ACE 관계 추출과 특징화 과정에서 성능 향상을 위한 새로운 방법(1)

김경덕*° 김석환* 이근배* 차정원**
*포항공과대학교 컴퓨터공학과 지능소프트웨어연구실
{getta, megaup, gblee}@postech.ac.kr
**창원대학교 컴퓨터공학과
jcha@changwon.ac.kr

A New Method for Improving Performance in ACE Relation Detection and Characterization

Kyungduk Kim*°, Seokhwan Kim*, Gary Geunbae Lee* and Jeongwon Cha**
*iSoft Lab., Dept. of Computer Science and Engineering,
Pohang University of Science and Technology
**School of Computer information & Communication,
Changwon University

요 약

텍스트 기반 문서의 급증으로 인해 정보 추출 기술이 더욱 중요해지고 있다. 특히 최근에 활발한 연구가 진행되고 있는 개체 간 관계 추출 기술은 정보검색과 질의응답 등 많은 분야에 걸쳐 활용될 수 있는 기술이다. 본 논문은 기존의 자질 기반 관계 추출 시스템의 재현율을 향상시키기 위해 WHISK 알고리즘을 도입한 시스템에 관한 것이다. WHISK 알고리즘은 문장으로부터 관계에 참여하는 개체 쌍을 추출하는 규칙을 자동으로 학습한다. 그리고 시스템은 최대 엔트로피 모델을 이용하여 WHISK에 의해 추출된 개체 쌍에 적합한 관계 유형을 파악해 낸다. 본 논문은 시스템에 사용된 WHISK 알고리즘과 최대 엔트로피 모델에 대해서 알아보고, 실제로 WHISK 알고리즘을 도입하여 관계를 가지는 개체 쌍을 추출하여 문제를 해결했을 때 어느 정도의 성능 향상이 있는지 알아본다.

1. 서 론

최근 인터넷의 보급과 디지털 미디어의 확산으로 텍스트 기반의 문서가 급증하고 있다. 그에 따라 텍스트로부터 유용한 정보를 추출하는 정보 추출 기술이 더욱 중요해지고 있다. NIST의 Automatic Content Extraction (ACE) 프로젝트는 이러한 정보 추출 기술의 발전의 장려를 위해 크게 세 가지의 연구 목적을 가지고 프로젝트가 진행되고 있다. Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), Event Detection and Characterization (EDC) 가 그것이다.

RDC는 두개의 개체 사이에 관계가 있는지 여부를 밝혀내고, 관계가 있을 경우, 그것이 어떤 유형의 관계인지를 파악해 내는 문제를 말한다. 현재 ACE에서는 다음과 같이 크게 다섯 가지의 유형으로 관계를 분류하고 있고, 세분화 하여 24가지의 유형으로 분류하고 있다.

- Role : 조직 내에서 개인의 역할에 관련된 관계
- Part : 부분과 전체의 관계
- At : 장소에 관련된 관계
- Near : 상대적인 장소와 관련된 관계
- Social : 개인의 사회적 관계

실제로 개체 간 관계의 추출은 정보검색과 질의응답을 비롯한 많은 분야에 도움을 줄 수 있다. "국회의사당은 어디에 있지?"라는 질의가 있을 때, 국회의사당이라는 개체와 "At" 이라는 관계를 가지는 개체를 찾아내면 해당 질의의 답이 될 수 있다.

(1)본 연구는 산업자원부 21C 프론티어 연구 "인간생활자원 지능로보" 과제의 지원을 받아 수행되었음

이 논문은 ACE 프로젝트의 RDC 문제의 해결을 위한 방법을 다루고 있다. 기계 학습 기반의 알고리즘인 WHISK 알고리즘을 이용하여 개체 간 관계의 유무를 밝히고, 최대 엔트로피 모델을 이용하여 개체 간 관계의 유형을 파악하였다.

2. 관련 연구

개체 간 관계 추출은 최근에 들어 주목을 받고 있는 정보 추출 분야이다. 현재 이 분야에 대해 많은 연구가 진행되고 있으며 다양한 각도로 문제에 접근하고 있다.

[6]은 가장 최근에 발표된 논문 중의 하나로, 문장으로부터 여러 언어학적 속성을 자질로 추출하여 지지 벡터 기계 (Support Vector Machine)로 관계의 유형을 파악하는 연구에 관한 것이다. [3]는 문장으로부터 추출한 언어학적 속성을 자질로 하는 최대 엔트로피 모델을 이용하여 개체 간 관계를 추출하였다. [6]은 [3]가 사용한 자질에 텍스트 단위화(chunking) 정보 및 단어 의미망 등을 이용한 몇 가지의 의미적 자질을 추가하여 더 나은 성능을 보여주고 있다. [6]와 [3]의 연구는 현재 가장 높은 성능을 나타내고 있는 시스템들이나 재현율이 낮다는 한계를 가지고 있다.

[4]는 문장의 문법 파스 트리에 개체와 그것들 사이의 관계에 대한 의미 정보를 추가하여 확장된 형태의 파스 트리를 이용해 개체 간 관계를 추출 하였다. [2]는 문장의 의존 트리간의 커널을 이용하여 문제에 접근하였다. 하지만 이와 같이 문법 파스 트리나 의존 트리를 사용하여 관계를 추출할 경우, 시스템 전체의 성능이 문장의 파서 성능에 크게 좌우될 수 있다는 문제점이 있다.

이 논문에서는 [3]와 같이 문장의 언어학적 속성을 자질로 가지는 최대 엔트로피 모델을 이용하여 문제를 해결하고자 하였으며, 낮은 재현율을 극복하기 위해 기계 학습 기반 알고리즘의 일종인 WHISK 알고리즘을 이용하여 관계를 가지는 개체 쌍을 추출할 수 있는 규칙을 학습하였다.

3. 시스템 구조

3.1 전체 구조

시스템은 크게 두 단계로 나누어 관계 추출을 수행한다. 첫 번째 단계에서 개체 쌍 추출 규칙을 이용하여 관계에 참여하는 개체 쌍을 추출해 낸다. 그리고 두 번째 단계에서는 최대 엔트로피 모델을 이용하여 앞 단계에서 추출된 개체 쌍에 대하여 어떤 유형의 관계가 적합한지를 예측한다.

관계 추출을 위한 학습과정은 개체 쌍 추출 규칙 학습과 관계 유형 분류모델의 학습으로 이루어진다. 사람이 직접 손으로 개체 쌍 추출 규칙을 만드는 일은 시간과 노력이 많이 소요되는 작업이다. 따라서 본 논문에서는 자동으로 개체 쌍 추출 규칙 만들어 내는 알고리즘인 WHISK를 도입하였다. WHISK는 지도학습(supervised learning)을 이용하는 기계 학습 알고리즘의 일종으로, 관계가 태깅된 문장들로부터 자동으로 개체 쌍 추출 규칙을 학습한다. 최대 엔트로피 모델은 정답이 태깅된 문장으로부터 획득한 언어학적 속성들을 자질함수로 하여 관계 유형 분류 모델을 만들어 낸다.

그림 1은 전반적인 시스템을 구조를 나타내고 있다.

3.2 관계를 가지는 개체 쌍 추출을 위한 WHISK 알고리즘

문장에서 관계를 발견해 내는 문제는 특정 개체 쌍 사이에 관계가 있는지 없는 지를 판단하는 문제와 같다. 이는 곧 관계가 있다고 판단되는 특정한 개체 쌍을 문장으로부터 추출해 내는 문제라고 볼 수 있다. WHISK 알고리즘은 단일 슬롯뿐만 아니라 다중 슬롯의 추출이 가능한 알고리즘으로 이러한 특정 개체 쌍 추출에 유용하게 사용될 수 있는 알고리즘이다[5]. WHISK 알고리즘은 사람이 손으로 태깅한 정답 문장들로부터 관계에 참여하는 개체 쌍을 추출할 수 있는 일련의 규칙들을 학습한다. 관계 추출 규칙은 정규표현식의 형태로 이루어져 있으며, 이 정규표현식을 이용하여 문장으로부터 관계에 참여하는 개체들을 추출한다.

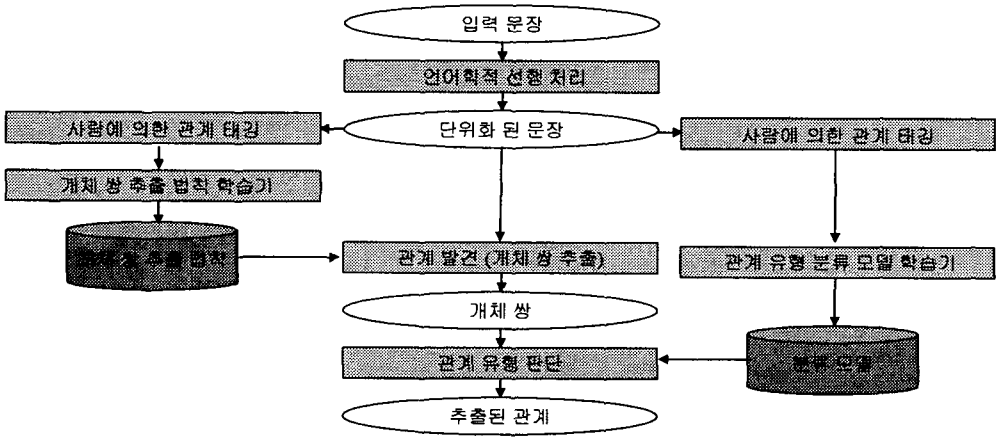

그림 1 시스템의 전체 구조

```
Pattern:: * (ORG) * in * (GPE) *
Output:: AT-Located {argument1} {argument2}
```

그림 2 WHISK로 학습된 개체 쌍 추출 규칙

그림 2은 WHISK 알고리즘을 통해 학습된 개체 쌍 추출 규칙의 예제이다. WHISK에 의해 학습된 개체 쌍 추출 규칙은 정규표현식과 형태가 비슷하나 약간의 차이가 있다. 그림 1의 패턴에서 괄호에 싸인 부분은 실제로 개체가 추출되는 부분으로, 문장에서 괄호 안의 해당 개체 유형이 나타나면 그 개체가 관계에 참여하는 개체로 추출이 된다. 예제에서 ORG는 조직을 나타내는 개체 유형 의미하고, GPE는 지리학적으로 정의된 지역을 나타내는 개체 유형을 의미한다. 그리고 위의 패턴에서 첫 번째 "*" 은 조직에 해당하는 개체가 나타날 때까지 모든 글자를 건너뛰라는 것을 의미하고, 마찬가지로 두 번째 "*" 은 "in" 이 나타날 때까지 모든 글자를 건너뛰라는 것을 의미한다. 대상 문장에 이와 같은 정규표현식을 적용하여 괄호에 해당하는 개체를 찾아내어 관계에 참여하는 개체의 쌍을 추출한다.

WHISK의 개체 쌍 추출 규칙은 (그림 3)과 같은 알고리즘에 의해 학습된다.

```
WHISK (Reservoir)
  RuleSet = NULL
  Training = NULL
  사용자의 요구에 따라 반복 수행
     Reservoir에서 새로운 데이터 집합인 NewInst를 선택
     (사용자가 NewInst의 정답 관계를 태깅)
     Training에 NewInst를 추가
     NewInst에 적용했을 때 오류가 있는 규칙은 제거
     Training 집합 내부의 각각의 데이터 Inst에 대해
        Inst가 가지는 각각의 정답 관계 Tag에 대해
           만일 Tag가 RuleSet에 의해 커버되지 않을 경우
              Rule = GROW_RULE(Inst, Tag, Training)
  RuleSet을 간결하게 다듬는다.
```

그림 3 top-level에서 본 WHISK 알고리즘

WHISK 알고리즘은 사용자가 정답 관계가 태깅되지 않은 문장들(Reservoir)로부터 새로운 학습 데이터(NewInst)를 입력할 때마다 반복을 수행한다. 현재 가지고 있는 규칙 집합으로 학습 데이터의 문장에서 정답 개체 쌍을 추출해 낼 수 없으면 그 문장으로 새로운 개체 쌍 추출 규칙을 학습한다.

```
GROW_RULE(Inst, Tag, Training)
  Rule = empty rule (단어들이 * 로 치환된 규칙)
  각각의 추출 슬롯 i에 대해
     ANCHOR(Rule, Inst, Tag, Training, i)
  추출 규칙을 학습데이터에 적용 시 오류가 나지 않거나
        laplacian 값의 변화 가 없을 때까지 반복
     EXTEND_RULE(Rule, Inst, Tag, Training)
```

그림 4 개체 쌍 추출 규칙을 만드는 Grow Rule 부분

WHISK 알고리즘은 빈 규칙(empty rule)으로부터 규칙을 키워 나간다(그림 4). 빈 규칙에서 추출 슬롯을 고정하여 가장 일반적인 규칙을 만들어야 하는데, WHISK 알고리즘은 두 가지의 추출 슬롯 고정 방법 중 더 많은 학습 데이터를 커버할 수 있는 방법으로 슬롯을 고정한다. 첫 번째 방법은 규칙의 추출 슬롯에 정답 개체 쌍이 속하는 개체 유형을 더함으로써 추출 슬롯을 고정하는 것이고, 두 번째는 규칙의 추출 슬롯에 인접한 앞뒤 두 단어를 더함으로써 추출 슬롯을 결정하는 것이다. 이 두 가지 방법을 추출 슬롯의 개수(개체 쌍이므로 두 개)만큼 수행하여 규칙의 추출 슬롯을 고정한다(그림5)(그림6).

```
ANCHOR(Rule, Inst, Tag, Training, i)
  Base_1 = Rule + i 번째 추출 슬롯 내부의 단어
  Base_1을 i 번째 슬롯까지 학습 데이터에 적용시켜 본다.
  Base_1이 정답을 커버하지 못할 때까지
    EXTEND_RULE(Base_1, Inst, Tag, Training)
  Base_2 = Rule + i 번째 추출 슬롯 직전, 직후의 단어
  Base_2를 i 번째 슬롯까지 학습 데이터에 적용시켜 본다.
  Base_2가 정답을 커버하지 못할 때까지
    EXTEND_RULE(Base_2, Inst, Tag, Training)
  Rule = Base_1
  만일 Base_2가 Base_1에 비해 더욱 많은 학습 데이터를
        커버할 경우
    Rule = Base_2
```

그림 5 개체 쌍 추출 규칙의 추출 슬롯을 고정하는 ANCHOR 부분

```
문장 : they fought the civil war on nearly 700 ships of
       the union navy.
개체 : ships (VEH), union (GPE), navy (ORG)
정답 개체쌍 : ships - navy
empty rule : * (*) * (*) *

추출 슬롯 1 고정:
Base_1: * (VEH)
Base_2: 700 (*) of

추출 슬롯 2 고정:
Base_1: * (VEH) * (ORG) *
Base_2: * (VEH) * GPE (*) .
```

그림 6 개체 쌍 추출 규칙의 추출 슬롯 고정 과정의 예

그리고 EXTEND_RULE 과정을 통해 한번에 해당 문장의 한 단어씩을 규칙에 추가하여 규칙을 확장시킨다(그림 7).

```
EXTEND_RULE(Rule, Inst, Tag, Training)
  Best_Rule = NULL
  Best_L = 1.0
  Rule의 laplacian값이 오류 허용 범위 이내일 경우
    Best_Rule = Rule
    Best_L = Laplacian of Rule
  Inst의 각 단어 Term에 대해
    Proposed = Rule + Term
    규칙 Proposed를 학습데이터에 적용시켜 본다.
    Proposed의 Laplacian < Best_L 일 경우
      Best_Rule = Proposed
      Best_L = Proposed의 Laplacian 값
  Rule = Best_Rule
```

그림 7 개체 쌍 추출 규칙 확장 부분

개체 쌍 추출 규칙을 확장할 때 기준이 되는 것은 규칙의 laplacian 값이다. 규칙의 laplacian 값은 (그림8)과 같이 계산되며 그 값이 작을수록 오류가 적은 좋은 규칙임을 의미한다. 규칙의 Laplacian 값은 규칙 학습 후 그 규칙의 점수가 된다.

$$Laplacian = \frac{e+1}{n+1}$$

n : 규칙이 학습 데이터에서 적용되어 개체 쌍을 추출한 횟수
e : 개체 쌍을 추출했을 때 오류가 난 추출 횟수

그림 8 규칙의 Laplacian 값

3.3 개체 간 관계 유형 분류를 위한 최대 엔트로피 모델

개체 간 관계의 유형을 파악하기 위해, 여러 자연언어 처리 문제에서 뛰어난 성능을 보여주었고 많은 분야에서 적용되고 있는 최대 엔트로피 모델을 사용하였다. 최대 엔트로피 모델은 여러 가지의 후보 해(solution)가 있을 때 특정한 단서가 없으면 한 해가 다른 해와 같은 가능성을 가져야 한다는 직관을 구현한 지수 모델이다[1][7].

자질	설명
Words	M1, M2 자체에 나타나는 단어들
	M1, M2 사이에 나타나는 단어들
Entity Type	M1과 M2의 개체 유형
Mention Level	M1과 M2의 언급 정도 (명사, 대명사, 명사 상당어 중 하나)
Overlap	M1, M2 사이에 나타나는 단어의 개수
	M1, M2 사이에 나타나는 개체의 개수
	M1과 M2가 같은 명사구에 속하는지 여부
POS tag	M1 직전 단어의 품사 태그
	M2 직후 단어의 품사 태그

표 1 관계 유형 분류를 위한 최대 엔트로피 모델에 사용되는 자질 함수를 구하기 위한 언어학 속성

최대 엔트로피 모델을 이용하기 위해서는 텍스트로부터 어떤 자질을 선택 할 것인가에 대한 문제를 해결해야 한다. 관계에 참여하는 개체 중 문장에서 먼저 나타나는 개체의 언급(mention)을 M1, 그리고 나중에 나타나는 개체의 언급을 M2라고 하자. 본 논문에서는 최대 엔트로피 모델의 자질 함수를 구성하기 위해서 (표1)와 같은 언어학적 속성들을 사용하였다.

예를 들어 문장에서 다음과 같은 부분이 나타났다고 하자.

문장 : ..... 700 ***ships*** of the ***union*** ***navy*** . .....
품사 : CD NNS IN DT NN NN PUNC.

각각 단어의 아래에 있는 품사가 그 단어의 품사 태그이다. 밑줄이 그어져 있고 이탤릭체로 표시된 단어는 개체를 의미한다. 여기에서 관계에 참여하는 두 개체는 굵은 글씨체로 표기된 "ships"와 "navy"이다. 이때에 이 문장에서 이 두 개체의 관계를 분류하기 위해 사용되는 자질은 다음과 같다.

Words : $ships_{m1}$, of_{bw}, the_{bw}, $union_{bw}$, $navy_{m2}$
Entity Type : $VEHICLE_{m1}$, $ORGANIZATION_{m2}$
Mention Level : $NAME_{m1}$, $NAME_{m2}$
Overlap : three-words-apart, one-mention-in-between, in-same-noun-phrase
POS tag : $CD_{m1\text{-}before}$, $PUNC._{m2\text{-}after}$

본 논문에서는 위와 같은 언어학 속성을 이용하여 자질 함수를 구성하였고, 이를 이용해서 최대 엔트로피 모델을 훈련하였다.

4. 실험 및 분석

4.1 실험 준비

실험을 위해 사용된 코퍼스는 ACE-3 pilot 코퍼스이다. (표2)는 ACE-3 pilot 코퍼스에서 나타나는 개체 간 관계의 유형과 관계의 출현 빈도를 나타내고 있다.
ACE-3 pilot 코퍼스는 CNN, ABC 등의 방송사와 신문사의 뉴스로 이루어져 있으며 총 27개의 뉴스 전문에 개체 및 관계가 태깅되어 있다. 총 487문장으로 이루어져 있으며, 1898개의 개체와 355개의 개체 간 관계가 태깅되어 있다.

4.2 실험 결과

실험은 크게 두 가지의 경우를 비교하는 형식으로 행하였다. 첫 번째는 최대 엔트로피 모델만을 이용하여 관

유형	세분화된 유형	출현빈도
VERT	Staff	42
	Leader	29
	Citizen	11
	Owner	15
	Member	14
	Family	2
	Management	6
	Client	4
	Subordinate	2
HORI	Family	11
	Other	4
	Associate	2
AT	Located	87
	Near	12
	Base-In	13
	Residence	3
	Part-Whole	37
	Other	14
DISC		47

표 2 ACE-3 pilot 코퍼스의 관계 유형 및 출현빈도

계 발견 및 유형 파악을 하는 경우이고, 두 번째는 WHISK 알고리즘을 이용하여 관계가 있다고 판단되는 개체 쌍을 추출한 뒤, 그 개체 쌍에 대해 최대 엔트로피 모델을 적용해서 적합한 관계 유형을 파악한 경우 이다. 관계의 첫 번째 인자와 두 번째 인자, 그리고 그 두 개체 사이의 관계의 유형(세분화된 유형)이 정답과 같을 때 바르게 추출되었다고 평가하였다. 코퍼스의 크기가 작기 때문에 학습용으로 사용된 데이터의 크기도 적었다. 작은 크기의 학습 데이터에 대해서 outer test는 큰 의미가 없으므로 inner test를 수행하였다. 시스템의 성능은 재현율(recall)과 정확률(precision)을 이용한 F-score로 비교하였다. 실험 결과는 (표3)와 같다.

	Rule Score Threshold	Recall (%)	Precision (%)	F-Score
WHISK + ME	0.8	83.38	78.72	80.98
	0.85	87.61	78.93	83.04
	0.9	90.14	78.24	83.77
	0.95	**90.42**	**78.10**	**83.81**
	1.0	90.42	77.53	83.49
ME only		**85.07**	**74.38**	**79.37**

표 3 ACE-3 pilot 코퍼스에 적용한 실험 결과

Rule Score Threshold는 WHISK에 의해 학습된 규칙 중 개체 쌍 추출에 사용할 규칙의 최소 점수를 말한다. 이 값은 개체 쌍 추출 규칙의 laplacian 값으로 각 규칙들은 0과 1 사이의 값을 가지며 0에 가까울수록 더 많은 학습데이터를 커버 할 수 있는 좋은 규칙이다. Rule Score Threshold의 값이 1.0이라는 것은 학습된 모든 추출 규칙을 사용하겠다는 의미이고, 그 값이 0.8일 때에는 0.8 이하의 laplacian 값을 가지는 규칙들만 개체 쌍 추출에 사용하겠다는 의미이다. 너무 많은 규칙을 사용하면 정답이 아닌 개체 쌍이 추출될 가능성이 커지므로 적당한 개수로 규칙의 개수를 줄여야 한다.

WHISK 알고리즘과 최대 엔트로피 모델을 모두 이용하여 관계를 추출한 경우 F-Score의 값이 최대 83.81이었다. 반면 최대 엔트로피 모델만으로 관계를 추출한 경우 F-Score의 값이 79.37이었다. 그리고 WHISK를 도입한 경우가 최대 엔트로피 모델만을 사용한 경우에 비해 재현율은 5.37(%), 정확률은 3.72(%) 상승하였다. WHISK를 활용한 개체 쌍 자동 추출이 전체 시스템의 성능 향상에 기여했다고 볼 수 있다.

WHISK 알고리즘과 최대 엔트로피 모델을 모두 사용하여 수행한 실험의 경우에서 Rule Score Threshold 값에 따라 정확률의 값은 큰 변동이 없으나 재현율의 값은 변화가 있음을 알 수 있다. 이는 Rule Score Threshold의 값이 커짐에 따라 개체 쌍 추출에 사용되는 규칙의 수도 증가하여 보다 많은 수의 개체 쌍이 추출되기 때문이다. 하지만 그 값이 0.95일때에 비해 1.0일 때에는 적합하지 않은 규칙까지 추출에 사용되므로 잘못된 개체 쌍들을 추출하여 정확률의 값이 감소하였음을 볼 수 있다.

5. 결론 및 향후 계획

본 논문에서는 텍스트로부터 개체 간 관계를 추출하는 새로운 방법을 제시하였다. 현재 존재하는 자질 기반의 시스템이 가지는 낮은 재현율의 한계를 극복하기 위해 자동으로 개체 쌍 추출 규칙을 학습할 수 있는 WHISK 알고리즘을 도입하여 관계를 가지는 개체 쌍을 추출해 내었다. 실험 결과에서도 볼 수 있듯이 자질 기반의 시스템만을 사용하여 개체 간 관계를 발견하고 관계 유형을 파악하는 시스템에 비해 WHISK를 도입한 시스템이 더 큰 F-score를 보여준다. 특히 이전의 자질 기반 시스템의 한계를 어느 정도 극복하여 상승하게 된 재현율이 시스템의 성능 향상에 많은 부분 기여 한 것을 알 수 있다.

하지만 실험한 코퍼스는 그 크기가 작게 한정되어 있었기 때문에 더욱 큰 데이터에 대해 결과를 내어 보아야 할 것으로 생각된다. 보다 큰 코퍼스에 대해 실험을 할 경우 더욱 일반적이고 많은 문장을 커버할 수 있는 개체 쌍 추출 규칙을 학습할 가능성이 커지게 된다. 따라서 더욱 큰 코퍼스에 대해 실험할 경우에도, 기존의 자질기반 시스템에 비해 재현율의 상승을 기대해 볼 수 있다.

또한 개체 간 관계 유형 파악을 위한 자질 함수로 사용되는 언어학 속성을 좀 더 다양화 할 필요가 있다. 최근에 높은 성능을 얻고 있는 파서를 이용한 자질을 추가할 필요가 있으며, 문장 각 부분의 의미적 역할에 관한 정보도 개체간 관계 유형 파악을 위한 자질 함수로 사용될 수 있다.

참고문헌

[1]Berger A., Pietra S., and Pietra V. 1996. A

Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics,* Vol. 22, No. 1, 1996

[2]Culotta A. and Sorensen J. 2004. Dependency tree kernels for relation extraction. In *Proceedings of 42th Annual Meeting of the Association for Computational Linguistics.* 21-26 July 2004. Barcelona, Spain.

[3]Kambhatla N., 2004. Combining lexical, syntactic and semantic features with Maximum Entropy models for extracting relations. In *Proceedings of 42th Annual Meeting of the Association for Computational Linguistics.* 21-26 July 2004. Barcelona, Spain.

[4]Miller S., Fox H., Ramshaw L. and Weischedel R., 2000. A novel use of statistical parsing to extract information from text. In *1st Meeting of the North American chapter of the Association for Computational Linguistics,* pages 226-233, seattle, Washington, April 29-May 4 2000.

[5]Soderland S., 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34, 233-72

[6]ZHOU G. et al., 2005. Exploring Various Knowledge in Relation Extraction. In P*roceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 427-434.

[7]박성배, 장병탁 2001. 최대 엔트로피 모델을 이용한 텍스트 단위화 학습. 제 13회 한글 및 한국어정보 처리 학술 대회 논문집 pages 130-137.