언어학적 분석을 통한 개념의 특성 정보 인식

백혜승 강영수 최기선
전문용어언어공학연구센터, 한국과학기술원 전산학과
{hspaik, yskang}@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Identification of Characteristics of a Concept through Linguistic Analysis

Haeseung Paik Young-Soo Kang Key-Sun Choi
KORTERM, Dept. of Computer Science, KAIST

요 약

개념은 그 개념을 나타내기 위한 특성들이 결합된 지식의 단위이며 각 특성은 개념에 속한 개체들의 성질을 축약한 것으로 정의될 수 있다[4]. 이 논문은 백과사전 설명문 텍스트를 분석하여 개념을 구성하는데 필요한 정보를 몇 개의 대표적인 특성으로 분류하고, 이를 개념의 특성정보로 구축하였으며, 이를 관련 개념 문서에 적용하여 특성 정보를 인식하는 것을 보여준다. 본 연구는 백과사전이 세계 지식(world knowledge) 전반을 함축적으로 표현하고 있다는 가정에서 출발하였으며 적은 양의 데이터에 대한 수동 분석 결과를 통해 많은 양의 코퍼스를 분석한 것과 같은 의미있는 결과를 얻었다. 백과사전에 표현된 많은 개념 중 '질병"에 관하여 실험한 결과 평균 81%의 정확율로 질병의 특성 정보인 원인, 증상, 치료를 자동 인식함을 보여주었다. 개념의 요소 정보 인식은 정보의 요약이나 질의 응답과 같은 분야에 적용될 수 있다.

1. 서 론

사용자가 필요로 하는 정보는 모두 몇 개의 구성 요소를 포함하고 있다. 예를 들어 사건 에 대한 구성요인은 5W1H(Who, What, When, Where, Why, How)로 표시될 수 있으며, [9]와 같이 좀 더 세분화된 템플릿으로 나눠질 수 있다. 또한, 인과적 지식은 cause-effect 템플릿에 의하여 설명될 수도 있다[5]. 사용자가 원하는 정보란 어느 정도 세분화된 구성요소 정보를 채우는 것으로 이해될 수 있다. 그러므로 의미있는 정보의 생성은 템플릿을 세부화된 정보 요소로 채우는 것으로 해석될 수 있다. 세분화의 정도에 따라 간략한 구성 요소 정보를 채우면 요약이 되고, 좀 더 세분화된 요소 정보를 만족시킨다면 질의 응답에서의 구체적인 답이 된다.

이와 같이 정보의 구성 요소를 찾는 것은 사용자에게 무엇을 제공할 것인가의 문제와 함께 그 무엇을 어떤 요소들로 채울 것인가를 알기 위해 필요하다. ISO 1087-1[4]에 따르면 개념은 개념을 이루는 특성들의 유일한 조합에 의해 생성된 지식으로 정의된다. 특성의 형태는 개념 시스템을 만들 때 하위 분류의 기준이 되기도 한다. 본 연구에서는 질병 개념의 요소 정보를 질병의 특성 정보로 정의하고, 각 특성을 인식하는 방법을 조사하였다.

이 연구에서 개념이 어떤 특성 정보들로 이뤄지는가를 찾기 위해 백과사전 지식을 분석한다. 백과사전은 세계지식 전반에 관한 정보를 담고 있기 때문에 여러 가지 개념을 찾을 수 있고, 그 개념의 특성정보를 효과적으로 찾을 수 있다고 가정하였다.

이 논문은 백과사전의 여러 가지 개념 중에서 질병'을 다루었다. 질병'의 중요 요소 정보로서 세 가지 특성- 원인, 증상, 치료-를 찾았으며 각 특성을 특징짓는 언어적 특성을 파악하여 질병 개념 지식을 개념지도로 표현하였다. 이 개념 지도를 이용하여 임의의 질병관련 문서의 내용을 원인, 증상, 특성별로 구분하는 실험을 하여 평균 81%의 정확율을 얻었다.

2장에서 연구 코퍼스로 사용된 백과사전 구성을 살펴보고 용어에 대한 정의를 하며 3장에서는 개념 정보의 구성 요소인 특성정보 추출을 위한 방법을 설명한다. 4장에서는 질병 개념의 특성에 대응하는 언어 표현을 설명하고 5장은 평가로서 개념 특성을

자동으로 인식하는 프로토타입 시스템 「특성 인식기」를 이용하여 '질병'개념의 여러 특성 정보를 인식하는 실험 및 결과를 보여 준다. 그리고 6장에서는 본 연구의 현재까지 결론 및 향후 연구를 언급하겠다.

2. 백과사전 지식의 구성

일반적으로 백과사전은 세계지식을 총체적으로 모아 놓은 것이라고 할 수 있다. 많은 개념들이 포함되어 있으면서도 각 개념에 대해 중요한 요소 정보들을 포함하고 있다. 아래는 한국 브리태니커 온라인의 백과사전'에 관한 정의이다..

"백과사전의 역사와 범주에 대한 설명은
학문발달에 대한 지침이 되어왔는데, 이는
백과사전이 수세기 동안 당대 지식과 학문을
포괄적으로 기록해놓음으로써 학문 발달에 있어서
이정표 역할을 해왔기 때문이다."

사용된 백과사전은 ETRI 와 (주)계몽사에서 발간된 ETRIKEMONG SET 으로 "보판 계몽사 학생 백과 사전의 텍스트"를 포함하는 것으로 약 23,000 여개의 항목을 갖고 있다. 각 항목은 다음 예에서 보이는 것과 같은 형식을 하고 있다.

<id> 15103
<title> 위궤양 胃潰瘍
<contents>
위의 점막이 헐어서 점막 밑에 있는 조직이 파괴되는 병.
위궤양은 자극이 심한 음식을 지나치게 먹거나 정신적인 피로가 원인이 된다.
증세로는 식후에 위가 아프거나 대변에 피가 섞이기도 한다.
30~40세의 남자에게 많이 생긴다.
자극이 심한 음식물과 정신적인 피로를 피하고 감정이 상하지 않도록 주의한다.

<id> 는 백과사전에 수록된 항목의 일련번호로서 고유번호를 갖는다. <title>은 항목의 "표제어"이며, <contents>는 표제어에 관한 내용을 설명한 부분이다. <contents>의 내용을 살펴보면 첫번째 문장은 항상 서술명사구로서 표제어에 대한 한 마디의 정의를 하고 있으며, 명사구로 맨 뒤에 나오는 단어가 표제어를 나타내는 "주제어"로서 표제어의 상위 개념으로 표현되고 있다. 예를 들어 질병의 경우 병, 병증, 질병, 질환 등의 주제어로 나타나며 이들 주제어는 유의어 사전을 이용하여 '질병'이라는 하나의 개념으로 묶을 수 있다.

'질병' 개념은 원인, 증상, 치료, 예방, 발병 대상 등으로 분류되는 특성과 이들 특성을 설명하는 실질적인 내용이 있게 된다. 아래 그림은 질병과 질병을 이루는 내용의 특성을 보여주며 이를 질병에 대한 개념지도라고 명명하기로 한다. 이 때 원인, 증상, 치료 등은 개념 '질병'에 대한 특성(C)'라고 정의하며 개념 지도의 1차 연결 특성을 형성한다. 타원형의 노드가 개념을 표현하며 사각형의 노드는 부분-전체 관계(part-whole relation)로서 개념의 특성을 나타낸다.

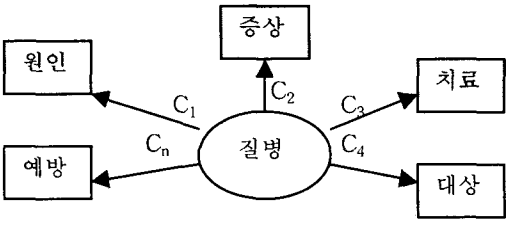

그림2-1. 질병 의 1차 개념지도

백과사전에서 한 개념에 관련된 항목들을 분석하여 각 특성과 관련된 정의문으로 개념지도를 확장할 수 있다. 본 연구에서는 백과사전의 질병항목 가운데 사람에 관한 질병 항목만을 선택하여 분석하였다. 2차 연결특성은 다음의 그림2에서 보는 바와 같이 특성을 설명적으로 표현하는 정의문들로 형성된다. $D_j(C_i)$는 특성 C_i를 설명하는 정의문(D_j)으로의 링크를 뜻한다.

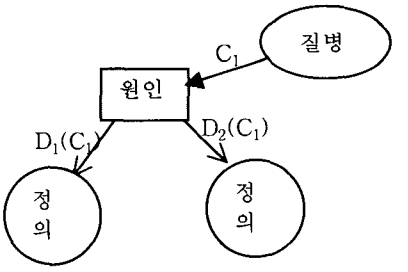

그림 2-2. 질병'의 원인특성이 확장된 2차 개념지도

본 연구에서는 이와 같은 개념지도를 형성하는 과정에서 분석된 각 특성별 언어적 특징을 파악하여 개념 지식을 형성한다. 그리고 이 지식에 근거하여 같은 개념의 문서 내용을 각 특성별로 인식할 수 있는지를 실험하였다.

3. 구성요소 정보의 추출

백과사전에서 추출한 질병' 개념에 대해 어떤 특성 정보들이 있는지를 파악하기 위해 35개의 관련 항목을 수동으로 분석하였다. 분석에 앞서 문장 구분과 비문에 대한 처리를 한 후, 각 특성별 문장

들을 추출하여 형태소 분석을 하였다. 복문을 단문으로 분할한 후 내용을 분석하여 질병'개념의 대표적인 특성에 나타난 언어적 특징을 발견하였다.

3.1 개념의 특성 분류

사전 항목에 포함된 내용의 분석은 질병' 개념의 중요한 요소 정보를 파악하여 질병에 대한 특성을 정의하기 위함이다. 이 분석은 언어학 전문가가 문장 내의 적절한 위치에 특성 범주 태그를 부여함으로써 이뤄졌으며, 이 과정에서 질병을 표현하는 여러 가지 특성에 해당하는 다음과 같은 태그를 정의할 수 있었다.

/DEF: definition (정의)
/CAU: cause (원인)
/SYM: symptom (증상)
/REM: remedy (치료)
/PRE: prevention (예방)
/PRO: progress (경과)
/OBJ: object (대상)
/SOR: sort (종류)
/IFT : if-then(결과)
/OTH: others (기타)

표 3-1 질병 개념에서의 특성 태그

본 논문에서는 위의 특성 범주 태그 가운데 원인, 증상, 치료(예방 포함), 경과, 대상, 종류, 결과의 7개를 질병 개념의 특성으로 정의한다. 아래의 표 3-2는 질병 항목 35개로부터 각 특성별 상대빈도(RF: Relative Frequency)를 계산한 결과이다. 개념 C의 특성 C_i 의 상대빈도는 다음 식을 이용하여 계산된다.

$RF(C_i) = \Sigma \{S(m,C_i)/S(m)\}$ $(m=1..35)$
C_i :질병 개념의 각 특성
$S(m,C_i)$:m번째 항목에서 C_i를 나타내는 단문의 수
$S(m)$:m번째 항목의 단문의 수

특 성 (C_i)	상대빈도(RF)
원인	10.1
증상	10.9
치료	3.48
경과	0.92
대상	0.96
종류	1.87
결과	1.21

표 3-2. 특성별 상대빈도

이 분석을 통해 질병을 설명하는데 가장 빈번히 나타나는 특성이 정의, 원인, 증상, 치료임을 알 수 있다. 이 세가지 특성이 질병' 개념을 설명하는 대표적인 구성 요소로서 질병의 1차 개념지도의 중요 특성 노드를 구성한다. 이 특성들은 질병에 관련된 문서의 요약 과정에서 요약에 포함될 주요 문장을 선택하는데 단서가 될 수 있다.

3.2 특성 관련 어휘의 추출

앞에서의 개념의 특성에 따라 추출된 문장들은 각각의 특성 문서를 구성하게 된다. 본 연구에서는 대표적인 특성만을 추출하여 원인 문서, 증상 문서, 치료 문서를 구성하고, 각 특성에서 나타나는 어휘 집합을 표 3-3과 같이 생성하였다. 어휘집합은 먼저 특성을 표현하는 술어집합으로 아래와 같이 나타난다.

1) 일반동사
2) 성상형용사
3) 동작성 명사+ 접사
4) 상태성 명사+ 접사

또한 명사집합은 특성의 서술어에 대한 논항으로 나타나는 명사를 추출하였다. 이를 위해 간단한 부분구문분석이 이뤄지기도 했다. 생성된 어휘집합은 임의의 질병 관련 문서를 분석할 때 요소 정보인 각 특성을 인식하는 실험에 사용되기 위해 좀 더 정교한 특성으로 분류되어 개념지식으로 이용된다.

특성	명사 집합	술어 집합
원인	원인, 때문, 결핍, 부족, 세균, 병원체, 자극,먼지,알레르기, 불균형, 피로	일어나다, 의하(다), 나타나다, 일으키(다), 생기(다) 걸리(다), 나빠지(다), 부족하(다)
증상	증상, 증세, 현기증, 통증, 갈증, 맥박, 숨, 호흡, 설사, 복통,열, 출혈	나타나(다), 일어나(다) 아프(다), 붓(다) 토하(다)
치료	치료, 처치,요법 휴식,안정,운동,식사 주사,수혈, 소독	내복하(다) 절제하(다) 피하(다) 쓰(다)

표3-3. 질병의 특성에 따른 어휘집합 예

4. 개념의 특성에 따른 언어적 특징

앞 장에서 개념의 각 특성에서 출현한 어휘지식을 분석한 결과 아래와 같은 특징을 알 수 있었

다. 원인문, 증상문, 비교문은 각각 원인 문서, 증상 문서, 치료 문서에 포함되어 해당 개념의 특성을 나타내는 문장을 뜻한다.

(1) 서술형 특징

각 특성별 문장에 따라 아래와 같이 특성을 서술하는 패턴을 찾을 수 있다. 먼저 원인을 서술하는 패턴은 아래와 같다.

- ~에[로] 의하(다)/말미암(다)/인하(다)
- ~때[경우]에 생기(다)/나타나(다)
- ~때문에 일어나(다)/나타나(다)

또한 증상을 나타내는 서술형은 아래와 같이 나타나는데 이 경우에 원인문에서와 같은 동사형이 함께 사용되고 있음을 알 수 있다.

- ~이 나타나(다)/생기(다)/일어나(다)
- ~게 되(ㄴ 다)
- ~ 수(것) (도) 있(다)

그러나 이 경우에는 서술어의 필수격 정보에 해당하는 앞 논항의 격정보를 하여 구분이 가능하다.

그리고 치료문의 경우는 다음과 같이 환자에게 요구사항을 권고형이나 피동형으로 표현하는 경우가 많았다.

- ~해야 하(ㄴ 다)
- ~시키(다)

위의 술어 패턴에 나타난 특징을 이용하여 개념의 특성을 구분하는 첫 번째 기준을 마련하였다.

(2) 술어 특징

질병의 세 특성에 나타난 특징을 살펴보면 원인과 증상사이의 서로 공통된 술어를 취함으로써 빈번한 술어에 대하여는 서술형 패턴을 정의하여 패턴 일치시도를 먼저 하도록 하였다. 그러기 위해 먼저 공통된 술어 집합을 구성하였다. 여기에는 '나타나다, 생기다, 일어나다, 일으키다, 오다, 발생하다' 가 있으며 모두 같은 의미의 동사이다. 또한 원인과 증상 그 자체를 설명하는 서술어를 분리하여 이 서술어에 가산 점수를 부여하였다. 예를 들면 '감염되다, 걸리다, 부족하다, 침해되다' 등의 술어는 원인지향 술어이고, '아프다, 붓다, 가볍다, 심하다' 등의 술어는 증상 지향 술어로 구분할 수 있다.

본 연구에서는 완전한 구문 분석이 이뤄지지 않고 부분 구문 분석만 이뤄졌다. 부분 구문분석에서는 각 특성을 설명하고 있는 서술어를 정확히 찾기 위한 규칙을 정의하였으며 각 규칙들은 찾고자 하는 특성 의존적이다. 다시 말하면 증상 문서에서는 서술어가 2어절 거리에서 두 번 나타나면 앞의 서술어를 "주 서술어"로 간주한다. 어절거리는 어절과 어절 사이의 공백 개수로 정의하였다. 주 서술어는 각 특성을 직접적으로 설명하는 서술어에 해당한다.

백과사전에 나타난 모든 술어는 발생빈도는 적지만, 그 발생자체에 의미를 둘 수 있는 어휘이므로 술어들을 유사한 몇 개의 그룹으로 분류할 수 있어 술어 생성시 어휘선택 단계에 유용한 정보를 제공할 수 있다.

(3) 명사 특징

각 특성에서 출현하는 명사들을 분석한 결과, 1군 명사와 2군 명사로 분류하여 특성인식시의 중요도를 달리 하였다. 먼저 1군 명사는 특성 자체를 명시하는 단어로서 원인문에서는 '원인, 이유, 때문'이고, 증상문에서는 '증상, 증세', 그리고 치료문에서는 '치료, 처치, 요법'등을 찾을 수 있다.

2군 명사는 특성을 의미적으로 설명하는 단어로서 원인문에서 '결핍, 부족, 감염, 병원체, 세균, 불균형, 알레르기', 증상문에서 '현기증, 통증, 빈혈, 설사, 발열, 출혈, 체중감소', 치료문에서 '휴식, 안정, 운동, 주사, 수술, 예방, 항생제, 약' 등을 구분하였다.

(4) 명사-술어 특징

술어와 같이 나타나는 명사들은 공기 정보를 갖고 질병의 특성 인식에서 보조적 역할을 감당하도록 하였다.

(5) 연결어미 특징

연결어미의 사용에 나타난 특징은 원인문에서 종속적 연결어미의 사용이 두드러지고 대등적 연결어미는 증상문에서 많이 나타나고 있다.

5. 평가

5.1 실험

백과사전에서 훈련 데이터에서 제외되었던 18개 다른 질병관련 데이터에 적용하는 실험을 하였다. 이 테스트 데이터는 훈련 데이터와 마찬가지로 사람이 각 특성에 대한 수동 태깅 후에 원인, 증상, 치료의 특성에 따른 특성 문서를 구성하였다. 그리고 각 특성별 특징을 나타낸 2차 개념지도의 정보를 기반으로 특성을 자동으로 인식하는 특성 인식기에 의해 인식된 결과를 사람이 태깅한 결과와 비교하게 된다.

자동 특성 인식기는 특성별 특성 정보에 따라 각 문장에 대한 원인값(M(CAU)), 증상값(M(SYM)), 치료값(M(REM))를 계산하여 가장 높은 값이 선택되며 문장은 해당 특성으로 인식된다. 이 계산을 위해 다음과 같은 값을 정의하였다.

- S: 문형 점수
- C: 단서어(clue word) 점수
- P: 술어 점수
- N: 명사 점수
- E: 연결어미 점수

문형점수(S)는 특성별 특성을 잘 나타내는 서술형 특징을 반영한 값이고, 단서어 점수(C)는 각 특성의 1군 명사들이 출현할 때 주어진다. 특성 지향 서술어나 명사가 출현할 때 마다 서술어 점수(P)와 명사 점수(N)가 계산된다. 또한 연결어미 점수(E)는 원인문에 두드러진 종속적 연결어미의 특징을 반영한 값으로서, 앞에서 고려된 특징이 나타나지 않아서 어떤 특징으로도 분류할 수 없을 경우에 대비하여 계산되므로 가장 낮은 가중치를 갖는다.

또한 본 연구에서는 특성 정보를 계산하는데 있어서 서술어만 고려한 경우와 명사 정보를 포함한 경우를 분리하여 실험하였다.1차와 2차로 나눠진 실험에서 <식1>과 <식2>가 따로 적용되었다.

$$M(C) = a\,S + c\,P + e\,E \quad \text{..... <식 1>}$$

$$M'(C) = a\,S + b\,C + c\,P + d\,N + e\,E \quad \text{..<식 2>}$$

$$C \in \{CAU, SYM, REM\}$$

가중치 인자는 원인과 증상 특성에서는 a > b > c > d 인 값으로 주어졌는데 이는 특성에 따른 언어적 특징의 중요도에 따라 결정된다. 위의 식에서 알 수 있듯이 문형 점수가 가장 결정적인 역할을 하는 것을 알 수 있다. 각 문장에 대해 원인값, 증상값, 치료값이 위와 같이 계산되며 최대값을 갖는 특성으로 분류된다.

5.2 결과

실험 오류를 줄이고자 형태소 해석 오류를 수정하였다. 제안한 방법이 질병의 개념 지식의 기반이 되는 35개의 학습 코퍼스에 대해서도 잘 적용된다는 것을 확인하였다. 정확율(Pr)은 전체 문장 가운데 처음부터 오분류 문장(*ill-classified* sentences)은 제외하고 특성을 제대로 인식한 비율이다. 잘못 분류된 문장은 한 특성 문서에 포함되었으나 실제로 해당 특성을 설명하지 않는 문장을 뜻한다. 오류율(Er)은 특성을 분류할 때 사람이 분류 오류를 일으키는 가능성으로 여기서는 두 사람의 특성 분류 의견이 일치하지 않은 경우에 발생한다. 각 특성 문서의 모든 문장에서 오분류 문장이 차지하는 비율이다. 정확율은 아래의 식과 같이 계산된다.

$$Pr = N(C) / (N(S) - N(I))$$

N(C) : 특성 인식이 옳게 된 문장의 수
N(S) : 특성 문서내의 문장의 수
N(I) : 특성 문서내의 오분류 문장의 수

표 5-1은 특성의 인식 정확율이 1차와 2차 실험에서 각각 평균 89.3%, 94.6%임을 알 수 있다. #(S)는 각 특성 문서에 포함된 문장의 개수이다. 1차와 2차 실험의 정확율을 비교하면 원인과 증상문서에서는 유사하나 치료문서에 큰 차이를 보임을 알 수 있다. 치료문서에는 명사 특징을 포함함으로써 크게 향상되었다.

특성	단문 문장수	실험	성공율	오류비율
원인	67	1차	0.98	0.05
		2차	0.97	
증상	112	1차	0.92	0.08
		2차	0.91	
치료	46	1차	0.78	0.04
		2차	0.95	

표 5-1 훈련 데이터에 대한 특성인식 결과

동일한 실험 환경에서 개념 지식을 만드는데 사용되지 않은 질병 항목을 사용하여 특성 인식기의 성능을 테스트하였다. 테스트 데이터에 대한 1차와 2차 실험 결과가 표 5-2에 나타나 있다. 치료문서에서 훈련 데이터에서와 같은 대조를 알 수 있다. 모든 개념 지식을 사용하였을 때 평균 정확율 81%로 각 특성을 인식하였다.

특성	단문 문장수	실험	성공율	오류비율
원인	26	1차	0.84	0.01
		2차	0.88	
증상	56	1차	0.76	0.02
		2차	0.78	
치료	26	1차	0.61	0.01
		2차	0.77	

표5-2 테스트 데이터에 대한 특성인식 결과

원인과 증상 문서에서는 문장 패턴과 서술어 특징만으로 질병 개념의 원인과 증상을 인식하기에 충분한 것으로 보인다. 그러나 명사적 특징이 치료의 특성을 인식하는데 중요함을 알 수 있다. 그 이유는 치료를 설명하는 문형과 서술어가 다른 특성에 비해 상대적으로 적고, 치료 문서가 여러 가지 약이나 유의어 명사들이 많이 출현하기 문이라고 할 수 있다.

본 논문의 접근 방법을 온라인 질병 백과사전인 Humedic.com (Samsung Life, 2000) 에 적용해 보았다. 휴메딕은 전문가적 의학지식으로 질병 전반에 관한 설명을 정의, 원인, 증상, 진행, 치료, 예방을 위한 안내, FAQ 등으로 나눠 설명해 주고 있다. 휴메딕에서 10개의 내과 질병 항목을 추출하여 실험한 결과, 원인, 증상, 치료 특성을 53%, 68%,

46%의 정확율로 인식하였다. 이 평가는 모든 언어 특징을 고려한 <식2>를 사용하였다. 이 결과는 형태소 분석 오류 수정과 오분류 문장에 의한 오류율을 고려하면 좀 더 향상될 것이라 기대된다. 또한 개념 지식에 포함된 어휘의 동의어들에 대한 불인식 문제가 있다. 온라인 의학 백과사전에서 추출한 문서의 특성 인식 실험에서 유의어를 추가하여 용어를 확장한 결과, 증상 특성 인식에서 17%까지 정확율이 향상되었다.

6. 결론 및 향후 연구

이 연구에서는 적은 양지만 개념의 요소 정보를 포함하고 있는 백과 사전 텍스트에 나타난 언어적 특성을 파악함으로써 개념을 구성하는 1차적인 특성 정보를 추출하였으며, 이 특성에 기반한 요소 정보를 개념 지식으로 구성할 수 있음을 보였다. 개념 지식을 만드는 과정에서 개념의 대표 특성을 찾고 그 특성에 기반한 개념 지도가 그려질 수 있음을 보였다. 그리고 개념의 대표 특성에 의존적인 언어적 성질을 찾아서 개념 지식을 구성하였다. 이 개념 지식을 이용하여 각 특성의 요소 정보를 자동으로 구분하여 인식함을 보여 주었다.

이 논문에서는 '질병' 개념의 특성 정보를 인식하는데 적용하여 평균 81%의 정확율로 질병의 특성 정보인 원인, 증상, 치료를 자동 인식함을 보여주었다. 이 과정을 통해 통계적 분석이 아닌 문형 패턴과 서술어 정보, 그리고 명사 정보에 의존하는 개념의 특성 정보를 이끌어 내었다. 동일한 특성을 표현하는 서술어나 명사 어휘 집합을 만듦으로써 관련된 지식 표현의 생성시에 어휘를 선택할 수 있는 기본어휘의 구성도 가능하다. 이와 같은 개념의 특성 정보 인식은 정보의 요약이나 질의 응답과 같은 분야에 적용될 것이다.

7. 참고 문헌

[1] Bae, Jae-Hak and Jong-Hyeok Lee (2001) "Topic Sentence Selection with Mid-Depth Understanding", Proc. of ICCPOL 2001, pages 199-204.

[2] Fujii, Atsushi and Tetsuya Ishikawa (2000) "Utilizing the World Wide Web as an Encyclopedia: Extracting Term Descriptions from Semi-Structured Texts", Proc. of the 38th Annual Meeting of ACL, pages 488-495.

[3] Hearst, Marti A. (1994) "Multi-paragraph Segmentation of Expository Text", Proc. of the 32nd Annual Meeting of ACL, pages 9-16.

[4] ISO 1087-1 (2000) Terminology Work – Vocabulary – Part 1: Theory and application.

[5] Khoo, Christopher S.G., Syin Chan and Yun Niu (2000) "Extracting Causal Knowledge from a Medical Database Using Graphical Patterns", Proc. of the 38th Annual Meeting of ACL, pages 336-343.

[6] McKeown, Kathleen.R., J.L. Klavans *et al.* (1999) "Towards Multidocument Summarization by Reformulation: Progress and Prospects", AAAI.

[7] Nakao, Yoshio. (2000) "An Algorithms for One-page Summarization of a Long Text Based on Thematic Hierarchy Detection", Proc. of the 38th Annual Meeting of ACL, pages 302-309.

[8] Novak, Joseph D., "The Theory Underlying Concept Maps and How to Construct Them", http://cmap.coginst.uwf.edu/info/.

[9] Radev, Dragomir R. and Kathleen R. Mckeown (1998) "Generating Natural Language Summaries", Computational Linguistics, 24(3):469-500.

[10] Saggion, Horacio (1999) "Using Linguistic Knowledge in Automatic Abstracting", ACL Workshop.

[11] Saggion, Horacio and Guy Lapalme (2000) "Concept Identification and Presentation in the Context of Technical Text Summarization", NAACL-ANLP Automatic Summarization Workshop.

[12] Samsung Life & Samsung Medical center (2000) Humedic.com. http://ss745.humedic.com:8080/HMD2/SilverStream/Pages/hmd2_dic.html .

[13] Yangarber, Roman, Ralph Grishman, *et al.* (2000) "Automatic Acquisition of Domain Knowledge for Information Extraction", COLING.