SVM 기계학습을 이용한 웹문서의 자동 의미 태깅†

(Automatic semantic annotation of web documents by SVM machine learning)

황운호*, 강신재**
(Woon-Ho Hwang, Sin-Jae Kang)

요 약 본 논문은 시맨틱 웹의 실현을 위해서는 필수적인 작업인 웹문서의 의미를 자동으로 태깅할 수 있는 시스템에 관한 것이다. 웹상의 방대한 자원을 일일이 사람이 수작업으로 의미를 태깅한다는 것은 사실상 불가능하기 때문에, 한국어 웹문서를 대상으로 대량의 학습 데이터를 수집하고 자연어처리 기법과 시소러스를 이용하여 특징을 추출한 후, SVM 기계학습을 통하여 개념분류기를 구축하였다. 한국어의 특징을 파악하여 의미 태깅에 필요한 특징 정보를 추출하기 위해서 형태소 분석과 구문 분석을 하였다. 추출된 특징 정보는 가도카와 시소러스의 의미코드를 이용하여 학습벡터로 구성되는데, 이는 유사한 단어나 구를 하나의 개념코드로 매핑하여 시스템의 재현율을 높이는 역할을 하게 된다. 실험결과 자동 의미 태깅 분야에서 본 접근방법의 가능성을 확인할 수 있었다.

핵심주제어 : 의미 태깅, 기계학습, 시소러스, 자연어처리

Abstract This paper is about an system which can perform automatic semantic annotation to actualize "Semantic Web." Since it is impossible to tag numerous documents manually in the web, it is necessary to gather large Korean web documents as training data, and extract features by using natural language techniques and a thesaurus. After doing these, we constructed concept classifiers through the SVM (support vector machine) learning algorithm. According to the characteristics of Korean language, morphological analysis and syntax analysis were used in this system to extract feature information. Based on these analyses, the concept code is mapped with Kadokawa thesaurus, which made it possible to map similar words and phrase to one concept code, to make training vectors. This contributed to rise the recall of our system. Results of the experiment show the system has a some possibility of semantic annotation.

Key Words : Semantic annotation, Machine learning, Thesaurus, Natural language processing

1. 서 론

사용자가 네트워크나 컴퓨터에 의존하지 않고 장소에 상관없이 자유롭게 네트워크에 접속하여 원하는 정보를 얻을 수 있는 유비쿼터스(ubiquitous) 시대가 머지않아 올 것이다. 마크 와이저[1]가 제안하였듯이, 유비쿼터스 컴퓨팅이 추구하는 목표는 시간과 장소의 제약을 받지 않고 일상생활에 광범위하게 적용될 수 있는 컴퓨팅 환경을 제공하는 것이라 할 수 있다. 컴

† 이 논문은 2006학년도 대구대학교 학술연구비 지원에 의한 논문임.
* 대구대학교 대학원 컴퓨터정보공학과 석사과정
** 대구대학교 정보통신대학 컴퓨터·IT공학부 교수

퓨터 환경에서는 사용자와 컴퓨터 사이에 네트워크의 연결을 전제로 하고 있지만, 사용자의 적극적인 개입 없이도 서비스를 제공할 수 있다. 이를 마크 와이저는 고요한 기술(calm technology)이라 정의하였다.

여기서 한 단계 더 나아가 미래의 사회는 스스로 정보를 분석하고 처리할 수 있는 웹을 요구하게 될 것이다. 현재의 웹은 사용자와 서버간의 단조로운 정보 전달로써의 역할을 벗어나지 못하고 있는 것이 사실이다. 또한 기하급수적으로 늘어나는 인터넷 상의 자원을 관리하고 제공하기에는 이미 현재의 웹은 한계를 드러내었다. 최근 이와 같은 현재의 웹을 사용자와 서버간 또는 서버와 서버간의 좀 더 활동적이고 동적인 형태의 통신을 가능하게 하며 스스로 움직이는 웹이라 불리는 시맨틱 웹(semantic web) 개념이 등장하였다. 이러한 차세대 웹이라 불리는 시맨틱 웹을 실현하기 위해서는 최우선적으로 필요한 것이 현재 웹에 존재하는 모든 정보들을 의미 태깅된 형태의 정보로 바꾸는 작업이다. 일단 웹상의 정보들이 의미 태깅되게 되면 온톨로지와 추론엔진을 통하여 인간의 개입 없이 자동화된 웹 서비스를 구현할 수 있게 되는 것이다. 현재 웹상에 존재하는 정보의 양은 가늠하기조차 힘들 정도로 방대하므로, 일일이 사람이 수작업으로 각 개념의 의미를 태깅한다는 것은 사실상 불가능에 가깝다. 따라서 이를 자동으로 해결할 수 있도록 하려는 연구가 필요한 것이다.

시맨틱 웹을 위한 웹문서의 의미 태깅은 기존의 자연어처리와 정보검색 분야에서 해오던 단어의미 중의성 해소와 개체명 인식 연구와 밀접한 관련이 있다. 단어의미 중의성 해소는 기계번역 등의 시스템을 위해 문장의 의미를 완전하게 분석하기 위한 절차로 의미 태깅을 위한 정답이라 할 수 있는 개념 분류가 개체명 인식의 경우보다 세부적이고 많기 때문에 좋은 결과를 얻기가 쉽지 않다. 하지만 개체명 인식의 경우는 정보추출 시스템의 목적에 따라 개념 분류의 수가 적게 정의되므로 상대적으로 처리하기가 쉽다.

본 논문에서는 문장의 완전한 의미를 분석하기보다는 중요 단어의 의미를 태깅하는 데에 목적을 두고 있으므로 개체명 인식연구와 보다 밀접한 관련이 있다고 볼 수 있다.

하지만 기존의 개체명 인식 연구들은 주로 패턴 인식이나 확률 통계, 기계학습 방법에만 의존하고 있으므로, 본 논문에서는 형태소/구분 분석과 같은 보다 심화된 자연어처리 기법과 시소러스를 활용하여 성능의 향상을 꾀하고자 하였으며, 향후 시맨틱 웹 환경에서의 응용을 위해 웹 문서를 대상으로 시스템을 구성하였다. 본 논문에서는 기 확보하고 있는 한국어 형태소/구문 분석기와 가도카와 시소러스, SVM 기계학습 알고리즘을 이용하여 한국어 웹 문서의 의미 태깅에 적합한 시스템을 제안하였다.

다른 측면에서 고려해야 할 또 다른 한 가지는 의미 태깅의 대상이 되는 언어의 고유한 특징에 따라 전산처리 방법이 상이하므로 영어와 같은 외국어를 대상으로 한 관련 연구는 간접적인 참고는 될 수 있으나 본 연구와 직접적인 비교를 하기에는 어려운 면이 있다. 따라서 한국어를 대상으로 한 관련 연구들을 중심으로 본 연구와의 특징과 실험결과를 논문의 마지막에 비교하였다.

2. 관련 연구

2.1 시맨틱 웹(Semantic Web)

시맨틱 웹[2],[3]은 웹의 창시자 팀 버너스 리(Tim Berners-Lee)에 의해 1998년에 제안된 차세대 웹의 이름으로 각종 회의와 연구를 통해 지속적으로 관련 규격과 기술이 개발되고 있다. 팀 버너스 리는 '시맨틱 웹은 현재 웹의 확장으로 잘 정의된 의미를 제공함으로써 사람과 컴퓨터가 좀 더 협력을 잘 할 수 있는 웹'이라고 설명하고 있다.

시맨틱 웹은 컴퓨터끼리 정보를 주고받을 때 잘 정리된 좀 더 많은 정보를 추가로 제공해 정보해석력을 높이고, 이를 통해 자동화처리를 향상시킨 웹이다. 이때 컴퓨터끼리 주고받는 추가정보는 사람 눈에는 보이지 않는 메타데이터(meta data)를 활용하게 된다. 기계끼리 일을 자동으로 처리하기 위해서는 기계끼리 서로 대화를 나누고 대화 내용을 이해할 수 있는 방법이 필요하다. 이를 위해 컴퓨터가 이해할 수 있는 언어로 웹을 구성해야 하는데, RDF (resource description framework), 온톨로지 (ontology), OWL(ontology web language) 등이 이를 위해 논의되고 있는 기술이다.

2.2 개체명 인식 관련 연구

본 논문의 주제와 관련 있는 연구로는 개체명 인식(named entity recognition) 분야를 들 수 있다.

김성원[4]은 개체명 인식을 위해서 개체명 경계 인식과 개체명 인식으로 작업을 구분하여 실행하였는데, 첫 번째 단계의 개체명 경계 인식에서는 개체명 범주별로 이미 분류되어 있는 단어 리스트들을 이용하였고, 두 번째 단계에서는 사용되는 말뭉치로부터 추출된 명사열을 학습 말뭉치로 사용하였다. 경계의 구분을 위해서 경계 레이블을 도입하고, 최대 엔트로피 모델의 빔 탐색 디코딩 방법을 이용하여 결정되는 경계 레이블들의 최대 확률 경로를 개체명의 경계로 인식하였다. 다양한 자질을 모델에 사용할 수 있는 최대 엔트로피 모델의 적용은 국내에서 처음 시도되는 연구로써 가치가 있다.

이기중[5]은 생의학 분야의 개체명을 인식할 수 있는 시스템에 관한 연구를 하였다. 기존의 사전을 이용한 사전 기반 접근 방식, 전문가가 수작업으로 작성한 규칙을 사용하는 규칙 기반 접근 방식 그리고 개체명이 표시된 학습 문서로부터 학습을 수행하는 기계학습 방식에 이르기까지 다양한 접근법이 시도되어 왔으며 최근에는 기계학습 방식에 대한 연구가 활발히 진행되고 있다. 위 논문에서는 자연어처리 분야에서 뛰어난 성능을 보여주었던 기계학습 접근 방식을 생의학 개체명 인식 문제에 적용하였다. 기존의 연구와 다른 점은 경계 인식과 의미 분류 단계를 분리한 분리 모델(2-phase model)을 사용하여 SVM학습 과정의 복잡도를 줄였고, 각 단계에 적합한 자질을 구분하여 사용함으로써 전체 인식 성능을 높일 수 있는 새로운 방법을 제안하였다.

황이규[6]는 개체명 인식 모델로 HMM(Hidden Marcov Model)을 이용하는 방법을 제안했다. 하나의 개체명 내의 여러 단어를 개체명 인접 단어로 정의하여 부개체형이라는 개체명 구성 단어 분류를 했다. 학습은 부개체 유형 인식 학습 과정과 개체명 경계 인식 학습이 동시에 진행되고, 이 과정에서 상태 천이확률과 어휘 확률이 학습된다. 이때, 개개의 단어에 부개체명 사전을 이용하여 트라이그램(trigram)과 어휘 확률을 추출하여 학습을 시킨다. 이 방법은 개체명 경계 인식을 HMM을 사용하여 인식하려고 한 것이 두드러지지만, 개체명 인식을 위해 부개체형 사전에 기반하기 때문에, 조사나 용언 정보와 같은 개체명 인식과 중의성 해소에 많은 도움을 주는 주변 어휘 정보를 고려하지 못하는 단점이 있다.

이승우[7]는 개체명 인식을 위한 오류제어 부트스트래핑 방식을 제안하였는데, 특히, 개체명의 세밀한 분류와 부트스트래핑 알고리즘 그리고 오류 제어 휴리스틱에 초점을 맞추고 있다. 기존의 인명과 지명 그리고 기관명만을 분류하는 것과는 달리, 이 논문에서는 지명에 대한 보다 세밀한 구분을 다루고 있는데, 이는 질의응답과 정보 추출과 같은 보다 정교한 응용에서 유용하게 사용될 수 있는 장점이 있다. 부트스트래핑 방식은 학습 도중에 유발되는 오류에 굉장히 민감하기 때문에 잘못 인식된 개체명이 학습을 방해하는 것을 막기 위해 여러 가지 오류제어 휴리스틱들이 적용되었다. 제안된 휴리스틱들은 언어적 지식에 더하여 웹과 워드넷, 백과사전, 고유명사사전과 같은 다양한 외부 지식원들이 적절히 활용되었다.

2.3 SVM(Support Vector Machine)

Vapnik[8]에 의해 개발되었으며 이진 분류에 사용된다. SVM은 분류 문제를 해결하기 위해 최적의 분리 경계면을 제공한다. 명백한 이론적 근거에 기반을 두므로 결과 해석이 용이하고 실제 응용에 있어서 인공신경망 수준의 높은 성과를 얻을 수 있고, 적은 학습 자료만으로 신속하게 분별학습을 수행할 수 있기 때문이다. 또한 기존의 기계학습 알고리즘은 학습 집단을 이용하여 학습오류를 최소화하는 경험적 위험 최소화 원칙(EMR : Empirical Risk Minimization)을 구현하는 것인데 비해 SVM은 구조적 위험(SRM : Structural Risk Minimization)을 최소화하려는 원칙을 구현한 것이다.

또한 SVM 알고리즘은 마진을 최대화시키는 초평면을 찾아주는 알고리즘으로 마진을 최대화시키는 초평면의 일반화 성능이 우수하다. 최적의 상태는 최대 마진을 가지는 것이며 최대 마진 초평면은 최적으로

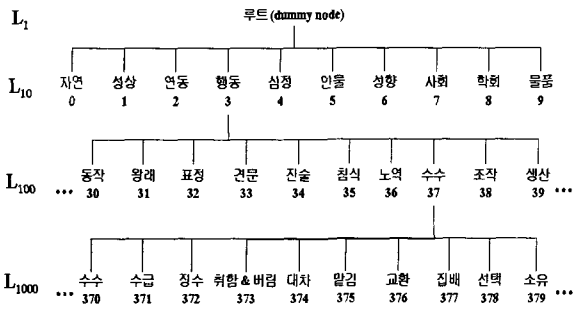
두개의 클래스를 분리할 수 있는 초평면이다. 패턴 집합이 선형이고 분리 가능한 경우에 SVM은 입력패턴의 학습을 통하여 +1과 -1의 두 클래스로 분류한다. 두 개의 클래스로 분류된 훈련 집단은 각 클래스에 포함된 훈련 패턴들을 분리하는 초평면을 결정하는데 사용된다. 초평면을 결정하는 입력 패턴을 지지벡터(support vector)라고 하고 적절한 초평면을 찾으면 오분류를 피할 수 있다.

통계적 학습기반의 자연언어처리 방식은 음성인식(speech recognition)이나 구문 태깅(syntactic tagging)과 같은 문제에서 좋은 결과를 얻게 되었고 언어학적인 지식에 의한 방법만으로는 구축하기 어려운 대규모의 자연언어처리 응용들이 출현하고 개발됨에 따라 더욱 인기를 얻고 있다.

2.4 가도카와(Kadokawa) 시소러스

한국어를 위한 시소러스는 여러 가지가 있으나 범용의 시소러스라기보다는 특정 도메인을 대상으로 한 것들이 일반적이며, 또한 그 분류체계에 따른 효과가 객관적으로 입증되었거나, 완전한 전자사전의 형태로 구축되어 전산처리가 즉시 가능한 형태로 제공되는 리소스를 찾기란 쉽지가 않다. 더욱이 이러한 리소스가 오픈 소스의 형태로 공개되어 있는 경우는 매우 드물다.

가도카와 시소러스는 총 1,110개의 개념과 4단계의 계층구조를 가지고 있으며, L1, L10, L100 레벨에 속해 있는 개념들은 각각 10개의 하위 개념들로 나뉜다[9]. 그림 1에 가도카와 시소러스의 구조를 나타내었다.



<그림 1> 가도카와 시소러스 계층도

가도카와 시소러스는 일본어를 대상으로 만들어지긴 하였으나, 개념 부류가 1,110개 정도이기 때문에 일본어에만 존재하는 특수한 개념 부류가 포함되었다고는 볼 수 없다. 만약, 개념 부류의 수가 더 많았다면 그러할 가능성은 높아질 것이다. 즉, 1,110개의 부류는 한국어와 같은 다른 언어에서도 존재하는 일반적인 개념들을 포함하기에 적당한 수준의 분류라고 볼 수 있다.

실제로 시소러스를 시스템에서 사용하려면 우리가 사용하는 모든 한국어 단어들과 시소러스 개념간의 사상(mapping) 관계가 전자사전이나 데이터베이스의 형태로 구축이 되어 있어야 한다.

가도카와 시소러스는 포스텍(POSTECH)에서 개발된 일한기계번역시스템(COBALT-J/K)[10]과 한일기계번역시스템(COBALT-K/J)[11]에서 활용되어 이미 그 성능을 입증 받은바 있는데, 특히 한일기계번역시스템에서는 한국어의 분석에서 본 시소러스를 사용하여 좋은 결과를 얻을 수 있었다[11]. 또한 그 결과로 한국어의 모든 단어들과 가도카와 시소러스의 개념코드간 사상 정보가 기계번역 전자사전에 코딩되었다.

따라서 본 논문에서는 한국어 분석에서도 성능이 확인되었고, 시스템 구현 시에도 추가의 사전 작업 없이 구축을 용이하게 할 수 있다는 장점으로 인해 가도카와 시소러스를 개념분류로 사용하게 되었다.

3. 자동 의미 태깅 시스템

본 장에서는 기계학습 기법을 이용하여 한국어 웹문서의 의미를 자동으로 태깅하는 시스템의 전반적인 구조와 방법론에 대해서 설명한다. 새로운 웹문서내의 단어의 의미를 시스템이 자동으로 이해하고 태깅하기 위해서는 사전에 기계학습을 통하여 개념별 분류기(classifier)를 생성하는 단계가 필요하다. 분류기란 사전에 여러 사실들을 분석하여 정보(또는 특징)를 추출한 후 이들의 관계 및 패턴을 학습해 놓는 것이다. 일단 분류기가 구축된 후, 새로운 웹문서의 의미 태깅을 위해서는 HTML 태그 제거 등의 전처리 과정을 거친 한국어 문장과 단어들로부터 분류기에 적용하기 위한 특징들을 추출하여 분류기에 적

용하는 절차를 거친다. 이 때 분류기가 새로운 입력을 기존의 정리된 사실들과 비교/분석 및 패턴 매치를 하여 확률적인 분류 결과 값을 리턴하게 되고 이를 기반으로 해당 단어의 개념 분류가 어디에 속하는지를 판단하게 된다.

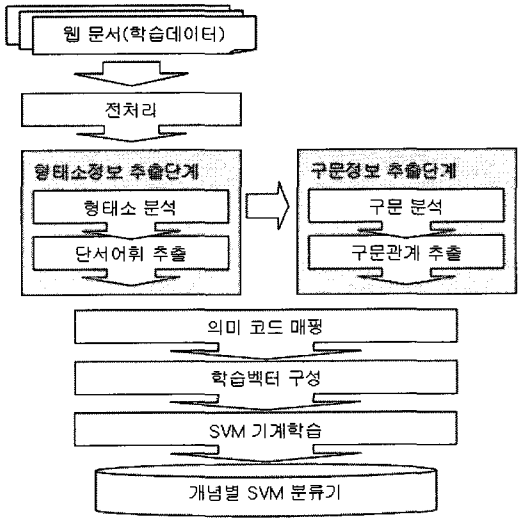
이를 위해서 시스템은 크게 기존 사실들을 학습해서 분류기를 생성하는 학습 단계와 새로운 입력을 분석하여 의미 태깅을 하는 적용 단계로 구성된다.

3.1 학습 단계

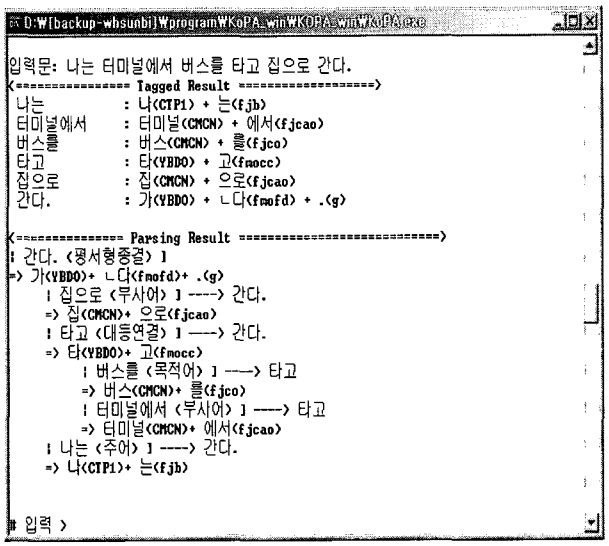
의미 태깅 시스템을 구축하기 위해서는 먼저 충분한 양의 학습 데이터들로부터 정보를 추출하여 SVM 분류기를 생성하여야 한다. 기계학습을 위해서는 학습 데이터들을 분석하고 학습벡터를 생성하는 단계가 필요하다. 그에 앞서 학습을 위한 웹문서에 존재하는 여러 HTML 혹은 다른 스크립트 태그들을 제거하여 순수 한국어 문장만을 추출하는 일련의 웹문서 정제 과정을 거친다.

한국어 문장의 특징을 추출하고 의미를 파악하기 위해서는 자연어처리 기법이 필요하다. 한국어 문장들에서 단어의 품사들을 추출하는 형태소 분석과 각 형태소들 간의 문맥 구문관계를 분석하는 구문 분석을 수행한 결과를 바탕으로 자질을 추출하여 사용한다. 형태소 분석을 통해 얻어진 정보를 의미코드에 매핑하는 부분과 구문 분석을 통해 얻어진 정보를 의미코드에 매핑하는 부분을 통합하여 학습벡터를 구축한다. SVM 기계학습을 위해서는 도르트문트 대학(University of Dortmund)에서 만든 SVM_light라는 도구를 활용하였다[12]. 학습을 위한 단계는 그림 2와 같다.

형태소 분석과 구문 분석을 위해서 본 시스템에서는 포스텍에서 개발하여 한국어분석에 대한 견고성과 실용성을 입증 받은 범용 형태소 분석기 KoMA (Korean Morphological Analyzer)[13]와 구문 분석기 KoPA(Korean Parsing Analyzer)[14]를 활용하였다. 그림 3은 형태소 분석과 구문 분석의 결과를 보여주고 있다.

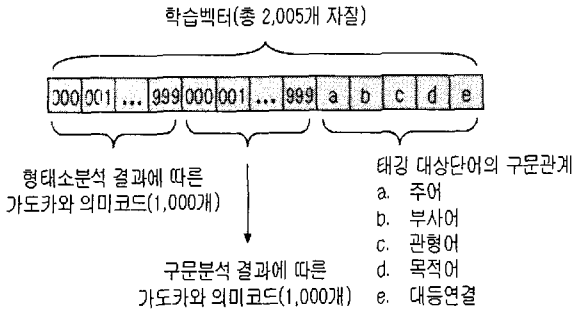


<그림 2> SVM 분류기 생성을 위한 학습과정



<그림 3> 형태소 분석과 구문 분석 예시

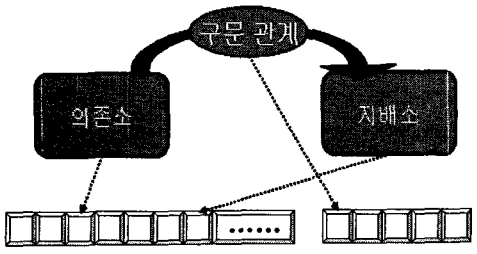
기계학습을 위한 학습벡터는 형태소 분석을 통해 얻어지는 1,000개의 자질(feature)과 구문 분석을 통해 얻어지는 1,005개의 자질을 하나로 합쳐서 총 2,005개의 자질로 구성된다. 그림 4에서 학습벡터의 전체 구성요소를 나타내었으며, 3자리 숫자는 가도카와 시소러스의 L1000 레벨의 의미코드이다.



<그림 4> 학습벡터 구성

학습을 위한 웹문서에서 한국어 문장들을 추출한 후, 특정 의미가 포함된 문장들끼리 모은 다음 이를 해당 의미를 태깅하기 위한 SVM 분류기 학습에 사용한다. 각 문장들을 형태소 분석하여 명사와 형용사, 동사들을 추출하는데, 추출된 단어들을 가도카와 시소러스의 L1000레벨에 존재하는 1000가지의 분류로 매핑하여 벡터를 구성한다. 이는 유사한 의미를 가지는 단어들을 하나의 범주로 묶음으로써 시스템의 적용 범위와 재현율을 높이기 위함이다.

또한 학습 문장을 구문 분석하여 어절 사이에 존재하는 구문 관계를 추출한다. 실험에 사용한 KoPA 구문분석기[14]는 의존문법을 사용하는데, 의존소와 지배소 사이에 존재하는 여러 구문관계 가운데 주어, 목적어, 부사어, 관형어, 대등연결의 관계를 학습에 사용하였다. 구문 분석된 결과의 예시는 그림 3에서 확인할 수 있으며, 구문 분석결과로부터 학습 벡터를 구성하는 과정이 그림 5에 나타나 있다.

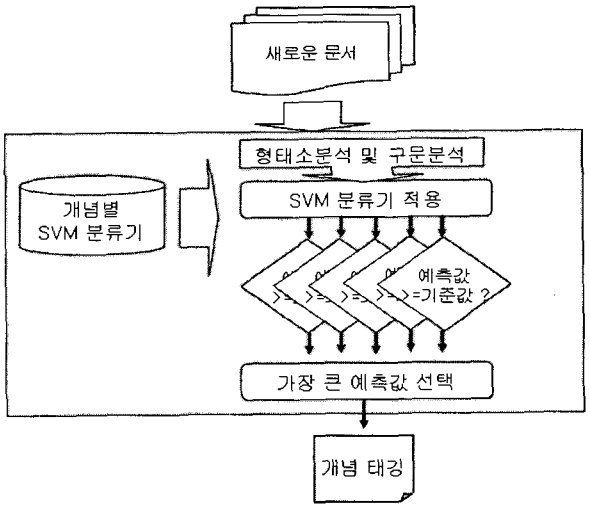


<그림 5> 구문분석결과의 학습벡터로의 매핑 예시

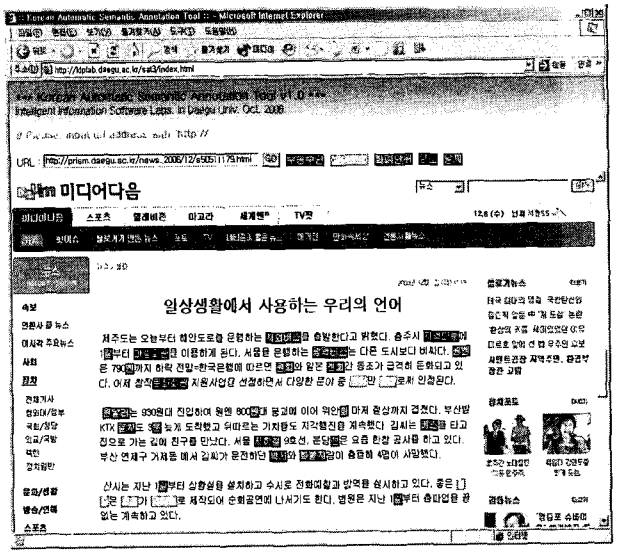
3.2 적용 단계

적용 단계는 학습 단계에서 만들어진 SVM 분류기에 새로운 웹문서를 적용하는 과정으로, 학습 단계에서 학습벡터를 만드는 과정과 비슷한 과정으로 진행된다.

의미 태깅할 새로운 웹문서를 입력으로 받아 전처리 과정을 거쳐 한국어 문장을 추출하고, 학습 단계의 자질 추출 방법과 동일한 방법으로 벡터를 구성하게 되는데, 학습 단계와 다른 점은 본 단계에서 만들어진 벡터가 어떤 의미 분류에 속할지를 모르기 때문에 학습 단계를 통해 이미 구축되어 있는 모든 SVM 분류기에 각각 적용시켜 보아야 한다는 점이다. SVM 분류기에 적용시키면 각 분류기마다 하나씩의 분류 예측 값을 되돌려주게 되는데, 각 분류기의 예측 값 수치가 높을수록 해당 분류일 확률이 높은 것이다. 여러 분류기에서 나온 예측 값들 중 제일 큰 예측 값을 갖는 분류를 선택하여 해당 의미로 의미 태깅하게 된다. 시스템의 적용 과정은 그림 6에 나타내었고, 그림 7은 웹 브라우저를 통해 본 시스템에 접속하여 자동 의미 태깅하는 모습을 보여주고 있다.



<그림 6> 의미 태깅 시스템의 적용 과정



<그림 7> 웹문서 의미 태깅 예시

4. 실험

4.1 실험 방법

영문 웹문서를 대상으로 자동 의미 태깅할 수 있도록 만들어진 대표적인 시스템으로는 KIM (Knowledge Information Management)[15],[16]이 있다. 이 시스템은 총 250개의 개념을 대상으로 의미 태깅을 수행한다. KIM 시스템은 상당기간동안 많은 연구자들에 의해 구축된 시스템으로 현재 실용단계에 이른 시스템이나, 본 논문은 한국어를 대상으로 웹문서의 자동 의미 태깅의 가능성을 보이기 위해 KIM의 분류 중에서 대표적인 5개의 개념분류만을 대상으로 실험을 하여 성능을 평가하였다. 본 방법론의 유효성 및 효율성이 입증되면 학습 문장의 수집 및 분류기 학습을 반복하는 과정으로 본 시스템을 확장시킬 수 있다.

실험을 위해서는 학습용 한국어 웹문서가 필요하나 한국어 웹문서를 개념별로 분류해 놓은 말뭉치가 확보되지 않아 학습과 테스트를 위한 말뭉치를 수작업으로 수집하였다. 각 분류마다 핵심 키워드 5개씩을 선정하고 각 키워드가 속한 문장을 100문장씩 수집한 총 2,500개의 문장을 학습을 위해 사용하였다. 그리고 학습 말뭉치의 10%에 해당하는 10개의 웹문서를 각 키워드별로 따로 수집하여 총 250개의 문장을 테스트를 위한 말뭉치로 구성하였다. 표 1은 학습과 테스트를 위해 사용된 한국어 문장의 수를 나타낸 것이며, 상세한 대표단어 정보는 표 2에 나타나있다.

<표 1> 실험 말뭉치의 분포

용도 / 분류	학습을 위한 문장의 수	테스트를 위한 문장의 수
수송수단	500	50
예술분야	500	50
화폐단위	500	50
장 소	500	50
시 간	500	50
계	2,500	250

<표 2> 실험 말뭉치 수집을 위한 대표 단어

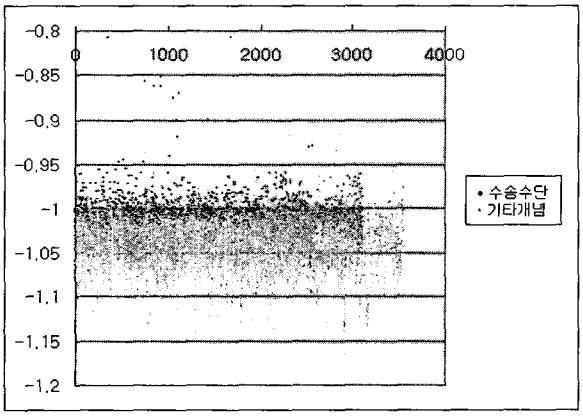
구분	대표 단어
수송수단	버스, 택시, 지하철, 기차, 비행기
예술분야	공연, 연극, 뮤지컬, 영화, 무용
화폐단위	원, 엔, 위안, 달러, 유로
장 소	레스토랑, 설악산, 공원, 학교, 호텔
시 간	년, 월, 일, 개월, 지난해

개념별로 수집된 학습 문장들은 형태소/구문 분석과 가도카와 시소러스로의 매핑 과정을 통하여 개념별 학습벡터로 추출되고, SVM 기계학습을 통하여 각 개념별 SVM 분류기가 만들어지게 된다.

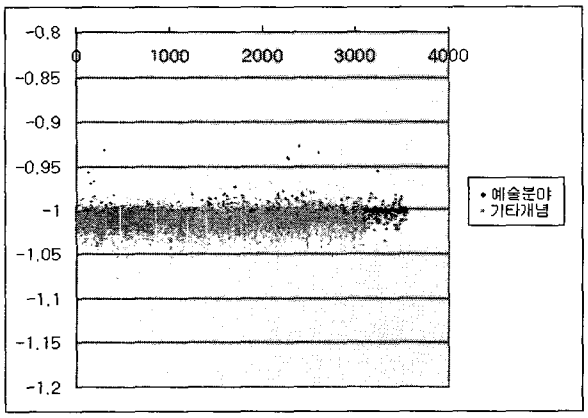
학습하지 않은 새로운 웹문서로부터 추출된 벡터를 SVM 분류기에 넣으면 해당 벡터에 대한 결과값을 반환하게 되는데 이 때 해당 단어가 해당 개념 분류인지 아닌지를 가리는 기준 값(threshold)을 정해야 한다. 반환 값이 기준 값 이상이면 해당 분류일 가능성이 높은 것으로 판단할 수 있다. 일반적으로는 0을 기준 값으로 사용하지만, 본 실험에서는 보다 정확한 기준 값을 설정하기 위해서 학습 과정에서 사용했던 학습데이터들의 반환 값 분포를 보고 이상적인 수준의 기준 값을 결정하여 시스템에 적용하였다. 기준 값은 학습벡터 헤드의 양수와 음수의 비율 혹은 하나의 벡터에서 나타나는 의미코드의 개수 등에 따라 달라지기 때문에 이와 같은 과정이 필요하며, 이를 통해서 최적화된 시스템을 만들 수 있다. 그림

8, 9, 10, 11, 12는 학습데이터를 각 SVM 분류기에 입력하여 나온 결과를 보여주고 있다. 그림에서 가로축은 실험말뭉치에서 추출한 학습데이터의 인스턴스 번호를 의미하고, 세로축은 각 학습데이터에 대한 해당 SVM 분류기의 반환값을 의미한다. 그림 8에서 회색으로 표시된 기타개념은 수송수단을 제외한 나머지 4개 개념, 즉 예술분야, 화폐단위, 장소, 시간에 관련된 학습데이터의 반환값을 의미한다. 이에 따라 학습데이타의 결과를 분석하고 각 분류기별로 기준 값을 설정하여 사용하였다.

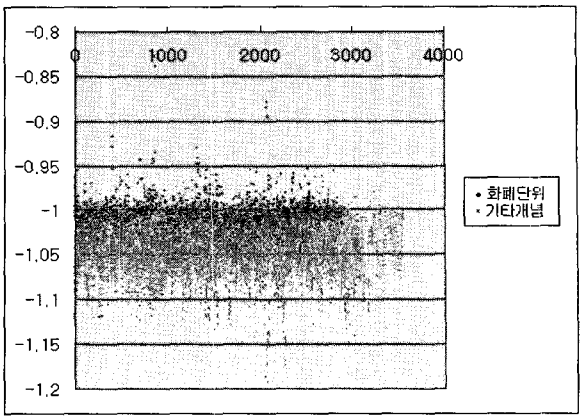
예술분야와 장소의 개념을 분류하는 SVM의 결과(그림 9, 그림 11 참조)를 살펴보면 분포가 몰려 있는 것을 볼 수 있는데, 이는 해당 개념을 분류하기 위한 자질이 다른 개념에 비해 상대적으로 부족하다는 것을 의미한다. 따라서 향후 연구로는 해당 개념을 위한 추가 자질을 고안할 필요성이 있다.



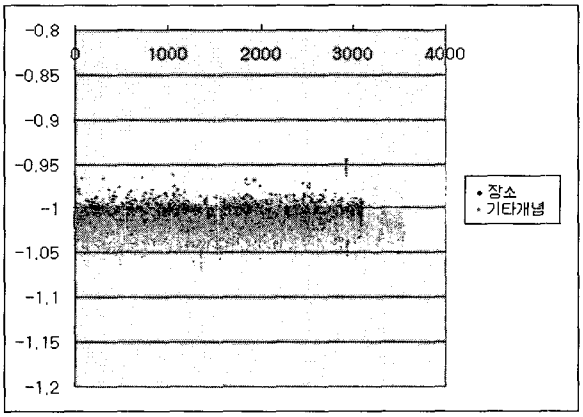
<그림 8> 수송수단 SVM분류기의 반환값 분포



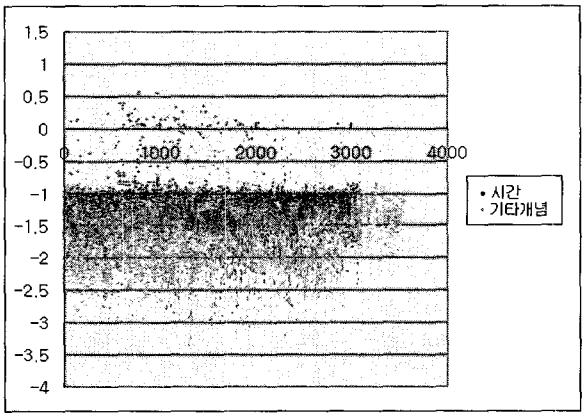
<그림 9> 예술분야 SVM분류기의 반환값 분포



<그림 10> 화폐단위 SVM분류기의 반환값 분포



<그림 11> 장소 SVM 분류기의 반환값 분포



<그림 12> 시간 SVM 분류기의 반환값 분포

4.2. 실험 결과

구축된 시스템의 성능 테스트를 위해 따로 수집해

놓은 250개의 문장을 입력으로 주어 얼마나 정확히 잘 분류하는지를 보고 각 개념 분류에 따라 정확률, 재현율 그리고 F-measure를 계산하였다.

본 시스템의 평가를 위해서는 정보 검색 분야에서 일반적으로 많이 사용하고 있는 정확률과 재현율, 그리고 F-measure를 이용한다. 표 3은 성능 평가 계산을 위한 기준 분할표이며, 각 평가척도의 수식은 다음과 같다.

<표 3> 성능 평가를 위한 분할표

시스템 / 실제	해당 분류	그 외 분류
해당 분류	a	c
그 외 분류	b	d

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

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

$$\text{F-measure} = \frac{(\beta^2+1) * \text{정확률} * \text{재현율}}{\beta * \text{정확률} * \text{재현율}} \quad (\beta=1)$$

정확률은 시스템이 의미 태깅한 결과 가운데 실제로 정답인 것의 비율을 의미하고, 재현율은 존재하는 정답, 즉 찾아내어야 할 정답 가운데 실제로 시스템이 태깅하여 찾아낸 정답의 비율을 의미한다. F-measure는 정확률과 재현율의 수치를 이용하여 하나의 평가 수치를 만들어 낸 것으로, 수식에서 β는 정확률에 대한 재현율의 가중치를 의미한다. 본 실험에서는 동일한 가중치를 부여하기 위해 β에 1을 사용하였다.

<표 4> 자동 의미 태깅 시스템의 성능평가 결과 (단위:%)

	정확률	재현율	F-measure
수송수단	84.8	90.3	87.5
예술분야	92.1	90.2	86.0
화폐단위	90.7	86.8	88.7
장 소	83.6	87.9	85.7
시 간	81.8	90.0	85.7
평균	**84.6**	**89.0**	**86.8**

본 시스템의 성능평가 결과는 다음의 표 4와 같다. 정확률과 재현율은 반비례적인 관계에 있으며 평균적인 의미의 F-measure로 전체적인 시스템 성능을 가늠할 수 있다.

실험 결과 5개 분류의 평균 정확률은 84.6%, 재현율은 89.0% 그리고 F-measure가 86.8%로 나타났다. 이는 본 연구와 유사하다고 할 수 있는 기존 개체명 인식 관련 연구들과 비교해 봤을 때 대체로 비슷한 성능을 보임을 확인할 수 있다. 표 5에 유사 연구들과의 특징 및 성능 평가 비교표를 제시하였는데, 물론 적용 도메인 및 학습 데이터와 개념 분류의 수가 다르기 때문에 평가 결과를 단순히 수치 비교할 수는 없으나 대략적인 참고는 할 수 있다. SVM 기반의 생의학 개체명 인식을 제외하고는 거의 비슷한 성능을 보임을 확인할 수 있다. 생의학 개체명 인식의 경우 많은 분류 개수를 대상으로 시스템을 구현하여 문제의 난이도가 높음을 감안하면 결과가 그렇게 나쁜 것은 아니라고 볼 수 있다.

본 시스템은 기존 연구들에서 시도된 적이 없는 구문분석과 시소러스의 의미코드를 이용하여 의미 태깅 시스템의 성능, 특히 재현율의 향상을 꾀하였으며, 또한 웹문서를 대상으로 한 점이 기존 연구와의 차이점이라고 할 수 있다.

5. 결론 및 향후 연구과제

본 연구는 유비쿼터스 환경 구축과 맞물려 최근 이슈화되고 있는 시맨틱 웹을 실현하기 위해서 필수적으로 필요한 웹문서의 자동 의미 태깅에 관한 연구이다. 시맨틱 웹의 기본이 각 자원 및 에이전트(agent)들 간의 자유로운 의사소통이라는 점에서 각 자원들의 의미를 사람이 아닌 에이전트 즉 프로그램이 이해할 수 있도록 꼬리표를 달아주는 작업이 반드시 필요한 것이다. 따라서 현재의 웹에 존재하는 방대한 자원을 자동으로 의미를 파악하여 태깅할 수 있도록 해주는 시스템을 구축하는 연구가 필요한 것이다.

이를 위해서 본 연구는 형태소 분석과 구문 분석 같은 자연어처리 기법과 최근 좋은 성능을 입증 받고 있는 SVM 기계학습 알고리즘을 이용하여 웹문서 의미 태깅 시스템을 구현하였다. 기계학습의 학습 단계에서는 가도카와 시소러스를 이용하여 비슷한 의미

<표 5> 관련 연구들과의 특징 및 실험결과 비교

	2단계 최대 엔트로피 모델 (연세대)	SVM 생의학 개체명 인식 (고려대)	HMM 한국어 개체명 인식 (ETRI)	오류제어 부트스트래핑 방식 (포항공대)	본 시스템
특 징	ETRINEDIC 적용	Bi-gram, 10-fold cross validation	Tri-gram, 다단계 Back-off, 부개체형	비지도식 기계학습 적용	구문분석, 시소러스 적용
말뭉치	두산동아 백과사전 1,000문서	GENIA corpus 논문 초록 2,000문서	경제 분야 기사 100문서	New York Times 21,000문서	대표 키워드로 문장 수집 2,750문장
알고리즘	Maximum Entropy Model	Support Vector Machine	Hidden Markov Model	Bootstrapping Model	Support Vector Machine
분류 수	3개	22개	9개	12개	5개
개념 분류	인명, 지명, 기관명	DNA, RNA, atom, ...	인명, 지명, 기관명, 날짜, 시간, 금액, 퍼센트, 수량, 전화번호	person, organization, country, state, city, ...	수송수단, 예술분야, 화폐단위, 장소, 시간
정확률, 재현율, F-measure	88.3% 82.4% 85.2%	66.4% 67.0% 66.7%	84.5% 91.9% 87.6%	80.7% 93.3% 86.6%	84.6% 89.0% 86.7%

의 단어들을 하나로 묶어서 처리함으로써 재현율의 향상을 도모하였다.

실험결과, 평균적으로 정확률은 84.6%, 재현율은 89.0%로 나타났으며 F-measure는 86.8%의 성능을 보였다.

향후에는 시스템의 범용성 및 성능의 향상을 위해 보다 많은 양의 학습데이터를 수집하는 것과, 태깅 가능한 의미 분류의 개수를 늘리는 것이 필요하며, 이를 통해 시스템의 성능이 만족할 만한 수준에 도달하게 되어 온톨로지 구축, 추론엔진 적용 등의 과정을 거친다면 머지않아 일반인이 사용가능한 시맨틱 웹 환경이 구현될 것으로 기대된다.

참 고 문 헌

[1] M. Weiser, "Ubiquitous Computing," IEEE Computer, vol. 26, no. 10, pp.71-72, 1993.

[2] 김중태, 웹 2.0 시대의 기회 시맨틱 웹, 디지털미디어 리서치, 2006.

[3] T. Berners-Lee, J. Henderler, and O. Lassila, The Semantic Web, Scientific American Magazine, 2001.

[4] 김성원, 2단계 최대 엔트로피 모델을 이용한 한국어 개체명 인식, 연세대학교 대학원 석사학위 논문, 2004.

[5] 이기중, SVM기반 분리 모델을 이용한 생의학 개체명 인식, 고려대학교 대학원 석사학위 논문, 2004.

[6] 황이규, HMM에 기반한 한국어 개체명 인식, 정보처리학회, 제10-B권, 2호, pp.229-236, 2003.

[7] 이승우, 오류제어 부트스트래핑 방식의 세밀한 개체명 인식, 포항공과대학교 대학원 박사학위 논문, 2005.

[8] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[9] S. Ohno, and M. Hamanishi, New Synonyms Dictionary, Kadokawa Shoten, Tokyo, 1981.

[10] 문경희, 이종혁, 김정인, 양기주, "일-한 기계번역 시스템 : 연어 패턴을 이용한 어휘 다의성 해소", 정보과학회논문지(B), 제25권, 8호, pp.1270-1280, 1998.

[11] K. H. Moon and J. H. Lee, "Translation of Discontinuous Multi-Word Translation Units

in a Korean-to-Japanese Machine Translation System," Int'l J. of Computer Processing of Oriental Languages, vol. 15, no. 1, pp.79-99, 2002.
[12] http://www.cs.cornell.edu/People/tj/svm_light
[13] 권오욱, 정유진, 김미영, 류동원, 이문기, 이종혁, "음절단위 CYK 알고리즘에 기반한 형태소 해석기 및 품사태거", 제11회 한글 및 한국어 정보처리 학술대회 및 제 1회 형태소 분석기 및 품사태거 평가 워크숍, pp.76-88, 1999.
[14] 김미영, 이종혁, "S-절 분할을 통한 구문 분석", 정보과학회 논문지: 소프트웨어 및 응용, 제32권, 제9호, pp.936-947, 2005.
[15] B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov, KIM-Semantic Annotation Platform, Ontotext Lab, Sirma AI EOOD, 2004.
[16] http://www.ontotext.com/kim

황 운 호(Woon-Ho Hwang)

- 학생회원
- 2005년 : 대구대학교 전산 공학과 (공학사)
- 2007년 : 대구대학교 컴퓨터정보공학과 (공학석사)
- 2007년 ~ 현재 : (주)코난테크놀로지 연구원
- 관심분야 : Semantic Web, Natural Language Processing, Information Retrieval

강 신 재 (Sin-Jae Kang)

- 종신회원
- 1995년 : 경북대학교 컴퓨터공학과 (공학사)
- 1997년 : 포항공과대학교(POSTECH) 컴퓨터공학과 (공학석사)
- 2002년 : 포항공과대학교(POSTECH) 컴퓨터공학과 (공학박사)
- 1997년 ~ 1998년 : SK Telecom 정보기술연구원 주임연구원
- 2002년 ~ 현재 : 대구대학교 컴퓨터·IT공학부 조교수
- 2007년 ~ 현재 : 오스트리아 U. of Innsbruck, DERI 연구소 방문교수
- 관심분야 : Semantic Web, Social Web, Recommender System, Ontology, Natural Language Processing