문장으로부터 여러 단어로 구성된 여러 유형의 요소 추출

양선○, 고영중
동아대학교 컴퓨터공학과
seony.yang@gmail.com, youngjoong.ko@gmail.com

Extracting Multi-type Elements Consisting of Multi-words from Sentences

Seon Yang○, Youngjoong Ko
Department of Computer Engineering, Dong-A University

요 약

문장을 대상으로 특정 응용 분야에 필요한 요소를 자동으로 추출하는 정보 추출(information extraction) 과제는 자연어 처리 및 텍스트 마이닝의 중요한 과제 중 하나이다. 특히 추출해야할 요소가 한 단어가 아닌 여러 단어로 구성된 경우 추출 과정에서 고려되어야할 부분이 크게 증가한다. 또한 추출 대상이 되는 요소의 유형 또한 여러 가지인데, 감정 분석 분야를 예로 들면 화자, 객체, 속성 등 여러 유형의 요소에 대한 분석이 필요하며, 비교 마이닝 분야를 예로 들면 비교 주체, 비교 상대, 비교 술어 등의 요소에 대한 분석이 필요하다. 본 논문에서는 각각 여러 단어로 구성될 수 있는 여러 유형의 요소를 동시에 추출하는 방법을 제안한다. 제안 방법은 구현이 매우 간단하다는 장점을 가지는데, 필요한 과정은 형태소 부착과 변환 기반 학습(transformation-based learning) 두 가지이며, 파싱 혹은 청킹 같은 별도의 전처리 과정도 거치지 않는다. 평가를 위해 제안 방법을 적용하여 비교 마이닝을 수행하였는데, 비교 문장으로부터 각자 여러 단어로 구성될 수 있는 세 가지 유형의 비교 요소를 자동 추출하였으며, 실험 결과 정확도 84.33%의 우수한 성능을 산출하였다.

주제어: 정보 추출, 텍스트 마이닝, 여러 단어 요소, 변환 기반 학습

1. 서론

정보 추출(information extraction) 과제의 어려움 중 하나로 단어 하나로 된 구성된 요소뿐만 아니라 여러 단어로 구성된 요소(multi-word element. 이하 'MultiEL'로 표기함)를 자동으로 식별 및 추출하는 일을 들 수 있다. 실제로 사람들이 문장에서 추출하고 싶은 정보가 한 단어로 국한되는 경우보다는 둘 이상의 단어(구 단위, 혹은 그 이상)로 구성된 정보까지 포함해서 원하는 경우가 대다수이기 때문에, MultiEL의 자동 감지 및 추출은 자연어 처리 (NLP) 및 텍스트 마이닝 분야의 매우 중요한 주제 중 하나이다.

그 동안 여러 단어로 이루어진 표현 문구에 대한 연구가 계속해서 발표되어 왔다. 하지만 선행 연구들은 별도로 구축된 사전을 사용하는 경우가 많았다. 물론 사전의 사용은 기본적으로 어느 정도의 성능을 보장하는 매우 요긴한 방법이긴 하지만, 사전을 구축하는 일은 별도의 시간과 노력을 필요로 하는 고비용 작업이다. 파싱(parsing) 또한 선행 연구들에서 주로 사용된 방법 중 하나인데, 정확한 파싱 결과는 MultiEL 추출에 매우 핵심적으로 사용될 수 있음이 분명하다. 다만, 한국어는 언어 특성상 파싱이 상대적으로 어렵다고 알려져 있으며, 이로 인해 전처리 과정으로 파싱을 사용하였을 경우 최종 성능 면에서 영어 등의 타 언어에 비해 불리할 결과를 산출할 가능성을 배재할 수 없음 또한 사실이다.

본 연구에서는 어떠한 사전도 사용하지 않고 또한 파싱 과정도 거치지 않는 MultiEL 추출 기법을 제안한다. 전처리 과정으로 파싱 외에 청킹(chunking)을 고려할 수 있으나, 본 연구에서는 전처리 과정을 최대한 생략함으로써 전처리 오류가 최종 성능에 미치는 영향을 최소화하고자 하였다. 따라서 제안 방법에서는 전처리 과정에서 형태소 부착(part-of-speech tagging)만을 수행하며, 그 후 변환 기반 학습 (transformaion-based learning, TBL)을 수행하여 MultiEL 추출을 진행한다.

또한 본 연구에서는 여러 유형의 요소를 동시에 추출한다. 실제로, 하나의 요소 자체도 여러 단어로 구성될 수 있지만, 추출해야하는 요소의 유형도 여러 가지로 다양하다는 점에 주목하였다. 감정 분석(sentiment analysis) 부문을 예로 들면, 하나의 요소만을 분석하는 것이 아니라, 화자(holder), 객체(object), 속성(sttribute), 강도(strength) 등의 여러 유형의 요소를 분석한다. 의견 마이닝(opinion mining) 등 다른 부문에서도 마찬가지이다. 이러한 여러 유형 요소를 추출하기 위해서는 전체적으로 추출 과정이 복잡해지거나 여러 단계를 거쳐야 하는 경우가 발생한다. 반면, 본 연구에서 제안하는 방법은 구현이 단순하면서도 동시에 여러 유형의 요소를 동시에 추출할 수 있다는 큰 장점을 가지는데, 이는 각 유형의 요소 태그도 패턴 자질에 포함되기 때문이다. 즉, 학습 수행 시 요소들의 배치 패턴 또한 학습에 필요한 자질 역할을 하게 된다. (자세한 설명은

이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2013R1A1A2009937)

3장 및 4장에 기술)

제안 방법의 성능 평가를 위해 텍스트 마이닝의 한 분야인 비교 분석(comparison analysis, 혹은 comparison mining) 실험을 수행한다. 제안 방법을 비교 문장(comparative sentence)에 적용하여 여러 유형의 비교 요소를 추출하는데, 본 실험에서는 세 가지 유형의 비교 요소를 추출한다. 아래에 하나의 비교 문장과, 그 문장에서 추출하게 될 세 유형의 비교 요소가 제시되어 있다. 이 예에서 볼 수 있듯이 하나의 문장에는 여러 유형의 요소가 있으며, 각 요소는 한 단어일 수도 있고 MultiEL일 수도 있다.

- 원본 문장: "A회사 회장이 B국가 대통령보다 더 부유하고 막강하다."
- 추출해야 하는 세 가지 유형의 요소:
  - 유형1 (EL1): A회사 회장 (비교 주체)
  - 유형2 (EL2): B국가 대통령 (비교 상대)
  - 유형3 (EL3): 더 부유하고 막강하다 (비교 술어)

본 연구를 요약 정리하면 다음과 같다.

1. 본 연구에서는 한 문장에서 여러 유형의 요소를 동시 추출하며, 이 때 각 요소는 한 단어거나 혹은 MultiEL 일 수 있다.
2. 제안 방법은 별도의 구축 비용을 요하는 사전을 사용하지 않는다.
3. 제안 방법은 형태소 부착을 제외한 파싱 혹은 청킹 등의 어떠한 전처리 과정도 거치지 않음으로써 성능 누수를 최소화한다.
4. 학습 기법으로는 TBL을 사용하며, 성능 평가를 위해 제안 방법을 비교 문장 말뭉치에 적용하여 세 가지 비교 요소(EL1, EL2, EL3)를 추출한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 소개하며, 3장에서는 제안 방법을 상세히 설명한다. 4장에서 실험 및 결과를 기술하며, 5장에서 결론 및 향후 연구에 대해 논한다.

2. 관련 연구

하나 혹은 둘 이상의 단어로 구성된 표현을 추출하는 연구가 계속 진행되어 오고 있다. [1]은 연어(collocation) 추출을 목표로 하였는데, 선행 연구에서 기 구축된 연어 사전을 기반으로 하였다. [2]는 여러 단어로 구성된 아랍어 표현 문구를 자동으로 감지하기 위한 연구를 수행하였는데, 다른 종류의 데이터 사전들을 기반으로 실험을 수행하였다. 그리고 [3]은 의존 파싱을 기본으로 독일어 관용구를 추출하였다.

비교 바이닝 관련한 선행연구는 다음과 같다. [4]는 영어 텍스트에서 비교 요소를 추출하였는데, 이 때의 비교 요소는 MultiEL가 아닌 하나의 단어로 제한하였다. [5,6]은 수작업 규칙과 지지 벡터 기계(support vector machine)를 이용하여 한국어 비교 요소를 추출하였다. [7]도 비교 요소 추출 관련한 연구를 수행하였는데, 웹에 영어로 등록된 비교 질문들에 국한하여 연구를 진행하였다.

TBL 관련한 연구는 다음과 같다. [8]에서 처음으로 이 기법을 소개하였으며, [9]는 이 기법을 이용한 청킹을 수행하였다. 그리고 [10]에서는 변형된 변환 기반 학습을 적용하여 NE(named entity) 식별 작업을 수행하였다.

3. 제안 방법

TBL은 다양한 NLP 응용 분야에서 사용되어 왔다. 이 기법은 학습 데이터로부터 오류 수정 규칙을 생성 후 오류를 줄이는 방향으로 계속해서 규칙을 적용하는 기법이다. 본 실험에서는 [8]에서 제시된 본래의 TBL을 약간 변형하여 사용하는데, 두 단계(초기 태깅 단계와 오류 보정 단계)로 진행되는 본래의 방식 대신 형태소 부착된 문장을 대상으로 바로 하나의 단계로 학습을 수행한다. [표 1]은 학습 데이터 중 하나의 문장을 예로 보여 준다.

표 1. 학습 문장의 예; pi는 문장 내 i번 째 위치의 품사 태그를 의미한다.

원본 문장	A회사 회장이 B국가 대통령보다 더 부유하고 막강하다
품사 부착	<A회사,p1> <회장,p2> <이,p3> <B국가,p4> <대통령,p5> <보다,p6> <더,p7> <부유하,p8> <고,p9> <막강하,p10> <다,p11>
학습 문장	<A회사+회장(p1+p2),**EL1**> <이,p3> <B국가+대통령(p4+p5),**EL2**> <보다,p6> <더+부유하+고+막강하(p7+p8+p9+p10),**EL3**> <다,p11>

위와 같이 정답이 태깅된 학습 데이터의 문장들로부터 변환 규칙을 생성하게 되는데, 학습 과정은 다음과 같다.

규칙 틀 정의 및 후보 규칙 생성: 학습 데이터의 요소 태그(EL1, EL2, EL3) 중심 좌우 길이 2 이내의 품사 패턴을 기준으로 변환 규칙들을 생성한다. 이 품사 패턴은 TBL에서 말하는 변환 규칙 틀에 해당된다. 규칙 틀은 [표 2]에서 볼 수 있듯이 크게 네 가지로 분류될 수 있다. 여기에서 '-1'은 좌측 1번째 거리를 의미하며 '-2'는 좌측 2번째 거리를 의미한다. 반대로 '+'는 우측을 의미한다. 그리고 P, W, TP, TW의 의미는 좌우 정보의 종류를 의미하는데 아래의 설명과 같다.

P: 품사 태그만 사용하는 경우
W: 실제 단어만 사용하는 경우
TP: 요소 태그와 품사 태그를 사용하는데 요소 태그가 품사 태그보다 우선된다.
TW: 요소 태그와 실제 단어를 사용하는데 요소 태그가 실제 단어보다 우선된다.

표 2. 변환 규칙 틀 41가지

설명	틀의 표기
1.현재 품사만 고려하는 경우	P0
2. 현재 품사 및 좌우 1번째 태그를 고려한 경우	P-1, P+1, P-1+1, TP-1, TP+1, TP-1+1, 그리고 P대신 W인 경우 6가지
3. 현재 품사 및 좌우 2번째 태그만 고려한 경우	P-2, P+2, P-2+2, TP-2, TP+2, TP-2+2, 그리고 P대신 W인 경우 6가지
4. 현재 품사 및 좌우 1번째, 2번째 태그 둘 다 고려한 경우	P-2-1, P-2+1, P-1+2, P+1+2, TP-2-1, TP-2+1, TP-1+2, TP+1+2, 그리고 P대신 W인 경우 8가지

구제적인 설명을 위해 [표 1] 예문으로부터 생성될 수 있는 규칙 몇 가지를 아래에 제시하였다.

- P0 틀
  - 현재 문장 내의 두 연속된 단어의 품사가 p1+p2이면 이 두 단어는 하나의 'EL1'으로 태깅한다.
  - 두 연속된 단어의 품사가 p4+p5이면 하나의 'EL2'로 태깅한다.
  - 네 연속된 단어의 품사가 p7+p8+p9+p10이면 하나의 'EL3'로 태깅한다.
- P-1+1 틀에서 나온 규칙 예
  - 두 연속된 단어의 품사가 p1+p2이고 좌 품사(P-1)가 없고(문장 시작) 우 품사(P+1)가 p3이면 하나의 'EL1'으로 태깅한다.
- W+1 틀에서 나온 규칙 예
  - 두 연속된 단어의 품사가 p4+p5이고 우 단어(W+1)가 '-보다'이면 하나의 'EL2'로 태깅한다.
- TP-2-1 틀에서 나온 규칙 예
  - 네 연속된 단어의 품사가 p7+p8+p9+p10이고 좌측 태그 혹은 품사(TP-1)가 p6이고 좌측 두 번째 태그 혹은 품사(TP-2)가 'EL2'면 하나의 'EL3'로 태깅한다.
- TW+1+2 틀에서 나온 규칙 예
  - 두 연속된 단어의 품사가 p1+p2이고 우 태그 혹은 단어(TW+1)가 '-가' 이고 우측 두 번째 태그 혹은 단어(TW+2)가 'EL2'면 하나의 'EL1'으로 태깅한다.

위 예에서 알 수 있듯이, 현재 관찰 중인 부분은 품사만 고려하고(P0) 좌우 정보는 품사(P), 실제 단어(W), 요소 태그 혹은 품사(TP), 그리고 요소 태그 혹은 실제 단어(TW)를 골고루 사용하였다. 그리고 요소 태그로 태깅되는 경우 그 문장에서 기존에 태깅되었던 동일 요소 태그는 취소시킴으로써 한 문장 안에 각각의 요소는 하나씩만 태깅되도록 한다.

최종 규칙 리스트 결정: 규칙 틀에서 생성된 규칙들은 후보 규칙 집합이 되며, 태깅 이전의 학습 데이터에 차례대로 적용하여 가장 높은 점수를 기록한 규칙을 찾는다. 점수 계산식은 각각의 규칙을 적용했을 때 오류에서 정답으로 바뀐 개수(C)와 정답에서 오류로 바뀐 개수(E)의 차이(C-E)를 구한 후, 최대값을 기록하는 규칙 하나를 최종 규칙 리스트에 등록한다. 등록된 규칙이 적용된 후 변환된 말뭉치를 기반으로 다시 규칙 틀로부터 후보 규칙 집합을 만들고 이 후보들 중 가장 높은 점수를 기록하는 규칙을 찾아 최종 규칙 리스트에 추가하는 과정이 반복된다. 그리고 더 이상 성능을 개선시키는 규칙이 발견되지 않는 순간 반복은 종료된다.

실험 데이터에 적용: 최종 규칙 리스트에 순서대로 등록된 규칙들을 차례대로 실험 데이터에 적용하여 성능을 측정한다.

4. 실험

실험을 위해 711개의 비교 문장을 수집하여 세 가지 비교 요소를 어노테이팅(annotating) 하였다. 이 비교 문장들은 온라인 뉴스, e-마켓 리뷰, 상품 비교 포럼 등에서 추출되었으며, 세 명의 어노테이터(annotator)에 의해 정답이 결정되었는데, 먼저 두 명이 별개로 작업을 수행하였고, 불일치한 부분에 대해서는 세 번째 어노테이터의 판단을 참고하였다. [표 3]은 데이터 분포를 나타낸다.

표 3. 데이터 분포

문장 수		711
MultiEL 비율	EL1	47.2%
	EL2	45.4%
	EL3	30.8%

위 표에서 알 수 있듯이, 둘 이상의 단어(정확히는 둘 이상의 형태소)로 구성된 요소가 매우 높기 때문에 요소 추출에서 MultiEL를 반드시 고려해야 함을 알 수 있다. 그리고 평가 도구는 정확도(accuracy)를 사용하며, 평가 방법은 5-fold cross validation, 그리고 t-test (http://www.graphpad.com/quickcalcs/ttest1.cfm)를 수행하여 통계적 유의미 여부를 확인하였다.

먼저 규칙 틀이 잘 정의되었는지를 판단하기 위해 틀의 종류를 나누어서 성능을 비교하였다.

표 4. 규칙 틀 평가

틀 종류 (P0틀은 공통 부분임)		정확도
품사 중심	P, TP	82.20%
실제 단어 중심	W, TW	79.49%
모두 사용	P, W, TP, TW	**83.84%**

위 표에서 알 수 있듯이 모든 틀을 다 사용했을 때 가장 좋은 성능을 얻을 수 있었으며($p<0.01$에서 통계적 유의미), 이는 41가지 규칙 틀이 잘 정의되었음을 보여준다.

다음으로 학습이 종료되는 조건에 대한 실험을 수행하였다. [표 4]의 성능은 C-E>0인 경우의 규칙에 대해서만 실험한 경우인데, C-E값의 임계치에 변화를 주어 성능을 비교해 보았다. [표 5]는 임계치별 성능을 나타내는데, 이 표에서 알 수 있듯이 임계치가 1인 경우 성능이 가장 우수하였다. 다만, 임계치가 0인 경우와 1인 경우를 비교했을 때, $p<0.01$은 물론 $p<0.05$ 레벨의 통계적 유의미도 발견되지 않았다. 반면 임계치가 2인 경우는 성능이 현저히 저하되었다. 결론적으로 1 또는 0인 임계치가 적합하다고 판단할 수 있다.

표 5. 학습 종료 임계치 비교

학습 종료 임계치	정확도
C-E>0 인 규칙이 없는 경우 종료	83.84%
C-E>1 인 규칙이 없는 경우 종료	**84.33%**
C-E>2 인 규칙이 없는 경우 종료	80.21%

또한 비교 마이닝 분야에서 다른 선행 연구[4,5]와 본 실험의 결과를 비교해 보았다. [표 6]은 그 결과를 나타내고 있다.

표 6. 비교 요소 추출 실험 비교

연구	성능	특징
Jindal외[4]	F1-score 72%	- 영어 - MultiEL 제외 (한 단어 요소로 제한)
양선외[5]	정확도 86.81%	- 한국어 - MultiEL 허용 - 수작업 규칙 사용 - 우열 비교 및 최상급 비교 문장만 사용.
제안 방법	정확도 84.33%	- 한국어 - MultiEL 허용 - 수작업 규칙 없음 - 우열 비교, 최상급, 유사 비교 등 다양한 비교 문장으로 실험 - 세 요소 동시 추출.

위 연구들 중 [4]는 영어 기반 연구로 언어 자체가 다르고, 무엇보다도 MultiEL을 제외한 실험이므로 본 연구와 직접적으로 비교하기에는 적합하지 않다고 판단된다. 반면 [5]는 같은 한국어 연구이고 MultiEL을 허용한다는 공통점이 있으며, 무엇보다도 최종 성능 자체가 제안 방법보다 약 2.5% 높았다. 하지만 [5]는 비교 문장의 특징을 관찰하여 수작업 규칙을 별도로 구축한 경우이며, 반면 본 연구는 그러한 과정 없이 한 번의 학습으로 세 요소를 동시에 식별하였다는 큰 장점이 있다. 따라서 감정 분석/의견 마이닝 등 다양한 타 분야로의 이식성은 제안 방법이 더 적합하다고 판단하고 있다.

또한 본 연구는 TBL을 사용한다는 면에서 Ramshaw외 [9]의 연구와도 공통점이 있다고 볼 수 있는데, 실제로 [9]는 본 연구에 동기 부여를 해 준 선행 연구 중 하나이기도 하다. 하지만 [9]는 전처리 과정 중 하나인 청킹 그 자체를 위한 연구로서 baseNP만을 식별하거나 (precision 92%) 혹은 문장을 NP/VP로 분할하는 실험을 수행한 경우이기 때문에(precision 88%), 특정 응용 분야의 최종 정보 추출 단계를 위해 TBL을 사용한 제안 방법과 직접 비교하기는 어렵다고 판단된다.

5. 결론

본 논문은 정보 추출 과제에 대한 연구로서, 문장으로부터 여러 유형의 요소를 동시 추출하는 기법을 제안하였다. 특히 각 요소를 한 단어로 제한하지 않기 때문에 MultiEL 추출이 가능하다는 특징이 있으며, 구축 비용이 드는 사전이나 수작업 규칙을 사용하지 않았고, 형태소 부착을 제외한 다른 모든 전처리 과정을 생략하여 성능 누수를 최소화하였다. 그리고 비교 요소 추출 분야에 제안 기법을 적용하여 84.33%의 우수한 성능을 산출하였다.

향후 계획으로는 성능 향상을 위해 다양한 관점에서 연구를 지속할 것이며, 제안 방법을 비교 마이닝 외의 다른 부문에 적용하여 성능을 평가할 계획이다. 또한 영어 등 다른 언어에 대해서도 실험을 수행할 계획이다.

참고문헌

[1] V. Seretan, L. Nerima and E. Wehrli, "Extraction of Multi-Word Collocations Using Syntactic Bigram Composition", Proc of Recent Advances in NLP (RANLP'03), pp. 424-431, 2003.

[2] M. Attia, A. Toral, L. Tounsi, P. Pecina and J. Genabith, "Automatic Extraction of Arabic Multiword Expressions", Proc of Workshop on Multiword Expressions: from Theory to Applications(MWE'10), pp. 18-26, 2010.

[3] M. Weller and U. Heid, "Automatic Extraction of Arabic Multiword Expressions", Proc of Language Resources and Evaluation (LREC'10), 2010.

[4] N. Jindal and B. Liu, "Mining Comparative Sentences and Relations", Proc of Advancement of Artificial Intelligence(AAAI'06), pp. 1331-1336, 2006.

[5] 양선, 고영중, "한국어 비교 마이닝을 위한 비교 요소 자동 추출", 정보과학회논문지:소프트웨어 및 응용, 제38권, 제 12호, pp. 689-696, 2011.

[6] S. Yang and Y. Ko, "Extracting Comparative Entities and Predicates from Texts Using Comparative Type Classification", Proc of Association for Computational Linguistics (ACL'11), pp. 1636-1644, 2011.

[7] S. Li, C. Lin and Y. Song, "Comparable Entity Mining from Comparative Questions", Proc of Association for Computational Linguistics (ACL'10), pp. 650-658, 2010.

[8] E. Brill, "Transformation-based Error-Driven

Learning and Natural language Processing: A Case Study in Part-of-Speech tagging", Computational Linguistics, pp. 543-565, 1995.
[9] L. A. Ramshaw, and M. P. Marcus, "Text Chunking using Transformation-Based Learning", Proc of Natural Language Processing Using Very Large Corpora (NLP/VLC'95),82-94, 1995.
[10] W. J. Black and A. Vasilakopoulos, "Language-Independent named Entity Classification by modified Transformation-based Learning and by Decision Tree Induction", Proc ofComputational Natural Language Learning (CoNLL'02), vol. 24, pp. 1-4, 2002.