분야연상어의 수집과 추출 알고리즘

이 상 곤† · 이 완 권†

요 약

인간은 문서전체를 읽지 않고 대표적인 단어를 보는 것만으로 정치나 스포츠 등의 분야를 정확히 인지할 수 있다. 문서전체를 대상으로 하지 않고 부분텍스트에서 출현하는 소수의 단어정보에서 문서의 분야를 정확히 결정하기 위해 분야연상어의 구축은 중요한 연구과제이다. 인간이 미리 분야체계를 정의하고, 각 분야에 해당하는 문서를 인터넷이나 서적을 통해 수집한다. 본 논문은 수집문서의 분야를 정확히 지시하는 분야연상어를 수집하는 방법을 제안한다. 문서의 분야결정 시점을 고려하여 분야연상어의 수준과 안정성 랭크에 대하여 논의한다. 학습데이터에서 분야연상어 후보의 각 수준을 자동으로 결정하고, 컴퓨터가 제시하는 분야연상어의 수준, 안정성 랭크, 집중률, 빈도정보를 이용하여 단일 분야연상어를 수집하는 방법을 제안한다.

Collection and Extraction Algorithm of Field-Associated Terms

Samuel Sangkon Lee† · Wan Kwon Lee†

ABSTRACT

Field-associated term is a single or compound word whose terms occur in any document, and which makes it possible to recognize a field of text by using common knowledge of human. For example, human recognizes the field of document such as <baseball> or <politics>, a field name of text, when she encounters a word 'pitcher' or 'election,' respectively. We proposes an efficient construction method of field-associated terms (FTs) for specializing field to decide a field of text. We could fix document classification scheme from well-classified document database or corpus. Considering focus field we discuss levels and stability ranks of field-associated terms. To construct a balanced FT collection, we construct a single FTs. From the collections we could automatically construct FT's levels, and stability ranks. We propose a new extraction algorithms of FT's for document classification by using FT's concentration rate, its occurrence frequencies.

키워드 : 분야연상어(Field-associated Term), 수준별 분야연상어(Level of FTs), 안정성 랭크(Stability Rank), 정보추출(Information Extraction), 문서분류(Document Classification), 정보검색(Information Retrieval)

1. 서 론

컴퓨터에서 이용하는 전자화된 문서의 증가에 따라 문서의 자동분류에 관한 연구개발이 대단히 활발하다. 문서전체의 정보를 이용하여 유사도를 계산하는 벡터모델[6]이나 확률모델[2] 등의 기술이 확립되어 있으나, 실제로 문서에는 복수의 화제나 분야가 혼합되어 있으며, 사용자가 검색을 원하는 내용은 문서 일부분(단편)에 존재하는 경우가 대부분이다.

인간은 문서전체를 읽지 아니하여도, 문서에서 대표적인 단어를 보는 것만으로 <정치>나 <스포츠> 등의 문서분야를 정확히 인지할 수 있다. 따라서, 문서단편 내의 소수의 단어정보를 이용하여 분야를 정확하게 결정하기 위한 분야연상어의 구축은 중요한 연구과제[12]이다.

한편, 인간은 자신의 상식지식으로 특정분야를 인지할 수 없는 경우에도 문서에서 처음으로 출현하는 몇 개의 단어들을 이용하여 연상되는 연상정보를 인지적으로 인식하고 문서의 내용을 읽어감에 따라 문서에 해당하는 분야를 연상하거나 추측할 수 있다. 또한 문서의 이전내용에서 애매성이 발생하여도 문서의 뒤에서 출현하는 단어에 의해 이전 문서내용의 애매성을 해소해 나갈 수 있다. 이와 같이 문서단편 내에 몇 개의 단어정보를 이용하여 문서가 포함되는 분야를 정확하게 결정할 수 있는 단어를 "분야연상어"라 정의하고, 상식적인 분야연상어의 구축, 유사문서(문장) 검색, 문서요약 등의 기초연구를 수행한다. 문서전체의 정보를 이용하는 대부분의 정보검색 모델은 문서를 대표하는 잘못된 중요어의 추출을 막을 수 없다. 그러나, 본 논문에서 제안하는 방법은 문서의 처음부분에서 잘못된 분야연상어가 추출되어도 이후에 나타나는 올바른 분야연상어에 의

※ 본 연구는 한국과학재단 목적기초연구(과제번호 : R05-2003-000-10690-0) 지원으로 수행되었습니다. 재단의 연구지원에 깊은 감사를 드립니다.
† 정 회 원 : 전주대학교 정보기술컴퓨터공학부 교수
논문접수 : 2003년 4월 30일, 심사완료 : 2003년 5월 29일

해 정확한 문서의 분야를 추적할 수 있는 장점을 가진다.

문서를 대표하는 적당한 키워드 추출 연구에서 복합어 처리 연구의 목적은 문서내용을 보다 명확하고 간결하게 표현하는 단어의 추출에 있기 때문에 분야연상어로서 복합어의 추출이 적당하다. 예를 들면, 어떤 신문기사 표제 "선동렬투수 1,000 삼진 달성"에서 분야 <야구>에 대해 적당한 키워드의 추출은 "선동렬투수", "삼진" 등은 적합한 추출로 볼 수 있다.

다음에 기존의 연구들과 본 논문에서 제안하는 방법의 차이점을 살펴보자.

- 키워드 추출이나 문서분류 등을 포함한 정보검색의 연구는 중요어 결정에 통계정보를 이용하여 단어의 중요도(가중치)를 결정하는 방법[7-9]과 단어의 중요도를 학습하는 방법[7], 시소러스를 이용하여 단어의 개념이나 의미정보를 이용하는 방법, 의미적으로 관련이 있는 명사에 관련성 링크를 연계하는 방법[10] 등이 제안되어 있다. 그러나, 이 방법들은 본 논문의 목표인 높은 정확율(쓸모 없는 키워드의 과잉추출이 극히 적음을 의미)은 실현하지 못하고 있다.
- 시소러스 등의 분류체계를 이용하는 방법은 단어의 통계정보에 기반 한 방법에 비하여 정밀도 향상을 기대할 수 있으나, 시소러스 구축이 매우 어렵고, 비용이 많이 드는 문제가 있다. 분류체계와 단어간의 대응관계를 미리 구축하여 문서의 특징을 학습하는 방법[3]은 데이터 희소성(data sparseness) 문제가 있으며, 충분한 정밀도의 향상은 얻지 못하고 있다.
- 분류체계의 특징을 규칙으로 학습하는 방법[1]은 높은 정밀도를 실현하고 있으나, 실험 데이터의 규모가 적고, 해석도 복잡하여 실용성이 낮다. 또한, 미리 인간이 정의한 분류체계에서 분야를 결정하는 규칙기반 기계학습 방법은 문서분류의 정밀도가 Break-even Point (재현율과 정확율이 일치하는 수치)에서 최고 0.80까지 향상하고 있으나, 본 연구가 목표로 하는 정확도는 달성하지 못하고 있다. 또한,
- 복합 분야연상어의 결정과 관련된 연구에서는 복합어 키워드 추출방법[4-5] 등이 있으나, 사람의 지식으로 수집된 단일어로 된 키워드를 이용하여 복합어로 된 키워드를 자동으로 결정하는 방법은 아직 논의된 바 없다.

본 논문에서는 인간이 결정한 분야체계와 수집된 학습데이터를 이용하여 분야연상어를 수집한다. 또한, 잘못된 분야연상어가 추출되는 비율이 몇 퍼센트 이하가 되는 추출방법을 제안한다. 단일 분야연상어의 수는 유한하기 때문에 인간이 수집한다. 사람이 직접 분야연상어의 적합성 평가를 하였기 때문에 양질의 분야연상어가 구축될 가능성이 높아진다. 또한 단일어의 연상정보를 이용하여 무한히 만들어지는 복합어에 대한 분야연상어를 자동 구축한다. 본 논문의 논의는 아래와 같이 전개한다.

먼저, 제 2장에서는 분야연상어를 단일과 복합 분야연상어로 나누어 정의하고, 형태소 사전과의 관계를 설명한다. 분야체계에 의해 분야트리를 정의하여 각 분야연상어의 수준을 정의한다. 분야연상어가 시간의 경과에 의하여 변화하는 것에 주목하여 안정성 랭크를 정의한다. 단일어로 된 분야연상어는 그 길이가 최단거리이며, 개수도 유한하기 때문에 분야연상어 후보를 사람이 직접 판단하여 선별하고, 무한히 만들어지는 복합 분야연상어의 길이를 자동적으로 단축하고, 수를 줄이는 방법의 기초연구를 한다.

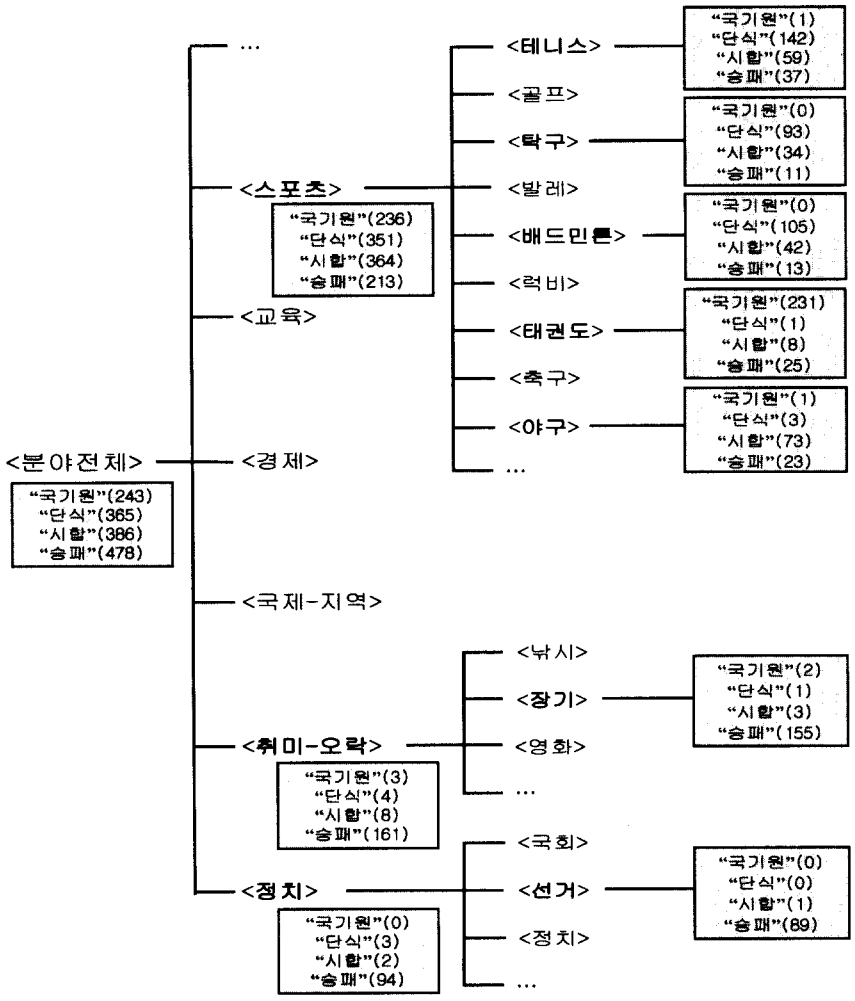
제 3장에서는 학습데이터에서 분야연상어 후보어와 수준을 자동적으로 결정하는 알고리즘을 제안한다. 단, 학습데이터가 각각의 종단분야에 균일하게 수집된다고 볼 수 없기 때문에 종단분야에 출현하는 모든 단어의 총 빈도를 계산하여 각 분야의 단어빈도를 정규화 한 빈도정보에 의해 분야연상어가 특정분야에 집중하는 기준율을 계산한다. 알고리즘에 의해 분야트리의 루트가 되는 전체분야에서 기준율 이상에 집중하는 특정분야를 우선 탐색하고, 그 조작을 하위분야로 진행하여 종단분야에 도달하면 그 단어를 완전한 분야연상어로 결정한다. 이 조작에서 기준율 이하로 출현하는 분야연상어가 존재하면 수준 2, 수준 3, 수준 4의 분야연상어 후보탐색으로 제어를 변경한다. 이 알고리즘은 한 개의 분야연상어 후보에 대하여 분야트리를 상위에서 하위로 한번 탐색하여 정확한 분야를 결정할 수 있다.

2. 분야연상어

본 장에서는 분야연상어를 단일과 복합 분야연상어로 나누어 정의하고, 형태소 사전과의 관계를 설명한다. 분야체계에 의해 분야트리를 정의하고, 분야트리 내에서 분야연상어의 수준을 정의한다. 분야연상어가 시간의 흐름에 의해 변화하는 점에 주목하여 안정성 랭크를 새롭게 정의한다. 예를 들면, 분야 <야구>에 대하여 분야연상어 '투수' 혹은 '포수' 등은 비교적 안정한 분야연상어이지만, 고교야구에서 우승고나 선수명은 변화하는 불안정한 분야연상어이다. 또한 단일어로 된 분야연상어(단일 분야연상어)의 길이는 최단이며, 그 개수도 유한하기 때문에 분야연상어 후보를 사람이 직접 선별하고, 무한히 만들어지는 복합어로 된 분야연상어(복합 분야연상어)의 길이를 최소화하고, 분야연상어의 개수를 줄이도록 시도한다.

2.1 단일과 복합 분야연상어

더 이상 분할이 불가능한 의미를 가진 최소단위를 단어



(그림 1) 분야트리와 분야연상어의 예

라 하고, 형태소 사전에 등록되어 있는 단어를 "단일어"라 부른다. 두 단어 이상의 단일어로 구성되는 단어를 "복합어"라 부른다. 이들을 (그림 1)에 표시한 바와 같이 기호 "과 "내에 기술한다. 단일어와 복합어로 구성된 분야연상어를 각각 단일 분야연상어와 복합 분야연상어라 기술한다. 단, 미등록어는 분야연상어의 대상으로 하지 않는다.

한 개의 접사와 명사로 구성되는 일반적인 복합어 "소비세", "핵연료", "온난화" 등은 세분화함으로서 분야정보를 잃어버리기 쉽기 때문에 단일어로서 취급한다. 또한, 고등학교명 "군산상고"나 회사명 "한국통신" 등의 고유명사(인명은 제외)는 단일어로 취급한다. 인명에 대한 고유명사 중 "김응용"과 같이 성명으로 문서 중에 존재하는 경우는 단일어로 취급한다. "김응용감독"과 같이 인명과 보통명사가 함께 결합하여 복합어로 나타나면, 각 구성요소 "김응용"과 "감독"을 독립된 단일어로 취급한다. 참고문헌[11]를 전자화하여 세 개 이상의 단일어로 구성된 복합어(약 1,000개)의 분석 조사에서 새로운 분야연상어가 그다지 많이 검출되지 않았기 때문에 이후에는 두 개의 단일어로 구성되는 복합 분야연상어를 논의의 대상으로 한다.

2.2 분야트리

본 논문에서는 부록 A에 표시한 "분야체계"(이후, "분야트리"라 부른다)를 사용한다. 분야트리의 단말노드에 상당하는 분야를 "종단분야"라 부르고, 종단분야 이외는 모두 "중간분야"라 부른다. 이 분야트리의 전체 분야수는 180개이며, 중간 분야수는 22개, 종단분야 158개(깊이 2와 3의 종단분야는 각각 122개, 36개)이다. 어떤 분야와 직접 인접한 상위 혹은 하위분야를 각각 "부모분야"와 "자식분야"라 부른다. 분야의 지정은 분야명의 패스 <*Path*>로 기술하나, 루트에 상당하는 <전체분야>는 생략하고 종단분야 만으로 기술하는 것을 원칙으로 한다. 예를 들어, 분야패스 <*Path*> = <스포츠/태권도>는 <스포츠>의 하위 종단분야 <태권도>를 표시한다. 특히, 모순이 생기지 않는 경우에 한하여 상위의 패스지정은 모두 생략한다. '<' 과 '>'내의 분야명과 구분하기 위하여 의미 혹은 개념 명은 기호 '[' 과 ']'내에 기술한다. 앞의 (그림 1)에 분야트리의 예를 표시하였다.

2.3 분야연상어의 수준별 랭크

분야연상어는 분야를 결정할 때 여러 개 존재할 수 있고, 연상되는 범위는 단일, 복수 혹은 상위, 하위 분야를 한정하는 등 여러 경우가 존재할 수 있다. 따라서 분야체계 내에서 연상되는 분야의 범위에 제약을 가하여 분야연상어의 수준을 다음과 같이 정의[12]한다.

[정의 1] 분야연상어의 수준(Level)

(수준 1) 완전 분야연상어 : w는 유일한 종단분야만을 연상한다.

(수준 2) 준완전 분야연상어 : w는 같은 부모분야를 갖는 종단분야 중에서 한정된 복수 개의 종단분야만을 연상한다.

(수준 3) 중간 분야연상어 : w는 완전 분야연상어, 준완전 분야연상어가 아니고, 하나의 중간분야를 연상한다.

(수준 4) 다분야연상어 : w는 완전 분야연상어, 준완전 분야연상어, 중간 분야연상어가 아니고, 다수의 중간분야와 다수의 종단분야를 연상한다.

(수준 5) 비연상어 : w는 위의 수준 1~4 이외이고, 어떠한 특정분야도 연상하지 않는다.

예를 들어, 다음의 <표 1>에서와 같이 수준 1의 완전 분야연상어 "국기원"은 종단분야 <태권도>를 오직 하나의 뜻으로 한정한다. 수준 2의 준완전 분야연상어 "단식", "복식"은 부모분야 <스포츠>내의 복수의 종단분야 <테니스>, <탁구>, <배드민턴> 등을 한정한다. 수준 3의 중간 분야연상어 "시합"은 어떠한 종단분야도 한정하지 않으나, 한 개의 중간분야 <스포츠>를 한정한다. 수준 4의 다분야연상어 "승패"는 복수의 종단분야 <취미·오락/장기>, <정치/선거> 등과 중간분야 <스포츠>에 속하는 복수의 종단분야를 한정한다. 마지막으로, 수준 5의 비연상어는 "경우", "사용"과 같이 분야트리 내의 어떠한 특정분야도 한정하지 않는 단어를 비연상어로 정의한다.

〈표 1〉 분야연상어의 연상분야와 수준

분야연상어	연 상 분 야	수 준
국 기 원	<스포츠/태권도>	1
단식, 복식	<스포츠/테니스> <스포츠/탁구>	2
시 합	<스포츠>	3
승 패	<취미·오락/장기> <정치/선거> <스포츠>	4
경우, 사용	-	5

2.4 분야연상어의 안정성 랭크

본 절에서는 분야연상어가 시간경과에 의하여 변화하는 점에 주목한다. 예를 들면, 분야 <야구>의 경우 "투수", "포수" 등은 시간이 흘러도 변화하지 않는 안정한 분야연상어이지만, 고교야구의 우승 고나 선수 명은 문서에서 비교적 단기간에 출현하였다가 없어지는 불안정한 분야연상어이다. 또한, 프로야구의 팀 명이나 유명선수 명은 고교야구만큼 그 변화기간이 짧지 않으나, 영원히 불변하는 분야연상어는 아니다. 이와 같이 안정성이 낮은 분야연상어는 고유명사에 많고, 특히 인명의 안정성은 대단히 낮다고 생각된다. 따라서 각 분야연상어에 대해 안정성 랭크를 다음과 같이 정의한다.

[정의 2] 안정성 랭크(Stability Rank)

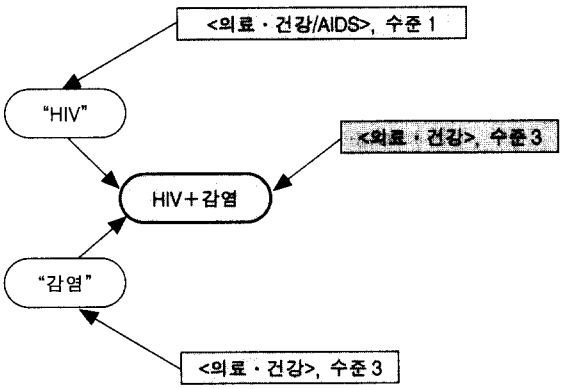
안정성 랭크는 랭크의 순위가 높은 순으로 보통명사를 a로, 인명이외의 고유명사를 b로, 인명에 해당하는 고유명사를 c로 할당한다. 다음의 <표 2>에 수준 1의 분야연상어와 안정성 랭크의 예를 표시하였다.

단일 분야연상어의 길이는 의미를 형성하는 가장 짧은 길이(최단길이)이며, 그 개수도 유한하기 때문에 분야연상어 후보를 사람의 상식지식을 이용하여 선별한다. 무한히 만들어지는 복합 분야연상어는 자동적으로 단어의 길이를 짧게 하고 그 수를 줄이는 방법이 필요하다.

예를 들어, (그림 2)와 같이 "HIV감염"은 <건강-의료/병명/에이즈>의 수준 1의 완전 분야연상어이지만, 이 복합어의 구성요소 중 "감염"은 <건강-의학> 분야의 수준 3의 중간 분야연상어이고, 복합어 "HIV감염"은 분야 <에이즈>의 수준 1의 완전 분야연상어 "HIV"를 구성어로 포함하고 있다. 이와 같이 복합어의 각 구성어가 동일한 분야를 중복하여 포함하고 있는 단어를 "분야 중복(redundancy) 연상어"라 부른다. 이에 대해서는 (그림 2)와 (그림 3)을 통하여 설명한다.

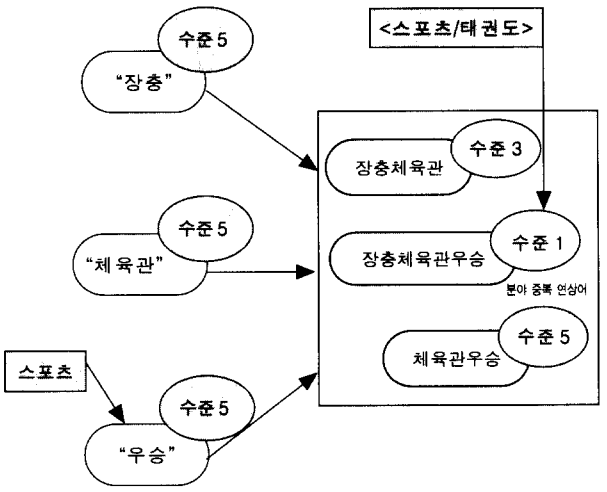
〈표 2〉 분야연상어 후보와 안정성 랭크

수 준	안정성 랭크	분야연상어 후보	분 야
1	b	해태타이거스	<스포츠/야 구>
1	a	투 수	<스포츠/야 구>
1	b	자 이 언 트	<스포츠/야 구>
1	a	야 구	<스포츠/야 구>
1	a	홈 런	<스포츠/야 구>
1	c	김 응 용	<스포츠/야 구>
1	a	선 구 안	<스포츠/야 구>
1	a	baseball	<스포츠/야 구>
1	a	홀 아 웃	<스포츠/골 프>
1	b	PGA	<스포츠/골 프>
1	a	티 샷	<스포츠/골 프>
1	a	단 련 대	<스포츠/태권도>
1	b	송 후 섭	<스포츠/태권도>
1	a	금 강 막 기	<스포츠/태권도>



(그림 2) HIV감염의 예

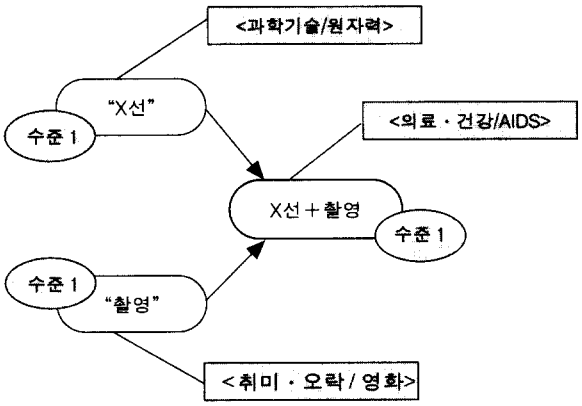
(그림 3)과 같이 동일분야 내에서의 분야연상어 수준이 변화하는 복합명사로써, "장충체육관우승"을 생각해보자. 단일어 "장충"은 비연상어이며, "체육관"과 "우승"은 <스포츠>의 중간 분야연상어이다. 이들 각 구성요소가 복합어 "장충체육관", "체육관우승", 혹은 "장충체육관우승" 등이 되면, "장충"이 수준 5의 비연상어이고, "체육관", "장충체육관"은 수준 3의 중간 분야연상어, "장충체육관우승"은 <태권도> 분야의 완전 분야연상어가 된다.



(그림 3) 장충체육관우승의 예

(그림 4)는 단일 분야연상어가 복합어를 구성하여 전혀 다른 분야의 분야연상어가 되는 경우의 예이다. <과학기술-학문/원자력>의 수준 1의 완전 분야연상어 "X선"과 <취미-오락/영화>의 수준 1의 완전 분야연상어 "촬영"으로 구성된 복합어 "X선촬영"은 각 단일어의 연상분야와 전혀 다른 <건강-의료>에의 수준 1의 완전 분야연상어로 변화한다.

이상의 분석으로 분야연상어 구축 시 먼저 단일 분야연상어를 결정하고, 복합어의 분야연상어 후보의 수준이 그것을 구성하는 단일 분야연상어의 수준보다 상위하는 경우에만 복합어의 연상범위를 재결정하도록 한다.



(그림 4) X선촬영의 예

2.5 분야연상어 규칙

이 절에서는 분야연상어의 결합규칙에 대해 간략히 알아본다. 단, 이 규칙은 향후의 연구과제를 위한 기초연구로 논의하는 것이기 때문에 본 논문에서 충분한 평가와 고찰은 하지 않는다. 2.4절의 "장충체육관우승"(수준 1의 완전 분야연상어)의 예와 같이 지역 명을 [대학명]으로 바꾼 "용인대학교태권도체육관우승" 혹은 "용인대학교체육관우승"의 경우, "용인대학교"를 "잠실학생"으로 바꾼 "잠실학생체육관우승"도 여전히 <스포츠>를 연상하는 단어이다.

"체육관"과 관련 개념어 "연습장"으로 치환하면 "용인대학교연습장"이 되어 <스포츠>를 연상하는 연상정보는 완전히 소멸한다. 따라서 [지명, 국명] + 장소 등의 결합규칙을 정의한다. 이와 같이 분야연상어 규칙을 찾아내면 사전에 등록되는 분야연상어의 총 수를 줄일 수 있다.

지금까지 분야연상어를 단일과 복합 분야연상어로 나누어 정의하였다. 형태소 사전에 채택된 단일 분야연상어의 특징에 관하여 설명하였다. 그리고, 분야체계를 "분야트리"로 정의하여 분야트리에 의한 분야연상어의 세기(분야를 한정하는 정도)와 분야를 한정하는 범위를 고려한 분야연상어의 수준을 정의하였다. 분야연상어가 시간의 경과에 의해 변화하는 점에 주목하여 안정성 랭크를 정의하고, 문서의 분야를 연상하는 시간에 대한 지속성에 관하여 논의하였다. 또한, 단일 분야연상어의 길이는 최단이며 그 개수도 유한하기 때문에 분야연상어의 후보를 사람이 직접 선별하도록 하였다. 이 방법은 분야연상어를 서로 관계 있는 것끼리 분류하고, 이들을 안정성에 따라 구성하여 인간이 문서에서 특정 단어를 보고 사고하는 방식과 유사한 방법으로 지식을 처리할 수 있다. 그리하여, 무한히 만들어지는 복합 분야연상어의 길이를 줄여 본 논문의 목표인 복합 분야연상어의 수를 적게 하는 방법을 논의하였다.

3. 분야연상어의 결정

본 장에서는 학습데이터에서 분야연상어의 후보와 그 수

준을 자동으로 결정하는 방법과 알고리즘에 대하여 설명한다. 학습데이터는 각 종단분야에 대하여 균일하게 수집되었다고 볼 수 없기 때문에 전체분야에 대하여 특정한 분야연상어의 합계빈도를 계산하여 각 종단분야에 출현하는 단어 빈도로 정규화 한 값을 이용한다. 이 정규화 된 빈도정보에 의해 특정단어가 특정분야에 집중하는 집중률을 정의한다.

본 논문의 방법은 한 개의 분야연상어 후보에 대하여 분야트리의 상위에서 하위로 한번 탐색하여 그 분야연상어가 지시하는 분야를 빠르고 정확하게 결정한다. 분야트리의 루트(전체분야를 의미)에서 기준율 이상으로 집중하는 특정분야를 집중 탐색하여 그 조작을 하위분야로 진행한다. 조작이 종단분야에 도달하면 수준 1의 완전 분야연상어와 정확한 연상분야를 결정하고, 기준율 이하의 분야에서 출현하는 분야연상어이면 조작을 수준 2, 수준 3 혹은 수준 4의 분야연상어 탐색으로 알고리즘의 제어를 변경한다.

3.1 수준 결정 알고리즘

인간이 수집한 분야연상어가 각각의 종단분야에 균일하게 수집되었다고 보기 어렵기 때문에 종단분야 <T>에 출현하는 모든 단어의 합계빈도 *Total_Frequency*(w)를 계산하고, 종단분야 <T>에 출현하는 단어 w의 빈도를 *Frequency*(w, <T>)라 하면, 다음의 식 (1)과 같이 정규화 된 빈도 *Normalization*(w, <T>)를 정의한다.

$$Normalization(w, \langle T \rangle) = \left\{ \frac{Frequency(w, \langle T \rangle)}{Total\_Frequency(w)} \right\} \times \gamma \quad (1)$$

여기서, γ(정수) 값에 대하여 설명하면, 스포츠 분야 전체에서 출현하는 분야연상어 "야구"의 전체 합계빈도 *Total_Frequency*("야구")는 약 100,000 단어 정도 출현하고, 실제 <야구>에서는 약 1,000 단어 정도 출현한다. 이것은 빈도가 높은 단어의 정규화 된 값이 매우 적음을 의미한다. 따라서 적당한 값 γ을 계산하여 출현빈도를 조정한다. 이후에는 부록 A에서 제시하는 전체 분야트리에서 출현한 빈도정보를 이용하여 계산한 $\gamma = 10^5$의 값으로 논의를 진행한다.

어떤 분야 <P>를 <C>의 부모분야라 할 때, 분야 <C>에 대한 분야연상어 w의 집중률 *Concentration*(w, <C>)는 중간분야 <P>에 대한 정규화 된 빈도 *Normalization*(w, <P>)를 <P>의 하위에 존재하는 종단분야 <C>의 *Normalization*(w, <C>)로 나눈 다음 식 (2)와 같이 정의한다.

$$Concentration(w, \langle C \rangle) = \frac{Normailzation(w, \langle C \rangle)}{Normalization(w, \langle P \rangle)} \quad (2)$$

다음에 분야연상어를 각 수준에 따라 적당한 후보로 결정하기 위한 알고리즘을 표시하였다.

- 분야연상어의 수준결정 알고리즘

  입력 : ① 분야연상어의 후보어 w

  ② 분야 <C>를 연상한다고 추측되는 단어 w의 Normalization(w, <C>) 값

  ③ 분야트리

  출력 : 연상되는 분야와 수준

(순서 A1) [완전 분야연상어의 결정]

분야트리의 루트 <P>에서 그 자식분야 <P/C>에 대하여 단어 w가 집중되어 있는가 혹은 그렇지 않은가를 다음의 조건식

$$Concentration(w, \langle P \rangle) \geq \alpha \quad (3)$$

으로 판정하고, 조건식을 만족하면 <P>에 자식분야 <C>를 연결하여 <P/C>로 바꾸고, 다시 하위의 자식분야에 대해 동일한 판정을 수행한다. α는 0.5보다 큰 값을 채택하였기 때문에 이 조건식을 만족하는 분야연상어 w는 한 개 미만이다. 반복처리 결과 <P/C>가 종단분야가 되면, w를 분야 <P/C>의 "완전 분야연상어(수준 1)"로 결정한다. 만약 조건식을 만족하는 <P>의 자식분야 <P/C>가 존재하지 않으면 다음으로 진행하여 수준 2, 수준 3, 수준 4의 분야연상어 후보 판정을 진행한다.

(순서 A2) [준완전 혹은 중간 분야연상어의 결정]

분야연상어 w가 수준 1로 결정되지 않은 경우에 분야 <P>는 종단분야까지 도달하지 않았다는 것을 의미한다. 따라서 이 분야 <P>는 반드시 중간분야이며, 최소한 m(≥2)개의 자식분야들을 가지고 있다. 수준 2는 <P>의 복수 개의 자식분야에 집중하는 분야를 탐색하는 것이 목적이므로 집중률의 총합이 α이상이 되는 자식분야를 복수 개 가진다. 단, 복수의 자식분야 수를 분야 <P>의 모든 자식분야의 수 m에 가깝게 하면 수준 3의 중간 분야연상어와의 구별이 명확하지 않게 되므로 후보가 되는 자식분야수의 빈도는 형제(혹은 자매) 분야 중 평균빈도 이상의 분야를 선정한다.

$$Concentration(w, \langle P \rangle) \geq \frac{Normalization(w, \langle C \rangle)}{m} \quad (4)$$

식 (4)를 만족하는 자식분야 <P/C>를 추출하고, Concentration(w, <P/C>)의 값이 큰 순서부터 누적가산하고, 누적 가산치가 최초에 α를 넘으면, k(1 < k < m)개의 자식분야를 결정한다. 이 때 k개의 자식분야 <P/C>가 모두 종단분야이면, w를 분야 <P/C>의 "준완전 분야연상어(수준 2)"로 결정한다. 모두 종단분야가 아니면, 다음 순서로 진행하여 다분야연상어 판정을 수행한다. 차례차례 누적 가산한 누적치가 α를 초과하지 않으면, w를 <P>의 "중간 분야연

상어(수준 3)"로 결정한다.

(순서 A3) [다분야연상어의 결정]

k개의 자식분야에서 종단분야 <P/C>를 추출하고, w를 분야 <P/C>의 다분야연상어로 결정한다. 종단분야를 제외한 자식분야 <P/C>를 부분트리의 루트 <P>로 수정하여 (순서 A1)과 (순서 A2)를 다시 실행하면 복수 개의 중간분야와 종단분야가 얻어진다. w를 분야 <P>의 "다분야연상어(수준 4)"로 결정한다.

(알고리즘 종료)

덧붙여 말하면, (순서 A2)에서 언급한 식 (4)는 자식분야의 후보를 평균빈도 이상으로 제한하는 것을 의미한다. 이것은 낮은 빈도의 자식분야를 후보로 포함하면 <P>의 대다수 자식분야 <P/C>가 준완전 분야연상어가 되어 <P>의 중간 분야연상어와의 구별이 모호하게 될 수 있다.

3.2 수준의 결정 예제

이 절에서는 분야연상어 "국기원", "단식", "시합", "승패" 등의 수준별 결정 예를 설명한다. 분야트리에서 출현하는 각각의 분야연상어 빈도 수를 (그림 1)의 괄호 안에 표시하였다. <분야전체>의 자식분야 수는 12, <스포츠>, <취미·오락>, <정치>의 자식분야 수는 각각 19, 13, 14개 이고, 임계값 α는 수 십 차례의 실험을 통하여 얻은 값 $\alpha = 0.92$를 사용한다.

(순서 A1)에서, w = "국기원", $<P>$ = <분야전체>에 대하여, 단어 "국기원"이 가장 많이 출현하는 분야 <스포츠>를 선택하여 $<C>$ = <스포츠>에 대한 집중률을 다음과 같이 계산한다.

$$\begin{aligned} &Concentration(\text{"국기원"}, \langle \text{스포츠} \rangle) \\ &= \frac{Normalization(\text{"국기원"}, \langle \text{스포츠} \rangle)}{Normalization(\text{"국기원"}, \langle \text{분야전체} \rangle)} \\ &= \frac{236}{243} \fallingdotseq 0.97 \geq \alpha (= 0.92) \end{aligned}$$

이 되어 "국기원"은 분야 $<P/C>$ = <분야전체/스포츠>에 집중한다. 다음의 <스포츠> 분야에서 "국기원"의 빈도가 가장 높은 분야 <태권도>를 선정하여 $<P>$를 <스포츠/태권도>로 고쳐서 하위의 분야 $<C>$ = <태권도>에 대하여 판정하면,

$$\begin{aligned} &Concentration(\text{"국기원"}, \langle \text{태권도} \rangle) \\ &= \frac{Normalization(\text{"국기원"}, \langle \text{태권도} \rangle)}{Normalization(\text{"국기원"}, \langle \text{스포츠} \rangle)} \\ &= \frac{231}{236} \fallingdotseq 0.98 (\geq \alpha) \end{aligned}$$

가 되고, 현재의 분야 <태권도>가 종단분야이므로 "국기원"은 <태권도> 분야에 대한 완전 분야연상어(수준 1)로 결정된다.

다음에, w = "단식"은 $<P/C>$ = <전체분야/스포츠>에서 다음의 집중률이 얻어진다.

$$\begin{aligned} &Concentration(\text{"단식"}, \langle \text{스포츠} \rangle) \\ &= \frac{Normalization(\text{"단식"}, \langle \text{스포츠} \rangle)}{Normalization(\text{"단식"}, \langle \text{분야전체} \rangle)} \\ &= \frac{351}{365} \fallingdotseq 0.96 (\geq \alpha) \end{aligned}$$

그러나, 분야 <스포츠>의 하위분야 중에서 α이상의 값으로 집중하는 분야가 다음의 계산결과와 같이 존재하지 않는다.

$$\begin{aligned} &Concentration(\text{"단식"}, \langle \text{테니스} \mid \text{탁구} \mid \text{배드민턴} \mid \text{태권도} \mid \text{야구} \rangle) \\ &= \frac{Normalization(\text{"단식"}, \langle \text{테니스} \mid \text{탁구} \mid \text{배드민턴} \mid \text{태권도} \mid \text{야구} \rangle)}{Normalization(\text{단식}, \langle \text{스포츠} \rangle)} \\ &= \frac{142 \mid 93 \mid 105 \mid 1 \mid 3}{351} \\ &\fallingdotseq 0.405 \mid 0.265 \mid 0.299 \mid 0.003 \mid 0.009 \end{aligned}$$

따라서, 앞의 알고리즘에서 (순서 A2)로 진행하여 <분야전체/스포츠>에서의 수준 2와 수준 3의 분야연상어로 결정되는 절차를 수행한다. 3.2절의 서두에 분야 <스포츠>의 자식분야수(m)는 19로 정의하였기 때문에 식 (4)의 우변은

$$\frac{Normalization(\text{"단식"}, \langle \text{스포츠} \rangle)}{m} = \frac{351}{19} \fallingdotseq 18.74$$

이다.

<스포츠>의 하위분야 중에서 이 수치보다 빈도가 높은 자식분야는 빈도 142, 93, 105를 갖는 각각의 분야 <테니스>, <탁구>, <배드민턴>이며, 집중률 $Concentration(w, <P/C>)$을 각각 계산하면,

$$\begin{aligned} Concentration(\text{"단식"}, \langle \text{테니스} \rangle) &= \frac{142}{351} \fallingdotseq 0.41 \\ Concentration(\text{"단식"}, \langle \text{탁구} \rangle) &= \frac{93}{351} \fallingdotseq 0.27 \\ Concentration(\text{"단식"}, \langle \text{배드민턴} \rangle) &= \frac{105}{351} \fallingdotseq 0.30 \end{aligned}$$

가 되고, 이들 후보 중 가장 큰 값을 갖는 최초 두 개의 자식분야 <테니스>와 <배드민턴>의 집중률을 가산하면

$$\begin{aligned} &Concentration(\text{"단식"}, \langle \text{테니스} \rangle) \\ &+ Concentration(\text{"단식"}, \langle \text{배드민턴} \rangle) \\ &= 0.41 + 0.30 = 0.71 \end{aligned}$$

이 되지만, 여전히 기준 집중률(α)로 사용된 임계값(0.92)를 초과하지 않는다. 따라서, 다음 후보분야 <탁구>와 가

〈표 3〉 분야 〈야구〉에서 수준 1에 해당하는 단일 분야연상어의 예

수 준	안 정 성	분야연상어 후보	분 야	집 중 률	빈도
1	b→a	해 태	<스포츠/야구>	0.99	944
1	a	투 수	<스포츠/야구>	0.99	703
1	b	박 찬 호	<스포츠/야구>	1.00	697
1	a	야 구	<스포츠/야구>	0.96	692
1	a	사직구장	<스포츠/야구>	1.00	442
1	c	김 응 용	<스포츠/야구>	0.99	231
1→5	a	중전안타	● <스포츠/야구>	1.00	74
1→5	b	성균관대	● <스포츠/야구>	0.93	64
1→3	a	신임감독	<스포츠/야구>→<스포츠>	0.92	22
1	a	선 구 안	<스포츠/야구>	1.00	3
1	a	baseball	<스포츠/야구>	1.00	2

산하면 $0.67 + Concentration(\text{"단식"}, \langle \text{탁구} \rangle) = 0.41 + 0.30 + + 0.27 = 0.98$이 되어 비로소 0.92를 초과한다. 따라서, "단식"은 분야 <스포츠>의 세 개의 자식분야 <C>=<테니스>, <배드민턴>, <탁구>에 집중하여 출현하고, 이들은 모두 종단분야이므로 "단식"은 준완전 분야연상어(수준 2)가 된다.

위의 예와 비슷하게 "시합"의 경우, <P>=<분야전체>의 유일한 하위분야 <C>=<스포츠>에 집중하지만, 그 하위의 분야에 집중하는 분야는 존재하지 않는다. 따라서, (순서 A2)에서 분야 <분야전체/스포츠>에 대한 수준 2와 수준 3의 분야연상어 결정을 수행한다. "단식"의 예와 동일하게 "시합"은 분야 <스포츠>의 자식분야 수 $m(=19)$에 의해 식 (4)의 우변은

$$\frac{Concentration(\text{"시합"}, \langle \text{스포츠} \rangle)}{19} = \frac{364}{19} \fallingdotseq 19.16$$

이므로 이 수치보다 높은 빈도를 갖는 자식분야는 <테니스>, <탁구>, <배드민턴>, <야구>이고, 각각 59, 34, 42, 73의 빈도를 갖는다((그림 1) 참조). 따라서 각각의 *Concentration* (w, <P/C>)는 다음의 수치로 계산된다.

$$Concentration(\text{"시합"}, \langle \text{테니스} | \text{탁구} | \text{배드민턴} | \text{야구} \rangle)$$
$$= \frac{59 | 34 | 42 | 73}{364} \fallingdotseq 0.16 | 0.09 | 0.12 | 0.2$$

이들의 수치를 모두 가산하여도 0.57이 되어 $\alpha(=0.92)$를 초과하지 않는다. 따라서 네 개의 자식분야 중 어떤 분야에도 집중적으로 출현하지 않는다고 판정하고, "시합"은 중간분야연상어(수준 3)로 결정된다.

다른 예로, w="승패"는 <P>=<분야전체>의 어느 하위분야에도 기준치 $\alpha(=0.92)$ 이상으로 집중되지 않으므로 (순서 A2)에 의해 집중하는 복수분야가 결정된다. <분야전체>의 자식분야 수 m은 13이므로 식 (1)에 의해 다음과 같이 계산된다.

$$\frac{Normalization(\text{"승패"}, \langle \text{전체분야} \rangle)}{m} = \frac{478}{13} \fallingdotseq 36.77$$

이 값 이상의 빈도를 갖는 자식분야는 <스포츠>, <취미·오락>, <정치>이며, 각각 다음의 집중률을 얻는다.

$$Concentration(\text{"승패"}, \langle \text{스포츠} \rangle) = \frac{213}{478} \fallingdotseq 0.45$$
$$Concentration(\text{"승패"}, \langle \text{취미·오락} \rangle) = \frac{161}{478} \fallingdotseq 0.34$$
$$Concentration(\text{"승패"}, \langle \text{정치} \rangle) = \frac{94}{478} \fallingdotseq 0.19$$

여기서, 이들 세 분야의 누적 가산치 $0.98(\fallingdotseq 0.45+0.34+0.19)$은 기준 집중률 α를 초과하지만, 분야 <P/C>=<스포츠>, <취미·오락>, <정치>는 모두 종단분야가 아니다. 따라서, "승패"는 수준 2의 준완전 분야연상어가 될 수 없으며, <C>=<스포츠>, <취미·오락>, <정치> 등을 부분 트리로 하여 (순서 A1)과 (순서 A2)를 반복 실행하면 "승패"는 분야 <스포츠>에 집중하지만, 그 하위의 자식분야에는 집중하지 않는다. 분야 <취미·오락>, <정치>에 대해서는 그 자식분야 <취미·오락/장기>, <정치/선거>에 집중함을 알 수 있다. 따라서, "승패"는 중간분야 <스포츠>와 종단분야 <취미·오락/장기>, <정치/선거> 등에 해당하는 다분야연상어(수준 4)가 된다.

3.3 단일 분야연상어의 결정

위의 알고리즘이 제시하는 단일 분야연상어의 후보어, 수준 및 안정성 랭크, 연상분야, 집중률, 빈도정보를 이용하여 연상되는 분야와 랭크열을 사람의 상식지식으로 판단한다. <야구>에 대한 단일 분야연상어 중에서 수준 1의 후보를 <표 3>에 표시하였다. 기호 "→"는 분야의 변경을 의미하고, ●는 삭제될 분야연상어를 의미한다. 예를 들면, 후보어 "중

〈표 4〉 〈스포츠〉에 대한 수준별 단일 분야연상어의 예

수준 2	안 정 성	분야연상어 후보	분 야 A	분 야 B	분 야 C
2	a	선 발	<스포츠/야구>	<스포츠/축구>	-
2	a	선제공격	<스포츠/야구>	<스포츠/축구>	<스포츠/럭비>
2	a	승 점	<스포츠/야구>	<스포츠/축구>	<스포츠/요트>
2→3	a	처녀출전	<스포츠/야구> → <스포츠>	●<스포츠/축구>	●<스포츠/골프>
2→5	a	기 어	●<스포츠/야구>	●<스포츠/축구>	●<스포츠/배구>
2→1	a	구 단 측	<스포츠/야구>	●<스포츠/축구>	-
2	c	김 병 연	<스포츠/야구>	<스포츠/축구>	-
수준 3	**안 정 성**	**분야연상어 후보**	**분 야 A**	**분 야 B**	**분 야 C**
3	a	시 합	<스포츠/야구>	<스포츠/축구>	<스포츠/배구>
3	a	리 딩 히 터	<스포츠/골프>	<스포츠/축구>	<스포츠/요트>
3	b	바르셀로나	<스포츠/야구>	<스포츠/검도>	<스포츠/육상>
3→5	a	출 전	●<스포츠/야구>	●<스포츠/농구>	●<스포츠/럭비>
3→5	a	중 반	●<스포츠/야구>	●<스포츠/복싱>	●<스포츠/육상>
3	a	선 수	<스포츠/축구>	<스포츠/야구>	<스포츠/배구>
3→5	a	오 른 발	●<스포츠/축구>	<스포츠/태권도>	●<스포츠/야구>
수준 4	**안 정 성**	**분야연상어 후보**	**분 야 A**	**분 야 B**	**분 야 C**
4	a	감 독	<스포츠/축구>	<스포츠/야구>	<오락-취미/영화>
4	a	승 부	<스포츠>	<취미-오락/장기>	<오락-취미/경마>
4→5	c	최 희 석	●<스포츠/야구>	●<스포츠/럭비>	●<정치/한국정치>
4→1	c	박 찬 호	<스포츠/야구>	●<스포츠/취미-오락/장기>	●<오락-취미/TV>
4→2	b	성균관대	●<스포츠/야구>	<교육/학교행사>	<교육/외국어교육>
4	a	트래이드	<스포츠/야구>	<경제/주식-채권>	<경제/세계경제>
4	a	밧 데 리	<스포츠/야구>	<환경문제/환경-에너지>	<과학-기술/우주개발>

전안타", "성균관대"의 분야는 삭제되고, "신임감독"은 <스포츠>로 변경되어 있다. "해태(구단명)"의 안정성 랭크 b는 a로 변경[1]되어 있다. 분야정보에서 수준의 변경은 자동적으로 결정된다. <표 4>는 단일 분야연상어 중에서 수준 2~4의 후보어를 표시하고 있으나, <표 3>과는 달리 복수분야가 제시[2]되고 있다. 사람의 수정에 의해 수준 2의 "처녀출전"은 수준 3으로 변경되어 있으며, 수준 2의 "구단측"과 수준 4의 "박찬호"는 수준 1로 변경되어 있다.

이상과 같이 학습데이터를 각 종단분야에 대해 균일하게 수집하도록 종단분야에 출현하는 모든 단어의 합계빈도를 이용하여 정규화하였다. 이 정보를 이용하여 단어가 특정분야에 집중하는 집중률을 정의하였고, 분야연상어의 수준을 결정하는 알고리즘을 제안하였다. 분야연상어의 수준결정은 분야트리의 하위에서 각 분야에 집중하는 분야연상어를 결정하며 1, 2, 3, 4의 순으로 분야연상어의 수준이 결정된다. 수준 2와 3의 구별을 위해 빈도의 평균치를 넘을 때까지 누적 가산하였는데, 이 평균치의 결정방법에 대해서는 후속 논문을 통해 깊이 고찰한다.

4. 기대효과

본 논문의 연구내용을 통해 다음과 같은 기대효과를 얻을 수 있다.

4.1 다국어간 분야번역과 문서요약에 응용

상식적 분야체계에 대한 다국어(multilingual)의 학습데이터를 통해 각 언어별로 분야연상어를 구축하면 분야체계에 대한 다국어정보를 구축할 수 있다. 다국어 분야정보를 이용하여 보편적인 분야정보를 추출하면 다국어 문서의 요약과 번역도 가능할 것으로 기대한다. 분야정보는 문서가 내포하는 상세한 의미나 정보를 인간에게 제공할 수는 없으나, 인간과 컴퓨터의 초기적인 통신 수단으로 이용되는 초점정보가 될 수 있다. 그러나, 표층적인 단어를 번역하는 간단한

1) 구단명 이외에도 "김응용", "데이비드 존슨(미국 LA Dodgers 팀의 감독명)", "이종범" 등은 고유명사이지만, 안정성이 높은 분야연상어로 분야 <야구>에 대한 일반성을 갖고 있기 때문에 사람이 직접 판단하여 랭크 a로 결정한다. 안정성 랭크 a는 다음에 언급하는 분야 중복 연상어를 제거하는 중요한 초점 정보가 된다.

2) 각 분야연상어는 <스포츠>에 대한 분야연상어이며, 사람이 직접 판단하여 하위분야를 표시하였다. 집중률과 빈도정보가 분야마다 다르게 제시되지만, 예제에서는 생략하고, 분야수도 세 종류로 한정하여 표시하였다. 또한, 분야정보가 없는 경우는 "-"으로 표시하였다.

처리만으로 적당한 분야를 결정할 수 없는 경우도 있다. 예를 들면 <야구>에 대한 분야연상어 "김응용감독"이나 "나가시마감독"을 영어로 번역하더라도 반대로, 야구분야의 분야연상어 "Babe Ruth"를 한국어나 일본어로 번역하더라도 분야연상어로 적당하다고 볼 수 없기 때문에 분야정보 <야구>만이 유용한 정보가 된다. 따라서 각 나라의 문화와 각 나라 고유의 분야정보는 나라별 분야연상어로 독립적으로 구축되어야 할 것이다.

4.2 분야체계의 확장

부록 A의 분야체계는 문서 수집의 용이성을 우선적으로 고려하였기 때문에 상식적인 분야체계로 사용하기에는 부적절한 분야도 간혹 존재한다. 따라서, 향후에 분야체계를 계속 보완, 수정, 확장하여야 한다. 저자의 실험실에서 수집한 한국어 분야체계는 수집한 문서데이터를 기반으로 작성하였다. 이에 의해 본 연구와 관련된 후속논문에서 실험평가를 하고 그 유효성을 정확하게 고찰하고 싶다. 확장된 분야체계 중 <스포츠> 분야 전체를 부록 B에 표시하였다.

지금까지 내용을 요약하면 논문에서 구축한 분야연상어는 어휘자체가 분야정보를 갖고 있기 때문에 인간의 뇌에서 일어나는 인지작용에 의해 분야의 연상작용이 가능하다. 특히, 단편문서에 대해서 인간은 소수의 분야연상어로 분야를 결정할 수 있기 때문에 본 논문에서는 초기단계에 분야결정에 관하여 실험을 수행하여 분야연상어의 유용성을 평가하고자 한다.

5. 결 론

본 논문에서는 분야연상어를 정의하고, 단일어에 대한 분야연상어 정보를 이용하여 일상생활에서 끊임없이 생성되는 복합 분야연상어를 효율적으로 결정하는 방법을 제안하여 180분야의 학습데이터를 토대로 그 유효성을 평가하였다. 구축된 분야연상어가 단편문서의 분야결정에 유효한 것인가에 대해 논의하였다.

분야연상어를 단일과 복합 분야연상어로 분류하여 단일 분야연상어를 형태소사전에 등록된 표제어와 일치하도록 한정하였다. 이것은 단일 분야연상어의 분야정보를 형태소사전에 그대로 등록하기 위한 실용성을 고려한 것이다. 또한, 본 연구에서는 분야체계를 미리 정의한다고 하였으나 분야연상어 구축은 어떠한 분야체계에도 손쉽게 적용될 수 있으므로 보편성은 충분하다. 다음은 향후에 수행할 과제에 대하여 정리한다.

- 분야연상어가 시간경과에 의해 변화하는데 주목하여 안정성 랭크를 새롭게 정의하였으나, 이 랭크에 대해서는 보다 상세한 분석에 의해 정확한 랭크의 정의가 존재할 가능성이 있기 때문에 향후에도 계속 연구하고자 한다.
- 인명에 해당하는 고유명사에 대해서는 "김대중대통령"의 실체를 이해하는 일이나 이 단어와 "박찬호투수"를 구별하는 것이 애매하므로 문서에서 독립된 단어의 실체를 이해하는 방법의 도입은 계속하여 연구할 중요한 과제이다.

본 논문에서는 학습데이터에서 분야연상어의 후보와 그 수준을 자동적으로 결정하는 알고리즘을 제안하였다. 학습데이터의 불균형성에 대해서는 상대빈도를 이용하여 빈도를 정규화하고, 분야연상어가 특정분야에 집중하는 기준율 α를 정의하였다. 이 연구에 대하여 다음의 과제가 생각된다.

- 적은 수의 문서가 수집된 분야에서 정규화 한 빈도가 대단히 크게되는 점을 보완할 필요가 있다.

제안한 알고리즘은 분야트리의 루트가 되는 <전체분야>에서 기준율 α이상에 집중하는 특정분야를 탐색하고 그 조작을 하위의 분야에 진행하여 종단분야에 도달하면 수준 1의 분야연상어와 그 특정분야를 결정하였다. 조작 중에서 언제나 기준치 α값을 이용하였으나 다음 문제를 보완하여야 한다.

- 기준율 α를 만족하지 않으나 α에 대단히 가까운 분야연상어의 취급은 어떻게 할 것인가? 이 점은 실제로 수집된 분야연상어 데이터를 충분히 분석하여 연구하여야 할 것이다.

참 고 문 헌

[1] M. J. Blosseville et al., "Automatic Document Classification : Natural Language Processing, Statistical Analysis, and Expert System Techniques Used Together," Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '92), pp.51-58, 1992.

[2] Norbert Fuhr, "Models for Retrieval with Probabilistic Indexing," Information Processing & Management, Vol.25, No.1, pp.55-72, 1989.

[3] Fumiyo Fukumoto et al., "Automatic Clustering of Articles Using Dictionary Definition," Transactions of Information Processing Society of Japan, Vol.37, No.10, pp.1789-1799, 1996(in Japanese).

[4] Masami Hara et al., "Keyword Extraction Using a Text Format and Word Importance in a Specific Field," Transactions of Information Processing Society of Japan, Vol.38, No.2, pp.299-309, 1997(in Japanese).

[5] Yoshitaka Hayashi, et al., "Efficient Method for Extracting Keywords of Compound Words Using Pattern Matching Machines," Transactions of Information Processing Society of Japan, Vol.38, No.4, pp.815-825, 1997(in Japanese).

[6] Naoyuki Nomura, "ConceptBase-A NL-based IT Solution Core," Proceedings of the 1999, the 18th International Conference on Computer Processing of Oriental Language (ICCPOL '99), pp.235, 1999.

[7] Salton, G., "Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer," Addison-Wesley Publishing Company, 1989.

[8] Salton, G. and McGill, M. J., "Introduction of Modern Information Retrieval," McGraw-Hill Book Company, 1983.

[9] Tokunaga, T. and Iwayama, M., "Text Categorization based on Weighted Inverse Document Frequency," Natural Language Processing, Vol.100, No.5, 1994.

[10] Mochizuki, H., Makoto, I., and Okumura, M. "Passage-Level Document Retrieval Using Lexical Chains. Journal of Natural Language Processing," Vol.6, No.3, pp.101-126, 1999(in Japanese).

[11] 남영신, 우리말 분류 사전, 성안당, 2001.

[12] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회논문지B, 제10권 제1호, pp.57-66, 2003.

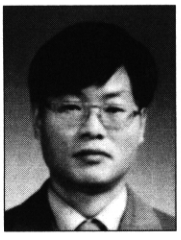
부록 A. 분야체계와 수집데이터의 양

분야체계의 루트에 바로 아래에 있는 자식분야를 다음에 기술한다. 분야명은 < >내에 기술하고 괄호내에는 파일 수와 KByte를 표시하였다. 이 분야의 하위분야는 괄호로 표시한다.

<분야전체(15,435 ; 42,092)>
<스포츠(1,856 ; 5,527)>
골프, 야구, 배드민턴, 농구, 유도, 야구, 레슬링, 배구, 테니스, 태권도, 쓰모, 복싱, 축구, 럭비, 수영, 검도, 동계스포츠(스키, 스케이트, 스키점프, 봅슬레이), 육상(포환던지기, 해머던지기, 원반던지기, 100m, 마라톤, 삼단뛰기), 모터스포츠(F1, 모터크로스, 보드)
<취미-오락(1,680 ; 4,891)>
애니메이션, 희극, 장기, 컴퓨터게임, 여행, 영화, 경마, 요리-식음료, 예술, 독서, 음악, TV
<과학기술-학문(735 ; 7,074)>
우주개발, 해양개발, 군사기술, 건축, 원자력, 전기전자, 재료, 화학, 수학, 물리학, 고고학, 언어학, 생물학-바이오, 컴퓨터(S/W, H/W)
<자연(102 ; 517)>
지구과학, 지진-화산, 천문학, 기상학
<건강-의료(514 ; 3,708)>
진단, 병명(O-157, 아토피성피부염, 에이즈, 암, 당뇨병, 간 질환), 건강(다이어트, 스트레스, 콜레스테롤, 혈압)
<환경(1,618 ; 2,782)>
인구증가, 공해, UN정책, 자연재해, 오존파괴, 쓰레기문제, 에너지, 온난화, 도로-교통
<교육(1,622 ; 4,102)>
교육기관, 교사-교수, 수험-입시, 외국어교육, 교육장소, 학교행사, 자격, 교육교재, 학벌, 교육문제(왕따, 장기결석)
<사회(1,104 ; 1,824)>
광고, 유행, 문화활동, 전쟁-분쟁, 저널리즘, 사건-사고, 재해(화재, 지진, 수해, 태풍)
<생활(998 ; 1,742)>
주택, 식생활, 여성, 보험, 연금, 가족-가정, 복지, 세금
<국제관계(2,179 ; 3,991)>
아시아, 중국, 일본, 한국, 북한, 오세아니아, 유럽, 미국, 캐나다, 중동, 남미, 아프리카, 중동, 소련, 북극-남극
<정치(2,026 ; 4,910)>
사법, 국회, 압력단체, 지방자치, 외교, 헌법, 정당, 정치이론, 선거, 한국정치, 국제정치, 세제, 행정-내각, 방위
<경제(1,011 ; 4,024)>
세계경제, 노동, 외화, 금융, 미국경제, 일본경제, 한국경제, 경영, 재무회계, 금융일반, 재정, 무역, 농업, 어업, 경기-물가, 주식-채권, 마케팅

부록 B. <스포츠>의 분야체계를 확장한 예

M. 스포츠
M.0 기타
M.1 구기
M.1.0 기타
M.1.1 테니스
M.1.2 배구
M.1.3 비치발리볼
M.1.4 농구
M.1.5 축구
M.1.6 럭비
M.1.7 미식축구
M.1.8 핸드볼
M.1.9 골프
M.1.10 야구
M.1.11 탁구
M.1.12 라켓볼
M.1.13 배드민턴
M.1.14 볼링
M.1.15 소프트볼
M.1.16 소프트테니스
M.1.17 당구
M.1.18 코프볼
M.1.19 네트볼
M.1.20 스쿼시
M.1.21 하키
M.2 동계스포츠
M.2.0 기타
M.2.1 스키
M.2.2 스피드 스케이팅
M.2.3 피겨 스케이팅
M.2.4 아이스하키
M.2.5 스키 점프
M.2.6 봅슬레이
M.2.7 커를링
M.2.8 알파인 스키
M.2.9 스노우보딩
M.2.10 루게
M.3 수상경기
M.3.0 기타
M.3.1 오픈 워터 스위밍
M.3.2 핀 스위밍
M.3.3 싱크로나이즈
M.3.4 수영
M.4 마린스포츠
M.4.0 기타
M.4.1 스쿠버다이밍
M.4.2 요트
M.4.3 세일링
M.4.4 카누
M.4.5 보트
M.4.6 수상스키
M.4.7 서핑
M.4.8 바디보딩
M.4.9 롱보드
M.5 육상경기
M.5.0 기타
M.5.1 마라톤
M.5.2 트랙경기
M.5.3 조깅
M.5.4 포환던지기
M.5.5 해머던지기
M.5.6 스프리팅
M.5.7 풀로볼트
M.6 스카이스포츠
M.6.0 기타
M.6.1 스카다이빙
M.6.2 글라이딩
M.6.3 에어로노틱스
M.6.4 에어로바틱스
M.6.5 파라추팅
M.7 인도어 스포츠
M.7.0 기타
M.7.1 인도어 암벽오르기
M.8 모터 스포츠
M.8.0 기타
M.8.1 F1
M.8.2 모터크로스
M.8.3 파워보우팅
M.9 격투기
M.9.0 기타
M.9.1 유도
M.9.2 검도
M.9.3 궁도
M.9.4 합기도
M.9.5 쓰모
M.9.6 가타
M.9.7 쿠미테
M.9.8 태권도
M.9.9 복싱레슬링
M.9.10 아쉐리
M.9.11 펜싱
M.10 짐네스틱
M.10.0 기타
M.10.1 리듬체조
M.10.2 기계체조
M.10.3 에어로빅
M.10.4 트렘폴린
M.11 등산
M.12 사이클
M.13 밴디
M.14 보디빌딩
M.15 슈팅
M.16 트라이아손
M.17 역도



이 상 곤

e-mail : samuel@jeonju.ac.kr

1996년 전북대학교 컴퓨터과학과(학사)

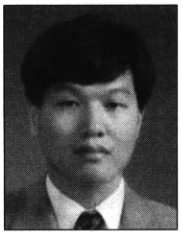
1998년 전북대학교 전산통계학과(이학석사)

2001년 일본 도쿠시마대학교 지능정보공학과(공학박사)

2001년~2002년 원광대학교 음성정보 기술산업 지원센터 연구원

2002년~현재 전주대학교 정보기술컴퓨터공학부 전임강사

관심분야 : 한국어 정보처리, 한글공학, 정보검색, 문서분류



이 완 권

e-mail : wklee@jeonju.ac.kr

1987년 서울대학교 컴퓨터공학과(학사)

1990년 한국과학기술원 전산학과(공학석사)

2000년 한국과학기술원 전산학과(공학박사)

1994년~현재 전주대학교 정보기술컴퓨터공학부 부교수

관심분야 : 소프트웨어 유지보수, 디자인 메트릭, 객체지향 모델링, 정보 검색 등