정보통합을 통한 생물/의학 분야 전문용어의 자동 추출

오종훈°, 최기선
한국과학기술원 전자전산학과/전문용어언어공학연구센터/언어자원은행
{rovellia°, kschoi}@world.kaist.ac.kr

Recognizing Biomedical Terminologies through Integration of Heterogeneous Information

Jong-Hoon Oh°, Key-Sun Choi
Department of EECS
Korea Advanced Institute of Science and Technology/KORTERM/BOLA

요 약

전문용어란 전문분야의 개념이 언어적으로 표현된 형태이다. 전문분야마다 분야 특성적인 개념이 사용되므로, 전문용어는 전문분야를 특성화하는 단위로 사용된다. 따라서 전문분야문서에 대한 자연언어처리에서 전문용어를 효과적으로 처리하는 것은 매우 중요하다. 전문용어 추출은 분야 특성적인 전문용어를 해당 분야 문서에서 파악하는 작업을 말한다. 본 논문에서는 기계학습방법을 이용한 전문용어 자동 추출 기법을 제안한다. 본 논문의 기법은 전문분야 사전과 전문분야 문서를 이용하여 문서에서 나타나는 전문용어의 특성을 파악하고 이를 이용하여 전문용어를 추출한다. 본 논문의 기법은 GENIA 2.01 문서에 대하여 86%의 정확률과 90%의 재현율을 나타내었다. 또한 기존연구보다 최고 21%의 성능향상을 나타내었다.

1. 서론

전문용어란 전문분야의 개념을 지칭하는 언어적 표현이다. 전문분야마다 분야 특성적인 개념이 사용되므로, 전문용어는 전문분야를 특성화하는 단위로 사용된다. 따라서 전문분야문서를 처리할 때 전문용어를 효과적으로 인식하고 처리하는 것은 매우 중요하다. 하지만 전문용어는 계속적으로 생성되는 특성이 있기 때문에 분야 사전에만 의존적인 전문용어의 인식에는 그 한계가 있다. 이러한 이유로 최근 전문용어 추출에 대한 연구가 활발히 진행되고 있다 [1,2,3,4].

전문용어 추출이란 분야 특성적인 전문용어를 해당 분야 문서에서 파악하는 작업으로 정의된다. 일반적으로 전문용어추출은 문서에서 전문용어가 될 수 있는 언어적 단위를 파악하는 **전문용어후보추출** 과정과 전문용어후보 중 올바른 전문용어를 파악하는 **전문용어 파악** 과정으로 구성된다. 전문용어후보 추출 과정은 주로 구문 규칙이나 용어의 형성 패턴 등의 언어적 지식을 이용하기 때문에 **'언어적 필터링** (linguistic filtering)' 과정이라 한다. 반면 전문용어 파악 과정은 문서 내에서의 전문용어후보의 빈도수, 문맥정보, 의미정보 등과 같은 통계적 지식을 이용하기 때문에 **'통계적 필터링** (statistical filtering)' 과정이라 한다.

전문용어 추출 시스템은 시스템이 문서로부터 추출한 전문용어 중 문서에서 나타나는 적합한 전문용어를 얼마나 포함하고 있는지를 나타내는 재현율 (recall)과 추출 결과로 제시된 전문용어 중 적합한 전문용어의 비율이 얼마나 되는지를 나타내는 정확률 (precision)으로 평가할 수 있다. 효과적인 전문용어 추출 시스템은 문서에 나타나는 가능한 많은 전문용어를 (높은 재현율) 정확하게 제시해주는 (높은 정확률) 시스템이다. 이러한 측면에서 언어적 필터링 과정은 높은 재현율을 위한 과정이며, 통계적 필터링 과정은 높은 정확률을 위한 과정이다.

최근의 전문용어 추출 연구들[1,2,3,4]에서는 언어적 필터링 과정보다는 통계적 필터링 과정의 성능을 향상시키기 위한 연구가 활발히 진행되어 왔다. 이들 연구를 **통계적 전문용어 추출 기법**이라고 한다. 통계적 전문용어 추출 기법에서는 전문용어의 대부분을 차지하는 명사구만을 전문용어후보로 추출한 후, 다양한 통계적 기법을 이용하여 비교적 높은 정확률을 가지는 전문용어추출 결과를 제시하였다. 본 논문에서는 기존의 통계적 전문용어 추출 기법의 한계를 극복하기 위한 새로운 통계적 전문용어 추출 기법을 제안한다..

기존의 방법들은 크게 네 가지 한계점을 가진다. 첫째 기존의 방법은 대부분 다중단어 전문용어만을 추출 대상으로 한다. 대부분의 전문용어가 이러한 형태로 나타나지만, 하나의 단어로 구성된 전문용어도 많다. 특히, 의학분야의 경우 단일단어 전문용어도 많은 비중을 차지한다[5]. 따라서 의학분야 전문용어를 추출할 때, 다중단어 전문용어 뿐만 아니라 단일단어 전문용어도 추출 대상이 되어야 한다. 둘째, 대부분의 기존 연구들은 해당분야에 이미 존재하는 전문용어 정보를 사용하지 않고 전문용어를 추출한다. 새로운 전문용어는 기존의 전문용어를 기반으로 생성되는 경우가 많기 때문에 기존의 전문용어의 정보는 전문용어 추출에 매우 중요하다. 셋째, 기존의 연구들은 일반적인 전문용어의 특성에 기반한 점수함수로 전문용어를 추출한다. 이로 인해 전문분야마다 또는 전문분야 문서마다 전문용어가 나타나는 양상에 대한 학습이 없이 일반적인 통계적 특성을 이용하기 때문에 전문용어를 추출하는데 한계가 있다. 넷째, 기존 연구들에서는 새로운 특성을 나타내는 점수함수를 추가하기 위해 선형결합 (linear interpolation)을 이용하였다. 하지만 선형결합은 각 점수함수를 나타내는 특성간의 의존성을 고려하는데 한계가 있으며, 선형결합되는 각 점수함수에 대한 최적의 가중치를 파악하기도 어렵다. 즉, 새로운 정보에 대한 정보의 통합이 어렵다.

본 논문에서는 이러한 문제점들을 해결하기 위하여 기계학습방법을 이용한 전문용어 자동 추출 기법을 제안한다. 첫째, 다중단어 전문용어 뿐만 아니라 한단어로 구성된 전문용어를 추출 대상으로 한다. 둘째, 기존의 전문용어를 포함하는 전문분야 사전을 이용하여 문서에서 나타나는 전문용어를 파악한다. 셋째, 전문분야 문서에 나타나는 전문용어 및 일반용어의 특성을 파악하고, 학습을 통하여 전문용어를 추출한다. 넷째, 전문용어의 특성은 여러 가지 특성자질으로 표현되며, 이를 효과적으로 통합할 수 있는 기계학습기법을 이용하여 전문용어를 추출한다.

2. 정보의 통합을 통한 전문용어의 추출

2.1 시스템 구조도

제안하는 전문용어 추출 시스템의 구조는 그림 1과 같다. 전문용어추출 시스템은 **전문용어 추출**과정과 **학습을 통한 전문용어**

후보 분류과정으로 구성된다. 전문용어후보 분류는 통계적 필터링의 한 방법으로 주어진 전문용어후보가 전문용어인지 비전문용어인지를 판별하는 문제로 정의된다 전문용어 후보 분류과정은 전문용어후보의 특성자질을 추출하고, 전문용어와 일반용어가 등재된 사전을 씨앗지식으로 초기분류를 수행하는 초기화 과정과 분류결과를 이용하여 분류패턴을 구축하고 분류패턴을 이용하여 전문용어후보를 다시 분류하는 학습과정으로 구성된다. 학습과정은 더 이상 새로운 분류패턴이 나타나지 않을때까지 반복한다. 본 논문에서는 특성자질로 형태적 특성자질, 어휘/구문적 특성자질, 사전적 특성자질을 이용하며, 최대엔트로피 모델 (Maximum entropy model)을 이용하여 학습을 수행한다.

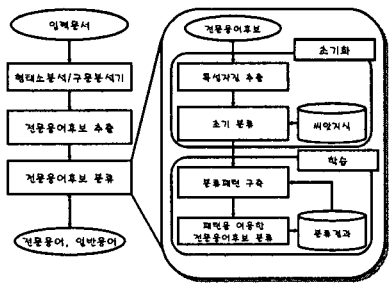

그림 1 시스템 구조도

2.2 전문용어후보의 추출

전문용어의 많은 부분은 명사구 형태로 나타나며, 이러한 특성으로 인하여 기존의 많은 연구에서는 전분분야문서에서 나타나는 명사구를 전문용어후보로 추출하였다. 본 논문에서도 전문용어후보를 명사구로 한정하고 전문용어후보를 추출한다. 본 논문에서 추출한 명사구의 형태는 다음과 같은 정규표현으로 표현된다. 정의된 정규표현으로 "EGF receptor"와 같은 명사구를 파악할 수 있다. 여기에서 BNP는 명사구를, Adj는 관형사를, Noun은 명사를 각각 나타낸다.

- `BNP := (Adj|Noun)* Noun`

2.3 전문용어후보 분류에 기반한 통계적 필터링

2.3.1 특성자질

전문용어후보에 대한 분류를 위해서는 분류패턴이 필요하다. 분류패턴은 전문용어와 일반용어의 차이를 표현해주기 위한 특성자질의 집합으로 표현되며, 본 논문에서는 형태적 특성자질, 어휘/구문적 특성자질, 사전적 특성자질을 사용하였다.

첫번째 자질인 어휘/구문적 특성자질은 어휘자질과 구문자질로 구성된다. 전문용어후보를 표현하기 위한 가장 일반적인 방법은 전문용어후보를 구성하는 중심어 어휘와 수식어 어휘를 사용하는 것이다. 만약 전문용어후보들이 같은 중심어로 표현되면, 해당 전문용어들은 유사한 의미를 가진다 (e.g. insulin *receptor* , estrogen *receptor*). 또한 같은 중심어정보를 가진 전문용어 후보들에 추가적인 수식어가 나타나면, 추가적인 수식어를 가진 전문용어후보가 다른 전문용어후보보다 전문적인 의미를 가지는 경향을 나타낸다 (e.g. estrogen receptor vs. *leucocytic* estrogen receptor). 따라서 전문용어후보를 구성하는 어휘적 특성(수식어 및 중심어)은 전문용어후보를 표현하는 중요한 특성자질로 사용될 수 있다. 의학분야에서 *bind, affect, activate*와 같은 동사는 의학분야지식을 표현하기 위해 많이 사용되므로, 해당 동사들의 논항들은 전문용어일 가능성이 높다 (e.g. "shed receptor *binds* interleukin-2", "tumour suppressor gene *affects* E2F-mediated regulation"). 본 논문에서는 전문용어후보와 주어 및 목적어 관계를 가지는 동사를 구문적 특성자질로 정의하고 이를 이용하여 전문용어후보를 표현한다.

두 번째 특성자질인 형태적 특성자질은 소문자/대문자/숫자/기호와 관련된 특성자질과 그리스/라틴어원과 관련된 특성자질로 구성된다. 의학분야 전문용어의 많은 부분을 차지하는 단백질과 유전자의 명명법으로 인하여 의학분야 전문용어는 일반용어와 구별되는 형태적 특성을 나타낸다. 단백질을 나타내는 새로운 의학분야 전문용어는 소문자로 구성된 단일단어용어보다는 대문자/소문자/숫자/기호로 구성된 단일/다중단어 전문용어의 형태로 나타난다[5, 6] (e.g., IL-1 responsive kinase). 따라서 전문용어를 구성하는 단어에 대한 소문자/대문자/숫자/기호들에 대한 형태적 정보가 일반용어와 전문용어를 구별하는 정보로 사용될 수 있다. 사용된 특성자질은 대문자/소문자/숫자와 관련된 'all uppercase' (e.g. *DNA*), 'all lowercase' (e.g. *motif*), 'alphanumeral' (e.g. *p53*), 'including capitals' (e.g. *c-Rel*), 'starting with capital' (*Egfr*), 'numeral' (e.g. *12, IV*) 등이 있으며, 기호와 관련된 'including hypen', 'including apostropy', 'including parenthesis' 등이 있다. 많은 의학분야 전문용어들은 그리스/라틴어원을 가지기 때문에 의학분야전문용어와 일반용어를 분류하는 척도로 사용될 수 있다. 그리스/라틴어원을 가진 용어는 접사 및 어근으로 구성되면 각 접사 및 어근은 고유한 개념을 표현한다 (e.g. *blast/o*와 *-blast*는 *immature cell, embryonic cell, productive cell*을 나타낸다). 따라서 같은 접사 및 어근을 가진 전문용어후보는 유사한 개념을 표현한다. 본 논문에서는 이러한 그리스/라틴어원 정보와 접사 및 어근 정보를 이용하였다.

세번째 특성자질은 사전적 특성자질이다. 주어진 전문용어후보를 구성하는 요소들에 대한 분야 정보 및 용어정보를 파악할 수 있다면, 해당 전문용어후보에 대한 분류를 보다 효과적으로 수행할 수 있다. 예를 들어, 전문용어후보의 중심어가 의학분야 전문용어일 경우 해당 전문용어후보는 전문용어가 될 가능성이 높다. 본 논문에서는 분야 및 용어 정보를 사전에서 추출하고 이를 이용하여 전문용어를 표현한다. 예를 들어 *T-cell receptor*와 *cell surface*가 의학분야사전에 존재할 경우 *cell surface T-cell receptor*에 대한 사전 특성자질은 {*BM=cell surface, BH=T-cell receptor*}로 표현된다. 여기에서 *BM*은 의학분야 용어인 수식어를 나타내며 *BH*는 의학분야 용어인 중심어를 나타낸다.

본 논문에서는 이러한 세가지 특성자질을 이용하여 분류패턴을 학습하여 전문용어후보를 분류한다.

2.3.2. 최대엔트로피모델을 이용한 전문용어후보의 분류

최대 엔트로피 모델은 비동형 정보를 효과적으로 통합하는 확률적 모델링 기법이다. 최대엔트로피 모델에서는 확률적 사건을 하위사건으로 세분화함으로써, 하위사건으로부터 도출될 수 있는 중복되는 특성자질을 이용하여 유연한 확률적 모델링이 가능하다[7]. 확률적 사건 e=<te,he>로 표현되며, 여기에서 te는 대상 사건을 나타내고, he는 문맥사건을 나타낸다. e는 특성자질 함수 $f_i(e)$의 집합으로 표현된다. 특성자질함수는 사건 e에 나타나는 특성들의 존재유무를 나타내는 함수로서 이진값을 가진다. 사건과 특성자질 함수에 의해 최대엔트로피 모델 p_M은 식 (2)와 같이 모델링될 수 있다.

$$t = \arg\max\, p_M(te \mid he) = \frac{1}{Z_{he}} \prod_i \alpha_i^{f_i(te,he)} \qquad (2)$$

$$Z_{he} = \sum_{te' \in \tau(he)} \prod_i \alpha_i^{f_i(te',he)}$$

여기에서 $\tau(he)$는 he로부터 도출할 수 있는 대상사건의 집합을 나타내며, α_i은 특성자질 함수 $f_i(e)$에 대한 가중치를, te는 전문용어 또는 일반용어라는 대상사건을 나타낸다.

최대엔트로피모델은 주어진 특성자질함수와 학습데이터를 이용하여 우도(likelihood)를 최대화하는 확률분포를 도출한다. 본 논문에서는 2.3.1절에 기술된 특성자질을 이용하여 특성자질함수를 구성하고 이를 이용하여 학습과 분류를 수행한다. 예를

들어, 어휘/구문적 특성자질 함수 중 주어진 전문용어후보 t_i의 중심어가 *gene*인지를 나타내는 특성자질 함수 $f_{LSF(SYN_HEAD,gene)}(te,he)$는 식 (3)과 같이 표현된다. 본 논문에서는 자질들에 대한 가중치를 계산하기 위하여 Maximum entropy modeling toolkit [8] 을 이용하였다.

$$f_{LSF(SYN\_HEAD,gene)}(c_i, he) = \begin{cases} 1 & if\ te = c_i\ and\ he(SYN\_HEAD(t_i)) = gene \\ 0 & otherwise \end{cases} \quad (3)$$

3. 실험

3.1 실험데이터

본 논문에서는 영어의 의학분야 전문용어를 추출하기 위하여 670개 생물의학분야 영어 논문 초록을 포함하는 GENIA 2.01 코퍼스[9]를 사용하였다. 사전적 특성자질과 전문용어후보의 초기 분류를 위한 지식으로 의학분야 전문용어사전 *UMLS Specialist lexicon*[10] 와 Brill 태거[11]의 명사사전을 사용하였다.

실험은 각 특성자질 유무에 따른 비교실험과 기존연구와의 비교 실험을 수행하였다. 각 특성자질 유무에 따른 비교실험의 평가를 위하여 정확률, 재현율, F-값을 사용한다. 기존연구와의 비교실험에서는 제안하는 전문용어추출 기법과 [1,2,4]의 기법과의 비교평가를 수행한다. 제안하는 기법과의 비교를 위하여, 본 논문에서는 제안하는 전문용어추출결과를 전문용어후보가 전문용어로 분류될 확률값 또는 유사도 등으로 순위화하여 비교 평가하였으며, 정보검색분야에서 사용하는 11포인트 평균정확률 (11pt-avg)로 평가하였다.

3.2 실험결과

3.2.1. 특성자질에 따른 전문용어 추출 성능 비교

OF	LSF	DF	정확률	재현율	F-값
√			69.89%	92.86%	79.75%
	√		81.52%	85.71%	83.56%
		√	84.62%	88.70%	86.61%
√	√		83.71%	90.67%	87.05%
√		√	84.52%	86.76%	85.63%
	√	√	85.34%	88.98%	87.12%
√	√	√	86.59%	90.24%	88.38%

표 1. 특성자질에 따른 전문용어 추출 결과

표 1은 특성자질에 따른 실험결과를 나타낸다. 표 1에서 OF는 형태적 특성자질, LSF는 어휘/구문적 특성자질, DF는 사전적 특성자질을 나타낸다. 실험은 각 특성자질을 단독으로 사용한 경우, 두 개의 특성자질을 조합하여 사용한 경우, 세 개의 특성자질을 모두 사용한 경우의 결과를 나타낸다. 실험결과에서 각 특성자질만을 이용한 결과는 87.45%로 DF가 가장 높은 성능을 나타내었으며, 79.75%로 OF가 가장 낮은 성능을 나타내었다. 따라서 전문용어후보의 분류에서 DF가 가장 효과적인 특성자질임을 알 수 있다. 두 가지 특성자질만을 사용하였을 경우에는 DF와 LSF를 사용한 경우가 87.12%로 가장 좋은 성능을 나타내었다. 그리고 모든 특성자질을 사용할 경우 가장 좋은 성능을 나타내며, 이를 통하여 세 가지 정보가 효과적으로 통합되어 전문용어후보의 분류에 사용될 수 있음을 알 수 있다.

3.2.2 기존 연구와의 비교

표 2는 기존연구와 제안한 기법의 11pt-avg의 결과를 나타낸다. 11pt-avg이 100%에 가까울수록 전문용어를 효과적으로 추출한다고 판단할 수 있다. 기존 연구의 11pt-avg은 73%~77%를 나타내며, 본 논문의 기법은 92%를 나타낸다. 이를 통해 본 논문의 기법은 기존연구보다 16~21%의 성능향상을 나타냄을 알 수 있다.

	[1]	[2]	[4]	제안기법
11pt-avg	74.6%	72.8%	76.8%	92.3%

표 2. 기존연구와의 비교실험결과

4. 결론

본 논문에서는 전문용어후보의 분류기법을 이용한 전문용어추출기법을 제안하였다. 본 논문에서는 전문용어추출에서 통계적 필터링 문제를 전문용어후보에 대한 분류문제로 변환하여 전문용어를 추출하였다. 본 논문의 기법은 전문분야 사전과 일반분야사전의 표제어를 씨앗지식으로 사용하여 문서에서 나타나는 전문용어와 일반용어의 패턴을 학습하였다. 패턴의 학습을 위하여 본 논문에서는 형태적 특성자질, 어휘/구문적 특성자질, 전문용어적 특성자질을 정의하였으며, 이를 통하여 분류기에서 사용되는 파라메터를 학습하였다. 본 논문의 기법은 단일언어 전문용어 및 다중언어 전문용어를 파악할 수 있는 알고리즘으로 86%의 정확률과 90%의 재현율을 나타내었다. 또한 기존연구보다 최고 21%의 성능향상을 나타내었다.

감사의 글

이 논문은 과학기술부, 과학재단의 지원에 의하여 이루어졌습니다.

참고문헌

[1] Justeson, J.S. and S.M. Katz (1995) Technical terminology : some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1) pp. 9-27

[2] Frantzi, K.T. and S.Ananiadou (1999) The C-value/NC-value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3) pp. 145-180

[3] Maynard D. and Sophia Ananiadou. (2000) TRUCKS: a model for automatic term recognition, Journal of Natural Language Processing, December

[4] Nakagawa, Hiroshi and Tatsunori Mori, (2003), Automatic term recognition based on statistics of compound nouns and their components, *Terminology*, 9(2), pp. 201–219

[5] Fukuda, K. and A Tamura and T Tsunoda and T Takagi (1998). Toward information extraction: identifying protein names from biological papers. *PSB98*. 707-718.

[6] Antonarakis, S.E., (1998) Recommendations for a nomenclature system for human gene mutations. *Nomenclature Working Group. Hum Mutat*, 11(1): p. 1-3.

[7] Berger A., S. Della Pietra, and V. Della Pietra. (1996) A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39-71

[8] Zhang Le. (2004), Maximum Entropy Modeling Toolkit for Python and C++, http://www.nlplab.cn/zhangle/

[9] Ohta and Yuka Tateisi and Hideki Mima and Jun'ichi Tsujii, (2002) GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain *In Proceeding of he Human Language Technology Conference*

[10] NLM, (2003), Unified Medical Knowledge System (UMLS)

[11] Brill, E. (1995), Transformation-Based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics*