구문 분석 결과를 이용한 한국어 무제한 정보추출

김병수○, 유환조, 이근배
포항공과대학교 컴퓨터공학과
{bsmail90, hwanjoyu, gblee}@postech.ac.kr

A Syntax-Based Hybrid System for Korean Open Information Extraction

Byungsoo Kim○, Hwanjo Yu, Gary Geunbae Lee
Department of Computer Science and Engineering, POSTECH

요 약

무제한 정보추출은 주로 영어를 대상으로 연구가 진행 되었지만, 최근에는 영어가 아닌 다른 언어에 대한 적용이 시도되고 있다. 본 논문에서는 관계 어휘의 유형을 동사형과 명사형 2가지로 정의하고, 각 유형별로 구문 분석 결과 기반의 서로 다른 방법론을 적용하는 한국어 대상 무제한 정보추출 시스템을 소개한다. 동사형 관계 어휘에 대해서는 의존 관계 기반의 추출 규칙을 적용하고, 명사형 관계 어휘에 대해서는 대량의 말뭉치로부터 자동으로 학습한 의존 관계 구조 기반의 추출 패턴을 적용한다. 임의의 100개 문장에 대해서 수행한 결과는 산출된 전체 트리플에 대해 0.8이상의 정밀도를 보임으로써 본 논문에서 제안하는 방법의 효용성을 증명하였다.

주제어: 무제한 정보추출, 구문 분석, 한국어

1. 서론

무제한 정보추출은 입력 문장에서 개체 및 개체 간의 관계를 추출하는 작업이다. 무제한 정보추출 시스템에서는 이러한 정보를 <개체1; 관계; 개체2; 개체3; ...>과 같은 n-튜플(Tuple)이나 <개체1; 관계; 개체2>와 같은 트리플(Triple)의 형태로 추출한다. 예를 들어 '알베르트 아인슈타인은 1879년 5월 14일 독일 울름에서 태어났다.' 라는 입력 문장이 주어졌을 경우 무제한 정보추출 시스템은 <알베르트 아인슈타인; 태어났다; 1879년 5월 14일; 독일 울름>과 같은 n-튜플을 추출하거나 <알베르트 아인슈타인; 태어났다; 1879년 5월 14일>, <알베르트 아인슈타인; 태어났다; 독일 울름>과 같은 트리플을 추출한다.

무제한 정보추출이 기존의 전통적인 정보추출에 대해 갖는 차이점은 개체 간의 관계에 제한이 없다는 것이다. 예를 들어 ACE 2005, SemEval 2010 Task8 에서는 추출할 수 있는 개체 간의 관계를 사전에 정의하였다. 따라서 개체 간의 관계가 미리 정의된 관계 집합 내에 존재하지 않을 경우에는 정확한 관계성을 파악할 수 없다. 또한 모든 가능한 관계를 포함할 수 있도록 관계 집합을 구성하는 것도 불가능하다. 이에 반해 무제한 정보추출은 개체간의 관계를 개체가 포함된 입력 문장에서 찾는다. 따라서 미리 관계 집합을 정의할 필요 없이 관계의 집합은 인간이 사용하는 어휘의 집합으로 확장될 수 있고, 이는 웹(Web) 규모의 정보추출을 가능하게 한다.

다른 자연 언어 처리 관련 작업들과 마찬가지로 무제한 정보추출은 주로 영어를 대상언어로 하여 개발되어 왔다[1,2,3,4,5,6,7,8,9]. 하지만 최근에는 중국어[10,11]나 포르투갈어, 스페인어, 갈리시아어[12]와 같이 영어가 아닌 다른 언어에 대해 무제한 정보추출을 수행하려는 시도가 있다.

본 논문에서는 관계 어휘의 유형을 동사형 관계 어휘와 명사형 관계 어휘 2가지로 정의하고, 입력 문장의 구문분석 결과에 기반 하여 각 유형별로 서로 다른 방법을 적용하는 한국어 대상 무제한 정보추출 시스템을 개발한다. 임의의 100개 문장에 대해 무제한 정보추출을 수행한 결과는 산출된 전체 트리플에 대해 0.8이상의 정밀도를 보임으로써 본 논문에서 제안하는 방법의 효용성을 증명하였다.

본 논문의 구성은 다음과 같다. 2장에서는 무제한 정보추출과 관련된 기존의 작업에 대해 기술하며, 3장에서는 본 논문에서 추출하는 관계 어휘의 유형을 정의한다. 4장에서는 전체 시스템 구조를 제안하고 시스템에서 사용한 방법에 대해서 설명한다. 5장에서는 실험 설계 및 결과에 대해 서술하고, 6장에서는 결론 및 개선 사항에 대해서 논의한다.

2. 관련 연구

무제한 정보추출은 사전에 정의되지 않은 개체간의 관계를 찾기 위해 다양한 방법을 취했었다. TextRunner[1]는 의존 관계 트리에 기반 하여 비 어휘적인(un-Lexicalized) 자질을 사용한 Naive Bayes 분류기를 학습시켰다. WOE^{pos}[2] 또한 분류기를 사용했지만 Wikipedia로부터 얻은 학습 데이터를 이용하여 보다 높은 정밀도와 재현율을 달성하였다. Reverb[3]는 간단한 품사 패턴으로 대부분의 관계성을 나타낼 수 있음을 보였다. [4]는 의미역 결정 결과를 무제한 정보추출의 트리플 형식으로 변환함으로써 정밀도를 향상시켰다. WOE^{parse}[2], OLLIE[5], ReNoun[6]은 사람의 직접적인 개입 없이 자동으로 수많은 학습 데이터를 구축한 후, 추출 패턴을 학습하는 자기 지도적인 방법을 적용하였다. KRAKEN[7], ClausIE[8], EXEMPLAR[9]는 의존 관계 트리

에 기반 한 추출 규칙을 정의하였다.

앞서 언급한 영어 대상 시스템 외에 최근에는 영어가 아닌 다른 언어에 대한 무제한 정보추출이 시도되고 있다. 중국어 대상 무제한 정보추출 시스템들은 입력 문장에 대한 구문 분석 결과를 이용하거나[10], 시맨틱(Semantic) 정보가 더해진 추출 패턴을 이용하였다[11]. [12]는 동일한 작업을 포르투갈어, 스페인어, 갈리시아어에 대해 수행하였다.

3. 관계 어휘의 정의

3.1. 동사형 관계 어휘

개체간의 관계를 나타내는 관계 어휘가 동사적인 성격을 갖는 경우이다. 각 관계에 대해 2개 이상의 개체를 갖는 n-튜플을 형성하며, 무제한 정보추출 시 산출되는 관계의 대부분을 차지한다. 문장 내에서 나타나는 모든 동사는 동사형 관계 어휘의 후보가 된다. 예를 들어 "웅가로는 침실과 식당, 욕실에서 사용하는 갖가지 직물제품을 디자인하여 최근 파리의 갤러리 라파예트 백화점에서 '색의 컬렉션'이라는 이름으로 전시회를 열었다." 라는 문장이 있을 경우 '사용하는, 디자인하여, 이라는, 열었다'가 동사형 관계 어휘의 후보가 된다. 각 후보는 <직물제품; 사용하다; 침실과 식당, 욕실>, <웅가로; 디자인하다; 침실과 식장, 욕실에서 사용하는 갖가지 직물제품>, <이름; 이다; '색의 컬렉션'>, <웅가로; 열었다; 전시회; 최근; 갤러리 라파예트 백화점; '색의 컬렉션'이라는 이름>과 같은 n-튜플을 형성한다.

3개 이상의 개체를 갖는 n-튜플을 트리플로 변환하기 위해 본 논문에서는 관계 구를 구성하는 데에 필수적인 개체를 정한다. 예를 들어 <웅가로; 열었다; 전시회; 최근; 갤러리 라파예트 백화점; '색의 컬렉션'이라는 이름>에서 필수개체는 '전시회'이며, 트리플로 변환 시 <웅가로; 열었다; 갤러리 라파예트 백화점> 보다는 <웅가로; 전시회를 열었다; 갤러리 라파예트 백화점>가 문장내의 정보를 충분히 포함한다. 필수 개체가 트리플의 개체2가 될 경우에는 <웅가로; 열었다; 전시회>와 같이 n-튜플에서의 관계를 그대로 사용한다.

3.2. 명사형 관계 어휘

관계 어휘가 명사인 경우 각 관계에 대해 2개의 개체를 갖는 트리플을 형성한다. 동사형 관계 어휘를 갖는 n-튜플의 경우 개체1은 관계를 수행하는 주체이지만, 명사형 관계 어휘를 갖는 트리플의 경우 개체1은 개체2의 속성을 나타내며 "'개체1'은 '개체2'의 '관계'이다"라는 의미를 포함한다. 예를 들어 "웅가로는 침실과 식당, 욕실에서 사용하는 갖가지 직물제품을 디자인하여 최근 파리의 갤러리 라파예트 백화점에서 '색의 컬렉션'이라는 이름으로 전시회를 열었다."에서 추출될 수 있는 명사형 관계 어휘는 '백화점'이고, <라파예트; 백화점; 파리>와 같은 트리플을 형성한다. 이 트리플이 포함하는 의미는 "'라파예트'는 '파리'의 '백화점'이다."이며, '라파예트'는 '파리'의 '백화점'이라는 속성에 해당하는 값이다. 따라서 <라파예트; 백화점; 파리>는 유효한 트리플이지만 <파리; 백화점; 라파예트>는 틀린 트리플이다.

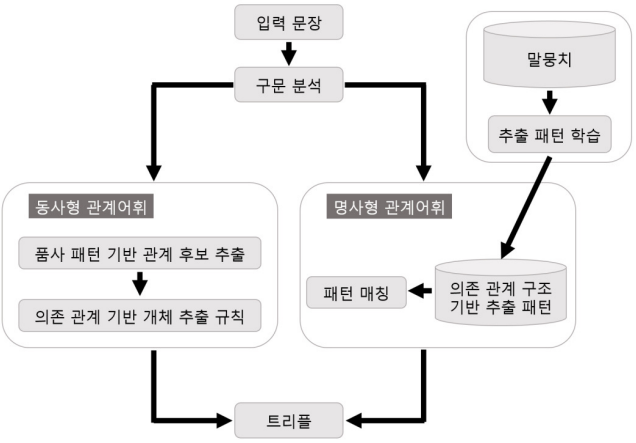

그림 1. 제안하는 전체 시스템 구조

4. 전체 시스템 구조

본 논문에서 제안하는 시스템은 관계 어휘의 유형에 따라 서로 다른 방법론을 적용하여 무제한 정보추출을 수행한다(그림 1). 이 시스템은 입력 문장에 대한 구문 분석을 수행하여 의존 관계 트리를 얻은 후, 의존 관계 기반 추출 규칙을 적용하여 동사형 관계 어휘를 갖는 n-튜플을 추출하고, 명사형 관계 어휘에 대해서는 대량의 말뭉치로부터 사전에 학습한 의존 관계 구조 기반의 추출 패턴을 적용한다. 관계 어휘의 유형 별 각 방법론에 대한 상세한 설명은 아래와 같다.

4.1. 의존 관계 기반 추출 규칙

동사형 관계 어휘를 갖는 n-튜플의 경우, 각 관계에 대한 개체들은 의존 관계 트리 상에서 관계와 자식 혹은 부모를 형성하는 어절들이다. 따라서 본 논문에서는 동사형 품사 패턴을 만족하는 어절의 연속을 동사형 관계 후보로 보고, 의존 관계 트리 상에서 각 후보와 자식 혹은 부모의 관계에 있는 어절 중 개체를 형성하는 것을 찾기 위해 의존 관계에 대한 규칙을 적용한다.

관계의 후보는 문장 내에서 동사형 품사 패턴을 갖는 어절의 연속이다. (그림 2)에서와 같은 입력 문장이 주어졌을 경우 '사용하는'의 품사는 '사용/NNG|하/XSV|는/ETM'인데 XSV에 의해서 동사적인 성격을 갖게 되므로 관계 후보가 된다. '디자인하여'의 품사는 '디자인/NNG|하/XSV|아/EC'인데 '사용하는'의 경우와 마찬가지로 XSV에 의해서 동사적인 성격을 갖게 되므로 관계 후보가 된다. '이라는'의 품사는 '이/VCP|라는/ETM'인데 VCP에 의해서 동사적인 성격을 갖게 되고, '열었다'는 품사 '열/VV|었/EP|다/EF'에서 VV로 인해 동사

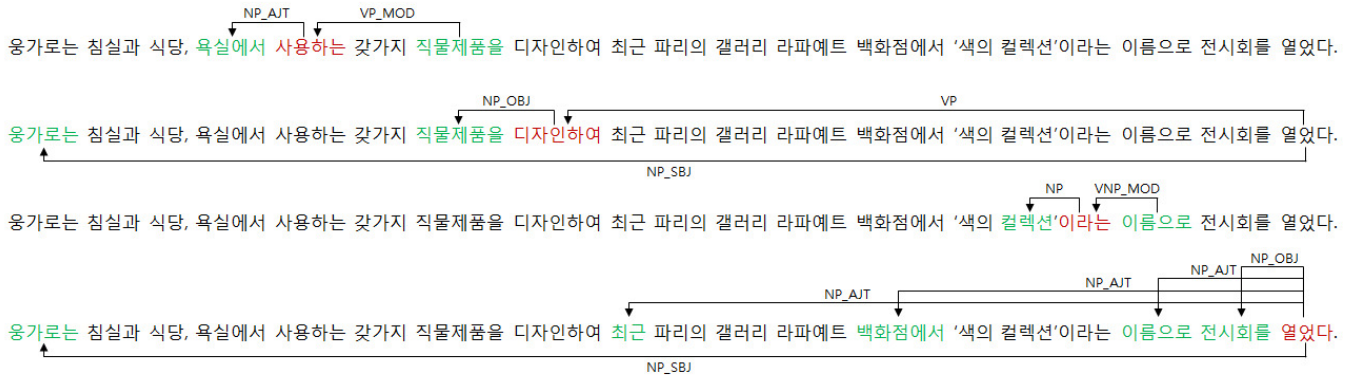

그림 2. 예시 문장 내의 각 동사형 관계 후보 및 개체와의 의존 관계

적인 성격을 갖게 되어 관계 후보가 된다.

관계의 후보가 정해지면 의존 관계 트리 상에서 각 후보가 자식 혹은 부모와 형성하는 의존 관계에 따라 개체를 결정한다. '열었다'의 경우 '웅가로는'과 NP_SBJ 의존 관계를 형성하여 개체1이 된다. '전시회를'은 NP_OBJ 의존 관계를 형성하여 개체2가 된다. 또한 OBJ는 개체1과 다른 개체들의 관계를 형성하는데 있어서 필수적인 정보를 제공하기 때문에 '전시회를'은 '열었다'의 필수 개체가 된다. '최근', '백화점에서', '이름으로'는 NP_AJT 의존 관계를 형성하여 개체3, 개체4, 개체5가 된다. 이로부터 추출되는 n-튜플은 <웅가로는; 열었다; 전시회를; 최근; 백화점에서; 이름으로>이며, 이를 트리플로 변환 시 <웅가로는; 전시회를 열었다; 최근>, <웅가로는; 전시회를 열었다; 백화점에서>, <웅가로는; 전시회를 열었다; 이름으로>가 된다.

'디자인하여'의 경우 '웅가로는'과 직접적으로 부모 혹은 자식관계를 형성하지는 않지만 부모인 '열었다'로부터 '웅가로는'과의 의존 관계를 상속 받아서 '웅가로는'이 '디자인하여'의 개체1이 된다. '직물제품을'은 NP_OBJ 의존 관계를 형성하여 개체2 및 필수 개체가 된다. 이로부터 추출되는 n-튜플은 <웅가로는; 디자인하여; 직물제품을>이며, 2개의 개체를 갖기 때문에 트리플은 n-튜플과 같다.

'사용하는'및 '이라는'의 경우 '사용/NNG|하/XSV|는/ETM', '이/VCP|라는/ETM'과 같은 품사를 갖는데, 이와 같이 ETM을 품사로 갖는 관계의 경우 부모가 개체1이 되고, 자식이 개체2가 된다. 이로부터 추출되는 n-튜플은 <직물제품을; 사용하는; 욕실에서>와 <이름으로; 이라는; 컬렉션>이며, 트리플은 n-튜플과 같다.

4.2. 의존 관계 구조 기반의 추출 패턴

자식 및 부모와의 의존 관계를 통해서 개체를 결정할 수 있는 동사형 관계 어휘의 경우와는 달리 명사형 관계 어휘의 경우 문장 내에서 그 관계가 유연하게 나타나기 때문에 규칙을 정의하는 것은 불가능하다. 따라서 본 논문에서는 명사형 관계 및 개체를 추출할 수 있는 의존 관계 구조 기반의 추출 패턴을 대량의 말뭉치로부터 학습하고, 학습된 추출 패턴을 적용하여 관계를 추출하는 방법을 취한다.

(그림 3)은 추출 패턴의 학습과 학습된 패턴을 적용하여 관계 및 개체를 추출하는 과정을 나타낸다. 추출 패턴의 학습은 씨드(Seed) 트리플을 추출하는 것으로부터 시작된다. 씨드 트리플을 추출하기 위해서 본 논문에서는 간단한 씨드 트리플 추출 패턴을 정의한다. 대량의 말뭉치에 씨드 트리플 추출 패턴을 적용하여 씨드 트리플이 추출되면 씨드 트리플의 개체1, 개체2, 관계를 포함하는 문장을 대량의 말뭉치에서 검색한다. 본 논문에서는 이렇게 검색된 문장이 씨드 트리플이 추출된 문장에서의 개체1과 개체2의 관계성을 포함한다고 가정한다. 따라서 '... 그의 일생은 여자 관계만 빼놓으면 한국의 이순신 장군과 비슷하다 ...'에서 <이순신; 장군; 한국>이 나타내는 관계성은 '... 한국의 역사적인 영웅 이순신 장군을 구하기 위해 ...'에서 <이순신; 장군; 한국>이 나타내는 관계성과 같다고 가정한다. 이와 같이 다른 문장 구조로 나타나는 같은 관계성을 추출하기 위해 검색된 문장에서 개체1, 개체2, 관계를 연결하는 의존 관계 트리 경로를 추출 패턴으로 학습한다. 학습된 추출 패턴 중 같은 의존 관계 트리 경로를 공유하지만 다른 관계 어휘를 갖는 패턴들은 일반화의 대상이 된다.

학습된 추출 패턴들은 대량의 말뭉치로부터 자동으로 구축되었기 때문에 오류를 포함할 수 있다. 따라서 본 논문에서는 각 추출 패턴의 정확도를 나타내기 위한 점수를 할당한다. 추출 패턴이 말뭉치 내에서 자주 나타났다면 해당 패턴은 유효한 트리플을 추출할 가능성이 높다는 것을 의미하며 이것은 추출 패턴의 일반성을 나타낸다. 이로부터 본 논문은 추출 패턴이 말뭉치 내에서 나타난 횟수를 기록한다. 하지만 추출 패턴이 말뭉치 내에서 자주 나타났다는 것은 그만큼 잘못된 트리플을 추출할 가능성도 많다는 것을 의미하므로 본 논문은 각 추출 패턴의 고유성에 대한 정보를 포함하기 위해 패턴 내의 관계 어휘들의 의미적 유사도를 측정한다. 의미적 유사도는 관계 어휘들을 벡터로 변환한 다음 각 벡터간의 코사인 유사도를 측정하여 구한다. 본 논문에서는 관계 어휘들을 벡터로 변환하기 위해 Word2Vec[1]을 이용한다. 일반성과 고유성을 통한 추출 패턴의 정확도는 추출 패

1) https://code.google.com/p/word2vec/

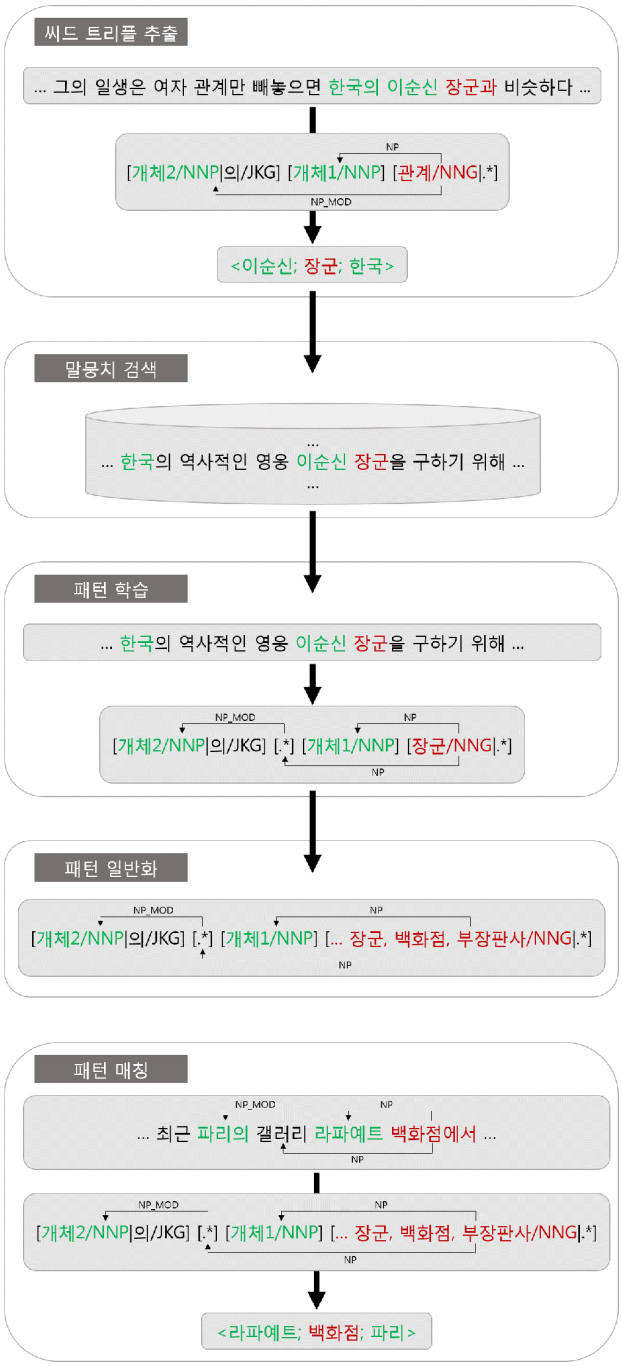

그림 3. 추출 패턴의 학습 및 적용

턴의 출현 빈도와 추출 패턴 내의 관계 어휘의 의미적 유사도간의 곱으로 나타난다. 본 논문에서는 일정 수치 이상의 정확도를 가지는 추출 패턴만을 실제 무제한 정보 추출에 이용한다.

4.3. 정보 확장 및 정규화

위의 각 과정을 거쳐서 추출된 트리플들의 관계 및 개체들은 하나의 어절만을 포함하며 문장 내에서 나타나는 형태 그대로 구성된다. 본 논문에서는 트리플들이 입력 문장 내의 모든 정보를 포함하며, 통일된 형식을 갖도록 정보 확장 및 정규화를 수행한다.

정보 확장은 의존 관계 트리에서 개체의 자손들을 모두 포함시킨다. (그림 2)에서 추출되는 트리플 중 <웅가로는; 디자인하여; 직물제품을>을 예로 들면 '웅가로는'의 자손들을 모두 개체1에 포함시키고, '직물제품을'의 자손들을 모두 개체2에 포함시킨다. 그 결과 정보 확장된 트리플은 <웅가로는; 디자인하여; 침실과 식장, 욕실에서 사용하는 갖가지 직물제품을>이 된다.

정규화는 개체 및 관계에 대해서 서로 다른 방법으로 수행되는데, 개체의 경우 끝에 붙는 모든 조사를 제거하며, 관계의 경우 어미를 모두 '다'로 바꾼다. 그 결과 정규화된 트리플은 <웅가로; 디자인하다; 침실과 식장, 욕실에서 사용하는 갖가지 직물제품>이 된다.

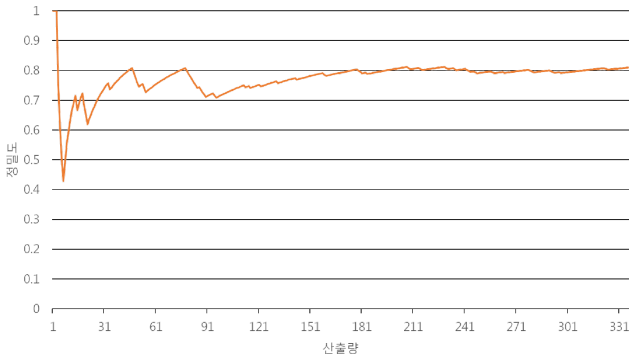

그림 4. 트리플의 산출량에 따른 정밀도

5. 실험 설계 및 결과

본 논문에서는 추출 패턴의 학습을 위해 인터넷 기사로부터 크롤링(Crawling)한 2천 4백만 개의 문장을 사용하였다. 또한 동일한 말뭉치에 대해서 Word2Vec의 학습을 수행하였다. 구문 분석을 위해서는 ETRI 언어 분석기를 사용하였다. 시스템의 테스트를 위해서는 세종 말뭉치의 한영 병렬 말뭉치 내의 한국어 문장 중 임의로 100개를 선택하여 테스트를 수행하였다. 각 테스트 문장에 대해서 추출된 트리플 내의 개체간의 관계가 입력 문장에서 나타날 경우 유효 트리플로 간주하였고, 만약 개체간의 관계가 잘못되었거나 혹은 개체가 잘못되었을 경우 틀린 트리플로 간주하였다. 100문장으로부터 전체 336개의 트리플이 추출되었으며 이중 272개의 트리플이 유효한 트리플로 간주되어 추출된 전체 트리플에 대한 정밀도는 0.81을 보였다.

(그림 4)는 트리플의 산출량에 따른 정밀도를 나타낸다. 무제한 정보추출의 특성 상 문장으로부터 추출되는 트리플의 정답을 얻기 어려우므로 본 논문에서는 각 트리플에 점수를 할당한 후, 점수에 따라 전체 트리플을 내림차순으로 정렬하여 산출량에 따른 정밀도를 측정하였다. 본 논문에서 제안하는 방법은 입력 문장의 의존

관계 분석 결과에 많이 의존하므로, 트리플의 점수는 입력 문장의 의존 관계 분석 점수를 사용하였다. 산출량에 따른 정밀도는 100개 지점까지는 상승과 하강을 반복하다가 그 이후 0.8로 수렴하는 경향을 보인다. 특히 10개 지점에는 0.45 정도의 정밀도를 보이는데 트리플에 대한 점수 할당 방법 개선의 필요함을 나타낸다.

6. 결론

본 논문에서는 관계 어휘의 유형을 동사형과 명사형으로 구분하여 서로 다른 방법론을 취하는 시스템을 제안하였다. 동사형 관계어휘에 대해서는 동사형 품사 패턴을 적용하여 관계 후보를 얻은 뒤, 의존 관계 기반 추출 규칙을 적용하여 각 관계 후보의 개체들을 파악하였다. 명사형 관계어휘에 대해서는 대량의 말뭉치로부터 자동으로 학습한 의존 관계 구조 기반의 추출 패턴을 적용하였다. 임의의 100개 문장에 대하여 무제한 정보추출을 수행한 결과는 0.8 이상의 정밀도를 보임으로써 본 논문에서 제안하는 방법의 효용성을 증명하였다.

* 본 연구는 산업통상자원부의 우수기술연구센터사업[10048448, 링크드 데이터 기반 대화형 질의응답 검색 프레임워크 개발]의 일환으로 수행하였음

참고문헌

[1] Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni, “Open Information Extraction from the Web.”, Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007.

[2] Fei Wu and Daniel S. Weld, “Open Information Extraction using Wikipedia.”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010.

[3] Anthony Fader, Stephen Soderland, and Oren Etzioni, “Identifying Relations for Open Information Extraction.”, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011.

[4] Janara Christensen, Mausam, Stephen Soderland, Oren Etzioni, “An analysis of open information extraction based on semantic role labeling”, Proceedings of the 6th international conference on Knowledge capture, 2011.

[5] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni, “Open Language Learning for Information Extraction.”, Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing, 2012.

[6] Mohamed Yahya, Steven Whang, Rahul Gupta, Alon Halevy, “ReNoun: Fact Extraction for Nominal Attributes”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014.

[7] Alan Akbik and Alexander Loser, “KRAKEN: N-ARY FACTS IN OPEN INFORMATION EXTRACTION”, In Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction, 2012.

[8] Luciano Del Corro, Rainer Gemulla, "ClausIE: Clause-Based Open Information Extraction", Proceedings of the 22nd International Conference on World Wide Web, 2013.

[9] Filipe de Sá Mesquita, Jordan Schmidek, Denilson Barbosa, “Effectiveness and Efficiency of Open Relation Extraction.”, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013.

[10] Yuen-Hsien Tseng, Lung-Hao Lee, Shu-Yen Lin, Bo-Shun Liao, Mei-Jun Liu, Hsin-Hsi Chen, Oren Etzioni, Anthony Fader, "Chinese Open Relation Extraction for Knowledge Acquisition", Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014.

[11] Likun Qiu, Yue Zhang, "ZORE: A Syntax-based System for Chinese Open Relation Extraction", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014.

[12] Pablo Gamallo, Marcos Garcia, "Dependency-Based Open Information Extraction", In Proceedings of the Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP, 2012.