두문자어 의미 태깅 방법

황명권*, 정도헌*#, 성원경*
*한국과학기술정보연구원 정보기술연구실, #교신저자
e-mail:{mgh, heon, wksung}@kisti.re.kr

A Method for Acronym Sense Tagging

Myung-Gwon Hwang, Do-Heon Jeong, Won-Kyung Sung
Korea Institute of Science and Technology Information

요 약

본 논문은 의미적 정보처리에서 걸림돌이 되는 두문자어(Acronym)의 의미처리를 위한 전체적인 구조설계를 포함하고 있다. 두문자어는 일반적으로 복합어에서 의미가 큰 단어의 첫 번째 문자들로 구성된다. 두문자어를 구성하는 복합어는 다른 일반 명사들과 달리 대부분 고유한 의미를 갖고 있기 때문에 정보처리에서 의미 파악의 핵심적인 역할을 수행할 수 있다. 본 논문은 문서에서 출현하는 두문자어의 정확한 의미를 판단하기 위한 방법을 제안하며 현재까지 진행된 결과에 대해 언급하도록 한다.

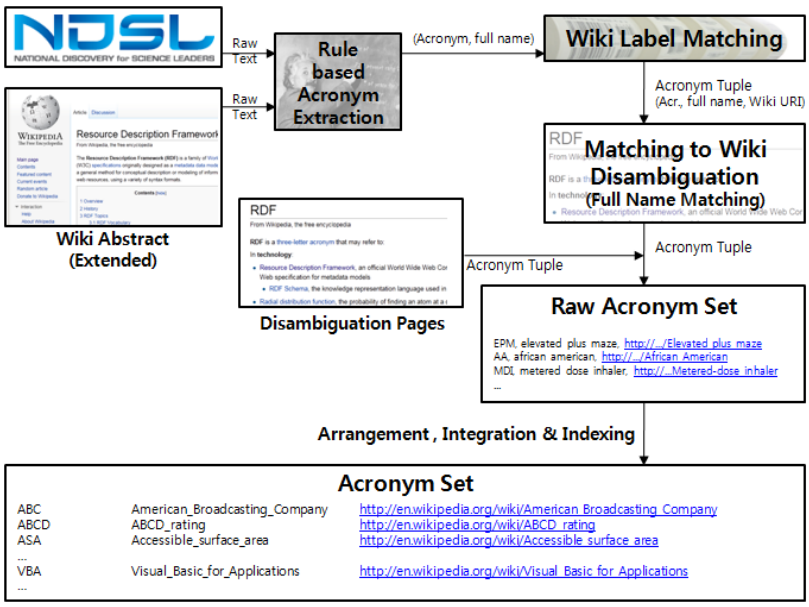
1. 서론

최근 의미적 정보처리를 통하여 비구조적인 문서로부터 지식을 습득하기 위한 연구들이 늘어나고 있다. 의미적 정보처리는 온톨로지(Ontology) 형태의 다양한 지식베이스(Knowledge Base)를 기반으로 하는데, 프린스턴 대학의 워드넷(WordNet)[1]이 가장 많이 활용되고 있다.[1] 워드넷은 일상생활에 사용되는 어휘들을 대부분 정의하고 있지만 전문용어나 두문자어에 대해서는 극히 일부만을 포함한다. 이에 전문용어(Technical Term 또는 Terminology), 개체명(Named Entity) 그리고 두문자어에 대해서는 처리하지 못하는 한계점이 존재하였다. 이를 극복하기 위해 도메인 문서집합으로부터 전문용어들을 식별하고 정의하기 위한 연구[2], 고유명사에 의미성을 부여하기 위해 관련된 어휘집합을 추출하기 위한 연구[3], 위키피디아(Wikipedia, 이하 위키)[2]의 문서제목을 개념화함으로써 워드넷의 특정 개념과 관계형성을 위한 연구[4], 문서 집합에서 자동으로 두문자어와 그에 해당하는 명사구를 찾기 위한 연구[5, 6] 등이 진행되었다. 본 연구 또한 이러한 연구 중의 하나로 두문자어를 자동으로 식별하고 그 두문자어의 정확한 명사구 파악 및 의미까지 판단하기 위한 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장은 전체 과정에 대해 상세히 설명하고, 현재까지 진행된 결과를 기술한다. 그리고 3장에서 본 논문의 결과와 향후 진행될 연구에 대해 제시한다.

2. 두문자어를 위한 의미 태깅 방법

문서에 출현하는 두문자어의 의미 태깅은 크게 두문자어 집합을 형성하는 과정과 형성된 두문자어 집합을 이용하여 의미를 태깅하는 과정으로 나누어진다. 본 장에서는 각 과정에 대해 상세히 설명하며, 현재까지 진행된 결과를 보이도록 한다.



(그림 1) 두문자어 집합 구성 과정

1) WordNet(A Lexical Database for English): http://wordnet.princeton.edu/
2) Wikipedia(The Free Encyclopedia): http://en.wikipedia.org/wiki/Main_Page

<표 1> 추출 결과

원천 문서 집합	(C) 규칙기반 추출	(D) 위키 레이블 매칭	(E)위키 구별집합 매칭	(G) 임시 집합	(H) 두문자-용어 집합
NDSL	565,401	60,991	64,266	64,266	133,751
Wiki Abstract	31,721	31,721	31,730	31,730	
Wiki Disambiguation	-	-	-	69,371	
Total	597,122	92,712	95,996	165,367	

2-1. 두문자어 집합 형성

두문자어의 의미 태깅을 위해서는 두문자어(예. HAC)가 대표하는 전체 용어(예. Herefordshire Aero Club, High Acid Crude oil 등) 집합을 형성해야 한다. 이후부터 두문자어와 전체 용어 쌍을 두문자-용어 쌍이라 하겠다. 두문자-용어 쌍 형성은 전문가에 의해 정의되는 것이 가장 정확하지만, 이는 많은 시간, 노력, 비용 및 의견 충돌을 유발할 수 있다. 이에 본 연구에서는 국가과학기술정보센터(NDSL)[3]가 보유하고 있는 논문 집합과 위키의 문서 집합을 이용하는 자동화된 방법을 제안한다. 두가지 문서집합에서 두문자-용어 쌍을 추출 및 형성하기 위한 전체 과정은 (그림 1)과 같으며, 다음은 각 과정을 구체적으로 기술한 것이다.

A. NDSL 논문 집합: NDSL은 국내외에 발표 및 게재된 논문들을 수집, 검색, 배포 등의 서비스를 수행하는 곳으로 한국과학기술정보연구원(KISTI)[4]에서 관리하고 있다. 본 연구에서는 NDSL에 보관된 영미권 논문의 초록 정보를 이용한다.

B. 위키 문서 집합: 위키의 문서들은 세계인들의 집단지성으로 만들어지고 지속적으로 정제된 고급 정보라 할 수 있다. 본 연구를 위해 위키의 초록(Abstract) 정보를 활용한다.

C. 규칙기반 두문자-용어 쌍 추출: 두문자-용어 쌍 추출의 가장 기본적인 형태는 HAC(Herefordshire Aero Club)와 같이 용어가 두문자 이후의 괄호 내에 포함된 것이다. 이 규칙을 이용하여 NDSL과 위키에서 두문자-용어 쌍을 추출한다.

D. 위키 레이블(Label) 매칭: C의 규칙을 이용하여 추출된 두문자-용어 쌍에서 용어(두문자어를 표현하는 전체 용어)는 작성자의 편의에 따라 임의로 작성된 형태가 많다. 이러한 용어들이 개념으로써의 가치가 있는지를 판단하기 위해 위키의 레이블 데이터에 매칭한다.

E. 위키 구별(Disambiguation) 집합 매칭: 두문자-용어 쌍에서 용어는 다양한 의미를 가질 수 있다. 예를 들어, EW로 이용되는 용어 'East West'는 위키에서 총 9개의 의미로 표현된다. 이러한 의미의 다양성을 확보하기 위해 위키 구별 집합으로의 매칭을 이용한다.

F. 위키 구별 페이지: C과정의 규칙기반에서 간과한 두문자-용어 쌍들이 존재한다. 예를 들어, OWL(Web Ontology Language), VBA(Visual Basic for Applications) 등의 형태가 자주 표현되지만, C과정에서는 이러한 형태를 추출하지 않는다. 두문자 커버리지를 높이기 위해 위키 구별 페이지를 활용한다.

G. 임시 집합(Raw Acronym Set): 앞에서 활용한 NDSL, 위키, 그리고 위키 구별 페이지에서 추출한 두문자-용어 집합을 모두 합쳐놓은 집합이다.

H. 두문자-용어 집합(Acronym Set): 임시 집합 F에서 중복된 것을 제거하고, 알파벳 순서로 정렬된 두문자-용어 집합을 저장한다.

본 과정을 통해 추출된 두문자-용어 집합은 다음 단계인 의미 태깅에서 활용되며, 현재까지 진행된 단계도 여기까지이다. <표 1>은 추출된 집합의 단계별 통계치를 보이고 있으며, 최종적으로 13만 여개의 두문자-용어 집합을 형성할 수 있었다.

2-2. 두문자어 의미 태깅 방법

일반 문서에 출현하는 두문자의 의미 태깅을 위해 앞의 과정에서 형성한 두문자-용어 집합을 이용한다. 이 방법은 현재 구조 설계를 완성하여 그 결과를 보기 위한 진행중에 있다. 두문자어 의미 태깅에 대한 전체 구조는 (그림 2)에서 보이며 본 장에서는 각 모듈을 설명한다.

A. 두문자-용어 집합(Acronym Set): 2-1에서 수행하여 최종적으로 추출된 두문자-용어 집합을 의미한다.

B. 위키 초록(Abstract): 어떤 문서에 출현한 두문자어의 전체 용어와 그 의미를 태깅하기 위해, A에 저장된 두문자-용어의 근거 데이터(문맥 정보)가 필요하다. 이 데이터를 추출하기 위해 본 연구에서는 위키 초록 집합을 이용한다.

C. 문맥 정보 추출기: B의 위키 초록에서 명사 유형을 추출하는 모듈이며, A 데이터에서 각각의 두문자-용어에 대한 문맥 정보를 확보한다.

D. 문맥 정보: A에 포함된 각 두문자-용어에 대한 문맥 정보를 저장하는 저장소이다.

E. 웹 문서, 두문자 탐지기, 의미 태깅: 최종적으로 확보된 두문자-용어의 문맥 정보를 이용하여 웹 문서에

3) 국가과학기술정보센터(NDSL, National Discovery for Science Leader): http://www.ndsl.kr/index.do

4) 한국과학기술정보연구원(Korea Institute of Science and Technology Information): https://www.kisti.re.kr/

서 출현하는 두문자를 추출한다. 이때, 웹 문서의 두문자 주변에 있는 정보와 D에 저장된 문맥 정보 사이의 유사도를 측정함으로써 가장 유력한 의미를 태깅한다.

위의 과정에서 의미 태깅을 위해 웹 문서의 두문자 주변 정보와 태거의 문맥 정보 사이의 유사도 측정 방법이 아주 중요하다. 현재는 VSM(Vector Space Model), LSA(Latent Semantic Analysis), 지식기반 방법 등을 상호 비교하고 있으며 그 결과는 차후에 보이도록 하겠다.

3. 결론 및 향후 연구

본 논문에서는 입력된 문서에서 두문자를 파악하고, 두문자가 표현하는 용어와 의미를 태깅하기 위한 방법을 제안하고 있다. 이를 위한 전처리 과정으로 NDSL 논문과 위키 문서를 이용하여 두문자-용어 집합을 형성하였고, 위키의 초록 정보를 이용하여 각 용어에 해당하는 문맥 정보를 추출하였다. 본 과정을 통해 13만 여개의 두문자-용어 집합을 형성하였으며, 현재는 문서에 출현하는 두문자의 의미 태깅을 위한 방법들을 분석중에 있다.

본 연구는 아직 완성되지 않았지만, 현재까지 추출된 데이터만으로도 충분한 가치를 지닐 수 있다. 다양한 연구자들에 의해 수행된 유사한 연구들은 두문자와 용어의 매칭에만 집중하거나 추출된 용어들의 다양성 및 신뢰성 검증이 부족한 반면, 본 연구는 대용량의 문서 집합을 이용하며 정확한 URI 부여 및 각 두문자-용어 쌍의 문맥정보까지 겸비하고 있기 때문이다. 지속적인 연구를 통해 얻어지는 결과는 향후에 보이도록 하겠다.

참고문헌

[1] Fellbaum, C., “WordNet: An Electronic Lexical Database,” MIT Press.

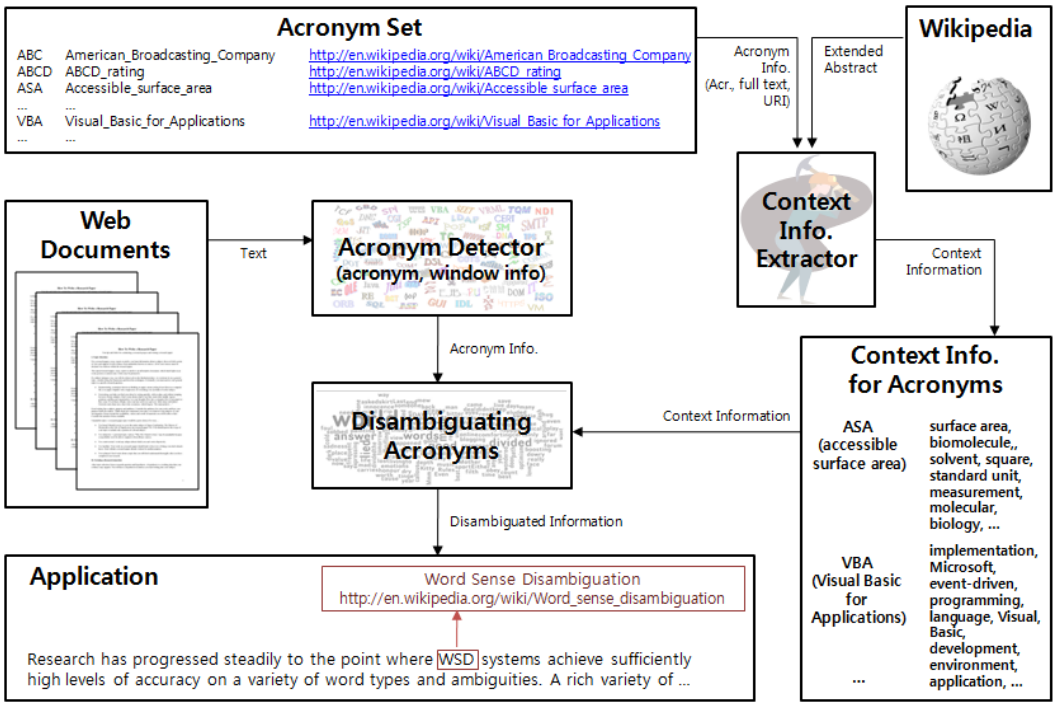
[2] Velardi, P., Cucchiarelli, A., and Petit, M., “A Taxonomy Learning Method and Its Application to Characterize a Scientific Web Community,” IEEE Transactions on Knowledge and Data Engineering, 19(2), pp. 180-191, Feb., 2007.

[3] Hwang, M.G. and Kim, P.K., “A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary,” International Journal on Semantic Web & Information Systems. 5(1), pp. 48-64, Jan.-Mar., 2009.

[4] Hwang, M.G, Choi, D.J. and Kim, P.K., ″A Method for Knowledge Base Enrichment using Wikipedia Document Information,″ Information - An International Interdisciplinary Journal, 13(5), pp. 1599-1612, Sep., 2010.

[5] Ji, X., Xu, G., Bailey, J. and Li, H, ″Mining, Ranking, and Using Acronym Patterns,″ Lecture Notes in Computer Science, Vol. 4976, pp. 371-382, 2008.

[6] Osiek, B.A., Xexéo, G. and de Carvalho, L.A.V., ″A Language-Independent Acronym Extraction From Biomedical Texts With Hidden Markov Models,″ IEEE Transactions on Biomedical Engineering, 57(11), pp. 2677-2688, Nov., 2010.



(그림 2) 두문자어 의미 태깅 방법