백과사전 기반 전문용어 태깅 시스템

배영준 최호섭 옥철영
울산대학교 컴퓨터정보통신공학과 한국어처리연구실
young4862@mail.ulsan.ac.kr {hoseop,okcy}@ulsan.ac.kr

Terminology Tagging System using elements of Korean Encyclopedia

Youngjun Bae Hoseop Choe Cheulyoung Ock
Korean Language Processing Laboratory,
Dept. of Computer Engineering and Information Technology,
University of Ulsan

요 약

지금까지 자연언어처리에서의 품사태깅(parts-of-speech tagging) 기술에 대한 연구는 활발히 진행된 반면, 전문용어에 대한 처리 기술은 미비한 점이 많았다. 전문용어에 관련된 연구는 대부분 구축, 표준화, 추출 등에 대한 연구가 많았으나 전문용어 태그 설정과 태깅 기술 연구는 부족한 상황이다. 본 논문에서는 전문용어 태그를 (분야정보:아이디) 순으로 설정하고 백과사전의 분류 체계를 이용하여 어떤 특정 분야 문서의 전문용어를 자동으로 태깅하는 시스템을 구축하였다. 전문용어 태깅 시스템은 형태소분석기를 사용하지 않고 문맥의 규칙과 조사·어미사전을 이용해 자동으로 태깅을 하게 된다. 이 시스템의 정확률 측정을 위한 정답말뭉치는 웹 상에 공개되어 있는 백과사전 html문서를 이용하였다. 우선 백과사전에 나와있는 용어는 전문용어라고 가정한다. 하나의 문서에는 '용어', '요약', '본문', '이미지', '분류', '참조항목' 등의 정보들이 있다. 이 중 '본문'에는 그 용어에 대한 자세한 설명이 있는데 특정 단어에는 <a>태그로 백과사전 내에 있는 단어를 찾아 볼 수 있게 링크 되어있다. 이 정보를 이용해 <a>태그로 되어있는 것을 설정한 태그로 바꾸고 단계별로 확장 태깅을 해서 정답말뭉치를 만든다. 태깅 시스템과 정답말뭉치를 비교해 정확률을 계산해서 시스템의 성능을 측정하였다.

1. 서론

그동안 품사 태그에 대한 연구는 있지만 전문용어 태그(terminology tag)에 대한 논의는 국내외적으로 거의 없는 실정이다. 자연언어처리뿐만 아니라 응용 분야에서도 전문용어에 대한 설정 논의보다도 전문용어 사전 구축, 전문용어 표준화, 전문용어 수동/자동 추출 등 전문용어 자체에 대한 논의가 일반적으로 연구되고 있다.

전문용어 태깅은 단순히 전문용어 사전을 기반으로 색인(indexing) 기술을 이용할 수도 있으나, 이는 전문분야 정보(분야태그 등)나 분야별 전문용어 정보(단어의 아이디 등)가 표시될 수 없는 단점을 가지고 있다. 또한 지식 정보의 공유 체제라는 측면에서는 단순한 전문용어 표지만으로는 지식 정보의 세밀화된 정보를 공유하기 어렵다. 이에 본 논문에서는 전문용어 태그 세트 설정을 위해 백과사전 분야 분류 체계를 이용하여 전문용어에 대한 태그 세트를 설정하고자 하였다. 그리고 태그 세트를 이용해 전문분야 문서를 태깅하는 시스템을 구현함과 동시에, 적절한 정답말뭉치 자동구축과 이 말뭉치를 이용한 전문용어 태깅 시스템 성능을 평가하고자 하였다.

2. 기존 연구

기존에 연구된 태깅 시스템의 태그에 관한 연구는 품사 태그(N, V, J 등)를 몇 가지로 볼 것이며 세부적으로 나눌 것인가 아니면 묶어서 하나의 태그로 볼 것인가 하는 연구였다. 기존 연구 중 "분야정보를 태그로 어떻게 설정할 것인가?" 그리고 "분야정보가 들어간 태

그로 전문용어를 어떻게 태깅 시킬 것인가?"에 대한 연구는 거의 없었다.

본 논문에서는 이러한 질문을 바탕으로 태깅 시스템을 구축하였다. 전문용어 추출을 위한 전문용어 태깅 시스템을 위해서 연구는 많이 있어 왔다. 본 논문의 시스템은 우선 전문용어에 대한 추출이 필요하다. 정보검색의 색인어 추출방법, 공기정보를 이용하는 방법, 기계학습 등 통계적 기법을 이용하는 방법이 있었다. 또는 특정 분야(domain)에만 나타나는 동사의 앞뒤 어절을 보거나 숫자, 특수기호, 대문자, 소문자의 형태, 문맥 정보를 이용하는 언어적 기법을 이용하는 연구가 있었다[2][4][5]. 그리고 이 두 가지(언어, 통계적 기법)를 접목시켜서, 전문용어후보추출 과정에선 언어적 기법을 쓰고 올바른 전문용어를 파악하는 데는 통계적 기법을 쓰는 등의 연구가 있었다[3].

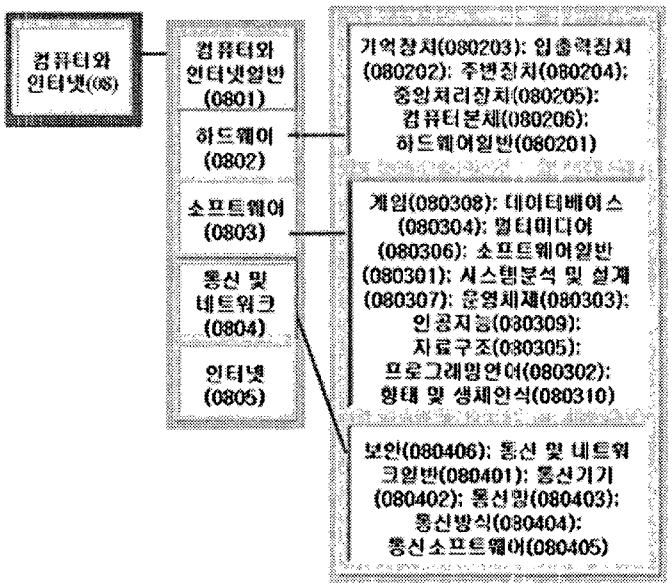
본 논문에서는 형태정보와 문맥정보를 이용해 규칙을 만들고 그리고 최장 일치 형태의 전문용어를 사전과 매칭시켜 전문용어를 추출하는 방법을 이용한다.

3. 분야 별 태그 설정

3.1 백과사전 태그 정보

백과사전에는 분야 정보가 최상위 13개 분야로 되어 있으며 각 분야 집합은 최대 5단계까지 세부 분류가 되어있다. 이 분류구조는 대략 3,500개 노드를 가지고 있다. 그리고 분야마다 쉽게 구분하기 위한 일종의 태그로 볼 수 있는 숫자가 할당 되어있다. 각 단계 마다 2자리의 자릿수가 할당되어 있다. 예를 들어 컴퓨터인터넷이라는 최상위(1단계)의 한 분야를 보면 (08)이라는 태그가 할당되어 있다.

그리고 이 숫자체계 태그는 단계가 내려가면 자신의 상위 숫자 태그와 또 자신의 고유 숫자태그를 합쳐서 하나의 숫자 태그가 된다. 예를 들면 데이터베이스라는 분야 숫자 04가 할당되고 데이터베이스가 속한 분야는 컴퓨터인터넷(08) 아래 소프트웨어(03) 아래 위치하기 때문에 데이터베이스는 (080304)라는 분야 태그를 가지게 된다. 물론 소프트웨어라는 분야도 컴퓨터인터넷 분야 밑이기 때문에 (0803)라는 태그 정보를 가지게 된다. 이처럼 태그세트는 자신의 상위 태그를 포함함으로써 한 눈에 어디에 속하는 분야인지 알아볼 수 있도록 되어있다.



[그림 1] 컴퓨터 인터넷 분야 태그 트리 형태

3.2 문서에 태깅될 태그 형식

본 논문에서 사용하는 태그 형식은 다음과 같다.

[그림 2]에 나타나 있는 백과사전의 HTML <a>태그 형태를 다음과 같이 바꾼다.

-형식 1:

3.1에서 살펴보았던 백과사전 태그 정보와 그 단어의 고유 아이디를 결합한 형식이다.

<분류태그:단어 아이디>단어</분류태그:단어 아이디>

예: <0801:770211>프로그램</0801:770211>

프로그램이라는 단어가 하나의 (0801)분야에 있는 단어 이며 770211이란 아이디를 가지고 있다는 뜻이다.

-형식 2:

만약 그 단어가 다의어라서 뜻이 여러 개면 아이디가 두개 이상이다. 이런 경우 태그가 겹쳐서 태깅 되도록 만들었다. 그리고 같은 뜻을 가진 단어지만 분야정보가 다르다면 태그가 겹쳐서 태깅되도록 하였다. 형식은 다음과 같다.

<분류태그1:단어아이디><분류태그2:단어아이디>단어
</분류태그2:단어아이디></분류태그1:단어아이디>

단어 '컴퓨터'는 분야 정보가 두 가지로 되어있는데 하나는 '과학(05)>기술과학(0502)>전자공학(050207)>전자공학일반(05020701)'으로 되어있고 또 하나의 분야는 '컴퓨터와 인터넷(08)>컴퓨터와 인터넷일반(0801)'의 두 분야 정보를 포함하기 때문에 태깅은 다음과 같이 된다.

예: <05020701:151218><0801:151218>컴퓨터
</0801:151218> </05020701:151218>

-형식 3:

그리고 태깅되어 있는 단어와 연결되는 단어일 때는 겹쳐서 태깅되도록 하였다.

<분류태그1: 단어1+단어2의 아이디> 단어1 <분류태그2: 단어2의 아이디> 단어 2 </분류태그2: 단어2의 아이디></분류태그1:단어1+단어2의 아이디>

예: <05020703:111557>고밀도<05020703:144199>집적
회로</05020703:144199></05020703:111557>

여기서 '집적회로'라는 단어가 한 번 태깅되고 그리고 '고밀도집적회로'라는 단어가 또 한 번 태깅된 형태이다.

4. 정답 말뭉치 자동 구축

전문용어 태깅 시스템의 정확률을 계산하기 위하여 정답말뭉치를 다음과 같이 구축하였다.

4.1 구축 방법

전문용어 사전이 없었을 때는 대부분 수작업으로 정답 말뭉치를 만들었다. 그리고 전문용어 사전이 만들어지고 난 후에는 사전에 있는 단어를 전문 문서에 매칭시켜서 일치되는 것들을 전문용어로 보고 나머지 명사들을 일반용어로 보는 형식이었다.

본 논문에서는 웹 상에 공개되어 있는 백과사전의 태그 정보를 이용해 전문용어 정답 말뭉치를 자동으로 생성한다. 이 방법은 세 단계로 나누어서 실행한다.

사용될 말뭉치는 웹 상에 공개되어있는 백과사전인데 백과사전 말뭉치에는 '용어', '요약', '본문', '이미지', '분류', '참조항목' 등의 정보들이 있다. 이 중 '본문'에는 그 용어에 대한 자세한 설명이 있다. 그리고 '본문'은 해당하는 카테고리 정보가 있는 용어의 설명이기 때문에 용어가 가진 분야의 전문문서라고 볼 수 있다. 그래서 본 논문에서는 하나의 '본문' 정보를 말뭉치로 선택해 한 문서로 본다. [그림 2]는 웹 상에 있는 HTML 말뭉치이다.

컴퓨터를 사용하려면 프로그램이 필요한데,
이 프로그래밍 기술은 고도(高度)의 것일수록 어려우며,
이 점이 일반인에게 컴퓨터를 쉽게 이용하여 그 효용을 누릴 수 있는 길을 가로막고 있다.

[그림 2] 백과사전 본문의 HTML 태그

첫번째 단계로 <a> 태그를 가진 단어들을 전문용어로 판단하고 분야별 태그를 넣어서 태깅한다.

말뭉치 내에는 백과사전에 나와있는 용어들이 <a>태그 형식으로 링크가 걸려있다. 이것들은 백과사전에 나와있는 단어이므로 전문용어라고 볼 수 있다. 그래서 이 부분을 분야정보가 포함된 태그로 태깅한다.

다음에 나오는 두 단계는 연속되는 전문용어들을 놓칠 가능성을 배제하기 위한 단계이다. 두 번째 단계로 <a> 태그가 되어 있는 단어와 그 단어의 앞 뒤 1~3어절을 포함하는 연결어구를 백과사전 용어리스트에서 찾아보고 태깅한다. <a> 태그 안의 단어를 포함한 앞뒤 어절이 전문용어가 될 가능성이 높다는 가정하에 실험하는 단계이다. 예를 들면 '인터넷마케팅'이란 용어가 있을 때 백과사전에는 인터넷마케팅이라는 용어가 존재하지만 <a>태그는 '인터넷' 과 '마케팅'이 따로 태깅되어있다.

이렇게 용어가 존재하지만 태깅이 되지 않은 부분들을 두번째 단계를 거치면 다음과 같이 묶어서 태깅된다.

<0805:275530>인터넷</0805:275530>^
<030407:59199>마케팅</030407:59199>

▼

<0805:754125><0805:275530>인터넷
</0805:275530>^
<030407:59199>마케팅
</030407:59199>**</0805:754125>**

마지막 단계로 태깅 되지 않은 단어들을 대상으로 띄어쓰기 공백을 없앤 뒤 좌우 최장일치법으로 백과사전 용어들과 매칭되는 단어들을 태깅한다. 이 단계에서 띄어쓰기 오류로 추출되지 못한 전문용어를 추출할 수 있다.

5. 전문용어 태깅 시스템

전문용어 추출방법으로는 정보검색의 색인어 추출방법, 공기정보를 이용하는 방법, 기계학습 등 통계적 기법을 이용하는 방법 등이 있다. 또는 형태, 문맥 정보를 이용하는 언어적 기법을 이용하거나 이 두 가지를 접목시켜서 전문용어후보추출 과정에선 언어적 기법을 쓰고

올바른 전문용어를 파악하는 데는 통계적 기법을 쓰는 등의 연구가 있었다.

본 논문에서는 빠른 전문용어 태깅을 위해 형태소분석기를 이용하지 않고 조사사전, 어미사전만을 이용하여 태깅한다.

5.1 자료

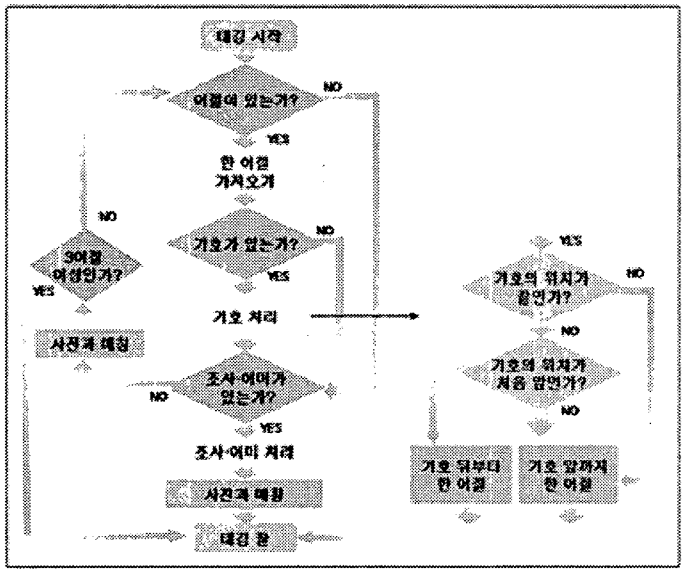
조사는 국립국어연구원에서 발행된 '현대 국어 사용 빈도 조사'[10]에서 조사된 고 빈도로 나타나는 72개의 조사만 이용한다. 그리고 어미는 고빈도로 나타나는 30개만 뽑아서 이용한다.

5.2 규칙

백과사전의 용어 리스트와 매칭되는 단어들을 뽑아서 태깅을 하였다. 일반적인 어절의 완전매칭이 아니라 조사나 어미 그리고 특정 기호의 앞·뒤 부분에 위치하는 구문을 좌에서 우 방향으로 가는 최장 일치법을 이용해서 태깅한다. 확장 태깅되는 범위는 최대 3어절까지로 한다.

그리고 전문용어에 '수'나 '이'(뜻: 중국철학 특히 정주학(程朱學)의 근본개념)같이 1음절의 용어들이 있다. 그러나 오류의 주요 원인이기 때문에 1음절인 단어는 태깅 대상에서 제외한다.

이 시스템은 띄어쓰기의 차이로 사전과 매칭이 되지 않는 용어나 띄어쓰기 오류로 인해 찾아지지 않는 용어들이 태깅되게 하였다.



[그림 3] 전문용어 태깅 시스템 순서도

위의 그림은 전문용어 태깅 시스템에 대한 대략적인 순서도이다. 그림에서 사용한 기호란 문장부호를 뜻한다. 기호가 나타나는 위치에 따라 다른 방식으로 처리했다.

제일 앞부분에 기호가 나오면 그 뒤부터 어절로 보고 조사 또는 어미가 있는지 확인한 뒤 사전과 매칭시켜 전문용어를 찾는다(아래 예1 참조). 그리고 중간에 기호가 나타나는 경우는 그 앞부분과 그 뒷부분을 각각 한 어절로 보고 태깅한다(아래 예2 참조).

예1> "광컴퓨터를 → 광컴퓨터를

예2> 기억장치(자기코어 등) → 기억장치, 자기코어

조사 또는 어미를 뽑아내기 위해서 어절의 뒷부분부터 차례로 조사사전 또는 어미사전과 비교해본다. 조사사전 또는 어미사전은 속도를 빠르게 하기 위해 ('에게' -> '게에')처럼 거꾸로 저장 하였다.

5.3 전문용어 태깅 도구

문서에 태깅을 하고 태깅된 결과를 살펴보기 위해서 도구를 만들었다. 도구의 주요 목적 첫번째는 문서의 자동태깅이고 두번째는 태깅된 문서를 보기 위한 목적이다. 세번째는 문서를 자동태깅한 후에 전문가들이 문서에 있는 전문용어에 태깅된 태그들을 판별하여 수정, 삽입, 삭제할 수 있도록 도와주는 관리 기능이다. 도구에는 자동태깅 기능이 있고 왼쪽에는 분류체계를 트리 형식으로 보여줌으로써 직접 드래그 앤 드롭으로 단어에 태그를 넣을 수 있다. 그리고 오른쪽 위쪽에는 전문용어 사전의 종류가 여러 가지 있을 수 있는데 이 중 선택할 수 있는 전문용어 사전을 보여주고 사전이 선택되면 찾고 싶은 단어를 검색 할 수 있게 된다. 검색된 단어는 리스트 형식으로 보여주고 여러 사전에서 동시에 검색이 되도록 하였다. 리스트에서 검색된 단어를 클릭하면 그 단어의 정보를 그 오른쪽에 보여준다. 리스트에서 더블 클릭을 하면 문서에 태깅된다.

자동태깅 방법은 3가지를 선택할 수 있다. 첫 번째는 단순히 어절 정보만을 이용한 전문용어 사전 리스트와의 완전 매칭 태깅 방법이다. 두 번째는 본 논문에서 사용하고 있는 조사사전과 어미사전을 이용하여 단어를 뽑아낸 뒤 태깅하는 방법이다. 세 번째는 형태소분석기 HAM을 이용해 형태소 분석을 한 후에 명사상당어구만 태깅하는 방법이다. 세 번째 방법은 두 번째 방법과 비교해 보기 위해 구현 중이다. [그림 4]은 전문용어 태깅 도구를 보여주고 있다.

[그림 4] Term Tagger

6. 실험

6.1 실험데이터

문서는 두산동아 백과사전에 있는 용어 중 한 용어의 본문을 하나의 문서로 보았다. 분야는 컴퓨터와 인터넷 분야로 한정했으며 이 분야의 문서는 998개이고 문서 전체 어절 수는 68,655개이다. 이 문서는 모든 html태그가 제거된 상태이다. 전문용어 사전은 백과사전의 용어들로 하였다. 그리고 마지막으로 태깅 시스템의 대상 문서로 전자신문의 기사 중 컴퓨팅 분야 문서 몇 개를 사용한다.

6.2 실험 결과

[표 1] 단계별 태그 개수

단계	태그 개수
A 태그	9681
정답말뭉치 1단계	13138
정답말뭉치 2 단계	14267
정답말뭉치 3 단계	**62463**
태깅 결과	**61189**

[표 1]는 정답말뭉치 구축 단계별로 태그된 개수를 보여준다. 1단계는 단순히 <a>태그를 설정한 태그로 바꾸는 단계였고 2단계는 <a>태그되어 있는 단어의 앞뒤 어절을 고려해 태깅하는 단계였다. 3단계는 2 단계까지 실행한 뒤 태깅 되지 않은 나머지 부분을 태깅하는 단계였다.

태깅이 되기 전 본문의 <a>태그 개수는 9,681개 이고 정답 말뭉치로 구축했던 용어의 태그 개수는 62,463개이다. <a>태그로 태깅이 안된 부분이 많다는 것을 알 수 있다. 전문용어 태깅도구로 태깅한 태그 개수는 61,189개 이다. 정답말뭉치와 태깅 시스템을 이용해 태깅한 결과가 다른 것은 총 1,758개 이다. 이 중 태깅도구에만 태그된 것은 127개이고 정답말뭉치에만 태그된 것은 73개이다. 나머지 1,558개는 둘 다 태깅은 되었지만 태깅된 결과가 다른 것들이라 볼 수 있다.

태깅 결과에 대한 정확률을 다음과 같이 계산하였다.

$$\text{정확률} = \frac{\text{정답문서의태그된개수} - \text{도구의태깅결과와다른태그개수}}{\text{정답문서의태그된개수}}$$

(1)의 식을 이용해 계산한 결과 97.2%의 정확률이 나왔다.

태깅된 결과가 정답말뭉치와 다르게 나온 오류 유형들을 보면 조사사전 또는 어미사전에 없는 단어가 사용되었을 경우에 찾지 못하는 것과 그리고 정답 말뭉치에는 태깅 되지 않았는데 태깅 시스템에서는 태깅하는 경우, 그리고 제일 많이 나타난 것으로 태깅이 되어있긴 하지만 태깅의 형태가 다른 경우이다. 태깅의 형태가 다르게 나타나는 이유는 우선 정답말뭉치는 <a>태그를 중심으로 주위의 단어들을 태깅시켰다. 그리고 나머지 태깅되지 않은 단어를 태깅시키는 순서 였기 때문에 처음에 <a>태그가 설정된 형태대로 단순, 다중 태깅이 되어 버리면 태깅 시스템과 다른 결과가 나올 수 있다.

예: – html 파일

자기 디스크

– 정답말뭉치 태깅 결과

<080203:131201>자기 디스크</080203:131201>

–도구의 태깅 결과

<080203:131201><0501020504:131189>자기 </0501020504:131189>디스크</080203:131201>

그리고 다음 예와 같이 1어절 이상이고 문장 부호가 포함되어 있는 <a> 태그의 경우는 '자기' 라는 단어는 태깅을 하지만 '자기 테이프' 전체를 태깅하지 못한다.

예 : – html 파일

자기(磁氣) 테이프

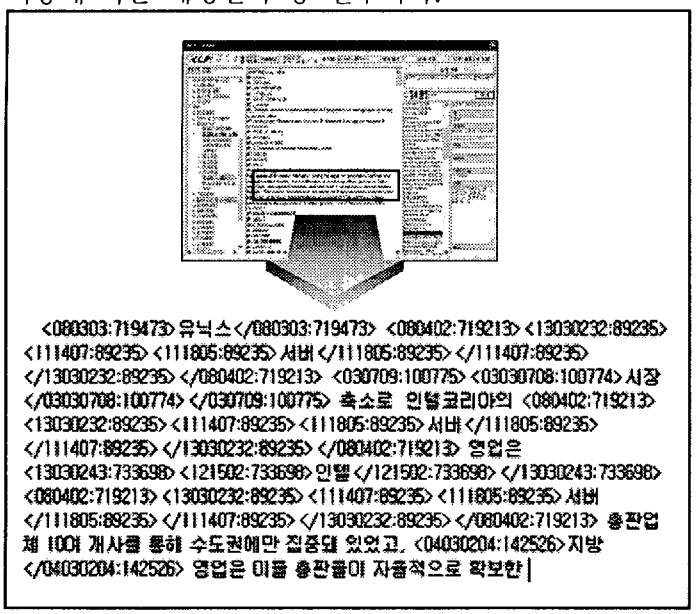
–도구의 태깅 결과

<0501020504:131189>자기

</0501020504:131189> (磁氣) 테이프

그리고 정답문서에는 태깅 되어 있지 않지만 태깅 시스템의 결과로 나온 문서에는 태깅 되어 있는 경우도 있다. 이와 같은 이유로 약 3%의 오류가 나왔다.

[그림 5]은 태깅 시스템의 문서로 전자신문 기사를 이용해 나온 태깅결과 중 일부이다.



[그림 5] 전자신문 기사 태깅결과의 일부분

7. 결론 및 향후 과제

본 논문에서는 전문용어 태그를 <분야정보:아이디> 순으로 설정하고 백과사전의 분류 체계를 이용하여 어떤 특정 분야 전문문서의 전문용어를 자동으로 태깅하는 시스템을 만들었다. 시스템은 형태소분석기를 사용하지 않고 문맥상의 몇 가지 규칙과 조사사전, 의미사전을 이용해 자동으로 태깅을 하였다. 이 시스템의 정확률 측정을 위해 정답말뭉치를 자동으로 구축하였고 문서는 웹 상에 공개되어 있는 백과사전 html문서를 이용하였다. 백과사전에 용어 리스트가 전문용어라고 가정하였고 하나의 문서에는 '용어', '요약', '본문', '이미지', '분류', '참조항목' 등의 정보들이 있다. 이 중 '본문'에는 그 용어에 대한 자세한 설명이 있는데 특정 단어에는 <a>태그로 백과사전 내에 있는 단어를 찾아 볼 수 있게 링크 되어있다. 이 정보를 이용해 <a>태그로 되어있는 것을 설정한 태그로 바꾸고 단계별로 확장 태깅을 해서 정답말뭉치를 만들었다. 태깅 시스템과 정답말뭉치를 비교해 정확률을 계산해서 시스템의 성능을 측정해 97.2%의 정확률을 얻었다.

본 태깅 시스템은 사전에 대한 의존도가 크기 때문에 미등록어가 나왔을 때 전문용어로 찾아내지 못하는 단점이 있다. 이런 요소는 미등록어 추정하는 부분이 필요하다.

지금 의미적인 요소까지 판단하는 기준이 없기 때문에 [그림 5]의 '지방'처럼 태깅되지 말아야 할 요소가 태깅되는 것을 볼 수 있다. 이런 요소를 처리하기 위해 단어의 빈도 정보나 일반용어 사전을 이용하는 방법 등의 연구가 필요하다.

태깅 시스템을 이용하여 문서 clustering에 대한 연구를 진행할 것이다.

참고 문헌

[1] 오종훈, 이경순, 최기선, "분야간 유사도와 통계기법을 이용한 전문용어의 자동추출", 한국정보과학회논문지, pp.258-269, 2002

[2] 오종훈, 김재호, 최기선, "EM 알고리즘을 이용한 전문용어의 자동 추출" 한국정보과학회 가을학술발표논문집(1), pp.487-489, 2003

[3] 오종훈, 최기선, "정보통합을 통한 생물/의학 분야 전문용어의 자동 추출", 한국정보과학회 가을학술발표논문집(1), pp.775-777, 2004

[4] 류법모, 최기선, "공기정보를 이용한 전문용어 관련 세그먼트의 자동 추출"

[5] 류법모, 배선미, 최기선, "구성정보와 문맥정보를 이용한 용어의 전문성 측정 방법" 한국정보과학회 봄학술발표논문집(B), pp.906-903, 2004

[6] 박정오, 황도삼, "전문용어 추출시스템", 한국정보과학회 - 봄 학술발표논문집, pp.381-383, 2002

[7] 니콜라이M.조슈티스, C++Standard Library, 인포북

[8] 강승식, p184~189"한국어형태소분석과 정보검색"

[9] 국립국어연구원, p1183 ~ 1187 "현대 국어 사용 빈도 조사", 2002-1-17

[10] Ido Dagan, Ken Church, "Termight: Idntifying and Translating Techical Terminology", Proceedings of the fourth conference, 1994

[11] Bernth,A., McCord,M., Warburton,K., "Terminology extraction for global content management", Terminology (Netherlands) Terminology, 9(1), pp.51-69, 2003

[12] Goran Nenadic, Sophia Ananiadou, John McNaught, "Enhancing automatic term recognition through recognition of variation", Coling, 2004