문서분류용 목적으로 이용할 효율적인 연상정보의 추출방법

최 현[1] 황남선[2] 이상곤[3]
전주대학교 교육대학원 컴퓨터교육전공[1]
전주대학교 정보기술컴퓨터공학부[23]
{kaiana[1], samuel[3]}@jj.ac.kr and hunjuk@nate.com[2]

Extraction of Field-Associated Term for the Purpose of Document Classification

Hyun Choi[1], Namseon Hwang[2], and Samuel Sangkon Lee[3]
Graduate School of Education[1],
School of Information, Technology and Engineering[23],
Jeonju University[123]

요 약

분야연상어는 어휘자체가 분야정보를 가지므로 인간이 분야를 인지할 때와 유사하게 문서의 분야를 판단한다. 인간이 한국어와 일본어의 180분야로 분류한 약 15,000개의 문서뱅크를 수집하고, 수집된 문서에서 복합어로 구성된 분야연상어의 효율적인 추출 알고리즘을 제안한다. 제안된 알고리즘으로 자동구축된 분야연상어를 문서분류의 초기결정에 이용할 수 있다. 분야연상어를 이용하면 어떠한 분야체계에도 손쉽게 적용할 수 있으므로 문서분류용 목적으로 이용할 수 있는 보편성은 충분하다.

1. 서론

인간은 문서전체를 읽지 아니하여도, 문서에서 대표적인 단어를 보는 것만으로 정치나 스포츠 등의 문서분야를 정확히 인지할 수 있다. 따라서, 문서단편 내의 소수의 단어정보를 이용하여 분야를 정확하게 결정하기 위한 분야연상어의 구축은 중요한 연구과제이다.

인간은 자신의 상식지식으로 특정분야를 인지할 수 있는 경우에도 문서에서 처음으로 출현하는 몇 개의 단어들을 이용하여 연상되는 연상정보를 감각적으로 인식하고 문서의 내용을 읽어감에 따라 문서에 해당하는 분야를 연상하거나 추측할 수 있다. 또한 문서의 이전내용에서 애매성이 발생하여도 문서의 뒤에서 출현하는 단어에 의해 이전의 문서내용에서 이해하지 못했던 애매성을 해소해 나갈 수 있다. 이와 같이 문서의 단락 내에 몇 개의 단어정보를 이용하여 문서가 포함되는 분야를 정확하게 결정할 수 있는 단어를 "분야연상어[2, 3, 4, 5]"라 정의하고, 상식적인 분야연상어의 구축, 유사문서(문장) 검색, 문서분류 등의 연구를 수행하고자 한다.

2. 복합 분야연상어의 결정

단일 분야연상어[4]의 분야계승에 기초하여 본 논문에서는 복합분야연상어의 분석을 논의한다. 기존의 연구들을 살펴보면 복합어를 분석할 때 각 구성어의 통사적 구성에 대해서는 많은 논의가 있으나, 구성어의 의미적 계승에 대한 연구는 별로 활발하지 못하다. 일반적으로 복합어를 구성하는 단어 중 오른쪽 단어가 복합어의 문법적 주요어가 되며, 이 주요어가 전체의 품사를 결정한다고 알려져 있다. 따라서 이를 토대로 복합 분야연상어의 구성에 관계하는 요소를 분석하여 정리한다. 단일 및 복합 분야연상어의 계승랭크를 정의하고 안정성랭크[4]와 조합한 복합 분야연상어의 효율적인 결정 알고리즘을 제안한다.

2.1 분야계승에 기반 한 복합 분야연상어의 분석

2.1.1 의미계승에 의한 분야계승

복합명사의 통사적 구성에 대해서는 일반적으로 오른쪽 단어의 품사가 복합어의 문법적 주요어가 되며, 복합어 전체의 품사를 결정한다. 오른쪽 단어의 통사적 주요어가 어휘적 주요부와 일치할 때 이는 분류학(taxonomic class) 명사라 한다. 이 때, 왼쪽의 단어가 오른쪽 단어를 수식하는 경우가 대부분이므로 오른쪽 단어의 의미가 복합어 전체에 계승되어 분야정보의 의미계승에 관계한다. 예를 들면, (그림 1)은 '냄비'에 관한 의미계승을 나타낸다. "냄비"는 "요리도구"를 연상하며, 분야는 <취미-오락/요리-먹는 것>에 관한 분야연상어가 된다. 이 단어와 다른 단어가 결합하여 복합어 "압력+냄비"과 "칭기즈칸+냄비"인 경우 왼쪽의 단어 "압력"과 "칭기즈칸"은 단지 오른쪽의 단어 "냄비"의 의미를 한정하는 수식어이며, "냄비"의 의미를 계승하고 단일어 "냄비"와 동일한 분야 <요리-먹는 것>을 연상한다.

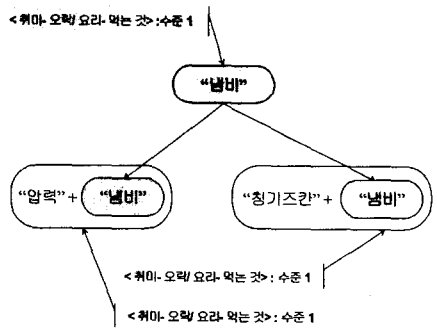

그림 1 "냄비"의 분야정보의 계승 예

2.1.2 은유적 전의에 의한 분야계승

복합 분야연상어의 오른쪽 단어가 은유적 전의(轉意)에 의하여 왼쪽의 단어가 분류학 명사가 되며, 연상되는 분야가 변화하는 경우가 있다. (그림 2)와 같이 단일어 "전쟁"은 본래의 분야는 <국제/지역분쟁>을 연상하고, 복합어 "걸프전쟁", "이라크전쟁", "6.25전쟁", "한국전쟁" 등은 전쟁의 본래의 의미를 계승하지만, 다른 예 "입시전쟁"에서 전쟁은 비유적으로 사용되고 있으며, 단어 자체의 의미인 "전쟁"을 의미하는 것은 아니다. 따라서 이 복합어에서 오른쪽의 "전쟁"은 통사적 주요어이긴 하지만, 어휘적 주요어는 아니다. 이런 경우는 왼쪽의 "입시"가 어휘적 주요부, 다시 말하면 분류학 명사이며 연상하는 분야는 <교육/수험-입시>가 된다. 이 경우 단일 분야연상어 "전쟁"은 두 가지의 분야를 연상할 수 있으며, 다른 단어와 결합하면 전혀 다른 분야를 연상할 수 있다. 이와 같이 단어의 의미적 전의(혹은 은유적 전의)는 일상생활에서 끊임없이 생성된다. 따라서 각 구성요소의 의미정보만으로 분야연상어를 결정하는 것은 바람직하지 못하다.

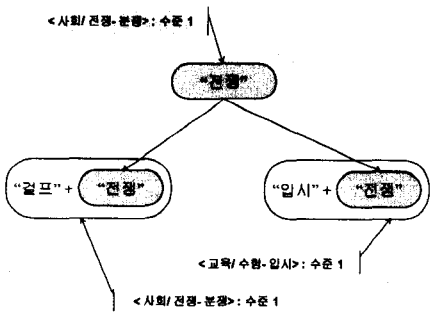

그림 2 "전쟁"의 분야정보의 계승 예

2.1.3 복합어의 왼쪽단어에 의한 분야계승

(그림 3)은 오른쪽의 단어가 은유적 전의가 발생하지 않고 왼쪽의 단어가 어휘적 주요부가 되는 경우의 예이다. 예를 들면, 일본어의 경우, 복합어 "장고냄비"는 "냄비"의 의미계승에 의해 <요리/먹는 것>의 분야를 연상하지만, 인간의 두뇌는 분야 <쓰모>를 연상한다. 어휘적 주요부가 되는 왼쪽의 구성어 "장고"의 분야는 <요리-먹는 것>과 <쓰모>를 모두 계승하고 있다. 이 단어가 "냄비"라는 단어와 복합 분야연상어로 구성되어도 동일한 분야를 동일한 수준으로 계승한다. 따라서 "장고냄비"와 같은 복합어는 쓸모 없이 길게 형성된 분야 중복 연상어로 간주하고, 본 논문에서는 더 이상 논의하지 않는다.

2.1.4 복합어의 왼쪽단어에 의한 분야계승

복합 분야연상어의 각 구성어들(왼쪽 혹은 오른쪽 구별없이)은 각각 독립된 분야를 계승하며, 때로는 유사한 분야를 연상하는 성질을 갖는다. 따라서 다음의 계승랭크를 정의한다.

■ 계승랭크(Inheritance Rank)

복합 분야연상어가 연상하는 분야를 <F>라 하고, 그 복합어의 구성어(x)가 연상하는 분야를 <F′>이라 하자. 두 분야 <F>와 <F′>이 다음의 세 가지 경우 중 하나에 해당하면 '유사분야'라 정의한다.

1) <F>와 <F′>이 일치하는 경우,
2) <F′>이 <F>의 상위분야인 경우, 혹은
3) <F>와 <F′>이 종단분야에서 동일한 부모분야를 갖는 경우,

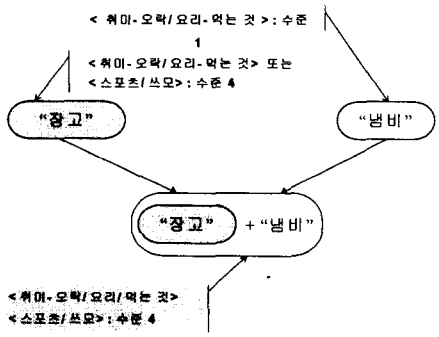

그림 3 "장고냄비"의 분야정보 계승 예

반대로, <F′>과 <F>가 전혀 다른 분야를 연상하면 '다른 분야'라 한다. 복합 분야연상어의 후보 w(w=xy라 하자)와 그 구성어 x가 한정하는 분야가 유사분야를 갖는 경우, 계승랭크가 가장 높으면 랭크열 A라 정의하고, x가 어떠한 유사분야도 갖지 않고 전혀 다른 분야를 갖는 경우의 계승랭크는 랭크 C(가장 낮은 랭크)로 정의한다. 구성어 x가 수준 5의 비연상어인 경우는 중간정도를 나타내는 랭크 B로 정의한다.

예를 들면, <야구>를 연상하는 복합어 w="김응용감독"에 대하여 x="김응용"과 "감독"은 연상분야가 모두 <야구>로서 유사분야를 갖기 때문에, w에 대한 x의 계승랭크 값은 A가 된다. 다른 형태의 예로서 분야연상어 후보 w="성균관대감독"을 생각해 보면 x="성균관대"는 <교육>의 하위분야에 해당하는 분야연상어이므로 계승랭크는 C가 된다.

2.2 우선순위의 결정

복합 분야연상어가 지시하는 정확한 분야를 결정하기 위해 계승랭크와 안정성랭크[4]를 이용한다. 이 두 가지 랭크열은 복합 분야연상어에 대한 각 구성요소들의 분야정보를 결정하는데 '우선순위'를 갖는다.

(1) 랭크열

복합 분야연상어 후보 w의 구성어가 x와 y라 하면 계승랭크와 안정성랭크를 연결시켜 '랭크열'을 생성한다. 예를 들어, <표 1>에서 표시한 바와 같이 <야구>에 대한 복합 분야연상어(김응용감독)에서 x="김응용"의 계승랭크는 'A'이고, 안정성랭크는 'c'이다. 다른 구성어 y="감독"의 계승랭크는 A이고, 안정성랭크는 a이다. 따라서 김응용감독(w)의 랭크열은 "김응용"의 계승과 안정성랭크를 결합하여 'Ac', "감독"은 'Aa'이다. 따라서 이를 조합하여 랭크열은 'AcAa'가 된다.

(2) 25단계의 판정기준

랭크열에서 분야연상어 후보(w)에 대한 분야연상어의 판정기준을 마련한다. 먼저 계승랭크(영문 대문자 사용)의

조합에 의하여 AA에서 CC까지 다음과 같이 다섯 단계 ① AA, ② AB(혹은 BA), ③ AC(혹은 BB 또는 CA), ④ BC(혹은 CB), ⑤ CC 등으로 우선순위를 정의한다. 단, AC는 A의 우성과 C의 열성이 서로 맞아 상쇄되기 때문에 우열이 없는 BB와 동일한 단계로 취급한다. 각 계승랭크의 다섯 단계에 대하여 다시 안정성랭크(영문 소문자 사용)를 aa에서 cc까지 다섯 단계(aa, ab(혹은 ba), ac(혹은 bb, 또는 ca), bc(또는 cb), cc 등)로 세분화하여 모두 25단계(5가지의 계승랭크 × 5가지의 안정성랭크)의 '판정표'를 결정한다.

(3) 기준빈도

분야연상어의 수준 1[4]에 대하여 분야 <F>를 연상하는 분야연상어 후보 w의 집합을 W_SET(L, <F>)라 정의하고, 그 집합 중에서 후보어의 최대빈도, 평균빈도, 최소빈도를 구한다. 최소빈도에서 평균빈도까지, 평균빈도에서 최대빈도까지를 각각 12개로 등분하여 평균빈도를 더한 25단계의 기준빈도를 대응시킨 판정표 Decision(L, <F>)를 정의한다. 이 판정표는 우선순위가 높은 후보일수록 제거되는 빈도를 낮게 하여 추출에서 제외될 가능성을 방지하도록 하는데 이용한다. 우선순위가 낮은 후보는 제거되는 빈도를 높게 설정하여 과잉추출 되는 것을 방지한다.

2.3 복합 분야연상어의 결정 알고리즘

참고문헌 [4]에서 논의한 단일어의 분야연상어 추출 알고리즘과 결합하여 알고리즘에 의해 복합 분야연상어의 후보가 결정되고, 아래의 알고리즘에 의해 최종적으로 복합 분야연상어로 선택된다. 무조건 계승랭크가 높은 후보를 선택하면 역으로 분야 중복 연상어를 선택하는 다음의 두 가지 모순이 발생한다. 첫째, 복합 분야연상어 후보(w)의 연상분야 <F>와 구성어 x의 연상분야 <F'>이 다른 분야인 경우. 둘째, 복합어 w의 모든 연상분야 <F>와 모든 구성어 x의 연상분야 <F'>이 유사한 분야이고, w의 수준이 x의 수준보다 높은 경우 등이다.

따라서 다음의 복합 분야연상어 결정 알고리즘을 이용하여 위의 두 가지 모순 점을 해결한다. w가 분야연상어 후보이면, 각 구성어에 의해 연상분야를 추축하지만, 분야 중복 연상어가 될 가능성이 있다. 다음에 언급하는 복합 분야연상어 결정 알고리즘의 (순서 B1)과 같이 계승랭크에서 발생하는 복합 분야 중복 연상어 후보를 먼저 제거한다. (순서 B2)에서는 계승랭크를 이용하여 판정표에 의한 복합 분야연상어를 우선적으로 결정한다. 분야연상어 w의 집합 W_SET(L,<F>)의 역 표현인 F_SET(w)는 분야연상어 w의 (L,<F>)를 요소로 하는 집합이다.

● **복합 분야연상어의 결정 알고리즘**

- **입력** : 복합 분야연상어 후보 w와 W_SET(L,<F>)
- **출력** : 복합 분야연상어 w의 연상분야와 수준

(순서 B1) [분야 중복 연상어의 제거]

분야연상어 w=xy에 대하여 다음을 실행한다.

(순서 B1-1) w의 수준 L가 1인 경우

F_SET(w)와 F_SET(x)가 같은 요소 (L,<F>)를 갖고 x의 안정성랭크가 a일 때, 혹은 F_SET(y)가 <F>와 다른 분야 <F'>이 되는 요소(1,<F'>)를 갖지 않으면, w를 W_SET(L,<F>)와 F_SET(w)에서 제거한다(y의 경우도 동일). 이것은 F_SET(w)에서 요소(L,<F>)가 제거되는 것을 의미한다. 수준 1의 분야연상어는 **유일한 종단분야를** 연상하는 중요한 분야연상어이므로 안정성랭크 a인 조건으로 한정하지만, 다음의 순서 (B1-2)에서는 분야연상어의 수준이 2~4이므로 추건 a를 붙이지 않는다.

(순서 B1-2) w의 수준 L가 2~4인 경우

F_SET(w)의 모든 분야 <F>가 F_SET(x)와 F_SET(y) 모든 분야 <F'>과 유사한 분야이고, w의 수준이 w와 y의 수준을 넘지 않으면 w를 W_SET(L,<F>)에서 제거한다.

(순서 B2) [판정표에 의한 압축]

이상의 처리로 얻어진 W_SET(L,<F>)에 대한 판정표 Decision(L,<F>)를 결정하고, w의 계승랭크열과 안정성랭크열에 대응되는 기준빈도보다 w의 정규화 빈도 Normalization(w,<F>)[4]보다 적으면, w를 W_SET(L, <F>)에서 제거한다.

(순서 B3) [수준의 최종결정]

이상과 같이 연상분야가 제거된 후보어 w는 수준 5로 변경하고, 연상하는 분야수가 감소한 후보어 w는 수준을 변경한다. (알고리즘 종료)

3. 결 론

본 논문에서는 단일어에 대한 분야연상어 정보를 이용하여 일상생활에서 끊임없이 생성되는 복합 분야연상어를 효율적으로 결정하는 방법을 제안한다. 구축된 분야연상어가 문서 내의 단락의 분야결정에 유효한 것인가에 대해 논의하였다.

복합어로 구성된 분야연상어를 형태소사전에 등록된 표제어와 일치하도록 한정하였다. 이것은 단일 분야연상어의 분야정보를 형태소사전에 그대로 등록하기 위한 실용성을 고려한 것이다. 본 연구에서는 분야체계를 미리 정의한다고 하였으나, 분야연상어 구축은 어떠한 분야체계에도 손쉽게 적용될 수 있으므로 보편성은 충분하다고 생각된다.

단일 분야연상어를 사람이 판단할 수 있기 때문에 각 기관의 이용 목적에 맞도록 분야체계를 결정하여야 한다. 본 논문은 명사 연속의 복합어를 대상으로 하였으나 용언이나 조사를 포함한 명사구, 명사와 용언의 조합 등과 같이 출현하는 공기정보와 분야연상어의 관계도 검사할 필요가 있다.

감사의 글

이 논문은 2003년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2003-003-D00415). 재단의 연구비 지원에 감사 드립니다.

참고 문헌

[1] 남영신, 우리말 분류 사전, 성안당, 2001.

[2] 이상곤, "분야연상어를 이용한 화제의 계속성과 전환성을 추적하는 단락분할 방법", 정보처리학회 논문지B, 제 10권, 제 1호, pp. 57-66, 2003.

[3] 홍성옥, 이상곤, "연상정보를 이용한 단락분할 방법", 2003년도 한국정보처리학회 춘계 학술발표 논문집(상), 제 10권, 제 1호, pp. 497-500, 2003.

[4] 이상곤, 이완권, "분야연상어의 수집과 추출 알고리즘", 정보처리학회 논문지B, 제 10권, 제 3호, pp. 347-358, 2003.

[5] 김숙영, 최창원, 이상곤, "한글문서분류용 분야연상어의 추출 알고리즘", 한국정보과학회 2003 가을 학술발표 논문집(Ⅰ), 제 30권, 제 2호, pp. 544-546, 2003.