기계가독형사전에서 상위어 판별을 위한 규칙 학습

최 선 화[†] · 박 혁 로[††]

요 약

기계가독형사전(Machine Readable Dictionary)에서 단어의 정의문에 나타나는 항목 단어의 상위개념을 추출하는 대부분의 연구들은 전문가에 의해 작성된 어휘패턴을 사용하였다. 이 방법은 사람이 직접 패턴을 수집하므로 시간과 비용이 많이 소모될 뿐만 아니라, 자연언어에는 같은 의미를 가진 다양한 표현들이 존재하므로 넓은 커버리지를 갖는 어휘패턴들을 수집하는 것이 매우 어렵다는 단점이 있다. 이런 문제점들을 해결하기 위하여, 본 논문에서는 구문적 특징만을 이용한 상위어 판별 규칙을 기계학습함으로써 기존에 사용되었던 어휘패턴의 지나친 어휘 의존성으로 인한 낮은 커버리지 및 패턴 수집의 문제를 해결하는 방법을 제안한다. 제안한 방법으로 기계학습된 규칙들을 상위어 자동추출과정에 적용한 결과 정확도 92.37% 성능을 보였다. 이는 기존 연구들보다 향상된 성능으로 기계학습에 의해 수집된 판별규칙이 상위어 판별에 있어서 어휘패턴의 문제를 해결할 수 있다는 것을 입증하였다.

키워드 : 상위어 추출, 의미분류체계, 의미계층구조, 시소러스 구축, 기계학습

Learning Rules for Identifying Hypernyms in Machine Readable Dictionaries

Choi, SeonHwa[†] · Park, HyukRo[††]

ABSTRACT

Most approaches for extracting hypernyms of a noun from its definitions in an MRD rely on lexical patterns compiled by human experts. Not only these approaches require high cost for compiling lexical patterns but also it is very difficult for human experts to compile a set of lexical patterns with a broad-coverage because in natural languages there are various expressions which represent same concept. To alleviate these problems, this paper proposes a new method for extracting hypernyms of a noun from its definitions in an MRD. In proposed approach, we use only syntactic (part-of-speech) patterns instead of lexical patterns in identifying hypernyms to reduce the number of patterns with keeping their coverage broad. Our experiment has shown that the classification accuracy of the proposed method is 92.37% which is significantly much better than that of previous approaches.

Key Words : Hypernym Extraction, Semantic Taxonomy, Semantic Hierarchy, Thesaurus, Machine Learning

1. 서 론

최근 자연언어처리 시스템의 처리 능력이 높아지고 다양한 응용분야에 적용됨에 따라 광범위한 어휘지식베이스의 중요성이 과거보다 더 강조되고 있다. 어휘지식베이스는 어휘사전, 시소러스, 온톨로지, 그리고 기계가독형사전(Machine Readable Dictionary) 등을 일컫는 말로 용어들의 목록과 용어들의 정의 그리고 용어들 간의 관계(동의어, 반의어, 상위어 등) 정보를 포함한다. 많은 연구자들은 용어들간의 다양한 관계 중에서도 의미계층관계가 기계번역, 정보검색, 단어의 의미모호성 처리 등과 같은 응용분야에서 다양한 추론과정에 사용될 수 있으므로 특히 유용한 관계로 여겨왔다[21, 4, 10].

의미계층관계를 구축할 경우 가장 확실하고 정확한 방법은 수동으로 구축하는 것이다. 그러나 이것은 많은 비용과 시간이 필요하다는 단점을 가지고 있다. 이러한 이유로 이미 구축된 많은 어휘 자원들을 이용하여 대량의 어휘지식베이스를 자동 혹은 반자동으로 얻어 내려는 많은 연구들이 있었다[14, 15, 17].

사전의 단어 정의문은 특별한 구조를 가지고 있어서 개념체계의 상-하위관계를 추출하는데 주로 이용되는 어휘 자원이다. 사전에서 단어의 정의문에 나타나는 항목 단어의 상위개념을 추출하는 대부분의 연구들은 사람에 의해 작성된 어휘패턴을 사용하였다[3]. 이 방법은 사람이 패턴을 수집하는 데 드는 비용이 크다는 점과 같은 의미를 지닌 다양한 표현이 존재한다는 자연언어의 특성 상 광범위한 어휘패턴을 수집하고 작성하기가 매우 어렵다는 단점이 있다. 따라

† 준 회 원 : 전남대학교 전산학과 Post Doc.
†† 종신회원 : 전남대학교 전자컴퓨터공학부 부교수
논문접수 : 2006년 1월 26일, 심사완료 : 2006년 3월 3일

서 사람에 의해 작성된 어휘패턴은 적용범위가 매우 한정적이며, 적용 도메인이 바뀌면 다시 작성해야 한다는 문제점이 있다.

본 논문에서는 수동 구축된 어휘패턴의 문제를 극복하기 위하여 상위어 판별에 적합한 구문적 특징들을 정의하고, 이 특징들을 이용하여 판별규칙을 기계학습을 통해 자동으로 생성하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구들을 소개하고, 3장에서는 어휘패턴을 이용한 상위어 자동 추출 방법과 그 방법의 문제점을 제시한다. 4장에서는 상위어 판별을 위한 구문적 특징들을 정의하고 이를 기계학습하는 과정을 설명하며, 실험 결과는 5장에서 보이고, 마지막으로 결론을 6장에서 맺는다.

2. 관련 연구

대량의 어휘 자원인 코퍼스에서 정보를 자동 추출하는 방법은 이미 널리 이용되는 방법으로 다양한 응용분야에서 활용되고 있다[10]. 이러한 연구 중 하나로, 코퍼스를 이용하여 개념체계를 자동으로 확장하려는 연구들은 단어 개념을 설명하는 용어들의 공기(co-occurrence)분포를 이용하는 연구[9, 5], 자유로운 텍스트에서 개념들 간의 의미계층관계를 얻기 위해서 어휘 또는 어휘의미패턴을 이용하는 연구[8, 1], 그리고 사전에 있는 단어 정의문의 특별한 구조를 이용하여 개념체계를 확장하려는 연구[17, 15]로 나뉘어 진다. 본 논문에서는 마지막의 경우인 사전 정의문을 이용하여 의미계층관계를 추출하는 문제를 다룬다.

사전에 있는 단어 정의문은 단어의 개념을 정확하게 표현하면서 일반 텍스트와 달리 특별한 구조를 가지고 있다. 이런 점을 이용해 의미개념체계의 상위개념을 추출하는 연구가 활발히 진행되고 있다[14, 15, 17]. 사전의 정의문은 주로 단어의 개념을 간결하게 기술하고, 단어에 대한 핵심적인 정보를 포함하고 있기 때문에 많은 연구들에서 주로 사용하는 유용한 어휘 자원이라고 할 수 있다[7].

Rigau외[16]은 기계가독형사전으로부터 대량의 시소러스를 정확하게 구축하기 위한 비교사 알고리즘(unsupervised algorithm)을 제안했다. 먼저 기계가독형사전의 항목들에서 의미계층 간의 링크를 자동으로 추출하고, 추출된 링크들에 대해 선택적으로 수작업을 허용하여 링크의 순위를 매겼다. 이 연구에서 저자는 상-하위어의 관계는 사전의 항목 용어와 중심어 사이에 나타난다는 사실을 소개하면서, 일반적으로 사전에서 용어를 정의하는 문장은 용어의 핵심을 나타내는 중심어와 다른 용어와 구별할 수 있는 의미론적인 특성을 가미한 구별어들의 조합으로 구성되어있다고 설명했다. 그는 몇 가지 상-하위어 관계가 나타나는 어휘패턴 예제들을 사용하여 간단한 휴리스틱 규칙을 정의하고 이를 이용하여 정의문에서 중심어를 찾아냈다.

Martin외[11]은 일반적인 사전 정의문에서 의미정보를 추출하는 방법을 제안했다. 그의 의미정보 추출방법 역시 정의문에서 중심어를 찾아내는 것으로 동사 그리고 명사의 정의문들에서 간단한 휴리스틱들을 이용해 중심어를 추출한다. 동사의 경우 정의문에서 'to' 다음에 나오는 원형동사가 해당 동사의 중심어가 된다는 휴리스틱 규칙을 적용하였고, 명사의 경우는 일반적으로 중심어의 왼쪽과 오른쪽에 주로 나타나는 어휘들을 사람이 미리 정의해 두고 이를 경계어휘로 사용했다. 정의문에서 경계어휘를 찾고 이것을 제거한 문자열이 바로 명사의 중심어가 된다고 간주했다.

Maria외[10]는 온라인 사전으로부터 개념들 간의 상위개념, 하위개념, 부분개념, 소속개념 등의 의미적인 관계를 표현하는 어휘 패턴을 학습하여 이미 존재하는 온톨로지 혹은 의미 네트워크에 새로운 관계를 확장하는 데 적용하였다. 어휘 패턴을 자동학습할 수 있는 알고리즘을 제시하였지만, 사전의 정의 문장이 간결하고 짧은 문장이어야 한다는 한계가 있다.

문유진[21]는 Rigau외[16]의 방법을 한국어에 적용한 대표적인 연구이다. 사전 정의문의 특별한 구조를 고려하여 사람의 직관에 의해 만들어진 여섯 가지의 어휘패턴 규칙을 소개했다. 상위어가 발생되는 정의문은 중심어와 구별어, 그리고 기능어로 조합된 특별한 구성형태가 있다고 주장하고, 상위어를 찾는 것은 중심어를 찾는 것이라고 규정했다. 어휘패턴 규칙들 중 정의문을 분석하여 적절한 규칙을 선택하고 선택한 규칙에 의해 중심어를 추출했다. 김민수외[18]은 문유진[21]에서 제시한 여섯 가지의 어휘패턴을 추가하고 보완하여 10개의 어휘패턴을 제시했다. 조평옥외[25]은 명사의 사전 정의문을 이용하여 상향식으로 한국어 명사 시소러스를 구축하였다. 최유미외[27]은 문헌정보학분야 용어사전에서 상위어를 자동 추출하는 알고리즘을 개발하였다. 상위어 추출을 위한 알고리즘 개발은 무작위 표본추출을 통하여 문헌정보학 용어사전에 기술된 문장의 구문적 특성을 분석한 후, 이 구문정보를 이용하여 수행하였다. 위의 모든 연구는 Rigau외[16]의 방법을 적용하여 한국어 명사의 시소러스를 구축하려는 연구들에 해당한다.

3. 어휘패턴을 이용한 상위어 자동 추출

이 장에서는 한국어 사전에서 어휘패턴을 이용하여 명사의 상위어를 자동 추출하는 방법에 대하여 간략히 설명하고 이 방법의 문제를 제시한다.

3.1 사전의 단어 정의문의 구조

사전에서 단어에 대한 정의는 보통 중심어, 구별어, 그리고 기능어로 구성되어 있다. 중심어(head word)란 상위어 의미를 갖는 단어를 말하고, 구별어(differentia)란 상위어에 의미론적 특성(semantic feature)을 가미한 단어들을 말한다. 기능어(functional word)란 상위어의 기능을 부연하여 설명해 주는 단어를 말한다. 예를 들어 '박사'의 정의가 국어사전에 "모든 일에 정통하거나 숙달된 사람을 비유하여 이르는 말"이라고 쓰여 있다면 중심어는 '사람'이고 구별어는 "모든

일에 정통하거나 숙달된"이며 기능어는 "을 비유하여 이르는 말" 이 된다.

3.2 상위어 자동 추출 방법

사전 정의문을 문자열과 문자위치에 의하여 '구별어', '중심어', '기능어'로 분석되는지를 검사한 후 분석된 문장의 유형에 따라 다음의 어휘패턴으로 상위어를 선택한다.

3.2.1 문장유형 : 구별어 + 중심어

명사 항목의 정의가 구별어와 중심어로 되어 있는 경우는 중심어를 그 명사의 상위어로 선택한다. 만약 중심어가 명사+"함"으로 끝난 경우에는 동사의 명사형을 제거한 중심어가 상위어로 선택된다.

3.2.2 문장유형 : 구별어 + 중심어 + 기능어

명사 항목의 정의가 구별어, 중심어, 그리고 기능어로 구성되어 있으면 중심어를 탐색하여 중심어를 그 명사의 상위어로 선택한다.

(그림 1)은 문유진[21]의 연구에서 중심어를 탐색할 때 사용하는 기능어들로 명사 정의에 사용한 것 중에서 찾아놓은 것이다[6, 20, 23]. 이 기능어 목록은 국어사전이 달라질 경우 기능어가 달라질 수 있으므로 그 국어사전에서 사용된 기능어들은 수시로 조사되어야 한다.

만약 명사 항목의 정의가 구별어, 중심어, 그리고 기능어로 구성되어 있고 구별어, 중심어에 따옴표 표시가 되어 있는 경우에는 구별어, 중심어를 3.2.1장에서 제시한 방법에 의하여 분석하여 상위어를 선택한다.

```
char rphrase[][64] =
  { "~을 부르던 이름", "~을 이름","~의 한 종류", "~의 덩이",
    "~의 총칭", "~의 통칭", "~의 하나", "~ 따위의 총칭", "~따위",
    "~부수의 하나", "~의 옛말", "~등", "~로 된 물질",
    "~을 비유하는 말", "~을 아울러 이르는 말", "~을 이르는 말",
    "~의 뜻을 나타내는 말", "~이라는 뜻을 나타내는 말",
    "~의 한 가지", "~을 비유하여 이르는 말", "~을 낮추어 하던 말",
    "~의 한 부분", "~의 높임말","~을 나타내는 말",
    "~으로 나타낸 것", "~로 나타냄", "~의 이름", "~의 원말",
    "~의 준말", "~의 힘줌말", "~의 변한 말" };
```

(그림 1) 기능어의 예

3.2.3 기타 유형

명사 항목의 정의에 접속사("~와", "~과", "~나", "~이나")가 포함된 경우에는 접속사 앞뒤 중심어가 상위어가 된다. 이 경우에는 문장의 파싱은 필수적이다. 또한, 항목의 정의가 "~일." , "~말." , "~것." 등으로 끝날 때는 명사의 정의를 탐색할 필요 없이 상위어는 수작업으로 정해준다.

3.3 어휘패턴을 이용한 상위어 자동 추출 방법의 문제점

어휘패턴을 이용한 상위어 자동 추출방법에서는 어떤 명사 정의문의 유형을 결정하기 위하여 어휘 분석과정이 필수적으로 이루어져야 한다[12]. 그렇지만, 어휘패턴을 이용한 한국어 상위어 추출 연구에서는 그 과정을 생략하고 어휘와 어휘가 문장에 나타난 위치만을 이용하여 문장유형을 결정하였다. 사전 정의문에서 단어의 상위어에 의미론적 특성을 가미하는 단어들인 구별어와 기능어를 판단하는 일은 정확한 상위어를 판별하는데 중요한 문제가 된다. 하지만, 구별어와 기능어를 판단하는데 있어서 단순히 어휘 비교와 어휘의 문장 내 위치만 가지고 판단하므로 정확한 판단이 어렵게 된다.

〈표 1〉 유사한 기능어 그룹

번호	종류	유사 어휘
1	명칭을 나타내는 어휘	"~을 부르던 이름" , "~을 이름" , "~의 총칭" , "~의 통칭" , "~ 따위의 총칭" , "~의 옛말" , "~의 높임말" , "~을 나타내는 말" , "~으로 나타낸 것" , "~로 나타냄" , "~의 이름" , "~의 원말" , "~의 준말" , "~의 힘줌말" , "~의 변한말"
2	부분 및 종류를 나타내는 어휘	"~의 한 종류" , "~의 덩이" , "~의 하나" , "~따위" , "~부수의 하나" , "~의 한가지" , "~의 한부분"
3	지칭하는 어휘	"~을 비유하는 말" , "~을 아울러 이르는 말" , "~의 뜻을 나타내는 말" , "~이라는 뜻을 나타내는 말" , "~을 비유하여 이르는 말" , "~을 낮추어 하던 말" ,
4	기타	"~등" , "~로 된 물질"

특히 정의문의 구성요소 중에서 상위어를 판별하는데 가장 중요한 요소는 기능어다. 하지만, 미리 국어사전에서 사용된 기능어들을 조사하여 목록으로 만들어 놓고, 정의문의 어휘들과 기능어 목록을 비교하여 기능어와 중심어를 판단하는 것은 국어사전이 달라질 경우 기능어가 달라질 수 있으므로, 국어사전에서 사용된 기능어들은 수시로 조사되어야 한다는 문제가 있다. 또한 동일한 의미를 가진 다양한 표현이 존재하는 자연언어의 특성을 감안할 때 방대한 어휘를 미리 조사하여 목록을 만들어 놓는다는 것도 한계가 있다. <표 1>은 (그림 1)에서 제시한 기능어들을 유사한 의미를 갖는 어휘들끼리 묶어서 보여준 예이다. 이 표에서도 알 수 있듯이, 명칭을 나타내는 어휘에는 이 표에서 제시된 어휘들 외에도 수 많은 어휘들이 존재하지만 일부 한정된 어휘들만 목록화되어 있다.

따라서, 상위어를 판별하는데 있어서 단순히 상위어 주변에 나타나는 어휘만을 비교하는 것은 정확성이 떨어질 수 있다. 이를 해결하기 위해서, 이 논문에서는 대량의 단어 정의문에 대해 어휘분석과정을 거치고, 상위어가 발생하는 구문적 특징을 추출한 후, 판별규칙을 기계학습하는 방법을 제시한다.

4. 구문적 특징을 이용한 상위어 판별규칙 학습

4.1 상위어가 발생하는 구문적 특징

상위어 판별 규칙을 학습하기 위해서, 한국어 문장의 특성과 상위어가 정의문에서 나타나는 구문적 특성을 정의하고 판별력 있는 특징들을 선택하는 문제에 관하여 설명한다.

4.1.1 상위어가 발생하는 위치

한국어의 경우 문장에서 중심이 되는 단어는 수식하는 단어들보다 뒤에 나타나는 특징을 가지고 있다. 그러므로, 중심어는 주로 문장의 끝에 위치하는 경향이 있다. 특히, 한국어 사전에서 단어를 정의하는 정의문에는 이러한 현상이 두드러지게 나타나고 있다. <표 2>는 학습 예제 정의문에서 상위어가 발생하는 위치를 조사한 결과다.

<표 2>에서도 알 수 있듯이 명사가 어떤 위치에 있느냐 하는 것은 그 단어가 상위어인지 아닌지를 결정하는데 중요한 정보를 주는 판별력있는 특징이라고 할 수 있다.

〈표 2〉 명사 정의문에서 상위어 발생위치

상위어 위치	비율
시작 부분	11%
중간 부분	7%
끝 부분	81%

4.1.2 상위어의 기능어 품사

한국어는 영어와 달리 한 어절이 한 개의 내용어와 하나 이상의 기능어로 구성되는 교착어에 해당한다. 내용어는 어절의 주된 의미를 내포하는 성분인 반면에 기능어는 문장에서 어절들 간의 문법적인 관계를 설명하는 것으로 어절의 문법적인 역할을 내포하고 있다.

사전의 정의문에서 상위어의 기능어 품사는 다양하지 않고 소수의 몇 가지로 한정되어 있다. 예를 들면, 명사화 접미사, 목적격 조사 등은 빈번하게 상위어에 부착되는 반면 여격 조사나 위치부사격 조사 등은 상위어에 부착되는 경우는 거의 발생하지 않는다. 따라서, 상위어의 기능어는 상위어를 판별하기 위한 학습시스템에 적절한 특징이 된다고 할 수 있다.

4.1.3 상위어의 문맥

• 상위어 문맥의 정보

기존 연구에서 상위어를 판별하는데 사용되었던 어휘패턴이 상위어를 판별하는데 영향을 주는 문맥정보를 좀 더 구체적이고 자세하게 표현할 수 는 있겠지만, 광범위한 어휘패턴을 구축하기 위해서는 거대한 학습데이터가 필요할 뿐

〈표 3〉 상위어 문맥의 구문패턴

유형	구 문 패 턴
A	~ [HT+관형격 조사] [명사]
B	~ [HT+목적격 조사] ~ [관형어구] [명사]
C	~ [HT+접속격 조사] [HT]
D	~ [HT+동사화/형용사화 접사+명사전성어미]
E	~ [관형어구] [HT]
F	~ [HT+동사화 접사+관형격 어미] [불완전명사]
G	~ [관형격 조사] [HT]
H	~ [관형어] [HT]

〈표 4〉 구문 패턴의 분포

구문 패턴 / HT의 발생위치	A	B	C	D	E	F	G	H	기타
시작	2,407	4,213				456	5		260
중간	602	1,053				196			64
끝			2,956	1,505	59,597		1,500	425	
합	3,009	5,266	2,956	1,505	59,597	652	1,505	425	324
비율	4%	7%	4%	2%	79%	1%	2%	0.6%	0.4%

만 아니라, 패턴의 양도 너무 방대하여 정확성이 떨어질 수 있는 문제를 안고 있다. 따라서 어휘패턴보다는 일반적인 개념으로 표현되는 구문패턴을 문맥의 정보로 규정하고 이를 학습에 이용한다.

정의문에서 상위어(HT: Hypernym Term)가 발생되는 주변의 문맥을 조사하여 구문 정보만을 추출한 결과 <표 3>과 같은 8가지 구문 패턴들을 발견할 수 있다.

<표 4>는 <표 3>에서 제시한 구문패턴들의 분포를 학습데이터를 대상으로 조사한 결과이다.

<표 4>는 상위어를 판별함에 있어서, 굳이 어휘 정보를 사용할 필요 없이 품사 정보만으로도 충분한 성능을 낼 수 있음을 보여주고 있다. 따라서 이 논문에서는 상위어를 판별하는 규칙을 기계학습하기 위해 상위어의 문맥의 정보로서 품사 정보만을 이용하였다.

• 상위어 문맥의 범위

상위어 판별에 영향을 주는 문맥의 범위를 결정하기 위해서, 위에서 제시한 구문패턴별로 상위어와의 거리를 계산하였고, 그 것들의 평균을 구하여 문맥의 범위로 결정하였다. 거리는 상위어와 구문패턴 사이의 어절 수로 계산하였다. 조사 결과, 상위어와 판별력 있는 구문적 요소 간의 평균거리는 1.32이다. 따라서, 상위어를 판별하는 규칙을 학습하기 위해서는 최소 두 어절은 살펴야 한다는 결론을 내릴 수 있다. 본 논문에서는 상위어의 문맥의 범위를 앞뒤 두 어절로 정하여 학습을 위한 특징을 수집하였다.

이상을 요약하면, 본 논문에서는 사전 정의문에 나타나는 어떤 명사가 항목 명사의 상위어인지 여부를 판단하기 위하여 해당 명사의 문장 내 위치, 부착된 조사 및 어미, 그리고 그 명사의 문맥정보를 자질로 사용하였다.

4.2 상위어 추출규칙 기계학습

4.2.1 상위어 판별문제 정의

결정트리(Decision tree)는 가장 널리 사용되는 학습방법 중 하나이며, 실용시스템에서 귀납적 학습을 위해 사용되는 실질적 방법에는 ID3, ASSISTANT, 그리고 C4.5[13]와 같은 알고리즘이 있다. Quinlan의 C4.5는 정보 이득에 근거하여 루트 노드에서 단말 노드 순으로 트리를 구성해 나간다. 각 노드에서는 이진 비교가 수행되며 단말 노드에 이르면 분류 결정이 완료된다. 이런 결정 트리는 그 결과의 가독성 및 해석력이 우수하기 때문에 널리 사용되고 있다.

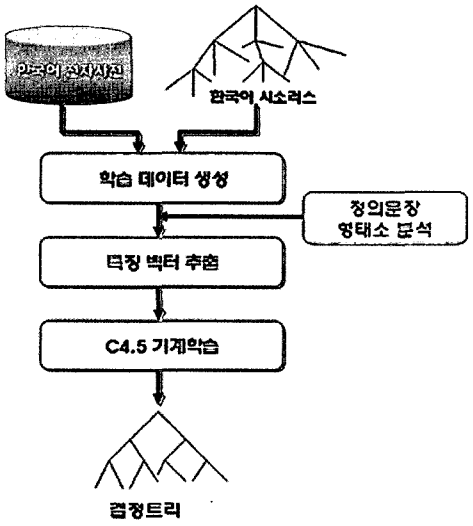
이 논문에서 풀어야 할 문제 역시 단어의 정의문에 나타나는 명사들이 상위어인지 아닌지를 판단하는 것으로 두 개의 카테고리를 갖는 분류문제로 모델링할 수 있다. 따라서 본 논문에서는 결정트리 학습 알고리즘인 C4.5를 학습과 실험에 이용하였다.

학습할 문제를 정의하면 다음과 같다.

- 작업 T: 임의의 명사가 항목 단어의 상위어인지 아닌지 결정
- 성능측정 P: 명사가 올바르게 분류된 비율
- 학습데이터 E: 사전 정의문에 나타난 명사들의 집합으로, 특징벡터와 목표 값(target value)을 지닌 집합

4.2.2 상위어 판별규칙 학습시스템 구성도

C4.5 알고리즘을 이용하여 사전의 정의문에서 임의의 명사가 상위어인지를 판별하기 위한 구문 패턴을 학습하는 기계학습과정은 (그림 2)와 같다.



(그림 2) 상위어 판별규칙 학습 시스템 구성도

학습 데이터를 수집하기 위해서 전문용어언어공학연구센터(KORTERM)에서 제공하는 한국어전자사전[24]과 한국전자통신연구소(ETRI)의 한국어 시소러스를 사용하였다. 사전에는 약 22만개 명사와 정의문을 포함하고 있는 반면에 시소러스는 약 12만 명사와 명사들간의 의미계층관계를 포함하고 있다. 사전의 약 46%의 명사가 시소러스에 존재하지 않는다는 사실은 사전을 이용하여 시소러스를 확장하는 연구의 필요성을 설명해 주고 있다.

시소러스와 전자사전을 이용하여 정의문에 상위어가 나타난 정의문만을 선택하고, 선택된 정의문에서 약 107,000개 명사를 추출하였다. 이 명사들의 70%를 학습 데이터로, 그리고 나머지 30%를 평가 데이터로 사용하였다. 명사의 상위어 판별을 위한 학습을 수행하기 앞서 다음 예제와 같이 세 개의 필드 "하위어", "하위어의 정의문", "상위어"로 구성된 학습 데이터를 구축하였다.

가경	[아름다운 경치]	경치
하위어	하위어 정의문	상위어

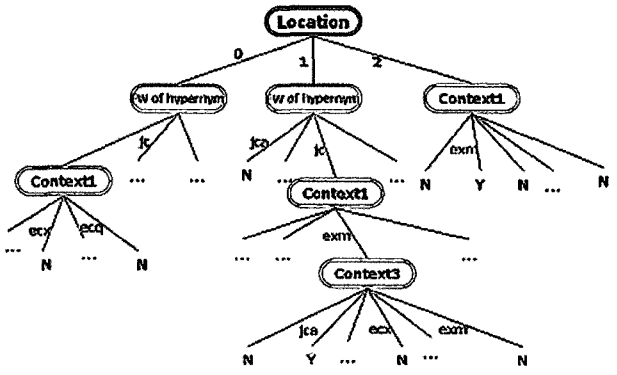
〈표 5〉 특징 벡터의 일부

Noun	Location	FW of a hypernym	context1	context2	context3	context4	IsHypernym
N1	1	jc	ecx	exm	nq	*	Y
N2	2	*	exm	ecx	jc	nq	Y
N3	2	*	exm	jc	nca	exm	Y
N4	1	exm	jc	jc	ecx	m	N
N5	1	jc	jc	ecx	m	jca	N
N6	1	jc	ecx	m	jca	exm	Y
N7	2	*	exm	exm	jca	exm	Y
N8	1	*	nc	jca	exm	jc	N
N9	1	jca	nc	nc	nc	jc	Y
N10	2	exn	a	nca	jc	nca	Y
..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..

단어의 정의문을 형태소 분석한 후 정의문에 나타난 명사들에 대해 4.1장에서 언급한 특징들을 추출하여 특징벡터로 변환한다. 단, 특징벡터는 목표 값을 포함해야 하며 목표 값은 해당 명사가 항목 명사의 상위어인지 아닌지를 나타내는 값이다.

<표 5>는 학습 데이터에서 추출된 특징 벡터의 일부를 보인 것이다. 여기서 속성 *IsHypernym*은 주어진 명사의 목표 값을 나타낸 것으로 Y 또는 N값을 가질 수 있다. 그러므로, 이 논문에서 수행하는 학습의 목적은 학습예제들에 나타나지 않았던 임의의 명사에 대해 목표 값을 정확하게 예측해주는 분류기를 만드는 것이다.

<표 5>에 *Location*은 명사가 정의문에 나타난 위치를 의미하는 것으로 0은 명사가 문장의 시작에 위치한 것을, 1은 문장의 중간에 위치한 것을, 2는 문장의 마지막에 위치한 것을 각각 의미한다. "*FW of a hypernym*"은 명사에 붙은 조사 즉, 기능어 품사를 의미하며, *Context1*에서 *Context4*는 명사의 왼쪽과 오른쪽에 나타난 어절의 기능어 품사를 나타낸다. '*'는 특징 값이 존재하지 않는 경우, 즉 해당 명사의 문맥정보가 없는 경우이다. 품사의 의미는 부록 A에서 자세히 설명하고 있다.



(그림 3) task T를 학습한 결과 트리

위와 같은 학습 데이터를 C4.5 알고리즘으로 학습한 결과를 트리 형태로 표현하면 (그림 3)과 같다. 트리에서 알 수 있듯이 속성 Location이 가장 높은 판별능력을 가지는 반면 Context는 상대적으로 낮은 판별능력을 가지고 있다는 사실을 알 수 있다.

5. 실 험

제안한 방법의 성능을 측정하기 위해 분류정확도(classification accuracy) 뿐만 아니라 정확율(precision), 재현율(recall), 그리고 F-measure을 이용하여 평가하였다. 평가척도들은 다음과 같이 측정된다.

$$\text{분류정확도}(classification\ accuracy) = \frac{a+d}{a+b+c+d}$$

$$\text{정확률}(precision) = \frac{a}{a+b}$$

$$\text{재현율}(recall) = \frac{a}{a+c}$$

$$F-Measure = \frac{2*\ precision * recall}{precision + recall}$$

성능평가 결과는 <표 7>과 같다. 단어의 문맥정보는 두 가지 접근방법으로 정의하고 각각 학습하였다. <표 7>에서 A라고 표기된 실험은 단어의 앞뒤 각각 두 어절의 기능어 품사를 문맥정보로 선택한 것이다. B라고 표기된 실험은 단어가 문장의 시작 또는 끝에 나타났을 경우는 오른쪽 또는 왼쪽 문맥만 존재하는 특수한 경우이다. 이 경우는 존재하지 않는 문맥이 많아져서 특징벡터에 '*'로 표기되는 건수가 증가하여 학습된 규칙에 판별력을 저하시킨다. 따라서 이 경우는 단어가 문장의 시작에 위치할 경우는 오른쪽 네 어절의 기능어 품사를 문맥으로 사용하고 단어가 문장의 끝에 위치할 경우는 왼쪽 네 어절의 기능어 품사를 문맥으로 각각 사용하여 학습하였다. 실험 결과 B 실험이 A 실험보다 분류 정확도가 약간 높게 나타났지만 유의한 차이를 보이지는 않았다.

평가 결과 C는 제안한 방법으로 학습된 규칙을 시소러스에 존재하지 않는 한국어전자사전의 46% 명사에 대해 평가한 결과이다. 분류 정확도는 약간 떨어지지만, 90% 정도의 분류 정확도를 보여 제안된 방법이 시소러스를 자동으로 확장하는데 효과적으로 적용될 수 있음을 보여주었다.

<표 6> 이진 분류기 평가를 위한 테이블

	Yes is correct	No is correct
Yes was assigned	a	b
No was assigned	c	d

<표 7> 평가 결과

표기	Classification accuracy	Precision	Recall	F-Measure
A	91.91%	95.62%	92.55%	94.06%
B	92.37%	93.67%	95.23%	94.44%
C	89.75%	83.83%	89.92%	86.20%

<표 8> 기존연구와의 성능 비교

	제안한 방법		김민수 [18]	문유진[21]		최유미 [27]
	A	B		C	D	
classification accuracy	91.91%	92.37%	88.40%	88.40%	68.81%	89.40%

<표 8>는 기존 연구들과 본 논문에서 제안한 방법의 분류 정확도를 비교한 것이다. 문유진[21]은 한국어 사전에서 상위어를 추출하는 방법으로 사람이 만든 어휘패턴을 제시한 대표적인 방법이다. 이 논문에서는 6가지 어휘패턴을 제시했으며, 이 어휘패턴으로 상위어를 추출한 결과 성능의 정확도는 C로 표기된 88.4%이다. 본 연구에서 제안한 방법보다 약 3.97% 떨어지는 성능이다.

제안한 방법과의 성능 비교를 좀 더 정확하게 하기 위해서 문유진[21]의 시스템을 재구현하여 [21]에서 제시한 어휘패턴을 가지고 본 연구에서 평가에 사용되었던 평가데이터를 대상으로 적용하여 상위어를 판별하는 성능을 측정하였다. 그 결과, <표 8>에 D로 표기된 정확도를 얻었다. 이것은 본 연구에서 제안한 방법보다 23.56%가 떨어지는 성능을 보였다. 또한 문유진[21]에서 제시했던 성능보다도 떨어지는 결과이다. 문유진[21]에서는 2,600개의 연속적인 소규모 표본을 대상으로 정확도를 평가하였지만, D는 본 논문의 평가데이터인 3만여 개의 데이터를 대상으로 정확도를 평가한 결과이다. 이는 어휘패턴의 경우 어휘에 의존하고 있어서 데이터 영역이 넓어지고 데이터의 수가 많아질수록 현저히 성능이 떨어진다는 사실을 보여준다.

6. 결 론

본 논문에서는 기계가독형사전에서 항목 단어의 상위어를 판별하기 위해서 단어의 정의문으로부터 항목 단어의 상위어가 나타나는 구문적 특징만을 이용하여 상위어 판별규칙을 기계학습하는 방법을 제안하였다.

기계가독형사전에서 상-하위관계를 추출하기 위한 기존의 방법들은 사람에 의해 수집된 어휘패턴을 사용하였다. 이 방법은 패턴 수집의 비용이 많이 들 뿐만 아니라, 어휘패턴의 어휘 의존성으로 인해 문서집합이나 응용분야의 도메인이 변경되면 그 패턴의 성능은 심각하게 저하된다. 따라서 본 논문에서는 사전 정의문에 상위어가 나타나는 구문적 특징들을 정의하고 이것을 이용하여 상위어 판별규칙을 기계학습하였다.

실험 결과에서도 알 수 있듯이 제안한 방법의 분류 정확도는 92.37%로 기존의 연구들보다 훨씬 우수한 성능을 보였다. 이로서 기계가독형사전에서 상위어를 추출 하기 위해서 어휘정보가 아닌 구문정보만을 이용하여도, 더 높은 성능으로 상위어를 추출할 수 있음을 알 수 있다.

참 고 문 헌

[1] Berland, M. and Charniak, E., "Finding Parts in Very Large Corpora," Proceedings of ACL-99, pp.57-64, 1999.

[2] Choi, S. H. and Park, H. R., "A New Method for Inducing Korean Dependency Grammars reflecting the Characteristics of Korean Dependency Relations," Proceedings of the 3rd Conference on East-Asian Language Processing and Internet Information Technology, pp.17-23, 2003.

[3] Choi, S. H. and Park, H. R., "Extracting Semantic Taxonomies of Nouns from a Korean MRD Using a Small Bootstrapping Thesaurus and a Machine Learning Approach," Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, pp.1-9, 2005.

[4] Choi, S. H. and Park, H. R., "Finding Taxonomical Relation from an MRD for Thesaurus Extension," Proceedings of the Second International Joint Conference on Natural Language Processing, pp.357-365, 2005.

[5] Faure, D. and Nedellec, C., "A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition," LREC workshop on Adapting lexical and corpus resources to sub-languages and applications, Granada, Spain, 1998.

[6] Guthrie, L., Brian, M.S., Wilks, Y., and Rebecca, B., "Is There Content in Empty Heads?," Proceedings of COLING'90, pp. 138-143, 1990.

[7] Harabagiu, S. and Moldovan, D.I., "Knowledge processing on an extended WordNet," WordNet: An Electronic Lexical Database, MIT Press, pp.379-405, 1998.

[8] Hearst, M. A., Automatic Acquisition of Hyponyms from Large Text Corpora, In Christiane Fellbaum(Ed.) WordNet: An Electronic Lexical Database, MIT Press, pp.132-152, 1998.

[9] Lee, L., Similarity-Based Approaches to Natural Language Processing, Ph. D. Thesis, Harvard University Technical Report TR-11-97, 1997.

[10] Maria, R. C., Enrique, A., and Pablo, C., "Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia," Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems, pp.67-79, 2005.

[11] Martin, S. C., Roy, J. B., and George, E. H., "Extracting Semantic Hierarchies from a Large On-Line Dictionary," Proceedings of the 23rd Conference of the Association for Computational Linguistics, pp.299-304, 1985.

[12] Montemagni, S. and Vanderwende, L., "Structural Patterns vs. String Patterns for Extracting Semantic Information from Dictionaries," Proceedings of COLING-92, pp.546-552, 1992.

[13] Quinlan, J.R., C4.5: Programs for Machine Learning, San Mateo, CA: Morgan Kaufman, 1993.

[14] Richardson, S. D., Dolan, W. B., and Vanderwende, L., "MindNet: Acquiring and Structuring Semantic Information from Text," Proceedings of COLING-ACL'98, Vol.2, pp. 1098-1102, 1998.

[15] Rigau, G., Automatic Acquistion of Lexical Knowledge from MRDs, Ph.D. Thesis, Universitat Politecnica de Catalunya, 1998.

[16] Rigau, G., Rodriguez H., and Agirre E., "Building Accurate Semantic Taxonomies from Mololingual MRDs," Proceedings of the 36th Conference of the Association for Computational Linguistics, pp.1103-1109, 1998.

[17] Wiks, Y., Fass, D. C., Guto, C. M., McDonald, J. E., Plate, T., and Slator, B. M., "Providing machine tractable dictionary tools," Journal of Computers and Translation, pp. 99-154, 1990.

[18] 김민수, 김태연, 노봉남, "국어사전을 이용한 한국어 명사에 대한 상위어 자동 추출 및 WordNet의 프로토타입 개발," 한국정보처리학회 논문지, 2(6), pp.184-156, 1995.

[19] 김영택 외 공저, 자연언어처리, 생능출판사, pp.52-54, 2001.

[20] 동아 새국어사전, 동아출판사, 1989.

[21] 문유진, 의미론적 어휘 개념에 기반한 명사 워드넷의 설계와 구축, 서울대학교 대학원 박사 학위논문, 1997.

[22] 문유진, 김영택, "한국어 명사의 Hypernym 자동 추출 방법," 한국정보과학회 학술발표대회 논문집, Vol.21, No.2, pp. 613-616, 1994.

[23] 엣센스 국어사전, 민중서림, 1993.

[24] 전문용어언어공학연구센터(KORTERM), : KAIST language resources http://www.korterm.or.kr/

[25] 조평옥, 안미정, 옥철영, 이순오, "사전 뜻풀이말에서 구축한 한국어 명사 의미계층구조," 한국인지과학회 논문지, 10(3), pp.1-10, 1999.

[26] 최선화, 박혁로, "한국어 확률 의존문법 학습," 제 30회 한국정보과학회 춘계 학술발표대회 논문집(B), pp.513-515, 2003.

[27] 최유미, 사공철, "상위어 자동추출 알고리즘 개발," 제 15회 한국정보관리학회 학술대회 발표 논문집, pp.227-230, 1998.

최 선 화

e-mail : csh123@dreamwiz.com
1991년 광주대학교 전자계산학과(학사)
2002년 전남대학교 전산학과(이학석사)
2006년 전남대학교 전산학과(이학박사)
1995년~1998년 송원백화점 정보시스템부
2003년~ 2005년 전남대학교 전자컴퓨터 정보학부 조교
2006년~현재 전남대학교 전산학과 Post Doc.
관심분야 : 자연언어처리, 정보검색, 정보추출

박 혁 로

e-mail : hyukro@chonnam.ac.kr
1987년 서울대학교 전산학과(학사)
1989년 한국과학기술원 전산학과(공학석사)
1997년 한국과학기술원 전산학과(공학박사)
1994년~1996년 연구개발정보센터 연구원
1997년~1998년 연구개발정보센터 선임연구원
1999년~2002년 전남대학교 전산학과 조교수
2002년 University of Maryland UMIACS Post Doc.
2002년~현재 전남대학교 전자컴퓨터공학부 부교수
관심분야 : 자연언어처리, 정보검색, 텍스트마이닝

부 록A

분류		태그	설명
체언	일반명사	nc	일반명사
		nca	동작성 보통명사
		ncs	상태성 보통명사
		nct	시간성 보통명사
	고유명사	nq	고유명사
	의존명사	nb	불완전명사
		nbu	단위명사
	수사	nn	수사
	대명사	npp	인칭 대명사
		npd	지시 대명사
용언	동사	pv	동사
	형용사	pa	형용사
		pad	지시 형용사
	보조용언	px	보조동사
수식언	관형사	m	관형사
		md	지시관형사
		mn	수관형사
	부사	a	일반부사
		ajs	문장 접속부사
		ajw	단어 접속부사
		ad	지시부사
독립언	감탄사	ii	감탄사
관계언	격조사	jc	격조사
		jca	부사격조사
		jcm	관형격조사
		jj	접속격조사
		jcv	호격조사
	보조사	jx	보조격조사
	서술격조사	jcp	서술격조사
어미	선어말어미	efp	선어말어미
	연결어미	ecq	대등적 연결어미
		ecs	종속적 연결어미
		ecx	보조적 연결어미
	전성어미	exn	명사 전성어미
		exm	관형격 어미
		exa	부사격 어미
	종결어미	ef	종결어미
접사	접두사	xf	접두사
	접미사	xn	접미사
		xpv	동사화 접사
		xpa	형용사화 접사