정보 검색을 위한 동의어/유의어 사전 구축

충북대학교 컴퓨터공학과
이 태 우† · 서 영 훈

A Synonym Dictionary Construction for Information Retrieval

Tae-Woo Lee,† Young-Hoon Seo
Department of Computer Engineering, Chungbuk National University, Chungbuk, Korea

요 약

본 논문에서는 많은 정보로부터 의미 있는 정보를 추출하기 위해 사용되는 정보 검색 시스템에서 이용이 가능한 동의어/유의어 사전을 구축하고 구축된 정보의 평가를 수행하였다. 사용한 자원으로는 미리 구축된 한-영 사전과 영-한사전을 이용하였다. 이들의 사용으로 다른 동의어사전과 달리 보다 많은 유의어 정보를 포함하는 이익을 얻었다. 본 논문의 시스템은 사전을 구축하기 위해 기본 자원을 이용하여 동의어/유의어 후보 목록들을 획득하고, 획득된 정보를 바탕으로 후보 목록의 빈도수와 사전의 위치 정보, 마지막으로 입력 명사 정보를 이용하여 동의어/유의어를 확정한다. 작성된 동의어/유의어사전은 한-영사전에 수록된 한국어 명사 64,630개를 대상으로 하였다. 작성된 사전을 문서 필터링 시스템에 추가하여 적용 전 보다 성능이 향상됨을 확인하였다. 또한 질의 색인어 확장에 이용하여 보다 정답을 추출하는데 추가적으로 확장된 유의어 정보가 정답을 추출하는 데 유용하게 사용됨을 확인하였다.

서 론

인터넷을 통한 정보 교류(information exchange)가 보편화되면서 사용자에게 제공되는 정보의 양은 많아졌으나, 그만큼 상대적으로 사용자가 원하는 가장 적절한 정보의 검색은 점점 더 어려워지고 있다. 이러한 문제를 해결하기 위해선 사용자가 방대한 양의 정보 속에서 정보의 습득 시간을 줄이고 검색된 정보의 내용이 찾고자 하는 정보와 일치하고 있는지를 사용자로 하여금 신속히 판단할 수 있는 방법이 필요하다. 이처럼 효과적으로 정보의 습득 시간을 줄일 수 있는 방법의 일환으로 많은 정보 검색 방법들이 있고 특히 이러한 정보 검색 방법들은 사용자의 요구에 만족하는 정보를 결정하기 위해 사용자의 질의에 의존한다. 따라서 질의에서 사용자의 요구를 나타낼 수 있는 정확한 색인어의 추출과 가중치 부여는 정보 검색 기법에서 매우 중요하다.

†E-mail : dickgap@white.chungbuk.ac.kr
E-mail : yhseo@cbucc.chungbuk.ac.kr

요즘 많이 연구되는 질의 응답 시스템[1] 및 문서 요약, 분류, 필터링(filtering)[2] 등과 같은 정보 검색에 유용하게 쓰일 수 있는 방법 등을 통해 정보의 유용성 여부를 판단하는 시간을 절약할 수 있다. 이러한 기법들은 일반적으로 색인어라고 불리는 핵심적인 키워드들을 이용하여 해당 과정을 수행한다. 한국어의 경우 색인어는 대부분 개념적 중요도가 높은 명사 위주로 추출된다. 예를 들어 질의 응답 시스템의 경우 사용자의 질의로부터 색인어를 추출하고 그 정보를 이용하여 검색된 정답 후보 가운데 가장 정답에 근사한 내용을 정답으로 추출한다. 또한 색인어를 확장시키고 그 정보를 이용하여 보다 정확한 정답을 구하고 있다. 이는 정답을 추출할 수 있는 정보를 많이 가지면 가질수록 보다 정확한 정답을 획득할 수 있기 때문이다. 이 같은 이유로 대부분의 정보 검색 기법에서는 워드넷(WordNet) 등과 같은 지식 베이스를 통해 색인어를 확장시키고 확장된 내용을 이용하여 결과를 얻는데 사용한다.

본 논문에서 제안하고 있는 동의어/유의어 사전 역시 색인어, 즉 핵심적인 내용을 담고 있는 키워드인 명사를 대상으로 동의관계이거나 유의관계에 있는 정보를 추출하여

정보 검색에서 보다 좋은 결과를 획득하기 위하여 제안되었다. 이 같은 동의어/유의어 사전을 구축하기 위해 해당 단어에 대하여 같은 뜻의 대역어 리스트가 존재하는 한-영사전과 영-한사전을 이용하였다. 작성된 동의어/유의어 대상 명사는 복합명사를 포함하고 있는 한-영사전에 수록된 명사를 대상으로 작성되었다.

일반적으로 정보 검색 기법에 사용되는 지식베이스를 구축하기 위해서는 많은 비용과 노력이 필요하다. 의미의 상-하위 관계 및 동의어 반의어 정보까지 포함한 하나의 워드넷을 만들기 위해서는 많은 인원과 기간이 소요된다. 하지만 본 논문에서 제안하고 있는 동의어/유의어사전의 경우 추가 정보를 제공하면서도 자동으로 구축이 가능하여 비용과 노력이 상대적으로 적다는 장점을 가지고 있다.

본 논문의 2장에서는 전체적인 시스템 구성도를 통해 본 논문에서 제안하고 있는 동의어/유의어사전을 구축하기 위해 사용된 자원과 구축 방법을 살펴보고, 3장에서는 사전의 구축 결과와 평가를 내린다. 4장에서는 작성된 사전을 실제적으로 정보 검색 기법에 사용하여 어느 정도의 성능 향상 효과가 있는지 살펴 보았으며, 그 결과에 따른 고찰 및 평가를 토대로 결론 및 향후 개선, 연구 과제를 5장에 기술한다.

시스템 구성

서론에서 언급했듯이 많은 정보 속에서 사용자가 원하는 정보를 획득하기 위해 많은 정보 검색 기법들이 존재하고 이 기법들은 색인어 확장을 통해 해당 작업을 수행한다. 이러한 색인어 확장을 위해 지식 베이스가 사용된다. 일반적으로 색인어 확장 등에 많이 사용되는 워드넷의 경우 해당 단어에 대하여 동의관계에 있는 정보뿐만 아니라 그 단어에 대한 상-하위 관계에 있는 정보를 포함하고 있는 경우가 많다. 이러한 정보를 구축하기 위해서 많은 비용과 노력이 필요하고 특히 동의/반의 관계에 있는 정보보다 상-하위 관계 정보를 구축하는데 더 많은 비용과 노력이 필요하다. 그러나 더 많은 비용이 드는 상-하위 정보의 경우 구축된 노력에 비해 정보 검색 기법에 사용되어 큰 성능 향상을 얻지 못하고 있다. 따라서 본 논문에서 제안하고 있는 동의어/유의어사전은 단어의 상-하위 정보가 아닌 해당 단어의 동의/유의정보에 초점을 맞추고 있다.

1. 사용 자원

동의어/유의어 사전을 구축하기 위해 기본자원으로 한-영사전과 영-한사전을 이용하였다. Fgi. 1과 2에서 보는

```
cost-of-living index;소비자 물가 지수;
cost;가격;원가;대가;비용;지출;경비;소비;희생;손실;고통;소송 비용;
costa;늑골;앞의 중특맥;중앙맥;
costar;공연 스타;공연자;
costard;큰 사과;&대가리;
coster;과일 행상인;생선 행상인;
costermonger;과일 행상인;생선 행상인;
costing;원가 계산;
costiveness;변비;께께함;동작이 둔함;우유 부단함;
costliness;값이 비쌈;비용이 많이 듦;희생이 큼;타격이 큼;호화로움;
costotomy;늑골 절제;늑골 절제술;
costume;복장;복식;의상;몸차림;차림새;여성복;슈트;수영복;
```

Fgi. 1. 영-한사전.

```
건망증;forgetfulness;short memory;loss of memory;amnesia;
건물;building;structure;edifice;dry foods;groceries;
건물상;grocery;grocer's;drysaltery;drysalter;
건반;keyboard;clavier;manual;
건반사;tendon reflex;
건반악기;keyboard instruments;clavier;
건밤;sleepless night;
건백;petition;representations;
건빵;hardtack;cracker;biscuit;
건삼;dried ginseng;
건설;construction;building;erection;establishment;
건설계획;construction plan;
```

Fgi. 2. 한-영사전.

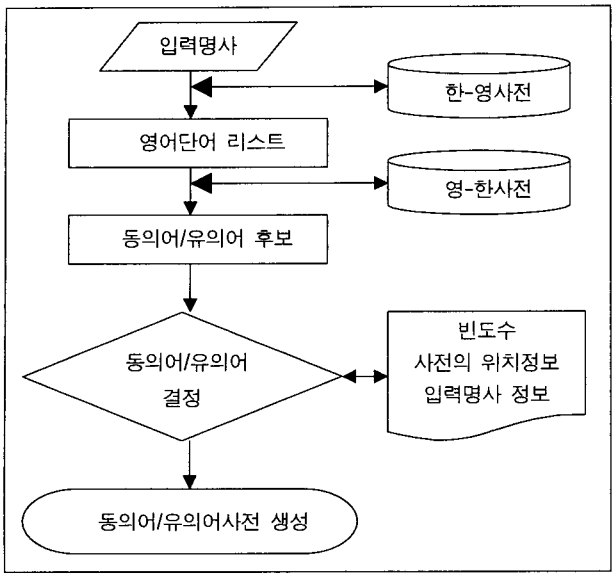
바와 같이 각 사전은 해당 단어에 대하여 같은 의미를 가지고 있는 대역어 단어들의 형태로 작성되었다. 동의어 정보와 확장시킨 유의어 정보를 포함시키기 위해 사전에 나타나 있는 동의관계에 있는 모든 대역어를 모두 리스트에 포함 하였다. 또한 일반적으로 사전을 구성할 때 가장 많이 쓰이면서도 중요한 단어가 맨 처음에 나타나는 경향을 보이는데 이를 반영하기 위해 나열 순서는 중요도 순으로 구성하였다.

이러한 기본 자원의 사용으로 몇 가지 장점을 얻을 수 있다. 영-한사전이 가지고 있는 특징으로 하나의 영어 단어에 대하여 비슷한 뜻을 지닌 여러 개의 한글 단어들이 존재하게 된다. 즉 하나의 단어를 통해 의미가 비슷한 많은 단어들을 획득할 수 있다. 이 같은 장점으로 단순히 동의어만이 아닌 개념을 확장시켜 유의어 관계에 있는 단어까지 사전에 포함시킬 수 있다는 큰 장점을 얻을 수 있다.

추가적으로 한국어에 나타나는 동음이의어의 경우는 각 의미에 대하여 개별적으로 서로 다른 엔트리로 구성하여 구축하였다.

2. 시스템 구성도

본 논문에서 제안하고 있는 동의어/유의어사전을 구축하기 위해 구축한 시스템은 크게 동의어/유의어 후보 생성 부분과 동의어/유의어 결정 부분으로 나누어진다. 첫 번째



Fgi. 3. 시스템 구성도.

단계인 동의어/유의어 후보 생성 부분에서는 기본 자원인 한-영사전과 영-한사전을 이용하여 사전의 엔트리에 존재하는 모든 단어들이 동의어/유의어 후보로 생성이 된다. 이렇게 생성된 후보들을 빈도수와 사전의 위치 정보, 색인어 정보를 가지고 동의어/유의어를 평가하고 평가된 결과를 사전으로 구축한다.

각 단계를 자세히 입력 명사에 대하여 동의어/유의어사전을 구축하는 과정은 다음과 같다. 우선 입력 명사에 대하여 한-영사전을 검색하고 그 단어와 같은 뜻을 가진 대역어 명사들의 목록을 획득한다. 영어단어 리스트 목록을 작성하기 위하여 기본 자원인 한-영사전을 사용하고 입력명사를 한-영사전에 나타나 있는 엔트리와 비교하여 같은 뜻을 지닌 대역어 단어를 획득하여 이 과정을 수행한다. 다음으로 시스템 구성도에 나타나 있는 동의어/유의어 후보 목록을 생성하는 과정에서는 영-한사전을 이용하여 수행한다. 전 단계를 거쳐 얻어진 영어단어 리스트를 대상으로 영-한사전 엔트리와 비교하고 일치되는 목록에 존재하는 한국어 명사들을 획득한다. 이렇게 획득된 모든 한국어 명사들을 동의어/유의어후보로 선택한다. 이 과정을 예를 들어 살펴보면,

'입력명사로 가격' 이라는 단어가 입력되었을 때 한-영사전의 엔트리에 존재하고 있는 '가격' 에 대하여 'price', 'cost', 'value', 'worth' 라는 영어 단어가 엔트리에 존재하고 있다. 위 네 개의 단어가 영어단어 리스트로 선택된다. 선택된 영어단어 리스트를 기본으로 하여 동의어/유의어 후보를 생성하는 부분에서는 모든 영어단어 리스트들을 대상으로 영-한사전을 검색한다. 각 단어를 영-한사전을

price:가격:대가:값:시세:물가:시가:대상:희생:건 돈의 비율:상금:현상금;
cost:가격:원가:대가:비용:지출:경비:소비:희생:손실:고통:소송 비용;
value:가치:유용성:진가:쓸모:고마움:가격:값:경제 가치:가격에 합당한 물건;
대가:똑바에 찾아낸 물건:값싸게 손에 넣은 물건:가치기준:진의:의의;
음표가 나타내는 길이:시간적인 가치:명암;
worth:가치:값어치:~값만큼의 분량:~어치:재산;

Fgi. 4. 가격에 대한 2차 후보 단어.

검색하여 얻어진 2차 후보 단어의 리스트는 아래 Fgi. 4와 같다.

일련의 과정을 거쳐 '가격' 이라는 단어에 대한 2차 동의어 후보 목록이 Fgi. 4 처럼 얻어졌다. 이렇게 선택된 2차 후보 단어들은 동의어/유의어 결정 단계를 거쳐 동의어/유의어사전이 생성된다.

3. 동의어/유의어 결정 방법

기본 자원인 한-영사전과 영-한사전을 이용하여 동의어/유의어 후보 단어까지 획득한 후 동의어/유의어 결정 단계를 거친다. 이렇게 획득한 다수의 동의어/유의어 후보들 가운데 적합한 단어를 선택하는 단계가 된다. 이 과정에서는 빈도수와 사전의 위치정보, 입력명사 정보 등 이 세 가지 결정 방법으로 실제적인 동의어/유의어를 결정한다.

첫 번째 빈도수를 이용하는 방법에서는 전체 동의어/유의어 후보 단어들에 대한 빈도수를 조사하여 가중치를 부여하는 단계를 의미한다. 기본적으로 동의어/유의어 후보들은 두 개의 대역어 사전을 통해 획득된 단어이므로 입력단어와 동의어/유의어 관계에 있는 단어들이 대부분이다. 따라서 여러 영어 단어의 엔트리에서 나타난 뜻이 일치 되는 단어인 경우 입력 단어와 보다 더 동의어/유의어 관계에 있다고 판단할 수 있다. 실제적으로 예로 살펴본 '가격' 을 통해 빈도수를 조사 하였을 때 빈도수가 2이상인 단어들로는 입력명사인 '가격' 을 비롯해 '대가', '값', '가치','희생' 으로 나타난다. 이렇게 선택된 단어들 중 '희생' 을 제외한 나머지 단어들은 '가격' 과 동의어/유의어 관계에 있다고 판단할 수 있다. 물론 '희생' 이라는 단어의 경우 영-한사전이 가지고 있는 특징으로 얻어지는 불필요한 정보이다.

두 번째 결정 방법은 사전의 위치정보를 이용하는 것이다. 여기서 말하는 사전의 위치정보란 사전에 나타나는 단어들의 순서정보를 의미한다. 일반적으로 사전을 구성할 때 가장 많이 쓰이고 보편적으로 가장 색인어와 일치하는 단어 순으로 단어들을 나열하게 된다. 즉, 사전의 맨 처음에 나타나는 단어가 색인어와 가장 밀접한 관계가 있다는 의미가 된다. 이 정보는 동의어/유의어를 판단하는데 중요

한 정보를 제공하고 본 시스템에서는 사전의 맨 처음에 나타나는 단어에 가중치를 부여하여 동의어/유의어사전을 결정한다. 사전의 위치정보가 '가격'이라는 입력단어에 대하여 선택된 동의어/유의어 후보 단어들 중 영어단어 엔트리 단어들의 사전 검색 결과에서 제일 처음 나타나는 단어들인 '가격', '가치'란 단어가 된다. 하지만 '가격'의 경우 입력 단어와 일치하므로 그 다음 단어인 '대가', '원가'등이 추가로 선택되어 가중치를 부여 받게 된다.

마지막으로 입력명사 정보부분은 사전의 검색 결과와 입력명사를 비교하여 해당 단어에 가중치를 부여하는 기법이다. 즉 영-한사전을 검색하여 얻어진 엔트리에서 가장 처음에 나타나는 단어가 입력명사와 같을 경우 그 엔트리에 있는 후보목록 중 실험을 통해 가장 높은 성능을 보였던 30%에 해당하는 명사에 추가 가중치를 부여하는 것을 말한다. 이 가중치 부여 기법은 앞에서 언급한 사전의 가장 처음 나타나는 단어의 중요성에 기인하고 있다. 사전의 가장 처음 나타나는 단어가 실제적으로 가장 보편적으로 사용되면서 가장 중요한 의미이므로 그 단어가 입력단어와 같다면 그 하위 단어들도 입력명사와 동의어/유의어 관계에 있다고 판단할 수 있다. 위의 '가격'이라는 단어에 대한 예를 살펴보면 'price'와 'cost' 엔트리의 가장 처음 나타난 단어가 입력명사와 일치하고 있다. 따라서 이 두 개의 엔트리에 존재하는 후보들 즉, 'price'의 경우 '대가', '값', '시세'의 단어에 'cost'의 경우 '원가', '대가', '비용'이라는 단어에 추가 가중치를 부여하게 되는 것이다.

사전 구축 결과와 평가

본 시스템을 통해 자동으로 구축된 동의어/유의어사전의 평가를 위해 여러 가지 방법으로 실험을 해 보았다. 평가 방법은 정보 검색 기법의 평가 방법으로 많이 사용되는 f-score방법을 채택하였다. 실험을 위해 10명의 테스트 인원을 선정하고 전체 동의어/유의어사전에 존재하는 단어 중 15개의 테스트 단어를 선정하였다. 또한 3.에 동의어/유의어 결정 방법에서 제안한 가중치 부여 기법에서의 가중치를 같게 한 경우와 서로 차등을 주어 부여한 결과에 대한 평가도 이루어졌다.

실험을 평가했던 방법은 보편 타당한 동의어/유의어를 선택하기 위해 10명의 테스트 인원 중 8명 이상이 동의어/유의어로 선택한 단어를 대상으로 재현률과 정확률을 계산하였다. 정확률의 경우 시스템을 통해 얻어진 동의어/유의어사전의 단어 중 8명 이상이 동의어/유의어라고 판단한 단어의 비율을 의미한다. 재현률의 의미는 대역어 사전에 포함되어 있는 실제 동의어의 포함 비율을 의미한다. 이를 위해 대역어 사전의 단어들을 대상으로 실제 동의어/유의어를 추출하고 생성된 동의어/유의어사전에 어떤 비율로 포함되었는지를 계산하여 재현률을 판정하였다. 이렇게 얻어진 재현률과 정확률을 기반으로 한 F-score의 경우 다음과 같은 식을 이용하여 구하였다.

$$F - score = \frac{2 * (recall * precision)}{recall + precision}$$

recall :재현률
precision :정확률

Table 1에 나타난 재현률과 정확률은 15개 테스트 단어에 대한 평균값이다. 첫 번째 실험에서는 세 가지 동의어/유의어 결정 방법의 가중치를 1로 동일하게 주고 가중치가 2이상인 단어들을 동의어/유의어로 결정 하여 실험하였다. 이 실험의 결과 Table 1과 같은 재현률은 획득하였지만 정확률은 떨어지고 있음을 확인할 수 있었다. 두 번째 실험은 실험 단어들에 대한 초기 정확률을 높이기 위하여 동의어/유의어 결정하는데 적용하는 방법들의 가중치를 각각 다르게 설정하고 실험하였다. 빈도수의 경우 1.0, 사전의 구조정보를 0.8, 입력명사 정보를 0.6으로 설정하고 가중치가 2.0이상인 단어에 대하여 동의어/유의어로 선택한 결과가 된다. 또한 두 번째 실험에서 동의어/유의어 선택 임계 가중치(break-even point)를 1.5, 1.8과 2.0으로 나누어 실험을 했지만 2.0을 제외한 1.5와 1.8의 경우는 첫 번째 실험과 같은 결과를 나타내었다. 마찬가지로 임계 가중치를 2.0이상으로 높였던 실험에서는 재현률이 현저히 떨어지는 것을 확인할 수 있었다. 위 실험을 통해 재현률은 만족할 만한 성과를 거두었지만 정확률은 재현률에 비해 낮은 성능을 보였다. 이 같은 원인은 실제적으로 동의어/유의어를 판단 내리는 대상인 영-한사전의 특성에 기인한다. 영-한사전의 특성상 하나의 영어 단어에 대하여 여러 개의 뜻을 지난 한국어 단어가 존재하게 되는데 여러 개의 뜻을 지닌 단어들 중 때로는 서로 판이하게 다른 뜻을 지닌 단어들이 같은 영어단어 엔트리로 속하는 경우가 많다. 따라서 3.에서 제시한 가중치 부여 방법을 적용하였을 때 위 '가격'의 예처럼 '값'이나 '가치'등과 같은 동의/유의 관계에 있는 단어 뿐 아니라 '희생'과 같은 단어들도 동의어/유의어 사전에 속하게 된다. Fgi. 4에서 보는 바와 같이 '가격'이라는 단어와 매치되는 'price'나 'cost'의 영어 단어 뜻 중 '희생'이라는 단어가 포함되었기 때문

Table 1. 차등 가중치 적용 시 정확률과 재현률

재현률	정확률	F-score
0.96	0.657	0.78

에 정확률은 재현률에 비해 떨어지게 된다. 정확률를 저하시키는 단어들을 제거하기 위해서 보다 효과적인 가중치 부여 기법에 대한 연구가 필요하고 기본 자원의 하나인 영-한사전을 구축 시 그러한 단어들을 제거할 수 있는 방법들이 필요하다.

정보 검색 기법에의 응용

본 논문에서 제안하고 구축한 동의어/유의어사전이 실제적으로 정보 검색 기법에 사용되었을 때 어떠한 성능 향상 효과가 있는지 알아보았다. 이를 위해 문서 필터링(filtering) 시스템[2]에 동의어/유의어사전을 추가하여 실험해 보았고 질의문에 나타나는 색인어를 확장하는데 이용해 보았다.

1. 문서 필터링 시스템에 응용

실험에 사용한 문서 필터링 시스템[2]은 범위가 경제 분야에 해당하는 문서에 대한 필터링 시스템이다. 이 시스템에서는 색인어 확장을 위해 네 가지 방법을 사용하고 있고, 네 가지 방법 중 지식 베이스의 일종인 '개념망'[3]을 이용한다.

본 논문에서 제안하고 있는 동의어/유의어사전에 대한 유용성을 판단하기 위해 본 실험에서는 기존의 시스템에 동의어/유의어사전을 부가 자원으로 사용하여 적용 전과 후의 결과를 비교하였다. 이러한 실험을 통해 본 논문에서 제안하고 있는 동의어/유의어사전의 응용 성능에 대하여 보다 쉽게 비교할 수 있었다. Table 2는 이 실험을 통해 얻어진 동의어/유의어사전 적용 전과 적용 후의 F-score를 보여준다.

위 실험에서 색인어의 복합명사 분해 결과의 각 명사에 대하여 동의어/유의어사전을 검색하여 색인어로 추가하여 실험하였다. 실험 결과에서 보듯이 적용 전보다 적용 후 시스템의 성능이 향상되었음을 확인할 수 있었다. 또한 이 결과를 통해 본 논문에서 제시한 동의어/유의어사전이 정보 검색 기법에 응용되었을 때 더 효율적이라는 사실을 확인할 수 있다.

2. 질의문 색인어 확장에 응용

다음으로 질의문에 나타나는 색인어를 확장하는데 본 논문에서 제안하고 있는 동의어/유의어사전이 어떤 성능 향상 효과가 있는지를 실험해 보았다. 이를 위해 특정 질의문에 나타나는 색인어에 대하여 동의어/유의어 사전을 이용 확장시켜 보고 확장 전과 확장 후의 정답 추출 가능성에 대하여 실험을 전개하였다. 유용성 여부를 판별 하기 위해 한글 2002에서 제공하고 있는 동의어사전의 결과와 비교하였다. 결론적으로 이 실험을 통해 단순히 동의어 정보를 이용한 것보다 본 논문의 기본 자원인 영-한사전을 이용하여 부가적으로 획득된 유의어 정보가 효율적으로 사용될 수 있음을 확인할 수 있었다.

Table 2. 문서 필터링 시스템에의 적용

색인어 확장 방법	적용 전	적용 후
개념망+2차 확장	0.705	0.712

이 실험에서는 네이버(www.naver.com)에서 제공하고 있는 사용자 질의문를 선정하여 이용하였다. 실험을 위해 '소설 동의보감의 작가는 누구입니까?'라는 질문을 이용하였다. 본 질의에서 제시된 정답은 문서 내에 '조선시대 허준의 이야기를 바탕으로 쓰여진 소설로서 저자는 이은성씨이다.'라고 존재하고 있다. 질문에서 추출되는 색인어로는 '소설', '동의보감', '작가'와 같은 세 가지 색인어를 얻을 수 있고 이를 한글 2002의 동의어 사전과 본 논문의 동의어/유의어사전을 통해 확장시킨 결과는 Table 3과 같다.

Table 3에서 보는 바와 같이 본 논문에서 제안하고 있는 동의어/유의어사전은 해당 어휘에 대하여 기본자원으로 사용된 영-한사전의 특성으로 획득된 유의어 정보까지 포함하고 있다. 실험에 사용된 질의 색인어를 동의어/유의어사전을 통해 확장함으로써 정답으로 추출 되는 문장내의 '이야기', '저자' 등과 같은 단어를 획득할 수 있었다. 이 같은 확장을 통해 얻어진 유의어 정보가 정답을 찾는데 유용한 정보를 제공하고 있음을 확인할 수 있었다.

결론 및 향후 과제

본 논문에서는 대역어 사전을 이용하여 자동으로 동의어/유의어사전을 구축하고 구축된 정보가 정보 검색 기법에 응용되어 성능 향상의 효과가 있는지를 실험해 보았다. 실험의 결과 본 논문에서 소개하는 동의어/유의어사전이 정보 검색 기법에서 색인어 확장 시 사용되는 부가 자원으로 이용하여 시스템의 성능을 높이는데 사용될 수 있음을 확인하였다. 또한 기본 자원으로 사용된 영-한사전의 특성에서 획득된 다수의 유의어 정보가 질의문에 나오는 색인어 확장 실험에서 단순히 동의어 정보를 이용한 것보다 정답을 찾는데 유용하게 사용될 수 있음을 확인하였다.

Table 3. 색인어 확장 결과

색인어	한글 2002 동의어 사전	본 시스템 동의어/유의어사전
소설	스토리, 소설책	이야기, 설화, 실화, 꾸민 이야기
동의보감		
작가	글쓴이, 소설가, 집필자, 제작자	저자, 필자, 지은이, 제작자, 소설가

마지막으로 본 시스템이 앞으로 수정, 보완해야 할 점은 첫째, 재현률에 비해 낮은 결과를 보이는 정확률을 높이는 연구가 필요하다. 앞에서 언급했듯이 영-한사전의 특성에 의해 뜻이 전혀 다른 단어가 엔트리에 포함되는 경우가 많다. 이러한 단어를 제거하기 위해 영-한사전을 구축 시 부가 정보를 이용하여 정확률을 높일 수 있는 연구가 필요하다. 둘째는, 본 논문에서 구축한 동의어/유의어사전의 경우 한국어 단일 명사로 한정이 되어 있다. 정보 검색에 사용되는 질의어의 경우 복합명사의 형태를 띠고 있는 경우가 많이 존재하고 있는데 이를 위해 복합명사형태의 동의어/유의어사전의 구축이다. 이를 위해 단일어 형태로 존재하는 동의어/유의어사전과 방대한 시소러스를 이용하여 연어정보 및 공기정보를 이용한 방법[7]으로 효과적인 복합명사 동의어/유의어사전 구축이 가능할 것이다. 마지막으로 명사뿐 아니라 구문 분석에 유용하게 사용될 수 있는 용언에 대한 동의어/유의어사전의 연구도 필요하다. 이러한 용언에 대한 동의어/유의어사전은 사용자의 질의 문장을 보다 효율적으로 해석하는데 중요한 자료가 될 것이다.

REFERENCES

1) 김수민, 백대호, 김상범, 임해창(2000) : "시소러스범주정보를 이용한 질의응답시스템". 한글 및 한국어정보처리 학술대회, *pp179-183, 10*
2) 정용교, 신승은, 오효정, 장명길, 서영훈(2002) : "*Answer set* 자동 구축을 위한 문서 필터링". 한글 및 한국어정보처리 학술대회, *pp253-258, 10*
3) 장문수, 장명길, 김현진, 오효정, 이재성(2000) : "인테넷 질의/응답을 위한 지식베이스 구축". 한글 및 한국어정보처리 학술대회, *pp198-202, 10*
4) Julio Gonzaol, Felisa Verdejo, Irina Chugur, Juan Cigarran(1998) : "*Indexing with WordNet synsets can improve text retrieval*"
5) Atsushi Fujii, Tetsuya Ishikawa(1999) : "*Cross-Language Information Retrieval for Technical Documents*"
6) 이주호, 배희숙, 김은혜, 김혜경, 최기선(2002) : "명사 워드넷과 단일어 사전을 이용한 한국어 동사 워드넷 구축". 한글 및 한국어정보처리 학술대회, *pp92-97, 10*
7) 장명길(1999) : "한영 교차언어 정보 검색에서 상호 정보를 이용한 질의 변환 모호성 해소 및 가중치 부여 방법". 한글 및 한국어정보처리 학술대회, *pp55-61, 10*
8) Brown PF(1991) : "*Word-sense Disambiguation using Statistical Method*". *In Proceedings of ACL 91*
9) 이 호(1997) : "최소한의 코퍼스 정보를 이용한 단어 의미 중의성 해결 기법". 한국정보과학회 '97봄 학술발표논문집, *pp467-470*