패턴생성을 통한 인터넷 문서의 한글-영문용어 추출[1]

강재호+, 김종성++ , 류광렬++
+동아대학교 지능형통합항만관리연구센터, ++부산대학교 정보컴퓨터공학부
{jhkang, kimjs1 , krryu}@pusan.ac.kr

Mining Korean-English Terminologies by Pattern Generation in Internet

Jaeho Kang+, Jongsung Kim++ and Kwang Ryel Ryu++
+Center for Intelligent & Integrated Port Management Systems, Dong-A University
++Division of Computer Science and Engineering, Pusan National University

요 약

전문용어의 가짓수가 많고 생성빈도 또한 높은 분야에서 고품질의 정보검색과 기계번역 결과를 얻기 위해서는 상당 분량의 번역용어사전의 확보가 필수적이다. 이러한 분야에서 번역용어사전을 수작업으로 구축하는 것은 큰 부담이 된다. 본 논문에서는 이미 알고 있는 용어(원어)와 번역용어를 말뭉치에서 함께 표기한 부분을 찾아 패턴화하는 작업과, 생성된 패턴으로 추가의 용어-번역용어를 추출하는 작업을 반복하여 수행함으로써 번역용어사전을 자동으로 구축하는 방안을 제안한다. 인터넷 문서를 대상으로 본 제안방법을 적용해 본 결과 상당분량의 유효한 한글-영문용어들을 추출할 수 있었다.

1. 서론

전문용어의 가짓수가 많고 생성빈도 또한 높은 분야에서 고품질의 정보검색과 기계번역 결과를 얻기 위해서는 상당 분량의 번역용어사전의 확보가 필수적이다[1][2]. 예를 들어 정보검색 결과문서들을 군집화하여 사용자에게 제시하고자 할 때 번역용어사전이 구축되어 있지 않다면 '인공지능'에 대한 검색결과로 'Artificial Intelligence'라는 영문용어를 중심으로 하는 군집이 생성될 수 있다[2]. 하지만 이러한 번역용어사전을 광범위하게 수작업으로 구축하는 것은 큰 부담이 된다. 본 논문에서는 이미 알고 있는 용어(원어)와 번역용어를 말뭉치에서 함께 표기한 부분을 찾아 패턴화하는 작업과, 생성된 패턴으로 추가의 용어-번역용어를 추출하는 작업을 반복하여 수행함으로써 번역용어사전을 자동으로 구축하는 방안을 제안한다.

본 연구와 관련된 기존연구로 괄호 안에 포함된 영문에 대응하는 한글대역어구(한글번역어구)의 범위를 인식하는 문제에 관한 연구가 있었다[3]. 이 연구에서 한글대역어구의 범위를 보다 정확히 파악하기 위하여 음운유사도와 복합어를 포함한 대역어 부분일치 방안을 병행하여 적용하였다. 본 논문에서 제안하는 방법이 기계번역의 측면에서 접근한 기존 연구와 다른 점은 정제된 말뭉치가 아닌 인터넷 상의 문서를 대상으로 하고 있다는 점과 괄호 이외에 다양한 형태로 나타날 수 있는 패턴들도 자동으로 생성하여 활용한다는 점이다. 또한, 기초사전의 도움 없이 사용자가 제시한 적은 수의 한글-영문용어 목록을 기반으로 패턴과 추가의 한글-영문용어를 생성하고 추출하는 방식은 데이터 마이닝[4]의 한 분야인 웹 마이닝[5] 측면에서의 접근이라는 데 그 차이가 있다.

이어지는 2장에서는 한글-영문 번역용어사전을 구축하는 방안을 설명하고, 3장에서는 인터넷 상의 문서를 대상으로 적용할 때의 고려사항과 생성된 한글-영문용어의 적절성을 평가하는 척도에 대하여 기술하였다. 4장에서는 본 논문에서 제시한 방법을 실험결과와 함께 분석하고 마지막 5장에서 결론과 향후 연구를 정리하였다.

2. 한글-영문용어 추출

본 장에서는 접근하고자 하는 방식을 먼저 예를 들어 설명하고자 한다. 그림 1은 사용자가 (인공지능, Artificial Intelligence)라는 한글-영문용어를 기존에 알고 있다고 할 때, 이를 질의로 하여 인터넷 문서를 검색한 결과의 일부이다. 검색결과에서 질의로 주어진 용어가 나타난 부분은 기울인 글씨로 표기하였다.

1	*인공지능* : *Artificial Intelligence* ... 따라서 인공지능은 공학적, 과학적 목표를 모두 가지고 있다 ... 인공지능의 연구 결과에 관심을 갖고 주시하는 사람들 중에는 최소한 네 가지 ... 인공지능이 직접적 연관성을 가지고 있는 또 하나의 분야는 심리학 (Psychology) 이다. ...
2	머신비전, 신경망에 관련된 종합정보 사이트 ... 머신비전, 신경망에 관련된 종합정보 사이트 - http://www.milab.co.kr 머신비전(machine vision), 신경망(neural network), *인공지능(artificial intelligence)*, 유도무기 기술 (guided-weapons technologies), 패턴인식(pattern recognition), ATR(automatic target recognition ...
3	교수님 소개 ... 패턴 인식(Pattern Recognition). 자연어 처리(Natural Language Processing). 한국어 정보 처리(Korean Information Processing). *인공 지능(Artificial Intelligence)*. 학회활동. 한국정보과학회 종신회원. 한국 정보처리학회 정회원. 한국 인지과학회 정회원. ...

그림 1. '인공지능 artificial intelligence' 검색결과의 일부

각각의 문서에서 질의로 주어진 한글-영문용어가 발견된 부분을 살펴보면 한글-영문용어의 표현형태가 서로간 유사한 경우가 많음을 알 수 있다. 따라서, 이러한 한글-영문용어의 표

[1] 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적 우리말 분석기술의 개발(M10203000028-02J0000-01510))의 지원을 받아 이루어진 것임.

[2] 현재 서비스되고 있는 클러스터링에 기반한 정보검색시스템에서 이러한 결과를 발견할 수 있다.

현형태를 패턴으로 생성하여 문서에 적용한다면, 추가적인 한글-영문용어를 추출할 수 있을 것이다. 예를 들어 문서3에서 '인공지능(Artificial Intelligence)'라는 구문을 발견할 수 있는데 이를 <셈표><한글용어><작은괄호-염><영문용어><작은괄호-닫음>과 같이 패턴화시킬 수 있다. 이 패턴을 문서3에 적용하게 되면 (한국어 정보 처리, Korean Information Processing)과 같은 추가의 한글-영문용어를 추출할 수 있다.

패턴은 그 표현력에 따라 다양하게 정의가 가능한데, 본 연구에서는 <선행 단어><용어1><가운데 단어들><용어2><후행 단어> 형태로 표현하였다. 표 1에는 그림 1에서 질의로 사용한 (인공지능, Artificial Intelligence)라는 한글-영문용어가 발견된 부분을 참고하여 생성한 패턴들을 보여주고 있다. 표 2는 생성된 패턴을 그림 1의 문서에 적용하여 추출할 수 있는 한글-영문용어들을 나열하였다.

표 1. 검색결과에서 생성된 패턴의 예

패턴	문서	선행	용어1	가운데	용어2	후행
1	1	Ø[3]	<한글>	:	<영문>	Ø
2	2	,	<한글>	(	<영문>	)
3	3	.	<한글>	(	<영문>	)

표 2. 생성된 패턴을 적용하여 추출한 한글-영문용어의 예

패턴	문서	한글용어	영문용어
2	2	신경망	Neural Network
2	2	유도무기 기술	Guided-Weapons Technologies
2	2	패턴 인식	Pattern Recognition
3	3	패턴 인식[4]	Pattern Recognition
3	3	자연어 처리	Natural Language Processing
3	3	한국어 정보 처리	Korean Information Processing

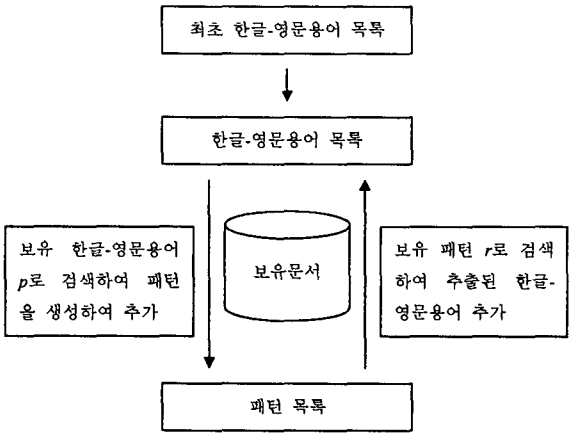

그림 2. 한글-영문용어 추출 흐름

[3] 제목이나 문서의 시작 또는 끝부분에는 <선행 단어>나 <후행 단어>가 없는 경우가 있다. 이러한 경우 해당 단어를 Ø로 표기하였다.
[4] 한글의 경우 '패턴 인식'과 '패턴인식'처럼 띄어쓰기가 다른 용어는 동일한 용어로 취급하였다. 영문은 대소문자 구분을 하지 않았다.

패턴을 생성하는 작업과 생성된 패턴으로 한글-영문용어를 추출하는 작업을 반복 수행하여, 번역용어사전을 구축하는 방안의 전반적인 작업흐름을 그림 2에 나타내었다.

3. 인터넷을 이용한 한글-영문용어 추출

앞에서 제시한 접근방법은 논문모음과 같이 정제된 말뭉치에는 그 적용이 용이하지만, 인터넷 상의 문서를 대상으로 적용하고자 하는 경우에는 다음의 두 가지 사항을 고려하여야 한다. 첫째로는 대부분의 검색시스템은 시스템 부하 등을 이유로 패턴형태의 질의를 허용하지 않는다는 점이다. 둘째로는 인터넷 문서는 오타의 발생빈도가 높은 편인데, 그 결과로 잘못된 한글-영문용어가 추출될 수 있다는 점이다.

패턴으로 검색이 허용되지 않는 경우에는 한글-영문용어 추출을 위한 문서로 인터넷 검색시스템의 검색결과를 활용할 수 있다. 구체적인 방법은 알고리즘 1에 기술하였다.

1. $P^+ \leftarrow \emptyset, R \leftarrow \emptyset, L \leftarrow \emptyset$
2. 적은 수의 한글-영문용어목록 P를 미리 제공한다.
3. P가 공집합이 아니면, 하나의 한글-영문용어 p를 골라 인터넷 검색엔진의 질의로 변환하여 검색한다. $P \leftarrow P-\{p\}$, $P^+ \leftarrow P^+ \cup \{p\}$, 검색결과 D를 검색결과의 모음 L에 추가한다.
4. D에서 p가 발견된 부분으로 찾아 패턴을 생성하고, 생성된 신규패턴은 R에 추가한다.
5. 신규패턴은 L를, 기존패턴은 D를 대상으로 패턴을 적용하여 추가의 한글-영문용어목록을 추출하고 이를 P에 추가한다.
6. 더 이상 추가의 한글-영문용어와 패턴이 발견되지 않을 때까지 3~5의 과정을 반복한다.
7. P^+의 한글-영문용어를 선호도에 따라 정렬하여 출력한다.

알고리즘 1. 인터넷을 활용한 한글-영문용어 추출 알고리즘

문서 내에서 한글-영문용어 관련 부분에 오타가 있거나, 잘못된 패턴이 있는 경우 해당 용어와 부적절한 번역용어가 짝을 이루는 결과가 발생할 수 있다. 이러한 문제를 해결하기 위해서는 추출한 한글-영문용어의 적절성을 평가할 수 있어야 한다. 본 연구에서는 표 3의 가정에 기반하여 한글-영문용어의 적절성을 추정하고자 하였다.

표 3. 적절한 한글-영문용어에 대한 가정

1. 빈번하게 발생한다.
2. 다양한 패턴에 의하여 지원된다.
3. 각각의 용어는 연관된 번역용어 중에서 비중이 높다.

위의 가정에 기반하여 한글-영문용어 p_m의 적절성 $w(p_m)$은 수식 1과 같이 추정할 수 있다. N은 발견된 한글-영문용어의 총 횟수이며, support(r_i, p_m)은 용어 k_m과 e_m으로 이루어진 한글-영문용어 p_m이 패턴 r_i에 의하여 발견된 횟수이다. $P(p_m|k_m)$은 k_m이 포함된 한글-영문용어 중에서 p_m이 발생한 비율이다.

$$w(p_m) = P(p_m \mid k_m) P(p_m \mid e_m) \times \left(-\sum_{r_i \in R} \frac{\text{support}(r_i, p_*)}{N} \log_2 \left(\frac{\text{support}(r_i, p_m)}{N} \right) \right)$$

수식 1. 한글-영문용어의 적절성 추정

4. 실험결과 및 분석

이상에서 제안한 접근 방법의 유효성을 확인하기 위하여 다음과 같은 조건으로 실험하였다. 먼저 최초 한글-영문용어 목록으로 하나의 용어 (기계학습, Machine Learning)을 사용하였다. 인접한 한글과 영문은 띄어 쓴 것으로 간주하였고 괄호, 콜론과 같은 기호는 한 글자씩을 단어로 취급하였다. <선행 단어>와 <후행 단어>는 최대 1단어로 제한하였으며, <가운데 단어들>의 수는 최소 1단어, 최대 3단어로 제약하였다. <영문용어>의 좌우로는 영문이 나타날 수 없으며, <한글용어>와 <영문용어>는 각각 최소 2자, 5자 이상으로 제한을 두었다. <한글용어>와 <영문용어>는 최대 3단어까지의 개수차이를 허용하였다. 인터넷 검색 시스템으로는 Google[6]을 이용하였다. 검색결과의 상위 100건 문서까지 패턴 생성과 용어추출을 위한 대상 문서로 활용하되 개별 페이지에 직접 접근하지 않고 검색시스템이 제공한 검색결과 페이지만을 이용하였다.

표 4에는 총 3,300번의 한글-영문용어 검색을 수행하여 추출한 한글-영문용어들을 수식 1의 적절성 평가 기준에 따라 정렬한 결과의 일부분을 보여주고 있다. 전문용어 이외에도 일반적인 용어나 회사 또는 인물에 대한 고유명사, 영문약어 등도 상당수 추출되었다. 표 5에는 생성된 총 1,840개의 패턴들을 수식 1에서 $P(p_m|k_m)$항과 $P(p_m|e_m)$항을 제외한 평가기준 즉, 다양한 한글-영문용어를 지원하는 패턴을 우선시하여 적절성을 평가한 결과의 일부분을 보여주고 있다. 이 중에서 작은 괄호를 포함한 패턴은 1,495개였다.

표 4. 인터넷 문서에서 추출한 한글-영문용어의 예

적절성순위	한글용어	영문용어
1	패러다임	Paradigm
2	야마하	Yamaha
4	인공지능	Artificial Intelligence
5	다형성	Polymorphism
9	한국과학기술정보연구원	KISTI
10	시스코 시스템즈	Cisco Systems
11	자막[5]	English
26	하이데거	Heidegger
604	최저임금제	Minimum-Wage Laws

표 5. 인터넷 문서에서 생성된 패턴의 예

적절성순위	선행	용어1	가운데	용어2	후행
1	Ø	<한글>	(	<영문>	)
2	.	<한글>	(	<영문>	)
6	)	<한글>	(	<영문>	)
7	의	<한글>	(	<영문>	)
18	Ø	<영문>	-	<한글>	Ø
19	(	<영문>	):	<한글>	Ø
47	Ø	<영문>	/	<한글>	Ø
54	Ø	<한글>	[	<영문>	]
157	새로운	<한글>	(	<영문>	)

그림 3은 추출한 한글-영문용어들이 유효한지 자세히 평가하기 위하여 시스템이 제시한 상위 1,000개의 한글-영문용어들을 대상으로 실제 일치 여부를 수작업으로 확인한 결과이다[6]. 최초 한글-영문용어를 단 하나만 제시하였으며 사전과 같은 언어적 정보는 사용하지 않는다는 점을 감안한다면 1,000개의 용어 중 85% 이상의 정확도는 상당한 것으로 평가할 수 있다. 순위가 내려갈수록 정확도는 조금씩 저하되는데 이는 <한글용어><나누기기호><영문용어>와 같이 발생빈도가 높으면서 오류발생 가능성 또한 높은 패턴에 의한 것으로 추정된다.

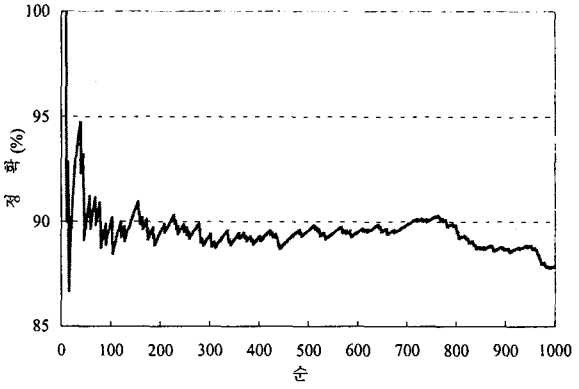

그림 3. 추출한 한글-영문용어들의 적절성에 따른 정확도

5. 결론 및 향후 연구

이상으로 적은 수의 한글-영문용어에 기초하여 유효한 한글-영문용어를 다량 추출할 수 있는 가능성을 실험적으로 확인하였다. 향후 패턴의 정확도와 언어적 정보를 반영하여 한글-영문용어를 보다 면밀하게 평가할 수 있는 방안과 사용자가 특정 용어에 대한 번역용어를 문의하였을 때 이를 효율적으로 처리할 수 있는 방안에 대한 연구가 요청된다.

참고문헌

[1] Hull, D. and Grefenstette, G. "Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval". *In Proceedings of the 19th Annual international ACM SIGIR 1996*, Zurich, Switzerland, pp. 49-57, 1996.

[2] Sheridan, P. and Ballerini, J. P., "Experiments in Multilingual Information Retrieval using the SPIDER System", *In Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval*, Zurich, Switzerland, pp. 58~65.

[3] 이재성, 서영훈, "한영 혼용문에서 괄호 안 대역어구의 자동 인식," *한국정보처리학회논문지* 제9-B권 제 4호, 2002. 8, pp. 445-452.

[4] Witten, I. H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* Morgan Kaufmann, 2000.

[5] Kosala, R. and Blockeel, H., "Web mining research: A survey", *SIGKDD Explorations*, Vol. 2, Num. 1, pp. 1~15, 2000.

[6] Google, http://www.google.co.kr

[5] 이러한 한글-영문용어가 생성된 이유는 DVD 관련 사이트에서 제품 설명에 '자막 : English'와 같은 내용이 포함되기 때문이었다.

[6] 복수의 번역용어가 존재하는 경우 이들 모두 올바른 것으로 처리하였다. 이러한 예로 '클러스터'와 '군집'이 혼용되는 경우를 들 수 있다.