KIPS Trans. Softw. and Data Eng.
Vol.7, No.8 pp.281~286 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2018.7.8.281

A Method for Extracting Relationships Between Terms Using Pattern-Based Technique

Kim Young Tae† · Kim Chi Su††

ABSTRACT

With recent increase in complexity and variety of information and massively available information, interest in and necessity of ontology has been on the rise as a method of extracting a meaningful search result from massive data. Although there have been proposed many methods of extracting the ontology from a given text of a natural language, the extraction based on most of the current methods is not consistent with the structure of the ontology. In this paper, we propose a method of automatically creating ontology by distinguishing a term needed for establishing the ontology from a text given in a specific domain and extracting various relationships between the terms based on the pattern-based method. To extract the relationship between the terms, there is proposed a method of reducing the size of a searching space by taking a matching set of patterns into account and connecting a join-set concept and a pattern array. The result is that this method reduces the size of the search space by 50-95% without removing any useful patterns from the search space.

Keywords : Ontology, Terms, Relationship, Extraction, Join-Set, Pattern

패턴 기반 기법을 사용한 용어 간 관계 추출 방법

김 영 태† · 김 치 수††

요 약

최근 정보의 복잡성과 다양성 및 방대한 양의 가용 정보가 증가함에 따라 대규모 데이터로부터 의미 있는 검색 결과를 추출하는 방법으로서 온톨로지에 대한 관심과 필요성이 증가하고 있다. 주어진 자연어 텍스트로부터 온톨로지를 추출하는 많은 방법이 제안되었지만, 현재 대부분의 방법은 온톨로지의 구조에 일치되도록 추출하지 못하는 실정이다. 본 논문에서는 온톨로지를 구축하기 위해 필요한 용어를 특정 도메인에서 주어진 텍스트와 구별하고 패턴 기반 방법을 기반으로 용어 사이의 다양한 관계를 추출하는 방법을 제안한다. 용어들 간의 관계를 추출하기 위해 일치 패턴 집합을 고려하고 조인 집합 개념과 패턴의 정렬을 연결하여 검색 공간의 크기를 줄이는 방법을 제안한다. 그 결과 이 방법이 검색 공간으로부터 유용한 어떤 패턴도 제거하지 않고 50-95% 정도로 검색 공간의 크기를 줄이는 결과를 보였다.

키워드 : 온톨로지, 용어, 관계, 추출, 조인 집합, 패턴

1. 서 론

온톨로지는 지식표현, 추론, 관계 정의를 위한 일종의 사전으로 볼 수 있으며, 일반적으로 관심 있는 특정 도메인에 대해 생성되고, 도메인의 개념에 대한 형식적인 표현이며 본질적으로 자연어와 관계가 있다[1]. 온톨로지를 특정 도메인에 적합하도록 구성하기 위해서는 수작업이 많이 필요하며, 이런 수작업 소요 비용 절감 및 효율성 향상을 위해서는 자동화 처리 방법이 필요하다. 주어진 자연어 텍스트로부터 온톨로지를 구성하는 여러 방법들이 제시되었지만, 현재까지 제시된 대부분의 방법은 온톨로지의 구조와 일치하도록 추출하지 못하고 있다. 특히 WordNet 등의 일반적인 온톨로지는 그리 효율적이지 않다. 특별한 도메인의 목적에 맞는 용어를 포함하지 않을 수도 있고, 관계 집합이 제한적인 이유로 인해 온톨로지 구축의 결과가 그리 만족스럽지 못하다. 따라서 특정 도메인에 대한 온톨로지를 효율적으로 구축하기 위한 방법으로 패턴 기반의 방법을 이용해 다양한 관계를 구별하는 방법을 고려해보는 것이 바람직하다.

본 논문에서는 자연어 텍스트로부터 온톨로지를 구성하는 데 필요한 용어를 식별하고, 패턴 기반의 방법을 이용해 용어들 간의 다양한 관계를 추출하는 방법을 제안한다. 용어들 간의 관계를 추출하기 위해 일치 패턴 집합을 고려하고, Join-set 개념과 패턴의 정렬을 연결하여 검색 공간의 크기를

† 비 회 원 : ㈜케이엔씨 기업부설연구소 부장
†† 종신회원 : 공주대학교 컴퓨터공학부 교수
Manuscript Received : February 26, 2018
First Revision : May 14, 2018
Accepted : June 5, 2018
* Corresponding Author : Kim Chi Su(cskim@kongju.ac.kr)

줄이는 방법을 제시한다. 수동으로 식별하기 어려운 관계를 찾기 위해 태그를 붙임으로써 패턴의 개발이 가능하도록 한다. 또한 Join-set 방법을 이용하여 텍스트의 특성에 기초하여 요소를 일치시키는 복잡한 규칙을 허용한다. 결과적으로 주어진 데이터로부터 패턴을 일반화하고, 잠재적으로 유용한 패턴을 유지하면서 검색 공간을 줄일 수 있는 방법을 제시한다.

2. 관련 연구

2.1 WordNet

WordNet의 목적은 인공지능 응용과 자동화된 분석을 뒷받침하기 위한 것이며, 보다 직관적으로 사용할 수 있도록 유사한 의미를 갖는 유의어 집합으로 용어를 분류하고, 계층 구조로 조직화한 시소러스와 사전의 조합을 제공한다.

WordNet을 새로운 도메인으로 확장하는 시도를 한 MedicalWordNet 프로젝트에서는 질병 명칭, 유전학 용어와 같은 의학 용어가 추가되었으며, 잘 확립되어 있고 대다수의 전문가가 동의하는 사실을 설명하기 위한 네트워크인 MedicalFactNet과 합의가 된 것은 아니지만 진단에 유용한 사실을 설명하기 위한 네트워크인 MedicalBeliefNet이 추가되었다[2]. WordNet은 명사에 대해서는 효율적이지만 동사에 유용한 정보는 많이 부족하다. 이러한 이유로 Baker는 “frame”의 용어에 대한 의미 정보를 인코딩하는 FrameNet을 제안했다. 이 프레임은 동사가 다른 용어와 상호작용하는 방법에 대한 많은 정보를 설명한다[3]. Fig. 1의 프레임은 “driving”의 동작에 관여하고, “driver”, “vehicle”, “cargo” 등의 동작을 암시하는 요소가 있음을 보여준다. 이를 통해 시스템은 동작을 보다 완벽하게 이해하고 해석할 수 있다.

```
frame(DRIVING)
inherits(TRANSPORTATION)
frame_elements(DRIVER(=MOVER), VEHICLE(=MEANS),
RIDER(s)(=MOVER(s)), CARGO(=MOVER(s))
scene(DRIVERstarts VEHICLE, DRIVERcontrols VEHICLE,
DRIVER stops VEHICLE)
```

Fig. 1. Example of Using Frame

2.2 온톨로지 응용

온톨로지에서 동의어 정보를 사용하려면 키워드를 포함하지 않는 문서, 키워드는 포함하지만 질의에 적합하지 않은 문서를 찾는 것이 가능해야 한다. Gonzalo는 문서의 인덱스를 만드는데 WordNet의 동의어 집합을 사용하여 기존 시스템 대비 48%에서 62%의 정밀도 향상을 보였다[4].

GO(Gene Ontology)는 유전자를 참조하기 위한 통합 시스템을 만들기 위해 만들어졌다. 유사한 프로젝트인 MeSH (Medical Subject Headings) 데이터베이스는 의료 분야의 모든 용어를 나열하고, 주제별 기사와 문서의 인덱스를 만들기 위해 고유 ID를 제공한다. 다른 프로젝트는 일관된 의료 정보를 구성하는 Snomed-CT로서 헬스케어 제공자간의 일관된 컴퓨터 기록 공유를 위해 설계되었다[5]. 이 데이터베이스는 후에 범위와 적용범위가 확장된 UMLS라는 메타 시소러스에 포함되었다[6].

바이오 의학 등의 전문적인 도메인에서는 많은 용어가 기존 온톨로지에 포함되어 있지 않아서 적용하고자 하는 도메인에 적합하도록 구조를 변경하거나 새로운 온톨로지를 생성하는 것이 필요하다. 그러나 두 경우 모두 관계를 추출하고, 제약사항을 정의하여 온톨로지로 구성하기 위해서는 많은 수작업이 필요하다. 따라서 다양한 도메인 범위에서의 적용과 효율성을 위해 패턴 기반 방법을 이용하여 관계를 추출하는 것이 필요하다.

2.3 시멘틱 관계 추출

기존의 관계 추출 방법은 자연어로부터 용어 식별, 분배 클러스터링, 용어 변화, 패턴기반 추출의 네 가지 유형으로 나누어 볼 수 있다. 패턴기반 추출은 컨텍스트에서 용어 쌍의 발견에 의존적이다. Hearst의 연구가 대표적인 패턴기반 추출 방법이며, 용어가 전형적인 패턴으로 서로 가까이에서 자주 발생한다는 점에서 주목할 만하다. Hearst는 Fig. 2의 상위어 관계 추출 패턴을 발견해 백과사전 텍스트에 적용했다[7]. 그러나 이 연구의 단점은 프로세스를 자동화할 방법이 없고, 다양한 패턴의 효과에 대한 비교를 제공하지 못한다는 것이다.

1. *NPh such as {NP, NP ... (or|and)} NP*
2. *such NPh as {NP, } * {(or|and)} NP*
3. *NP{, NP}*{,}? or other NPh*
4. *NPh {,} * {,}? and other NPh*
5. *NPh {,} including {NP, } * {(or|and)} NP*
6. *NPh {,} especially {NP, } * {(or|and)} NP*

Fig. 2. Patterns for Extracting Hypernym Relations

Cimiano와 Staab는 Hearst의 패턴과 자신들이 직접 만든 몇 가지 패턴을 적용하여 45.1%의 재현율과 62.3%의 정확도를 얻었다[8].

Yang은 두 용어 사이에 단어 몇 개를 포함하여 패턴을 만드는 단순한 방법을 사용했다[9]. 이 방법은 많은 패턴을 빠르게 추출할 수는 있지만 PMI(point-wise mutual information)라는 Equation (1)을 사용하기 때문에 패턴 수가 너무 많다.

$$pmi(x,y) = \log \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

Soderland는 텍스트 어휘 구조를 기반으로 하여 텍스트에서 정보를 추출하는 WHISK 시스템을 제시했다. WHISK 시스템은 와일드카드를 사용하는 기본 표현에서 출발한다[10]. 두 개의 slot을 추출하기 위한 패턴은 다음과 같다.

$$*\ (\ \text{slot1}\ )\ *\ (\ \text{slot2}\ )\ * \qquad \text{(패턴 1)}$$

이 규칙은 대부분 일치하여 의미가 없으므로 용어를 추가하고, 텍스트를 와일드카드 “*”로 대치하여 처리한다. 문장

"the price is $ 20"에서 "price"를 "slot1"로, "20"을 "slot2"로 표현하여 처리한다. 그러면 다음과 같은 (패턴 2)를 획득할 수 있다. 용어를 추가하고 와일드카드를 제거하는 과정을 반복적으로 수행하여 다양한 패턴을 얻을 수 있다.

(slot1) * "$" (slot2) * (패턴 2)

본 논문에서는 이러한 와일드카드 방법을 적용하여 단순한 패턴에서 출발하여 다양한 패턴을 생성한다.

3. 패턴 기반 관계 추출

본 논문에서는 온톨로지 구성에 필요한 용어를 식별하고, 이 용어들 간의 다양한 관계를 추출하기 위해 패턴을 생성하며, Join-set 개념과 패턴의 정렬을 이용하여 검색 공간의 크기를 줄이고자 한다. 기초 용어 쌍 집합인 base-set을 이용해 텍스트에서 일치하는 match-set을 찾고 선택된 용어를 와일드카드로 대체하여 패턴을 생성한다. 생성되는 경우의 수가 매우 많으므로 평가함수를 이용하여 많은 일치를 만드는 일반적인 경우만 패턴으로 생성한다.

3.1 패턴의 일반화

텍스트에서 정보를 추출하고 토큰화하는 문제는 일련의 문자에서 단어로 분리하는 것으로 시작한다. 특수한 경우가 아니라면 텍스트를 공백과 구두점으로 분리하는 간단한 토크나이저를 작성할 수 있다. 그리고 패턴은 토큰화 된 텍스트와 와일드카드의 도입으로 쉽게 정의할 수 있다. 패턴과 일치하는 토큰의 시퀀스를 찾음으로써 match-set을 정의할 수 있고, 패턴에 부분 순서가 생기게 된다. 와일드카드로만 구성되는 아주 단순한 패턴에서 시작하여 와일드카드를 용어로 대체하는 과정을 통하여 보다 구체적인 패턴을 개발한다.

1) *base-set*

관계의 예로 주어지는 기초 용어 쌍을 *base-set*으로 정의한다. 심볼을 σ, 심볼의 시퀀스를 s라고 할 때, 패턴의 *(m,n)-base set*은 다음과 같이 정의한다.

*(m, n)-base-set(σ) = {Si-m…Si-1*Si+1…Si+n|Si=σ}*

같은 방법으로 심볼 σ_1, σ_2 쌍의 *(l, m, n)-base set*도 다음과 같이 정의한다.

*(l, m, n)-base-set(σ) = {Si-m…Si-1*Si+1…Si+m*Sj+1…Sj+n | Si=σ1, Sj=σ2, j-i=m+1}*

2) *match-set*

*match-set*은 패턴 p와 일치하는 심볼의 시퀀스 s의 집합을 의미하며 본 논문에서는 다음과 같이 정의한다.

$$match-set(p)=\{s\in\Sigma^{'*}|p\,matchs\,s\}$$

두 패턴 p와 p′에 대해 $p\le p'$라면 $0\le j\le|p'|$인 모든 j에 대해 $p'_j=p_{i+j}$ 또는 p'_j=*이다. 이것은 임의의 요소가 "*"로 대체된 것만 제외하면 p′이 p의 서브시퀀스와 같아야 함을 의미한다.

용어를 와일드카드로 대체하여 패턴을 생성할 수 있도록 허용하기 때문에 와일드카드가 아닌 심볼의 수가 n개이면 2n개의 패턴이 생성될 수 있다. 이것은 매우 큰 수이므로 가장 유용한 패턴을 선택하는 것이 필요하다.

3) *extraction* 패턴

extraction 패턴은 적어도 두 개의 와일드카드 *가 있는 패턴이다. $*_1$, $*_2$ 등으로 표현하며 추출 와일드카드라고 한다.

(l, m, n)$_{1,2}$-extraction 패턴은 관련 용어를 추출하는데 사용되며, 형식은 다음과 같다.

$$\underbrace{*\ldots*}_{l\text{회}}\ *_1\ \underbrace{*\ldots*}_{m\text{회}}\ *_2\ \underbrace{*\ldots*}_{n\text{회}}$$

*l, m, n*은 심볼 길이의 상한으로 모든 패턴은 *l, m, n* 보다 작거나 같은 심볼의 길이를 갖는다. *(l, m, n)-extraction* 패턴에 대한 패턴 집합은 다음과 같이 주어진다.

{(i, j, k)1,2-extraction, (i, j, k)2,1-extraction

| 0≤i≤l, 0≤j≤m, 0≤k≤n,}

3.2 패턴 평가 함수

많은 수의 패턴을 생성하기 쉽지만, 효율적인 패턴인지 평가하기 위한 방법이 필요하다. 가장 단순한 측정방법은 단순히 정확히 일치하는 비율이 가장 높은 패턴을 선택하는 것이다. 이것을 패턴의 정확도(precision)라 한다.

$$e_{precision}=\frac{\text{일치하는 용어 쌍}}{\text{선택된 용어 쌍}}=\frac{tp}{tp+fp} \quad (2)$$

이 메트릭은 양호하지만 하나의 올바른 서브 시퀀스와 일치한다고 하면 정확도는 100%가 될 것이다. 반면 패턴의 재현율(recall)은 다음과 같다.

$$e_{recall}=\frac{\text{일치하는 용어 쌍}}{\text{전체 용어 쌍}}=\frac{tp}{tp+fn} \quad (3)$$

이 메트릭 역시 extraction 패턴이 높은 재현율을 갖게 되는 결함이 있다. 따라서 두 항목간의 균형적인 메트릭이 필요하다. 본 연구에서는 정확도와 재현율을 사용하는 F-Measure라는 가중치 조화평균(weighted harmonic mean)을 사용한다. F-Measure를 구하기 위해 precision과 recall에 대한 조화 평균에 가중치 α를 적용하면 다음과 같다.

$$e_{F-Measure} = \frac{1}{\alpha\frac{1}{P}+(1-\alpha)\frac{1}{R}} = \frac{\frac{1}{\alpha}PR}{(\frac{1}{\alpha}-1)P+R} \quad (4)$$

3.3 규칙 및 Join-set

1) 규칙

여러 단어 용어를 효율적으로 처리하기 위해서는 용어의 경계를 식별하는 것이 요구되며, 보다 복잡한 규칙 언어가 필요하다. 따라서 말뭉치에서 명사구를 찾고 품사 태그를 붙임으로써 용어를 찾는 방법을 사용한다. 본 논문에서는 규칙을 요소의 시퀀스로 정의하며, 규칙에 사용되는 요소는 다음과 같다.

- literal : "σ"로 나타내며, 이것은 단순히 심볼 σ와 일치한다.
- words : words(n, m)으로 나타내며, $n, m \in N$에 대해 $n \le |s| \le m$ 인 심볼 s의 시퀀스와 일치한다.
- entity : name()으로 나타내며, 엔티티를 구성하는 심볼의 시퀀스와 일치한다.

예를 들어 다음과 같은 (패턴 3)이 있다면,

*1 * such as *2 (패턴 3)

다음과 같은 규칙을 얻을 수 있다.

:- name() words(1, 1) "such" "as" name()

2) Join-set

두 규칙 r_1, r_2가 *base-set*과 같은 부분 순서 관계가 있다고 할 때 r_1, r_2가 각각 $r_1 \le rj$이고, $r_2 \le rj$이며, $r_1 \le r'$, $r_2 \le r'$, $r' \le rj$를 만족하는 다른 규칙 r'이 존재하지 않는다면 rj는 r_1과 r_2의 join이다. 본 논문에서는 r_1, r_2의 모든 join의 집합을 Join-set이라고 정의한다. 두 규칙 r과 r'의 정렬을 $1 \le i \le |r|$, $1 \le j \le |r'|$을 만족하는 쌍의 집합 $A = \{(i, j)\}$로 정의한다. $(i = i')$ 또는 $(j = j')$ 이거나 $(i > i'$ 이고 $j < j')$인 조건을 만족하는 $(i, j) \in A$과 $(i', j') \in A$ 쌍은 존재할 수 없다. 또한 ri가 엔티티 표현이면 어떤 j에 대해 $(i, j) \in A$이고, 마찬가지로 rj가 엔티티 표현이면 어떤 i에 대해 $(i, j) \in A$이다. 이것은 정렬이 엔티티가 일치하는 표현의 매핑이고, 교차 일치를 포함하지 않는다는 것을 의미한다. Fig. 3은 "also"가 교차 일치하는 잘못된 정렬의 예이다.

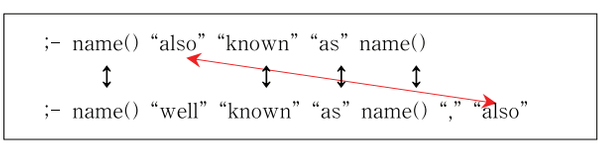

Fig. 3. Example of an Alignment

정렬된 집합 *A(r, r')*를 취하고, 그 외의 다른 규칙을 제거함으로써 *Join-set(r, r')*을 쉽게 찾을 수 있다. 두 규칙 간의 모든 유효한 정렬을 계산하여 일반화된 규칙을 생성할 수 있음을 의미한다. 따라서 두 규칙의 동일한 문장을 최소한으로 일치시켜야 한다. 먼저 r의 모든 *entity*에 대해 r'에 순차적으로 매치시켜 일치하는 *entity-base*를 얻는다. 다음으로 r의 각 *literal*에 대해 r'에서 같은 *literal*을 모두 찾는다. 이를 *literal-base*라고 한다. 그리고 다음과 같이 *find-aligns(entity-base, literal-base)*를 반환한다.

$$find-align(\{a_1, \dots, a_n\}, B) = \begin{cases} \{a_1, \dots, a_n\} & , B = \varnothing \\ find-align(valid(\{a_1, \dots, a_n\}, b), B - \{b\}) & , b \in B \end{cases}$$

$a_i = (i, j)$이고, $b = (i', j')$일 때 $i > i', j < j'$ 또는 $i < i', j > j'$을 만족하는 어떤 a_i도 없으면 $valid(\{a_1, \dots, a_n\}, b)$는 $\{a_1, \dots, a_n, b\}$이고, 그렇지 않은 경우는 $\{a_1, \dots, a_n\}$이 된다. 즉, 교차 매치되지 않아야 한다. 올바르게 추출되었다고 하면, 말뭉치의 올바른 일치 집합 *correct-matches(s)*를 정의할 수 있다.

각각의 $s' \in correct-matches(s)$에 대한 규칙 $r_{s'}$의 집합을 s의 *base-rules*로 정의한다. 이때 s'과 일치하고 $r' < r_{s'}$을 만족하는 r'은 존재하지 않아야 한다. 즉, 규칙은 문자와 *entity*로 만들어지는 가장 짧은 것이다. 또한 규칙 집합 $R = \{r_1, \dots, r_n\}$의 *completed-join-set*은 다음과 같이 재귀적으로 정의된다.

- $r_i \in completed-join-set(R)$이다.
- $r, r' \in completed-join-set(R)$이면, $join-set(r, r') \subseteq completed-join-set(R)$이다.

Fig. 4는 규칙 생성 알고리즘의 일부를 나타낸다.

Input : A corpus s, *correct-matches(s)*
Output : A set of rules R
1. Initialize Q = *base-rules(s)*, $R = \varnothing$
2. While $Q \neq \varnothing$
  ① Take some $r \in Q$, $Q \leftarrow Q - \{r\}$
  ② For each $r' \in Q$, $Q \leftarrow Q \cup join-set(r, r')$
  ③ $R = R \cup r$

Fig. 4. Algorithm for Rules Creation

4. 실험 및 평가

패턴 기반 관계 추출 방법을 평가하기 위해 테스트 데이터를 이용하여 실험을 수행하였다.

4.1 동의어 추출

온톨로지 자동 생성의 필요성을 해결하기 위하여 먼저 PubMed에서 “infectious disease”에 대한 검색결과의 상위 150개 초록을 수집했다. 이 초록에서 모든 의학 용어를 추출하고, 용어를 동의어 집합으로 분류했다. 용어가 유사하거나 거의 같은 의미이지만 조금 다른 면이 있기 때문에 완전히 일치하는 것은 아니다. 150개의 초록 문서에서 301개의 용어를 발견했고, 이 용어를 101개의 동의어 링크를 표현하는 221개의 동의어 집합으로 분류했다. Table 1은 검색된 초록 문서의 일부 목록이고, Table 2는 추출된 동의어 목록의 일부이다.

Table 1. Partial List of Abstract Documents

No	Abstract Documents
1	Why infectious diseases.
2	Climate change-related migration and infectious disease.
3	Extreme weather events and infectious disease outbreaks.
4	Impact of infectious disease consultation on the clinical and economic outcomes of solid organ transplant recipients admitted for infectious complications.
5	Dynamics of infectious diseases.
6	Incorporating pathology in the practice of infectious disease: myths and reality.
7	Trends in infectious disease mortality rates, Spain, 1980-2011.
8	Vaccination against infectious diseases: what is promising?
9	Merging economics and epidemiology to improve the prediction and management of infectious disease.
10	Emerging, evolving, and established infectious diseases and interventions.

Table 2. Partial List of Extracted Terms

No	Synonym list
1	Colorado tick fever, Tick fever, mountain fever
2	Crimean-Congo hemorrhagic fever, CHF Congo virus, Crimean hemorrhagic fever
3	Eastern equine encephalitis, Neuroinvasive Eastern equine encephalitis virus infection, EEE
4	West Nile fever
5	West Nile encephalitis, West Nile fever encephalitis
6	Dengue Fever, dengue, breakbone fever
7	dengue shock syndrome
8	Ebola hemorrhagic fever, Ebola virus disease
9	acquired immunodeficiency syndrome, Human immunodeficiency virus 1, HIV-1
10	Human immunodeficiency virus 2, HIV-2

그리고 이 용어들이 WordNet 및 두 특별한 도메인 시소러스인 MeSH와 UMLS에 잘 표현되어 있는지 살펴보기 위해 재현율과 정밀도를 구하고, F-Measure를 계산하였다. 또한 일반 지식 자료인 Wikipedia를 사용했는데 Wikipedia는 시소러스로 설계되지 않았기 때문에 정보 추출을 위해 약간의 처리가 필요했다. 추출이 얼마나 잘 일치되는지 살펴보기 위해 사용된 자원의 Precision, Recall, F-Measure, Coverage를 Table 3에 보였다.

Table 3. Results of Term Extraction

	Precision	Recall	F-Measure	Coverage
Wikipedia (redirect)	46.4%	18.8%	26.8%	54.1%
Wikipedia (search)	40.1%	24.6%	30.9%	56.1%
WordNet	100%	6.9%	13.0%	38.0%
Medline Encyc.	66.7%	4.0%	7.5%	28.1%
MeSH	61.6%	15.8%	25.2%	55.1%
UMLS	94.0%	46.5%	62.3%	79.7%

4.2 관계 추출

WordNet에서 “disease”의 하위어에 대한 용어 집합을 수집하여 총 1152개의 용어에 대한 동의어 및 상위어 목록을 작성하였다. 그리고 PubMed에서 용어의 집합을 포함하는 문서를 다운로드하여 문서로부터 말뭉치를 수집했다. 304개의 동의어 쌍과 299개의 상위어 쌍이 말뭉치에서 발견되었고, 111개의 동의어 쌍과 63개의 상위어 쌍에 대해서만 10번 이상 발견되었다. 패턴 생성 프로세스를 이용하여 동의어와 상위어의 기본 패턴을 생성하였고, Join-set을 계산했다. 적어도 3개 이상의 컨텍스트가 일치하고, 10% 이상의 정확도를 보이는 패턴만 취한 결과 동의어 10개, 상위어 45개로 패턴의 수를 줄였다. Table 4와 Table 5는 동의어 쌍 및 상위어 쌍의 출연빈도를 나타낸다.

Table 4. Frequency of Synonym Pairs from “Disease”

frequency	pairs	frequency	pairs	frequency	pairs
2	69	16	9	70	2
3	34	18	7	75	2
4	11	20	3	80	3
5	24	25	5	85	5
6	12	30	7	95	2
7	24	35	6	100	7
8	4	40	3	105	1
9	15	45	9	110	6
10	1	50	2	115	1
12	17	55	1	120	1
14	7	65	3	145	1

Table 5. Frequency of Hypernym Pairs from “Disease”

frequency	pairs	frequency	pairs	frequency	pairs
2	82	9	8	25	5
3	56	10	7	30	5
4	29	12	8	35	2
5	17	14	14	50	2
6	19	16	7	60	2
7	9	18	7	90	1
8	16	20	2	220	1

5. 결론 및 향후 연구

본 논문에서는 패턴을 기반으로 하여 자연어 텍스트로부터 동의어를 추출하는 문제를 제안했다. 일치하는 말뭉치의 컨텍스트 개수인 일치 패턴 집합을 고려하고, Join-set의 개념과 패턴의 정렬을 연결하여 활용하였다. 결과적으로 이 방법이 검색 공간으로부터 유용한 어떤 패턴도 제거하지 않고 50-95% 정도로 검색 공간의 크기를 줄이는 결과를 보였다. 또한 용어를 식별하기 위해 말뭉치에 태그를 붙이는 방법을 사용했다. 이전에 수동으로 식별할 수 없었던 용어 간의 관계를 찾을 수 있는 패턴의 개발을 가능하게하기 위한 것이다. Join-set 방법은 텍스트의 특성에 기초하여 요소를 일치시키는 복잡한 규칙을 허용한다.

향후 연구로는 추출된 관계로부터 온톨로지를 구성하는 방법을 연구하는 것이다. 추출 시스템을 절대적으로 신뢰할 수는 없으므로 온톨로지를 구성하는 효율적인 방법이 필요하다.

References

[1] Y. T. Kim, J. H. Lim, and C. S. Kim, “UML changes for efficient ontology development,” *Journal of the Korea Academia-Industrial Cooperation Society*, Vol.9, No2, pp.415-421, 2008.

[2] B. Smith, and C. Fellbaum, “Medical WordNet: a new methodology for the construction and validation of information resources for consumer health,” *In Proceedings of the 20th International Conference on Computational Linguistics*, pp.371, 2004.

[3] C. F. Baker, C. J. Fillmore, and J. B. Lowe, “The Berkeley FrameNet Project,” *In Proceedings of the 17th International Conference on Computational Linguistics*, pp.86-90, 1998.

[4] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran, “Indexing with WordNet synsets can improve Text Retrieval,” *arXiv preprint cmp-lg/9808002.*, 1998.

[5] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang, “SNOMED clinical terms: Overview of the development process and project status,” *In Proceedings of AMIA Symposium*, pp.662-666, 2001.

[6] O. Bodenreider, A. Burgun and T. C. Rindflesch, “Lexically suggested hyponymic relations among medical terms and their representation in the UMLS,” *In TIA'2001: Proceedings of Terminology and Artificial Intelligence*, pp.11-21, 2001.

[7] M. A. Hearst, “Automatic acquisition of hyponyms from large text corpora,” *In Proceedings of the 14th Conference on Computational Linguistics*, pp.539-545, 1992.

[8] P. Cimiano and S. Staab, “Learning by Googling,” *ACM SIGKDD Explorations Newsletter*, Vol.6, No.2, pp.24-33, 2004.

[9] X. Yang and J. Su, “Coreference resolution using semantic relatedness information from automatically discovered patterns,” *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp.528-535, 2007.

[10] S. Soderland, “Learning information extraction rules for semi-structured and free text,” *Machine Learning*, Vol.34, No.3, pp.233-272, 1999.

김 영 태

https://orcid.org/0000-0003-2103-8609

e-mail : ytkim1769@gmail.com

2000년 공주대학교 전자계산학과(학사)

2002년 공주대학교 전자계산학과(석사)

2012년 공주대학교 컴퓨터공학과(박사)

2002년~2016년 공주대학교 강사

2016년~2017년 엠에스티코리아(주) 기업부설연구소 팀장

2018년~현 재 ㈜케이엔씨 기업부설연구소 부장

관심분야 : 데이터통합, UML, 온톨로지

김 치 수

https://orcid.org/0000-0002-5675-1853

e-mail : cskim@kongju.ac.kr

1984년 중앙대학교 전자계산학과(학사)

1986년 중앙대학교 전자계산학과(석사)

1990년 중앙대학교 전자계산학과(박사)

2018년~현 재 공주대학교 컴퓨터공학부 교수

관심분야 : 소프트웨어 품질, 데이터 통합, 온톨로지