추출되는 정보를 이용하여 명사를 분류하는 방법을 제시한다. 또한 실제 코퍼스에서 추출된 명사들을 중심으로 의미적 집합으로 묶는 작업을 하고, 각 의미군마다 특정적인 구문 정보를 적용하여 자동 명사 추출에서 나타나는 모호성 해소에도 이용하였다. 용언의 구문관계 추출은 기존 연구된 용언 하위 분류 연구를 이용하였고, 코퍼스를 통해 얻은 명사와 용언을 이용하여 수정 및 보완하였다. 실험 코퍼스는 1만 문장 가량의 구문 구조가 부착된 코퍼스(Tree Tagged Corpus)를 이용하였다.

1. 서론

한국어 문장을 처리하다 보면 명사 각각의 쓰임 뿐만 아니라 명사를 의미적으로 유사한 것들의 집합으로 이용하는 것이 많이 필요하게 된다. 예를 들어 구문 분석이나 의미 분석과 같은 분야에서 모호성 해소를 위해서도 각 단어들의 유사군의 정보를 이용하기도 하고, 기계 번역에서 동사를 번역하는 문제에서도 적용할 수 있다.

이러한 명사 분류에 대해서 기존 연구를 살펴보면, 말뭉치에서 단어들의 분포 정보 또는 구문 관계를 이용하여 단어를 분류하는 기법을 쓰기도 했으며[Brown92:Hindle90:Pereira93:정연95], 명사의 공기 유사성(co-occurrence similarity)을 이용하여 기계 번역 등의 모호성 해소에 적용하기도 했다[양재95]. 그리고 명사 계층 관계를 구축하기 위해 사전의 풀이말을 이용한 연구도 있다[문유96:조평95:한영94].

본 연구는 문장에서 용언과 명사의 구문 관계로 추출되는 정보를 이용하여 명사를 분류하는 방법을 제시한다.

먼저, 각 용언의 문형 구성요소인 명사항에 나타나는 단어들에 대한 정보를 수집하였다. 그리고 의미적으로 유사한 자질을 가진 명사항들을 서로 묶어 하나의 의미군으로 분류하고, 이 정보를 이용하여 구문구조가 부착된 코퍼스(Tree Tagged Corpus)에서 같은 의미군으로 추정되는 명사들을 추출하였다.

또한 용언의 구문관계를 더 효율적으로 이용하기

위해 기존 코퍼스에서 추출된 명사들을 활용하는 방안을 제시하였다.

2. 용언의 구문관계

한국어 문장에서는 각 용언의 구문적 구성요소인 명사항에 올 수 있는 명사들이 한정되어 있다. 즉, 용언의 주격 명사로 가질 수 있는 명사들과 목적격 명사로 가지는 명사들이 제한되어 있다. 예로써 '느끼다'와 '달래다' 용언의 구문관계를 [표2-1]에 나타내었다.

[표2-1] '느끼다'와 '달래다'의 구문관계

용언	문형구조	명사자질
느끼다	N1(sub) N2(obj) 느끼다	N1: 사람명사 N2: 감정명사 감각명사
달래다	[문형1] N1(sub) N2(obj) 달래다	N1: 사람명사 N2: 감정명사 감각명사
	[문형2] N1(sub) N2(obj) 달래다	N1: 사람명사 N2: 사람명사

[표2-1]에서 나타나듯이 '느끼다'와 '달래다(문형1)'에서 N2자리에는 사람의 감정이나 감각을 나타내는 단어들이 공통적으로 나타남을 알 수 있다[홍재96].

이 표를 명사를 중심으로 다시 나타내면 [표2-2]와 같다. 즉, 감정명사와 감각명사가 문장에서 함께 나타나는 용언들이 추출된다.

[표2-2] 감정명사, 감각명사의 구문관계

명사	문법 기능	해당 용언
감정명사, 감각명사	Obj	느끼다, 달래다

여기서 '느끼다'와 '달래다'는 감정명사와 감각명사를 추출할 수 있는 특정 용언으로 분류된다.

본 논문에서는 명사가 용언에 대해 가지는 문법 기능을 조사의 쓰임을 중심으로 [표2-3]과 같이 6가지로 한정하였다[장석95]. 문법 관계는 문장에서 나타나는 명사와 용언과의 문법 기능 관계 중 가장 많이 나타나는 것을 중심으로 하였다.

[표2-3] 문법 기능

문법 기능 관계	해당 조사 목록	
주격(Sub)	이/가	
목적격(Obj)	을/를	
담화 기능(Col)	은/는, 도, 만, 조차 등	
부사격	방향,처소(Loc)	에, 에서, 에게, 한테 등
	도구,시발(Ins)	로서, (으)로
	공동격(Com)	와/과, 하고

이런 식으로 각 구문에서 수집할 수 있는 용언과 명사와의 구문관계를 이용하여, 명사들을 군으로 모을 수 있는 특정 용언과 문법 기능을 추출한다.

이 논문에서는 구문관계가 구축된 용언 612개를 중심으로 각 명사 또는 명사군과 그것을 추출하는데 쓰일 수 있는 특정 용언들을 분류하였다. 분류 결과의 일부를 부록의 [표 부-1]에 보였다. 결과를 보면 여러 용언에서 함께 나타나는 명사들을 특정한 의미군으로 명명할 수 있는 경우도 있고, 규정할 순 없지만 의미적으로 또는 그 쓰임이 비슷하므로 하나의 의미군으로 볼 수 있는 경우도 있다.

의미적으로 유사한 명사군을 만들어 낼 수 있는 용언들을 보면, 문법적인 성질이 유사한 경우도 있고, 용언들 사이에도 의미적으로 유사한 집합으로 모이게 되는 경우도 있었다. [표2-4]에서 보면, 유사한 명사군을 이루는 해당용언을 보면 문법적 성질 뿐만 아니라, 용언 사이에도 비슷한 의미를 나타내는 성질도 가지고 있음을 알 수 있었다.

[표2-4] 유사 용언군 형성

명사	문법 기능	해당 용언
'잘못', '실수'에 관계된 명사	Obj	깨닫다, 꼬집다, 꾸짖다, 따지다 등
출장, 순찰, 왕진, 감시, 강의 등	Obj	나가다, 나오다, 다녀가다, 다니다 등

3. 구문관계 조정

[표3-1]을 보면, 용언의 구문 관계에는 '지저귀다'처럼 용언의 의미가 단일해서 명사들의 의미가 유사한 명사군으로 한정할 수 있는 경우도 있지만, 동사 '싸다'처럼 동사 자체가 여러 가지의

의미로 쓰이므로, 명사를 수집하는데 어려움이 있을 수 있다. 또한 하나의 원형으로 형용사와 동사로 동시에 쓰이는 경우엔 더 복잡하게 되고, 명사 자질을 모으는데 모호성이 있을 수 있다. 이런 경우엔 그 명사군을 한정하는데 하나의 용언이 아니라 여러 용언의 집합을 이용하여 모호성을 가진 용언으로 인한 부족한 정보를 보충하도록 한다.

[표3-1] '지저귀다'와 '싸다'의 구문관계

용언	문형구조	명사자질
지저귀다	N1(sub) 지저귀다	N1: 조류 (종달새, 꾀꼬리 등)
싸다	[문형1] N1(sub) N2(obj) 싸다	N1: 사람명사 N2: 음식명사
	[문형2] N1(sub) N2(obj) 싸다	N1: 사람명사 N2: '분비물'에 관계된 명사
	[문형3] N1(sub) N2(obj) N3(Ins) 싸다	N1: 사람명사 N2: 구체물 N3: 도구명사

예로써, '싸다'의 [문형1]과 [문형2]의 N2에 나타나는 명사가 '음식명사'인지 '분비물'에 관계된 명사인지를 구분하려면 [표3-2]에서처럼 추가의 용언 정보를 제공한다.

[표3-2] '음식명사'와 '분비물에 관계된 명사'의 구분

명사	문법 기능	해당 용언
음식 명사	Obj	싸다, 맛보다, 굽다, 데우다, 먹이다, 묵히다 등
	Sub	돌구다, 맛나다, 먹히다, 묵다, 맛들다 등
'분비물'에 관계된 명사	Obj	싸다, 분비하다 등

이 경우 '싸다'의 N2에 나온 명사가 같은 집합의 다른 해당용언에 나타나는 지를 확인해서 구분할 수 있다.

이 논문에서는 '지저귀다'처럼 명사군을 한정할 수 있는 정보를 제공하는 용언을 충분조건 용언이라고 하고, '싸다'처럼 다른 용언과의 관계에 의해서 의미가 분별되는 용언을 불충분조건 용언이라고 구별하였다.

4. 구문구조부착 코퍼스에서의 명사 추출

2장과 3장에서는 각 용언의 명사항에 나타나는 단어들에 대한 정보를 이용하여 유사 명사군을 결정할 수 있는 요소를 구축하였다. 이 장에서는 실제 코퍼스에서 용언의 명사항에서 나타나는 명사를 대상으로 명사군을 분류했다.

본 논문에서 사용한 코퍼스는 구문구조부착 코퍼스[1]이며 약 10,000문장 100,000어절로 구성된다.

4.1. 유사 명사군 추출

구문구조부착 코퍼스에서 명사들을 추출하기 위해서 각 용언과 명사가 가지는 문법 관계(G)는 다음과 같이 6가지로 제한하였다.

$$G \equiv \{ Sub, Obj, Ins, Com, Col, Loc \}$$

$R_g(v)$는 코퍼스에서 각 용언(v)이 문법 관계($g \in G$)를 통해서 취하는 명사의 집합을 나타낸다.

여기서 유사 명사군으로 한정할 수 있는 용언의 집합은 3장에서 언급했듯이 충분조건 용언(V_t)과 불충분조건 요언(V_s)으로 구분되므로, $R_g(v)$는 다음과 같이 구해진다.

$$R_g(v) = R_g(v_t) + R_g(v_s)$$

$$R_g(v_s) = \bigcup_{i=1}^{n-1} \bigcup_{j=i+1}^{n} \{R_g(v_i) \cap R_g(v_j)\}$$

$$g \in G, V = V_t + V_s$$

즉, 충분조건 용언과의 문법 관계에서 나온 명사는 유사 명사군에 모두 속하게 하였고, 불충분조건의 용언에서는 다른 불충분조건의 용언과 하나 이상이라도 동시에 나오는 명사를 유사 명사군에 속하도록 하였다.

예를 들어 [표3-1]에서 '싸다'가 불충분조건 용언이므로 다음과 같이 다른 해당 용언과 한 번이라도 같은 용례로 나오는 명사를 유사 명사군에 넣도록

[1] 한국과학기술원에서 과학기술처의 step2000과제의 세부과제로 구축한 결과임.

하였다.

Vg(싸다) = {{싸다, 맛보다}, {싸다, 굽다}, {싸다,데우다}, {싸다,먹이다}, {싸다,묵히다} 등}

이렇게 불충분조건 용언인 경우 다른 용언들과의 관계를 설정해 두면, 코퍼스에서 자료가 부족해서 많은 명사를 모으지 못하는 단점은 있다. 그러나 용언의 의미와 용법이 다양함으로 인해 생기는 부적합한 명사의 추출은 막을 수 있다.

2장과 3장에서 구축한 용언의 구문관계를 이용한 유사 명사군에 대해서 구문구조부착 코퍼스를 통해 나온 결과의 일부를 보이면 [그림 부-1]과 같다.

결과를 보면, 유사한 문법 성분을 가지는 용언의 명사항에는 비슷한 의미의 명사들이 모임을 알 수 있었다. 특히 단일한 의미를 가지는 용언의 경우는 추출되는 명사들도 유사한 의미를 가지고 있다.

4.2. 코퍼스를 이용한 구문관계 조정

일반적으로 용언을 중심으로 구축한 구문관계를 보면, 용언의 하위 범주 정보로는 필요하지만, 명사를 추출하는 데는 별로 효과가 없는 의미 정보가 많이 들어 있다. 이런 경우는 명사를 추출하는데 오히려 모호성을 발생하는 요소가 되므로 조정이 필요하다.

예로써, '먹다'의 구문관계에서 명사항에 들어가는 의미 정보로는 '음식물', '돈, 뇌물 등의 명사', '꾸지람, 욕 등의 명사' 등 다양한 의미가 올 수 있다. 그러나 코퍼스를 통해서 얻은 명사항의 명사들을 보면 대부분이 '음식물'에 관계된 명사들이었다. [표4-1]에 동사 '먹다'를 코퍼스에서 추출한 결과를 보였는데, '나이를 먹다'의 '나이'를 제외하고는 대부분이 '음식물'에 관계된 명사였다.

[표4-1] '먹다'의 명사항 추출

명사항	관계	동사
고기, 깡통음료, 나이, 떡, 맥주효모, 먹거리, 먹이, 물개, 물고기, 바나나, 밥, 사과, 새, 쌀, 야채, 양, 얼음, 찰밥, 우유, 유제품, 율무, 음식, 잉어, 저녁, 주스, 찰떡, 특식, 현미	Obj	먹다

이런 결과로 볼 때, '먹다'의 명사항은 '음식물'로 한정하는 것이 명사를 추출하는데 유용하다. 그래서 본 논문에서는 유사 명사군을 결정하는 용언을 용언의 구문 관계에서 나온 결과와 실제 코퍼스에서 나온 결과를 비교하여 조정하였다.

5. 결론

본 연구는 기존 명사와 동사의 분포 정보만으로 이용한 방식과는 달리, 문장에서 용언과 명사의 구문 관계로 추출되는 정보를 이용하여 명사를 분류하였다. 또한 기존 용언의 구문 관계에서 발생하는 모호성을 줄이기 위해 코퍼스를 통해 추출한 결과로 조정하였다.

논문에서는 의미적으로 유사한 명사군을 결정하는데 필요한 용언 정보를 정보 제공의 가중치에 따라 충분조건 용언과 불충분조건 용언으로 나누고, 각 용언에서 추출되는 명사항을 모으는 방식도 다르게 하였다.

그러나 이 논문에서는 단순히 불충분조건 용언인 경우 다른 조건의 용언과 겹치게 나타나는 명사만을 모으므로, 추출되는 명사에서 모호성이 완전히 사라지진 않는 문제점이 나타났다. 따라서 각 용언에 해당 명사군을 결정하는 자질에 대해서 수치로 나타내고, 그 값에 의해서 가중치를 두는 방식도 요구된다.

본 논문은 명사를 의미적으로 나누는 데에 용언의 구문관계를 이용할 수 있음을 보여주고 있으며, 추출된 명사군에 상하위 개념을 도입하면, 명사의 의미 분류 체제 구축에도 활용할 수 있으리라고 기대한다.

참고문헌

[양재95] 양재형, *한국어 분석 모호성 해소를 위 명사의 공기 유사성*, 서울대학교 대학 컴퓨터공학과 박사학위논문, 1995.

[조평95] 조평옥, *한국어 명사의 의미계층 구조구축* 울산대학교 교육대학원 석사학위논문, 1995.

[정연95] 정연수, *개념분류기법을 적용한 한국 명사분류*, 한글및한국어정보처리학회, 1995.

[문유96] 문유진, *한국어 명사를 위한 WordNet 설계 구현*, 한국정보과학회 논문지, 1996.

[한영94] 한영균, *명사류 의미망 구축을 위한 사전 뜻풀이의 어휘구조분석*, 한글및한국어정보처리학회, 1994.

[홍재96] 홍재성, "*한국어 동사 구문 사전*", 두산동아, 1996.

[권재92] 권재일, "*한국어 통사론*", 민음사, 1992.

[장석95] 장석진, "*정보기반 한국어 문법*", 한신문화사, 1995.

[Brown92] Brown,P.F., *Class-based n-gram models of natural language*, Computational Linguistics, 1992.

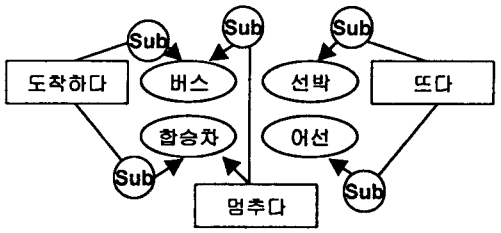
[Hindle90] D.Hindle, *Noun classification from predicate-argument structures*, Processing of 28th Annual Meeting of the ACL, 1990.

[Pereira93] Fernando Pereira, *Distributional clustering of English words*, Proceedings of the 31st Meeting of the ACL, 1993.

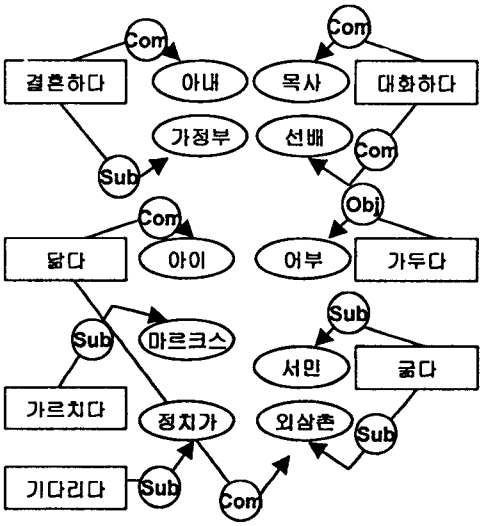
[Naohiko96] Naohiko Uramoto, *Corpus-based thesaurus-positioning words in existing thesaurus using statistical information from a corpus*, Information Processing Society of Japan, 1996.

[표 부-1] 유사 명사군을 결정하는 용언의 일부

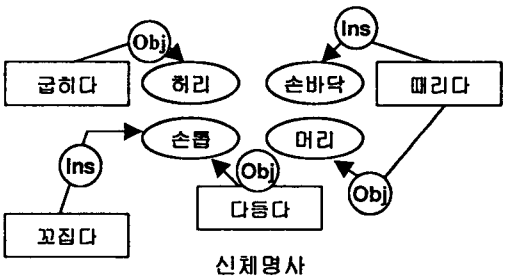
명사	문법 기능	해당 용언
교통수단	Sub	나르다, 늦다, 닿다, 도달하다, 멈추다, 멎다 등
	Ins	가다, 갈아타다 등
	Obj	갈아타다, 기다리다, 대다, 막다, 멈추다 등
인물	Com	가까이하다, 거래하다, 결합하다, 결혼하다, 대화하다, 닮다, 돕다, 마주보다, 마주치다, 만나다 등
	Obj	가두다, 가르치다, 간섭하다, 그리다, 그리워하다, 기다리다, 깨우다, 놀리다, 돕다 등
	Sub	가르치다, 각오하다, 간섭하다, 거래하다, 계몰하다, 계획하다, 굶다 등
신체	Obj	가리다, 굽히다, 기대다, 꼬집다, 다듬다, 대다, 만지다 등
	Sub	곪다, 타다 등
	Ins	가리다, 때리다 등



교통수단



인물명사



신체명사

[그림 부-1] 명사군 추출 결과의 일부