한국어 오픈 워드넷 (KWN) : 사전 기반의 반자동 구축

이인근, 황도삼, 함영균, 최기선
대구도시철도공사, 영남대학교 컴퓨터공학과, 한국과학기술원 전산학과
inkeunlee@gmail.com, dshwang@yu.ac.kr, hahmyg@kaist.ac.kr, kschoi@kaist.edu

Open Korean WordNet (KWN): Dictionary-based Semi-Automatic Development

In Keun Lee, Dosam Hwang, Younggyun Hahm, Key-Sun Choi
Daegu Metropolitan Transit Corporation
Department of Computer Science, Yeungnam University
Department of Computer Science, KAIST

요 약

본 논문에서는 사전자원에 기반한 한국어 워드넷(Open Korean WordNet: KWN)의 반자동 구축 방법을 제안한다. 제안한 방법에서는 각 전문분야별로 분류된 영어-한국어 대역사전, 일본어-한국어 대역사전을 이용하여 영어 워드넷(Princeton WordNet 3.0)과 일본어 워드넷(Japanese WordNet 1.1)의 어휘를 번역하였다. 그리고 번역 결과의 애매성을 해소하기 위하여, (1)영어와 일본어에 대한 한국어 대역어의 중복 여부, (2)사전의 분야 정보와 워드넷의 계층구조를 고려하였다. 제안한 방법으로 117,659 개의 워드넷 synset 중 63,221 개(약 54 %)의 synset에 대한 자동번역을 수행하여 한국어 워드넷을 구축하였다. 그리고 워드넷 synset의 정의문은 한국어 사전의 정의문을 참조하여 한글화 할 수 있도록 하고, 이 과정을 지원하기 위한 정의문 추천 알고리즘을 제안한다. 제안한 방법에 기반하여 전문가들이 상호 협력하여 한국어 워드넷을 구축할 수 있는 시스템을 개발한다.

주제어: WordNet, definition recommendation algorithm, Korean WordNet editor

1. 서론

워드넷(WordNet)[1]은 어휘 의미망으로써, 어의 중의성 해소나 정보추출 등과 같은 다양한 자연언어 처리에 폭넓게 사용된다. 그러나 기본적으로 워드넷이 영어로 작성되어 있어, 워드넷과 유사한 다양한 언어로 구현된 어휘망을 개발하기 위한 많은 연구가 수행되었다. 이를 통해 EuroWordnet[2], CoreNet[3], howNet[4], and BabelNet[5] 등이 구축되었고, 최근에는 Open Multilingual WordNet project를 통해 60 개 이상의 언어로 작성된 워드넷 다국어 워드넷이 개발되어 Global WordNet Association website[6]을 통해 무료로 공개하고 있다. 이러한 다국어 워드넷은 대역사전과 Wikipedia와 같은 전자화된 언어자원을 이용한 반자동 구축 방법을 통해 전문가들이 각 언어별로 워드넷을 구축하는 노력과 시간을 줄인다.

본 논문에서는 전자사전에 기반하여 한국어 워드넷(Open Korean WordNet: KWN)을 반자동으로 구축하는 방법을 제안한다. 제안한 방법에서는 각 전문분야별로 분류된 영어-한국어 대역사전, 일어-한국어 대역사전을 이용하여 영어워드넷(Princeton WordNet 3.0: PWN)[7]과 일본어 워드넷(Japanese WordNet 1.1: WN-JP)[8]의 어휘를 번역하였고, 번역 과정에서의 애매성을 줄이기 위해 워드넷의 계층구조를 이용하였다. 제안한 방법으로 117,659 개의 워드넷 synset 중 63,221 개(약 54 %)의 synset에 대한 자동번역을 수행하였다. 그리고 한국어 사전 정의문으로부터 워드넷 synset의 정의문을 결정하기 위한 정의문 추천 알고리즘을 제안한다. 또한, 제안한 방법에 기반하여 전문가들이 상호 협력하여 KWN을 구축할 수 있는 한국어 워드넷 구축 시스템(Open Korean WordNet Construction System: KWN-S)을 개발한다.

2. 방법

KWN의 구축 작업은 워드넷의 특정 synset의 의미에 적합한 한국어 어휘와 사전 정의문을 결정하는 것으로 볼 수 있다. 즉, 예를 들어, PWN의 “dolphin”(02068974, noun)은 WN-JP에서 “イルカ”, “海豚”, “ドルフィン”으로 표현되어 있고, KWN에서는 “돌고래”, “물돼지”로 표현할 수 있다. 또한 해당 synset에 대한 한국어 정의문은 “돌고래”에 대한 사전 정의문 중에서 의미적으로 적합한 정의문인 “이가 있는 돌고랫과의 포유류를 통틀어 이르는 말”을 연결함으로써 한글화가 가능하다. 다음은 KWN 구축 방법을 보인다.

2.1 워드넷 synset의 한글화

대역사전의 종류가 많을수록 특정 synset의 어휘에 대한 한국어 어휘의 수가 늘어나므로 적절한 어휘 선택은 어려워진다. 따라서 어휘 선택의 애매성을 줄이기 위해, 영어와 일본어에 대한 한국어 대역어의 중복 여부, 그리고 대역사전의 분야 정보와 워드넷의 계층구조를 고려한다. 본 연구에서는 34개의 분야(예: 의학, 역사, 공학, 음악 등)로 분류된 83개의 영어-한국어, 일본어-한국어

대역사전과 11개의 일반(대역)사전을 참조하였고, 다음 네 가지 경우에 대하여 한국어 어휘를 결정한다.

Case 1: 특정 synset의 영어나 일본어 어휘 중에서 한국어 대역 어휘가 하나만 존재하는 경우, 그 한국어 어휘를 선택한다.

	Synset 번호	영어	일본어	한국어
예)	02403003 (noun)	ox	-	황소

Case 2: 다수의 한국어 대역 어휘 중에서, 영어의 한국어 대역 어휘와 일본어의 한국어 대역 어휘가 동일한 경우, 그 한국어 어휘들을 모두 동의어로 간주한다.

	Synset 번호	영어	일본어	한국어
예)	01899593 (noun)	wool, fleece	羊毛	양모, 양털, 울

Case 3: 특정 synset의 어휘(예: arteria_renalis, renal_artery)와 그에 대한 상위 어휘(hypernyms)(예: arteria, artery, arterial_blood_vessel)가 동일한 분야의 대역사전에 존재하는 경우, 그 분야의 대역사전으로부터 번역한 한국어 어휘들(예: 콩팥동맥, 신동맥, 신장동맥)을 모두 동의어로 간주한다.

	Synset 번호	영어	일본어	한국어
예)	05354381 (noun)	arteria_renalis, renal_artery	-	콩팥동맥, 신동맥, 신장동맥

Case 4: 특정 synset의 어휘(예: roller)와 그에 대한 상위 어휘(예: wheel)가 다양한 분야(예: 기계, 의학, 조선)의 대역사전에 모두 존재하고, 이들 분야의 대역사전으로부터 공통된 한국어 대역어가 존재하는 경우, 그 한국어 대역어(예: 롤러)를 모두 동의어로 간주한다.

	Synset 번호	영어	일본어	한국어
예)	01899593 (noun)	roller	-	롤러 (기계, 의학, 조선)

2.2 워드넷 synset 정의문의 한글화

워드넷 synset에 대한 한국어 어휘를 결정한 후, 한국어 어휘의 사전 정의문 후보들 중에서 synset과 의미적으로 동일한 정의문을 선택해야 한다. 이 과정은 전적으로 전문가의 판단에 의존하며, 많은 시간과 노력을 필요로 한다. 따라서 다음 가정을 기반으로 특정 synset의 정의문을 추천한다. 즉, 특정 synset의 상위 한국어 어휘와 하위 synset의 한국어 정의문 정보를 이용하여, 정의문 후보들을 정렬하는 알고리즘을 식 (1)과같이 제안한다.

가정: "워드넷은 상·하위 개념에 해당하는 어휘들이 구조적으로 표현되어 있으므로, 어휘의 정의문에서 이러한 구조들을 유추할 수 있다."

$$S_k = n(L_{hyper} \cap D_k) + \sum_i \left(\alpha^i \times n\left(D_{hypo}^{i,j} \cap D_k\right)\right) \quad (1)$$

여기서 S_k는 k번째 한국어 정의문 후보에 대한 의미 적합도이다. $L_{hyper} = \{l_{hyper}^{i,j} | i = semantic\ distance, j \in N\}$는 대상 어휘에 대한 상위 어휘(hypermyns) 집합을 뜻하고, i는 해당 synset으로부터의 의미거리, j는 단순 자연수(N)로 표현된 일련번호를 뜻한다. $0 < \alpha \leq 1$ 이며 의미거리에 따라 가중치를 다르게 적용한다. $n(P)$는 함수 P에 포함된 요소의 개수로써, $n(L_{hyper} \cap D_k)$ 는 상위 어휘(hypernyms) 집합(L_{hyper})과 k번째 한국어 정의문 후보의 명사집합(D_k)의 교집합에 대한 요소의 개수를 의미한다. 그림 1은 식 (1)에서 정의한 기호의 예를 보인다.

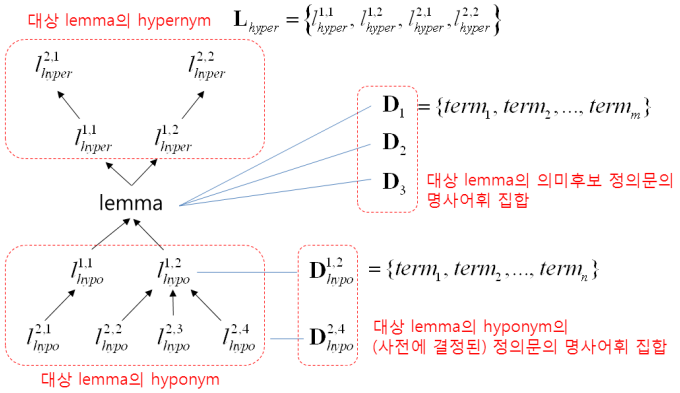

그림 1. 정의문 추천 알고리즘에서의 기호 예

예를 들어, 특정 synset "dolphin" (02068974, noun)의 한국어 대역어가 "돌고래"로 결정되었고, 한국어 정의문 후보, 상위 어휘, 하위 어휘에 대한 정의문이 그림 2와 같다고 할 때, 다음 계산을 통해 첫 번째 정의문이 대상 synset의 의미에 좀 더 적합하다고 판단한다.

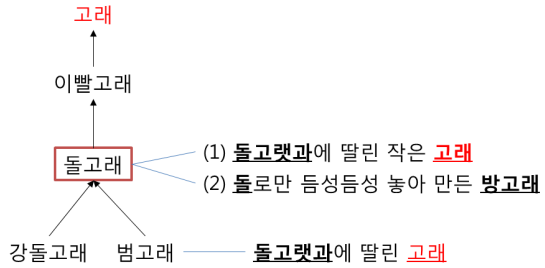

그림 2. 한국어 정의문 정렬 예

D_1={돌고랫과, 고래}, D_2={돌, 방고래}

L_{hyper}={고래, 이빨고래}, $D_{hypo}^{1,2}$={돌고랫과, 고래}

$S_1 = 1 + 0.9^1 \times 2 = 2.8$, $S_2 = 0 + 0.9^1 \times 0 = 0$

3. 한국어 워드넷 구축 과정 및 시스템

한국어 워드넷은 네 단계를 거쳐 구축된다. 즉, (1) synset의 한국어 어휘 자동구축, (2)첫 번째 단계에서 제외된 synset에 대해 전문가에 의한 한국어 어휘 및 정의문의 반자동 구축, (3)두 번째 단계에서 제외된 synset에 대해 전문가에 의한 한국어 어휘 및 정의문의 수동 구축, 마지막 단계에서는 (4)한국어 워드넷을 검증한다. 단계 (2)~(4)의 반자동/수동 구축 및 검증작업을 지원하기 위해 제안한 방법에 기반하여 한국어 워드넷 구축 시스템을 개발하였다. 개발한 시스템은 그림 3과 같이 다양한 사전자원을 참조하여 다수의 전문가가 협력하여 한국어 어휘와 정의문을 쉽고 빠르게 구축할 수 있는 환경을 제공한다. 즉, PWN과 WN-JP의 영어 및 일본어 어휘로부터 대역사전을 참고하여 한국어 대역 어휘 그룹을 생성하고, "어휘 추천기"를 통해 해당 synset에 적합한 한국어 대역어를 선택한다. 한국어 대역어가 선택되면 각종 전자사전으로부터 "사전정의문"을 추출하고, "명사추출기"를 이용하여 기 구축된 KWN의 어휘와 정의문으로부터 의미추천에 필요한 기본정보를 추출하여 "의미추천기"에서 적당한 정의문을 추천한다. 그리고 전문가는 KWN Editor를 통해 사전자원 및 KWN의 구축작업을 수행한다. 그림 4는 개발한 시스템의 인터페이스 화면을 보인다.

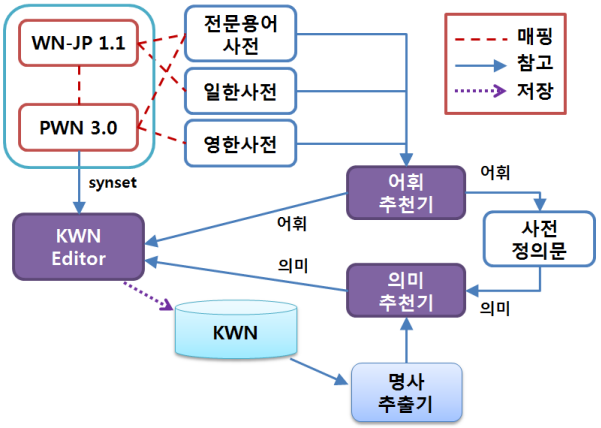

그림 3. KWN 구축 시스템과 자원의 관계

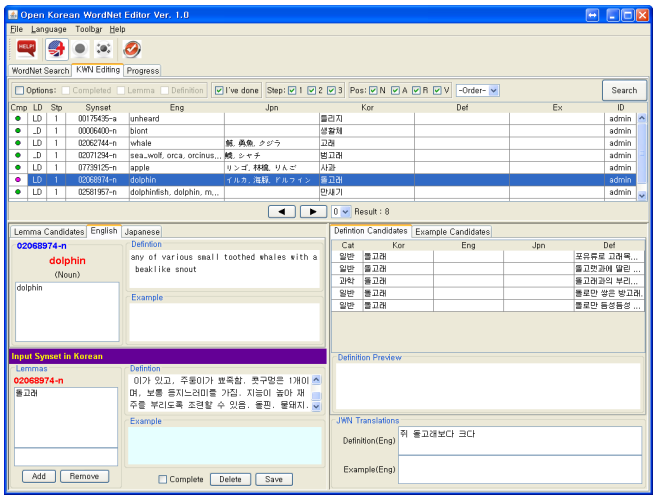

그림 4. KWN 구축 시스템 인터페이스 화면

4. 결과 및 결론

본 연구에서는 노동 집약적인 워드넷의 개발 과정을 지원하기 위해 다양한 분야의 대역사전에 기반하여 KWN을 반자동으로 구축하는 방법을 제안하였다. 그리고 한국어 어휘에 대한 다수의 사전 정의문으로부터 해당 synset의 구조정보를 이용하여 사전 정의문 후보들을 적합도에 따라 정렬하고 추천하는 정의문 추천 알고리즘을 제안하였다. 또한 다수의 전문가가 협력하여 KWN을 구축하기 위한 환경을 제공하기 위해 KWN 구축 시스템을 개발하였다. 제안한 방법에 기반하여 PWN 3.0의 synset 중 약 54 %에 대해 한국어 어휘를 자동으로 결정하여 KWN을 구축하였다.

본 연구에서는 한국어 워드넷의 개발 초기 단계로써 총 4 단계 중 1단계를 수행한 결과만을 보였다. 그러나 개발한 시스템을 이용하여 KWN의 개발 속도를 높일 수 있을 것으로 생각하며, 향후 KWN 구축 전문가 팀을 구성하여 나머지 단계의 작업을 수행할 예정이다. 따라서 제안한 방법에 대한 검증은 다음 연구로 남겨둔다.

사사

본 연구는 미래창조과학부 및 정보통신기술연구진흥센터의 정보통신•방송 연구개발사업의 일환으로 수행하였음 [10044494, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발]

참고문헌

[1] Christiane Fellbaum, Wordnet: An Electronic lexical Database. MIT Press, Cambridge, 1998.

[2] Piek Vossen, EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers Norwell, 1998.

[3] Key-Sun Choi and Hee-Sook Bae, "Procedures and Problems in Korean-Chinese-Japanese Wodnet with Shared Semantic Hierarchy," In Proceedings of Global WordNet Conference, pp.91-96, 2004.

[4] Dong Zhen Dong, "Knowledge Description: What, How and Who?," In Proceedings of the International Symposium on Electronic Dictionaries, Tokyo, Japan, 1998.

[5] Roberto Navigli and Simone Ponzetto, "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network," Artificial intelligence. Elsevier, 2012.

[6] Global WordNet Association, http://www.globalwordnet.org/

[7] WordNet: A lexical database for English, http://wordnet.princeton.edu/

[8] Japanese WordNet,

http://nlpwww.nict.go.jp/wn-ja/index.en.html