Wikipedia에서 온톨로지 개념 인식을 위한 핵심어 추출

고병규*, 김판구**
*조선대학교 대학원 컴퓨터공학과
**조선대학교 컴퓨터공학부
e-mail : rhqudrb135@gmail.com

Term Extraction for Ontology Concept Recognition in Wikipedia

Byeong-Kyu Ko*, Pan-Koo Kim**
*Graduate School of Computer Engineering, Chosun University
**Division of Computer Science and Engineering, Chosun University

요 약

최근 주목받고 있는 의미적 정보처리의 지식베이스인 온톨로지는 정형화된 표현을 통해 정확한 지식 처리와 추론관계를 명시해야 하기 때문에 온톨로지 확장에 대한 중요성 역시 강조되고 있다. 온톨로지 확장을 위한 기존의 방법들은 전문가를 통한 수작업 형태이거나 보편화된 사전이나 시소러스 집단의 분석을 통한 통계의 확률분포를 이용하는 반자동화된 방법들이 있다. 이에 본 논문에서는 Wikipedia에서 특정 도메인 문서들만을 수집한 후 중요문장 추출과정을 통해 해당 문서 내의 핵심어를 파악하여 이를 온톨로지의 개념 인식을 위한 정보로 활용할 수 있는 방안을 제시하고자 한다.

1. 서론

최근 주목받고 있는 의미적 정보처리의 지식베이스인 온톨로지는 정형화된 표현을 통해 정확한 지식 처리와 추론관계를 명시해야 하기 때문에 온톨로지 확장에 대한 중요성 역시 강조되고 있다. 온톨로지 확장을 위한 기존의 방법들은 전문가를 통한 수작업 형태이거나 보편화된 사전이나 시소러스 집단의 분석을 통한 통계의 확률분포를 이용하는 반자동화된 방법들이 있다. 수작업으로 생성할 경우 컨셉 추출과 관계 생성에 대한 정확성은 뛰어나지만 많은 시간과 비용을 필요로 하며 이를 해결하기 위한 반자동화된 방법에서는 텍스트 분석 시 태깅된 단어에 대한 해석의 차이점과 개념과 관계를 추출하기 위해 보편화된 사전이나 고차원적인 학습문서에 의존하는 경향이 존재한다. 이에 본 논문에서는 Wikipedia에서 특정 도메인 문서들만을 수집한 후 중요문장 추출과정을 통해 해당 문서 내의 핵심어를 파악하여 이를 온톨로지의 개념 인식을 위한 정보로 활용할 수 있는 방안을 제시하고자 한다.

2. 관련 연구

2.1 Wikipedia 문서를 이용한 정보 추출

Wikipedia는 웹기반의 다국적 언어의 자유로운 콘텐츠를 지향하는 백과사전 프로젝트로부터 시작되어 오늘날 가장 많이 접속하는 웹사이트 중의 하나로, 특정 도메인의 개념과 관계를 추출하는데 있어서 훌륭한 대상이다.

Wikipedia 문서의 구성은 주요 기사의 제목과 이를 설명하는 Text body와 그림, 표, 목차, Reference, Category들로 이루어져 있으며 중요한 특징으로 기사내용에 대한 분류항목을 지니고 있다. 목차에서는 기사내용을 순차적으로 기술하고 있고 문서 내부에 분류항목으로 infobox나 Navbox와 같은 Box형식의 표, 그림을 통해 세부 분류와 특징을 표현하고 있다. Infobox와 Navbox는 특정양식에 따라 사용자들에 의해 작성되는 것으로 전문적인 사전이나 서적, 연구문헌, 뉴스 등을 참고하여 기재된다.

(그림 1) Wikipedia의 Infobox

2.2 Wikipedia를 이용한 온톨로지 자동 구축

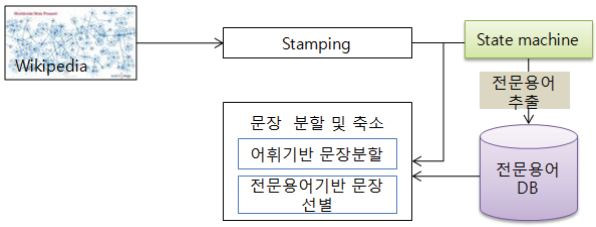
Harvesting Wiki Consensus[1]연구는 Wikipedia 자체가 온톨로지 체계를 구성하고 있음을 파악하고 Wikipedia

기반을 통해 온톨로지를 구성한다면 Wikipedia의 개체들을 재사용할 수 있기 때문에 일반 온톨로지에서 사용되는 개체의 생성 및 유지가의 어려움을 줄일 수 있다고 주장하였다. 또한 Wikipedia의 멀티미디어 개체들은 온톨로지에서의 의미명확화와 의미의 풍족화의 향상을 위해 사용할 수 있다고 주장하였다.

Robust Minimal Recursion Semantics(RMRS)[2]시스템은 Wikipedia의 biological내용으로부터 12,000개의 동물과 관련된 Wikipedia 페이지로부터 개념들의 관계를 추출하고 이를 통해 온톨로지를 구축 하였으며. [3]의 연구에서는 Wikipedia의 infobox의 class들을 SVM과 HMM등을 통해 IS-A관계 등으로 추출하고, wordnet[4]과의 매핑을 통하여 온톨로지를 구축하였다. 본 논문에서는 이와 같은 Wikipedia의 대중성과 발전성을 바탕으로 손쉽게 유지 보수가 가능한 온톨로지 확장 방법을 제안하였다.

3. Wikipedia 문서의 문장 분석 및 핵심어 추출

Wikipedia에서 도메인 관련 문서들만을 수집한 후 중요문장 추출과정을 통해 해당 문서 내에서 핵심어 처리과정을 통해 핵심어를 파악하고 핵심어를 바탕으로 한 중요문장을 추출한다. 그리고 문장의 복잡성을 증가시키는 관계어나 접속사를 어휘매칭을 기반으로 분할한다. 이를 도식화하면 다음 (그림 2)와 같다.



(그림 2) 시스템 구성도

3.1 Wikipedia 문서 전처리

본 절에서는 Wikipedia 도메인 관련 문서로부터 다양한 태그와 도표, 그림 등을 배제하고, Text 영역에 대한 데이터를 수집하는 과정을 다룬다. 문서를 처리를 하기 위한 과정은 다음과 같다.

① html 문서를 읽어온다.
② html문서에서 (p | table | from)tag를 포함하고 있는 문구를 읽어온다.
③ (div | bt | tr)tag를 삭제 후, html tag들과 유니코드, 특수문자 처리한다.
④ 각 단어들을 공백을 삭제 후, 줄바꿈 문자를 삽입한다.
⑤ 구문단위로 배열에 저장한다.

(그림 3)은 Wikipedia 본문에 대해 문장 추출과정을 통해 변환된 문서의 결과이다.

```
Further information: DNA-binding protein
Interaction of DNA with histones (shown in white, top). These
proteins' basic amino acids (below left, blue) bind to the acidic
phosphate groups on DNA (below right, red).
Structural proteins that bind DNA are well-understood examples of non
-specific DNA-protein interactions. Within chromosomes, DNA is held in
complexes with structural proteins. These proteins organize the DNA
into a compact structure called chromatin. In eukaryotes this
structure involves DNA binding to a complex of small basic proteins
called histones, while in prokaryotes multiple types of proteins are
involved.[74][75] The histones form a disk-shaped complex called a
nucleosome, which contains two complete turns of double-stranded DNA
wrapped around its surface. These non-specific interactions are formed
through basic residues in the histones making ionic bonds to the
acidic sugar-phosphate backbone of the DNA, and are therefore largely
independent of the base sequence.[76]
```

(그림 3) 태그를 삭제한 Wikipedia 본문

3.2 핵심어 추출을 위한 문장 분석

전처리 과정을 통해 Wikipedia로부터 추출된 본문들을 문장단위로 분류한다. 문장이라고 판단하는 기준은 각 paragraph의 끝인 줄 바꿈 문자가 포함된 부분과 마침표를 기준으로 판단하였다. 하지만 소수점 단위의 수치를 나타내는 표현 등을 문장으로 취급하는 것을 배제하기 위해 마침표 이후에 다른 문자가 있는지를 파악하고 공백일 경우에만 문장으로 취급하였다. 먼저 본문을 정규식 표현 과정을 거침으로서noise를 생성할 수 있는 문자열을 삭제한다. 즉 문장의 어미에 reference를 나타내는 기호인 [숫자]와 ( )나 물음표기호, 그리고 본문 추출 시 특수문자 처리 오류로 생성된 물음표 등의 기호를 모두 삭제하였다. 단 2-deoxyribose와 같이 '-'기호와 어퍼스트로피 문자는 문장의미 파악을 위해 삭제하지 않았다. (그림 4)는 Wikipedia로부터 전처리 과정을 거친 최종적인 문장 추출 결과를 보이고 있다.

```
Sentence 24: 'For instance, the largest human chromosome, chromosome number 1,
approximately 220 million base pairs long.'
Sentence 25: ' In living organisms, DNA does not usually exist as a single
molecule, but instead as a pair of molecules that are held tightly together.'
Sentence 26: ' These two long strands entwine like vines, in the shape of a
double helix.'
Sentence 27: 'The nucleotide repeats contain both the segment of the backbone
the molecule, which holds the chain together, and a base, which interacts with
the other DNA strand in the helix.'
Sentence 28: 'A base linked to a sugar is called a nucleoside and a base linke
to a sugar and one or more phosphate groups is called a nucleotide.'
Sentence 29: 'If multiple nucleotides are linked together, as in DNA, this
polymer is called a polynucleotide.'
```

(그림 4) Wikipedia 문장 추출

분할된 문장들로부터 핵심문장만을 획득하기 위해서 다음 절에서 각 문장으로부터 핵심어를 파악하게 된다.

3.3 핵심 문장 추출

핵심어를 추출하는 과정은 토큰화, 태깅, 핵심어 추출 과정으로 이루어진다. 핵심어란 주어진 도메인 안에서 의미를 가지고 있는 단어들의 집합으로 도메인 내에서 사용되는 개념을 표현하여 주제를 특성화해주는 어휘적 단위를 말한다. 이러한 핵심어는 하나의 도메인을 이해하는데 필요한 요소이기 때문에 특정 도메인에 대한 기계번역이나 정보검색을 보다 효율적이고, 정확히 수행하기 위해서 필요하다. 본 논문에서는 낮은 모호성(Ambiguity)과 높은 특정성(Specificity)을 지닌 핵심어를 이용하여 문서 내에서 중요성을 지닌 문장을 추출한다. 핵심어를 바탕으로 문

장을 추출하기 위해 가장 먼저 핵심어의 출현 형태를 분석하였다.

핵심어의 형태결합 방식은 매우 다양하다. 해당 도메인에 출현하는 대부분의 핵심어들은 약어이거나 복합명사 형태로 출현하며, 이를 분석한 결과 크게 두 가지의 결합 형태로 나눌 수 있다. 하나는 단일어절(Singleton Term) 즉, 띄어쓰기가 없는 한 어절로 나타나는 형태이고, 다른 하나는 다중어절(Multi-word Term)의 형태로 띄어쓰기가 나타나며 앞의 어절성분과 의미적으로 관련이 있는 두 어절이상으로 이루어진 복합명사이다. <표 1>은 이를 기반으로 핵심어의 출현 형태를 파악하고 있다.

< 표 1 > 핵심어 구조

Structure	example
1.Singleton Term(NN,NNS,NNP)	(chromosome, NN), (genes, NNS), (DNA, NNP), (strand, NN), (protein, NN)
2. multi-word Term(1+1, JJ + 1)	Ribonucleic acid, Nucleic Acids, Recombinant DNA, oxidative lesions

단일어절로 이루어진 전문용어들은 약어로 이루어진 경우와 일반 명사들로 이루어져있으며 약어일 경우 주로 고유명사인 NNP로 파악되는 경우와 단일명사(NN), 복수명사인 (NNS)로 파악되었으며, < 표 1 >의 2번과 같이 명사와 명사 또는 수식어(JJ)와 명사의 결합으로 전문용어의 완전한 표현이나 복합어를 통한 새로운 용어들이 이에 해당한다. 문장 추출을 위한 알고리즘은 다음과 같다.

① 핵심어 데이터베이스로부터 핵심어 선택.
② 문장 내에서 핵심어와 일치하는 단어 파악.
③ 일치 시 중복 문장의 여부 파악.
④ 일치하지 않는 경우, 다음문장 탐색.
⑤ 저장된 문장번호와 일치하는 핵심문장 추출.

4. 온톨로지 개념 인식을 위한 핵심어 추출

4.1 문장 분할

추출된 핵심문장의 길이를 조절함으로서 문장이 포함하고 있는 복잡성을 줄이는 과정이 필요하다. 문장 분할의 기준이 되는 어휘들을 파악하고 이를 바탕으로 문장 내에서 분할 기준어휘 이전의 명사구를 새로운 문장의 주어로 판단한다. 출현 빈도가 높은 어휘들을 파악한 결과, 분할의 기준이 되는 단어는 관계대명사, 접속사 등이 있다. 특정 어휘들이 발생했을 때 판단할 수 있는 패턴은 < 표 2 >와과 같다.

< 표 2 > 문장분할 패턴

패턴	경우
(1)	명사구 + 관계대명사 +verb
(2)	'comma' + 관계대명사 or (while,so)
(3)	It + verb(verb ≠ be동사)

4.2 핵심어 추출

핵심어 추출을 위해 품사 태깅된 각 단어들을 정해진 패턴에 따라 핵심어 후보를 생성한다. 용어의 JJ(형용사) 태그와 NN(명사) 태그를 기준으로 하여 Single Term(단일어절)과 Multi-word Term(복합어절)을 생성한다. 핵심어 추출을 위한 알고리즘은 다음과 같다.

① 태그가 NN 또는 JJ로 시작하는지를 파악.
② NN으로 시작할 경우 핵심어 후보에 추가한 후 Multi-word 탐색 시작.
③ Multi-word 탐색 중이며, 다음 태그가 NN일 경우는 현재 단어를 핵심어 후보에 추가한 후 현재까지의 용어를 저장하고 다음 단어의 tag를 파악.
④ ③의 과정 후 현재 단어의 태그가 NN이 아닐 경우 이전까지의 용어가 2어절 이상인지를 파악하고 핵심어 후보에 추가한다. NN일 경우는 ③의 과정 반복.
⑤ ①의 과정에서 JJ로 시작할 경우 현재 단어를 저장 후 multi-word 탐색.
⑥ JJ이후 단어의 tag가 NN일 경우 현재까지 용어를 핵심어로 추가하고 저장한 후 다음 단어 파악.

5. 실험 및 평가

Single Term인 경우 발생 빈도수가 4회 이상인 경우를 핵심어로 판단하였고, Multi-word인 경우에는 1어절 이상이며 4어절 이하로 구성된 Multi-word인 경우만 핵심어로 판단하였다. Multi-word Term의 제한은 태깅의 오류와 본문추출 과정에서 Wikipedia에 존재하는 표, 그림에 대한 주석들, 고유명사로 태깅된 목차, Reference의 고유명사들이 noise를 발생시켰다고 판단하여 다수의 어절을 가지는 Multi-word는 제외하였다.

< 표 3 > Single Term

Term	freq	Multi number
DNA	269	1
base	73	1
strand	69	1
sequence	69	1
protein	48	1
information	41	1
RNA	40	1
gene	32	1
structure	31	1
chromosome	29	1
enzyme	26	1
helix	25	1

본 논문에서는 Wikipedia의 DNA 관련 문서 중 300개의 문장에 대한 Single Term에 대해서 필터링을 적용한 후 총 850개의 핵심어 중 빈도수 내림차순에 의한 상위 14개의 결과를 보이고 있다. < 표 4 >는 빈도수 내림차순으로 정리한 multi-word Term에 대한 결과이다.

< 표 4 > Multi-word Term

Term	freq	Multi number
double helix	10	2
DNA replication	10	2
hydrogen bonds	9	2
genetic information	8	2
DNA strands	7	2
DNA sequence	7	2
base pairs	6	2
transcription factors	5	2
DNA nanotechnology	4	2
DNA-binding proteins	4	2

결과의 성능평가를 위해 Wikipedia의 Term Extraction 문서에 기술된 External link로 등록되어 있는 Translated LAB의 Term Extraction 도구와 비교평가를 수행하였다.

< 표 5 > Term Finder와의 비교

Term Finder	핵심어
hydrogen peroxide produce	hydrogen peroxide
including dna replication	dna replication
lambda repressor helix-turn-helix transcription	O
repressor helix-turn-helix transcription factor	O
regulating gene expression	gene expression
dna supercoil dna	O
methylated cytosines	O
codons signifying	X
artificial nucleic acid	nucleic acid
imprinting transcriptional	X
cytosine methylation	O
bind single-stranded	O
ethidium bromide	O
single-stranded telomere dna	O

6. 결론

본 논문에서는 Wikipedia에서 특정 도메인 문서들만을 수집한 후 중요문장 추출과정을 통해 해당 문서 내의 핵심어를 파악하여 이를 온톨로지의 개념 인식을 위한 정보로 활용할 수 있는 방안을 제시하였다.

제안한 방법을 이용하면 Wikipedia의 도메인 문서로부터 상하위어 및 관계를 추출할 수 있으나 이를 바로 적용하기에는 어려움이 있다. 따라서 정확성을 높이기 위해 Named Entity Recognition에 대한 연구와 Anaphora Resolution에 관한 연구가 필요하고, 향후 빠르게 변화하는 Wikipedia의 장점을 활용함으로써 지식베이스 확장과 의미적 정보처리에 큰 효과를 줄 것으로 기대한다.

Acknowledgements

"이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No.2009-0064749)."

참고문헌

[01] M Hepp, K Siorpaes, D Bachlechner. "Harvesting WikiConsensus:Using Wikipedia Entries as Vocabulary for Knowledge Management", IEEE InternetComputing 2007.

[02] A.Herbelotand A, Copestake. "Acquiring ontological relationships from Wikipedia using rmrs". Proceedings of Workshop on Web content Mining with Human Language Technologies, ISWC062006.

[03] F Wu, D Weld. "Automatically refining the Wikipedia infobox Ontology". portal.acm.org2008.

[04] http://wordnet.princeton.edu/.

[05] 황금하, 신지애, 최기선, "개념 및 관계 분류를 통한 분야 온톨로지 구축", 정보과학회논문지: 소프트웨어 및 응용 제 35 권 제 9호, 2008.9

[06] Feiyu Xu, Daniela Kurz, Jakub Piskorski, "Term Extraction and Mining of Term Relations from Unrestricted Texts in the Financial Domain, proceedings of BIS 2002, poznan, poland.

[07] 류법모, 최기선, "텍스트에서 IS-A 관계의 자동 추출 및 순위화", 2007년도 제 19 회 한글 및 한국어 정보처리 학술대회

[08] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, D. Mladenic. Triplet Extraction from Sentences. Ljubljana: 2007. Proceedings of the 10th International Multiconference "Information Society-IS 2007". Vol. A, pp. 218 - 222.

[09] K. Englmeier, F. Murtaghì, J. Mothe, Domain Ontology: Automatically Extracting and Structuring Community Language from Texts, IADIS Applied Computing, Spain, Espagne, 2007