웹 검색을 활용한 워드넷에서의 IT 전문 용어 확장

박경국, 이광모, 김유섭
한림대학교 컴퓨터공학과
{pkkstory, kmlee, yskim01}@hallym.ac.kr

Wordnet Extension for IT terminology Using Web Search

Kyeong-Kook Park, Kwang-Mo Lee, and Yu-Seop Kim
Dept. of Computer Engineering, Hallym University

요 약

본 연구에서는 기존 워드넷에 등록되지 않은 IT 전문 용어와 같은 신규 용어들을 웹 검색을 사용하여 워드넷에 추가 시켜 확장시키는 설계를 하였다. 워드넷은 단어 간의 관계를 표현하는 어휘 사전이지만 일반적인 단어들로 구성되어 있고 새로이 등장하는 전문 용어는 포함하지 않는 경우가 많아 이러한 용어들을 새로이 워드넷에 등록함으로써 워드넷을 확장해야 한다. 이 작업은 웹 검색 결과를 분석하여 이 용어와 관련 깊은 용어들을 찾아서 워드넷에 없는 용어들을 워드넷에 추가시킴으로써 이루어 진다. 웹 검색 결과 문서를 형태소 분석기를 사용하여 가중치가 높은 순으로 관련 단어들을 찾고 이들 중 워드넷에 등록되어 있는 단어를 찾아 해당 단어의 하의어로 신규 단어의 위치를 배치시킨다.

Abstract

In this paper, we designed a methodology to expand the WordNet. We added unknown terms like IT technical terms to the existing WordNet by using web search. The WordNet is an online taxonomy representing the relationships among terms, but it usually showed limitation to contain new technical terminologies. That's why we tried to expand the WordNet. Firstly, when we met unregistered terms in WordNet, we built a query of those terms for web search. Given a web search results, we tried to find out terms with a high-level relatedness with the unregistered terms. We used the Korean Morphological Analyzer to score the relatedness between terms and located the unregistered term as a hyponym of terms with high score of relatedness.

1. 서론

일반적으로 사전은 현재 생활에서 통용되고 있는 단어들을 토대로 만들어 진다. 하지만 사회의 급격한 변화로 수 많은 새로운 분야들이 발전하고 이에 따라서 새로운 단어들이 지속적으로 생겨나기 시작했다. 이러한 신규 용어들은 매우 빠른 속도로 발전하게 되고 이러한 발전 속도를 일반 사전들은 거의 반영하지 못한다. 따라서 이들 신규 용어들은 전문 사전을 통하여 뜻을 이해하게 되어 활용할 수 있다.

웹 검색 사이트에서 단어를 찾으면 그 단어와 관련된 단어들과 함께 결과가 나온다. 이중에는 관련 깊은 단어도 있고 없는 단어들도 있다. 하지만 단어 빈도수나 가중치로 단어의 값을 알아내면 검색 단어와의 연관성 정도를 알 수가 있다. 대부분 함께 나오는 단어들은 연관도가 높다고 본다.

워드넷(WordNet) [1] 은 단어 간의 관계를 나타내는 신뢰성이 높은 사전이다. 단어 간의 친밀도를 알 수가 있으며 비슷한 분류로 나눠져 있으므로 단어가 어느 분류에 속하는지 측정할 수가 있다. 하지만 워드넷은 보편적인 단어들로 구성되어 있는 반면에 새로이 정의되는 전문 용어는 정의되지 않는 경향이 있다. 따라서 워드넷에 전문 용어를 추가시킴으로써 워드넷을 확장할 필요가 있다. 전문 용어를 가지고 웹 검색을 하면 수많은 관련 단어들이 결과 페이지에 등장하며, 이를 토대로 이 전문 용어와 관련이 있는 일반 용어를 찾을 수가 있다. 이들 일반 용어 중에서는 해당 전문 용어와의 관련도가 매우 높은 용어도 있으며, 그렇지 않은 용어들도 있다. 이러한 관련도는 형태소 분석 엔진에서 제공하는 계산을 통하여 얻을 수 있으며, 관련도가 높은 일반 용어의 하의어로 전문 용어를 등록시킨다.

본 논문은 다음과 같은 구조를 가진다. 2절에서는 관련연구로 워드넷과 한국어 어휘의미망에 대하여

설명하고, 3절은 워드넷에 등록되지 않은 전문 용어와 관련이 있는 용어를 찾는 방법을 설명한다. 4절에서는 이를 토대로 한 워드넷 확장 방법을 설명하고, 5절에서 그 결과를 보여준다. 마지막으로 6절은 결론 및 향후 과제를 논한다.

2. 관련연구

2.1 워드넷과 한국어 어휘의미망

워드넷(Wordnet)은 1985 년 Princeton 대학 인지과학연구실의 심리학자, 언어학자, 전산학자 등이 중심이 되어 구축해온 단어 간의 관계를 표현하는 영어 어휘 데이터베이스이다. 기존의 사전이 가나다순으로 제작된 것과는 달리 워드넷은 개념을 바탕으로 네트워크를 구축한 대용량 지식베이스 이다[1]. 일반적인 시소러스에는 용어간 관계를 나타내기 위해 상하위 관계, 동등관계, 부분-전체관계, 연관관계, 사례관계 등을 이용하며 워드넷에서는 동의관계, 반의관계, 상의관계, 하의관계, 분의 관계, 양식관계, 함의관계를 이용하였다. <표 1>은 워드넷의 어휘들을 보여주고 있다.

<표 1> 워드넷 2.1 버전의 통계

구분	단어	신셋
명사	117097	81426
동사	11488	13650
형용사	22141	18877
부사	4601	3644

워드넷을 바탕으로 세계 각국에서 각국의 언어로 확장하는 연구가 진행되고 있으며 대표적으로 유럽은 유럽공동체가 개설한 유로넷[2]을 운용하고 있다. 국내에서는 유로넷 확장을 위한 방법[3]이나 부산대학교 한국어정보처리연구실에서 한국어 어휘의미망 (KorLex)을 구축하는 연구를 진행하고 있다[4]. 연구가 진행 중인 한국어 어휘의미망은 1 차로 영어 워드넷을 영-한 대역하여 구축한 다음 2 차로 문제점을 분석하고 유형화하였다. 3 차로 구축된 영-한 대역 어휘의미망을 정제하여 한국어 어휘 의미망을 구축하였다. 이렇게 구축된 어휘의미망은 여러 분야에서 활용될 수 있다. 정보검색, 자동번역, 문장 분석 등과 온톨로지를 구축하기 위한 기반으로 사용될 수 있다 [5].

본 논문에서는 부산대학교 한국어정보처리연구실에서 영-한 대역 어휘 의미망을 정제하여 만든 신셋(Synset)를 쓴다. 이 신셋은 상하위 관계를 갖는 워드넷의 명사를 번역하고 수정한 데이터이다. 신셋은 신셋번호, 영어, 상위번호, 하위번호, 번역어로 구성되어 있고 상하위 번호에 의하여 단어에 대한 깊이를 알 수 있다. 예를 들어, <표 2>를 보면 상하위 관계는 객체지향 프로그래밍언어를 기준으로 상위에는 '프로그래밍언어', 하위에는 '자바'로 구성된 것을 알 수 있다.

<표 2> 신셋의 상하위 관계

상위단어	기준	하위단어
프로그램밍 언어	객체지향프로그램	자바

그래서 신셋은 검색 키워드에 대한 관련 용어들을 찾는 데이터가 된다. 하지만 컴퓨터에 관련된 용어는 워드넷에 드물게 나타나기 때문에 IT 용어를 추가시켜 용어검색을 원활하게 만들어야 한다.

2.2 형태소 분석기

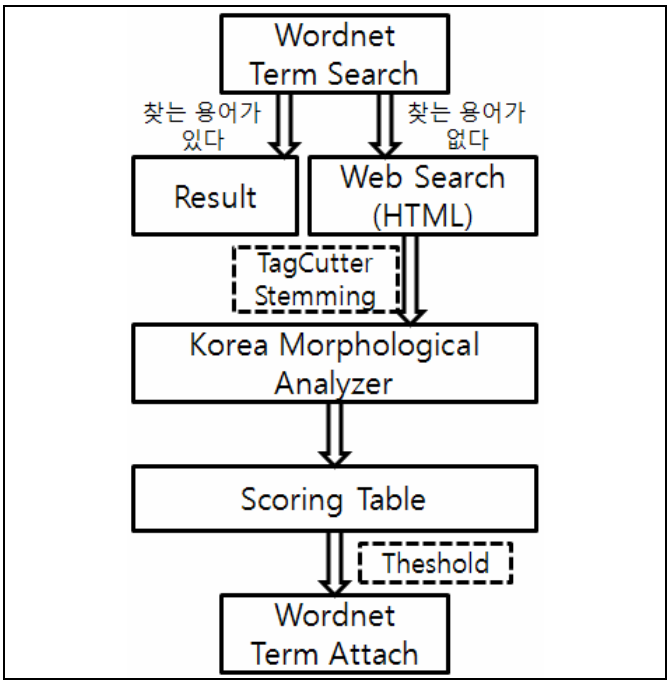
형태소는 일정한 음성에 일정한 뜻이 결합되어 있는 말의 가장 작은 단위이다. 형태소 분석은 주어진 문장으로부터 최소 의미 단위인 형태소를 추출하는 과정을 말한다. 형태소 분석기는 입력을 받아서 어절을 구성하고 있는 각 형태소를 인식하여 관련 정보를 사전으로부터 탐색 및 선정하여 결과를 적절한 구조로 출력시키는 프로그램의 총칭한다.

국민대학교 형태소 분석기에는 용어 가중치를 사용하는데, 용어 빈도에 관련하여 자주 출현하는 용어들의 분포를 이용하고, 그 이외에 품사 유형 및 어절 위치 등을 고려하여 문서내의 용어 중요도를 계산한다[6]. 이러한 용어 중요도는 주어진 어휘의 웹 검색 결과에서 질의어로 사용된 어휘와의 관련도를 나타낼 수 있다. 본 논문에서 쓰인 분석기는 국민대 한국어 형태소 분석기(KTL Version 2.1.0b)로 자동색인, 한글 맞춤법 검사/교정, 복합명사 분해, 한글 자동 띄어쓰기 기능을 갖추고 있다[7][8].

3. IT 용어 관련 검색

3.1 워드넷 검색

워드넷 확장을 위하여 워드넷에서 단어의 유무를 확인해야 한다. 단어가 있으면 현 상태를 유지하고 없으면 단어를 추가시켜야 하기 때문이다. 웹 검색을 하면 단어들에 대한 내용이나 관련 단어들을 많이 나온다. 웹에서 나온 단어들은 신뢰성이 부족한 것도 있지만 주로 쓰이는 단어들이 주를 이루므로 데이터로 쓰기에 부족하지는 않다. 예를 들어, 최근 웹 2.0 이 활성화 되면서 '태깅', '롱테일', '집단지성' 등 새로운 용어가 생겨났는데, 워드넷에는 이러한 최근 단어가 없다. 따라서 IT 사업이 발전함에 따라 신규 용어에 대한 추가가 필요한 것이다. 본 논문에서는 워드넷에 없는 IT 전문 용어를 찾고 없으면 추가시킨다. <그림 1>은 용어 추출 과정과 워드넷에 추가하는 것을 표현했다.



<그림 1> 웹에서 용어 추출 과정과 워드넷 추가

3.2 웹에서 단어 추출

웹 검색 창에서 IT 용어를 찾으면 수 많은 단어와 관련 사이트 내용들이 나온다. 페이지 형식은 HTML 형식으로 출력이 되는데 데이터로 쓸 수 있는 단어들만 추출하여야 한다. HTML 은 태그(Tag)로 만들어진 언어이다. 태그를 제거하여야 하고, 그 외에 스크립트 언어, 불필요 문자들이 있으므로 이것 또한 제거해야 한다. TagCutter 과정에서 HTML 태그와 불필요한 단어, 영문자, 특수문자, 숫자들을 제거 하여 한글단어만 남도록 하였다. HTML 문서는 검색시 URL 주소로 웹 문서 소스를 가져와서 쓰는데, 주요 검색 사이트인 DAUM, YAHOO, GOOGLE, NAVER 등에서 문서를 가져온다. <그림 2>는 HTML 문서에서 태그를 제거한 것이다.

<table border=0 cellpadding=0 cellspacing=0><tr><td class="j hc"><font size=-1>그런데 도대체 잘 된 태깅이란 무엇을 말하는 것일까? 혹은, 태깅을 잘 하려면 무엇이 필요한가? 1. 멀티 카테고리로서의 태그 ...	그런데 도대체 잘 된 태깅이란 무엇을 말하는 것일까? 혹은, 태깅을 잘 하려면 무엇이 필요한가? 1. 멀티 카테고리로서의 태그 ...

<그림 2> HTML 문서에서 태그 제거 후 문서

3.3 불용어 제거

TagCutter 과정에서 태그가 제거 된 단어들만 남지만, 남은 단어들 중에도 중복되거나 형용사, 관사, 특수문자 같은 단어가 있다. TagCutter 에서 완전하게 없앨 수는 없다. 다음 과정은 스태밍(Stemming)을 해야 한다. 일반적으로 스태밍은 불용어 제거와 원형 복원을 하여 단어를 추출하고 한글 같은 경우 형태소 분석을 하여 복합명사 분해, 단순명사 추출 자동 뛰어쓰기 등을 한다. 하지만 본 논문에서 글자에 대한 분석이 아니고 용어만을 중점을 두기 때문에 스태밍 과정은 TagCutter 에서 제거하지 못한 중복 단어나 불필요 단어들을 다시 없애고 순수하게 필요한 단어들만 남기도록 하였다.

3.4 형태소 분석

TagCutter 와 Stemming 과정이 끝나고 순수 단어들을 국민대 형태소 분석기로 실행을 시키면 단어 가중치에 대한 값이 큰 순서부터 출력이 된다.

```
==============================
 국민대 형태소 분석기에서 "태깅" 스코링
==============================
 번호    빈도    가중치   단어
 ------------------------------
1        22      1000     태깅
2        11      498      블로그
3        5       299      현재창
4        11      275      검색
5        9       265      사이트
6        4       220      검색결과
7        8       212      결과
8        3       128      동영상
9        4       122      태그
10       2       110      연합뉴스
 ------------------------------
```

<그림 3> 국민대 형태소 분석기 “태깅” 결과

Scoring Table은 번호, 빈도, 가중치, 단어, 위치정보가 나오는데 필요한 데이터는 가중치의 크기와 단어를 기준으로 워드넷에 추가하므로 위치정보는 제외되고 번호, 빈도, 가중치, 단어를 주요 데이터로 추출하고 가중치의 순위에 의하여 새로운 용어의 상위 단어를 결정 한다. <그림 3>는 “태깅”에 대한 국민대 형태소 분석기 결과와 위치정보를 제거 후 Scoring Table 이다. Scoring Table에 나온 가중치는 현 페이지의 단어 중요도를 나타내고 워드넷에 추가 될 단어들의 후보가 된다.

4. 워드넷 확장

새로운 용어는 형태소 분석기에서 나온 가중치가 높은 단어를 기준으로 하위에 추가시켜야 한다. 분석기에 1순위인 단어는 가중치가 1000인데 이것은

보통 찾고자 하는 단어가 되는 경향이 있다. 그래서 그 다음으로 가중치가 높은 단어를 찾아 하위단어로 추가하면 된다. 만약 없으면 다음 가중치가 높은 것을 찾아 그 단어의 하위 단어로 현 단어를 삽입을 시킨다. 가중치가 높은 단어가 새로운 명사의 상위어가 될 수 있는 것은 현 페이지에서 빈도수가 높다는 것을 나타내며 찾는 용어와 관련성이 가깝다는 것을 보여준다. 하지만 가중치가 낮아질수록 관련 없는 단어들이 나타나게 된다. 그래서 상위단어에서 후보단어를 찾아야 한다. 예를 들어, 워드넷에서 '태깅'을 찾으면 신생 IT용어이므로 없다. <그림 3>의 스코링 테이블에 나온 결과를 보면 '태깅' 다음으로 가중치가 높은 단어가 '블로그'이다. 이것은 '태깅'과 '블로그' 사이에 관계가 높다는 것을 의미한다. 그래서 '블로그'의 하위 단어로 '태깅'을 추가하면 되는 것이다. 실제로 블로그에서 태깅은 많이 쓰여지고 있고 웹의 주요 검색 대상이기도 하다.

4.1 워드넷 확장 구현

부산대에서 만들어진 워드넷의 신셋 형식은 신셋번호, 영어, 상위번호, 하위번호, 번역어로 구성되어 있다. 새로운 용어에 대하여 신셋 형식을 갖추기 위하여 신셋번호가 우선 만들어져야 한다. 현재 신셋번호는 워드넷의 상하관계를 의미간의 거리가 측정되어 생성되어 있다. 하지만 본 논문에서 IT용어만을 선정하여 추가시키는 것이므로 단어간의 거리 측정보다는 상하위관계를 중요시한다. 그래서 새로운 신셋번호는 마지막 신셋번호에서 10씩 카운트 하여 일정한 간격을 두었다. 찾는 새로운 용어에 대해서는 영어와 번역어를 같이 입력하고, 새로운 용어의 신셋 상하위 번호는 분석기에서 나온 스코어링 테이블에서 선택된 단어의 하위번호로 넣고, 새로 생긴 신셋 형식의 상위번호는 선택 단어의 신셋번호를 상위번호로 지정하였다. 예를 들어, '태깅'을 워드넷에 추가하려면 워드넷 신셋의 형식을 갖춘 후에 '태깅'의 신셋번호를 '블로그'의 하위 번호로 추가시키고 '태깅'의 상위번호로 '블로그'의 신셋번호로 설정하면 된다. <그림4>은 자바, 워드넷 신셋, 웹문서를 형태소 분석하여 나온 스코링 테이블을 이용한 실험의 결과로 워드넷에 새로운 용어가 추가된 것을 보여준다.

```
================================
 "태깅"  워드넷에 추가 필요
 "블로그" 하위단어로 " 태깅"  워드넷에 추가
 --------------------------------
 "블로그"  워드넷에 단어 유무 확인
 ================================
 "블로그"  워드넷  신셋 결과
 --------------------------------
```

```
=== 1 ===
SynsetID= 06004942
EngWord= web_log, blog
Hypernym= 06004789_n_0000
Hypornym=
KorWord= 웹로그, 블로그
--------------------------------
신셋에 새로 생긴 라인
14434750        tagging   06004942_n_0000          태깅
--------------------------------
-"태깅" 상위 번호(웹로그, 블로그 의 하위 번호): 06004942
-"웹로그, 블로그" 의 하위 번호 : 14434750
-워드넷에 새로 생긴 용어      : 태깅
```

<그림4> 워드넷에 새로운 용어 추가 결과

5. 실험 및 평가

실험에서는 워드넷에 추가 시킬 IT용어 중 Web2.0과 관련된 용어인 '롱테일', '웹2.0', '와이브로' 등을 대상으로 이루어졌다. 웹 문서에서 형태소 분석기를 사용하여 나온 결과들을 보면 대체적으로 찾는 용어와 단어들이 서로 관련이 있다는 것을 <그림 5>를 통해서 알 수가 있다. 웹 문서에서 단어의 빈도수와 형태소 분석기의 단어 점수를 계산하여 나온 결과이며 가중치가 높은 것은 찾는 용어와 밀접한 단어이며 낮아질수록 관련 없는 단어가 된다. 간혹 전혀 관련이 없는 결과가 나타나기도 하는데 이것은 형태소 분석기가 웹 문서에서 발생하는 단어에 대해 완전하게 분석할 수 없다는 것을 확인해준다. 형태소 분석기는 일반적인 텍스트 문서에서 효력을 발휘하지만 웹 문서는 전혀 관계없는 단어들로 인하여 다른 단어에 가중치가 높아져버린다. 본 논문은 웹 문서에서 반복해서 나타나는 불필요 단어를 제거했지만 어느 정도 한계가 있었다. 형태소 분석을 하면 300~600 이상의 단어가 추출되며 온전한 단어가 아닌 형태가 분리된 단어도 나오기 때문이다. 예를 들어, <그림 5>에서 "유비쿼터스"의 단어를 보면 "유비"와 "쿼터스"가 분리되어 관련 없는 단어가 되는데 아직 "유비쿼터스"와 "유비 쿼터스"처럼 정확한 단어가 정립되지 않아서 나타난 것이다.

용어	가중치	단어
웹2.0 (Google)	1000	서비스
	482	인터넷
	437	디지털
	340	기업
와이브로 (Google)	1000	인터넷
	444	통신
	414	서비스

	347	기술
롱테일 (DAUM)	1000 654 654 653	롱테일 경제 사회 고객
유비쿼터스 (DAUM)	1000 843 817 552	정보 유비 쿼터스 전자

<그림 5> 웹문서에서 형태소 분석기 결과

6. 결론 및 향후 과제

본 논문은 시대가 변함에 따라 새로운 IT전문용어들이 발생하고 기존의 워드넷에 용어에 대한 추가를 함으로써 워드넷의 확장을 위한 방법을 제안하였다. 새로운 용어는 웹 문서에서 나온 단어를 정제하여 형태소 분석기를 통해 나온 단어의 가중치로 워드넷에 추가시킬 단어를 선정하고 워드넷 신셋 형식을 갖추어 상하위 관계를 가지는 IT전문용어들을 추가시켰다. IT용어뿐만 아니라 다른 분야의 전문용어도 같이 확장이 되면 워드넷은 일반 단어뿐만 아니라 전문용어도 포함하는 전문 어휘 사전이 될 것이다. 하지만 웹 검색 문서는 각 검색사이트마다 매번 그 결과가 다르기 때문에 일정한 용어에 대한 데이터를 가져오기가 어렵고 형태소 분석기로 단어의 가중치를 구하는 방식은 일반적인 텍스트 문서에서 사용되므로 웹 문서를 이용하는 방법에는 다소 무리가 있었다. 효과적인 워드넷 확장을 위해서는 전문용어와 새로운 용어에 대한 분석을 할 수 있는 분석기를 통하여 워드넷에 추가시키는 방법에 대한 연구가 필요하다.

참고 문헌

[1]Miller,G.A.,"WordNet", http://wordnet.princeton.edu

[2]Vossen, p.(1998). “Introduction to EuroWordNet”, Computers and the Humanities, Special Issue on EuroWordNet, 73-89.

[3]오장근, “정보검색 활용을 위한 다국어 데이터베이스의 ILI(Inte-Lingual-Index) 방법론 연구 - 유로워드넷의 ILI 확장과 관련하여-“, 언어과학연구 제29집, 181-208, 언어과학회 2004.

[4]황순희, “어휘의미망의 이해와 명사 어휘의미망 구축”, 어휘의미망의 이해와 응용 세미나, Aug. 2005

[5]정홍석, “용어추천기능을 가진 온톨로지 편집기의 설계와구현”,http://klpl.re.pusan.ac.kr/graduates/hsjun g/

[6]강승식, 이하규, 손소현, 홍기채, 문병주, “조사 유형 및 복합명사 인식에 의한 용어 가중치 부여 기법”, 한국정보과학회 가을 학술발표논문집, Vol.28, No.2, pp.196- 198, 2001.

[7]이경찬, 강승식, "용어 가중치와 역범주 빈도에 의한 자동 문서 범주화", 제15회 한글 및 한국어 정보처리 학술발표 논문집, pp.14-17, 2003년 10월.

[6]HAM, 한국어 형태소 분석기와 한국어 분석 모듈, 국민대학교 자연언어 정보검색연구실, http://nlp.kookmin.ac.kr.