Conditional Random Fields를 이용한 세부 분류 개체명 인식

이창기 황이규 오효정 임수종 허정 이충희 김현진 왕지현 장명길
한국전자통신연구원 임베디드SW연구단 지식마이닝연구팀
{leeck,yghwang,ohj,isj,jeonghur,forever,jini,jhwang,mgjang}etri.re.kr

Fine-Grained Named Entity Recognition using Conditional Random Fields for Question Answering

Changki Lee Yi-Gyu Hwang Hyo-Jung Oh Soojong Lim Jeong Heo Chung-Hee Lee
Hyeon-Jin Kim Ji-Hyun Wang Myung-Gil Jang
Electronics and Telecommunications Research Institute (ETRI)

요 약

질의응답 시스템은 사용자 질의에 해당하는 정답을 찾기 위해서 세부 분류된 개체명을 사용한다. 이러한 세부 분류 개체명 인식을 위해서 대부분의 시스템이 일반 대분류 개체명 인식 후에 사전 등을 이용하여 세부 분류로 나누는 방법을 이용하고 있다. 본 논문에서는 질의응답 시스템을 위한 세부 분류 개체명 인식을 위해서 Conditional Random Fields를 이용한다. 개체명 인식의 과정을 개체명 경계 인식과 경계가 인식된 개체명의 클래스 분류의 두 단계로 나누어, 개체명 경계 인식에 Conditional Random Fields를 이용하고, 경계 인식된 개체명의 클래스 분류에는 Maximum Entropy를 이용한다. 실험결과 147개의 세부 분류 개체명 인식에 대해서 정확도 85.8%, 재현률 81.1%, F1=83.4의 성능을 얻었고, baseline model 보다 학습 시간이 27%로 줄고 성능은 증가하였다. 또한 제안된 세부 분류 개체명 인식기를 이용하여 질의응답 시스템에 적용한 결과 26%의 성능향상을 보였다.

1. 서론

질의응답 시스템은 사용자의 질문에 대한 문서 리스트 대신에, 실제 정답을 찾아주는 시스템이다. 많은 질의응답 시스템이 passage 추출 방법을 사용하고 있다. Passage 추출 방법은 사용자의 질의와 가장 관련이 있는 문장이나 passage를 추출하고, 추출된 문장이나 passage에서 개체명 인식 등의 NLP 기술을 이용하여 정답을 추출한다.

최근 BBN의 IdentiFinder 같은 시스템은 person, location, organization 등과 같은 대분류 개체명 인식에서 90% 정도의 성능을 보이고 있다. 이러한 대분류 개체명 인식이 자연어처리의 많은 분야에 유용하지만, 질의응답 시스템 같은 경우에는 세부로 분류된 개체명 인식이 더 많은 도움을 준다 [1, 2]. 이러한 이유로 많은 질의응답 시스템에서 세부 분류된 개체명을 추출하기 위해서 대분류의 개체명 인식기를 적용하여 얻은 개체명을 세분화된 개체명 사전을 이용하여 세부 개체명으로 분류한다 [6]. 대분류 개체명 인식에 대한 연구는 지금까지 많이 있었지만 세부 분류 개체명 인식에 대한 연구는 찾아보기 힘들다.

Fleischman과 Hovy는 person 카테고리에 대해 자동으로 8개의 세부 카테고리로 분류하는 방법을 연구했다 [1]. 이 연구에서는 로컬 컨텍스트와 글로벌 컨텍스트를 피처로 사용하는 교사 학습 (supervised learning) 방법을 사용했으나, 학습 데이터를 자동으로 얻기 위해서 간단한 bootstrapping 방법을 이용하였기 때문에 학습데이터 분포가 왜곡되어 성능이 좋지 않았다. 또한 그들은 person과 location 만을 대상으로 했다.

Mann은 질의응답 시스템에 사용하기 위해서 세부 분류의 고유명사 온톨로지를 구축하는 연구를 수행했다 [2]. 그는 일반 텍스트로부터 자동으로 고유명사 온톨로지를 구축하고 이를 질의응답 시스템에 사용하였다. 그러나 이 방법은 단순한 co-occurrence 패턴을 사용하였기 때문에 적용범위가 좁았다.

본 논문에서는 Conditional Random Fields (CRFs)를 이용하여 세부 분류의 개체명 인식을 수행하고 이를 질의응답 시스템에 적용한다.

2. CRFs를 이용한 세부 분류 개체명 인식

기존의 기계학습 기반의 개체명 인식에는 hidden Markov models (HMMs) 및 maximum entropy Markov models (MEMMs) 등이 많이 쓰였다. CRFs는 HMMs과 다르게 조건부 확률을 구한다는 특성 때문에 independence assumption이 필요한 HMMs 보다 성능이 뛰어나다고 실험적으로 증명되었으며, MEMMs과 비교하여 *label bias problem*이 해결되어 좀더 나은 성능을 보이고 있다.

본 논문에서는 질의응답 시스템의 정답이 될 수 있는 147개의 세부 분류된 개체명 카테코리를 정의한다. 이 147개의 세부 분류는 15개의 대분류에 속하는데, 이 대분류는 다음과 같다 (세부 분류는 appendix 참고).

> person, study field, theory, artifacts, organization, location, civilization, date, time, quantity, event, animal, plant, material, term.

대부분의 기계학습 기반의 대분류 개체명 인식기는 개체명 클래스에 경계 정보를 추가하여 개체명 인식의 문제를 classification 문제로 변환하고 있다. 예를 들어, 개체명 클래스가 person인 경우, 여기에 경계 정보 B, I 등을 추가하여 B-Person, I-Person 등으로 나뉘게 된다. 따라서 개체명 클래스의 수가 147개인 경우에는 총 295개의 클래스를 분류하는 문제가 된다 (147 개체명 클래스 * 2개의 경계 정보 + 1개의 개체명이 아닌 클래스).

그러나, CRFs는 클래스의 숫자가 많은 경우에 학습 속도가 급격히 떨어지기 때문에 이 경우에는 CRFs를 바로 적용할 수 없다. 왜냐하면 CRFs 모델의 파라미터의 수는 클래스의 숫자를 S라 하고 피쳐의 수를 O라 할 때 O(|S||O|+|S|^2)를 따르고, 학습 시에 forward-backward inference를 수행하는 데에 O(|S|^2 * T)의 시간이 걸리기 때문에 결국 총 학습 시간은 O(|S|^4)을 따르게 되기 때문이다.

본 논문에서는 이러한 문제를 해결하기 위해서 개체명 인식의 문제를 개체명 경계 인식(boundary detection)과 개체명 클래스 분류(NE classification)로 나누고, 경계 인식에 CRFs를 이용하고 개체명 클래스 분류에 Maximum Entropy (ME)를 이용한다.

입력 데이터 열을 $\mathbf{x} = \langle x_1, x_2, \ldots x_T \rangle$라 하고 이에 대응되는 개체명 클래스 정보와 개체명 경계 정보가 합쳐진 클래스 (예를 들어, B-Person, I-Person 등)를 $\mathbf{y} = \langle y_1, y_2, \ldots y_T \rangle$ 라 하면, 주어진 입력 데이터 열 $\mathbf{x}$에 대한 클래스 열 $\mathbf{y}$의 조건부 확률은 다음과 같이 정의된다.

$$P(\mathbf{y}|\mathbf{x}) = P(\mathbf{c},\mathbf{b}|\mathbf{x}) = P(\mathbf{b}|\mathbf{x})P(\mathbf{c}|\mathbf{b},\mathbf{x})$$

$$\approx P(\mathbf{b}|\mathbf{x})\prod_{i=1} P(c_i|b_i,x_i)$$

위 식에서 $\mathbf{b} = \langle b_1, b_2, \ldots b_T \rangle$는 개체명 경계 정보(B, I, O)의 열이고, $\mathbf{c} = \langle c_1, c_2, \ldots c_T \rangle$ 는 개체명 클래스 정보(Person, Location 등)의 열이다. 위 식에서 마지막 줄은 $P(\mathbf{c}|\mathbf{b},\mathbf{x})$ 값의 계산을 쉽게 하기 위해서 이미 경계 인식이 된 개체명 후보의 클래스 정보를 구하기 위해서는 다른 개체명의 클래스 정보와는 독립이다라는 가정을 한 것이다. 즉 개체명 경계 인식 과정에서 인식된 개체명 후보는 그 개체명 후보의 로컬 피쳐만으로 개체명 클래스를 구할 수 있다라는 가정이다.

개체명의 경계 인식을 위한 $P(\mathbf{b}|\mathbf{x})$는 다음과 같이 CRFs를 이용하여 정의한다.

$$P_{CRF}(\mathbf{b}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(b_{t-1},b_t,\mathbf{x},t)\right)$$

위 식에서, $f_k(b_{t-1},b_t,\mathbf{x},t)$는 k번째 피쳐 함수이고, λ_k 는 k번째 피쳐 함수의 학습된 가중치이다. λ 의 값이 클수록 그에 따른 피쳐가 많은 가중치를 갖게 된다.

개체명 경계인식이 된 개체명 후보의 개체명 클래스를 구하기 위한 확률 식은 Maximum entropy를 이용하여 다음과 같이 정의된다.

$$P_{ME}(c_i|b_i,x_i) = \frac{1}{Z(b_i,x_i)}\exp\left(\sum_{k}\lambda_k f_k(c_i,b_i,x_i)\right)$$

본 논문에서는 다음과 같은 피쳐를 사용한다.

- Lexical feature — (-2,-1,0,1,2) 위치에 해당하는 형태소 lexical 정보.
- POS tag — (-2,-1,0,1,2) 위치에 해당하는 형태소의 POS tag 정보.
- 개체명 사전 피쳐 — 147 개 세분류 개체명 사전 피쳐 및 15 개 대분류 개체명 사전 피쳐.
- 15개의 문자 단위 정규 표현식 – [A-Z]*, [0-9]*, [0-9][0-9], [0-9][0-9][0-9][0-9], [A-Za-z0-9]*, ...

3. 실험

본 논문에서는 세부 개체명 인식을 위해서 ETRI의 147개의 세부 개체명 태그셋으로 태깅된 백과사전 학습데이터 8037 문서를 학습데이터로 사용하고 100문서를 테스트 데이터로 사용하였다. CRFs와 ME의 학습을 위해 ETRI의 CRF_TOOL을 이용하였다. 이 툴은 C++로 구현되어 있으며 L-BFGS 알고리즘을 이용하여 학습을 수행한다. 학습 파라미터로 gaussian prior의 값을 ME는 1, CRFs는 10의 값을 주었으며 iteration 값은 1000을 사용하였다.

표 1. 개체명 경계 인식.

Sub-task	경계인식
Model	CRFs
# Class	3 classes
학습시간 (초/iter)	24.14
Pre.(%)	95.2
Rec.(%)	84.3
F1	89.4

표 1은 CRFs를 이용한 개체명 경계 인식의 성능을 나타낸다. 개체명 경계 인식의 경우 F1 값으로 89.4를 얻었다.

표 2. 개체명 클래스 분류 (개체명 경계 인식이 100% 성능이라 가정 했을 경우).

Sub-task	개체명 클래스 분류
Model	ME
# Class	147 classes
학습시간 (초/iter)	18.36
Pre.(%)	85.7
Rec.(%)	86.2
F1	86.0

표 2는 ME를 이용한 개체명 클래스 분류의 성능을 나타낸다. 이 성능은 개체명의 경계 인식의 결과가 모두 맞았다고 가정했을 때의 성능이다.

표 3은 147개의 개체명 카테고리에 대한 세부 분류 개체명 인식 결과이다. 1단계 개체명 인식은 일반 기계학습 기반 개체명 인식 방법으로 개체명 클래스 정보(Person 등)와 경계 정보(B, I)를 합쳐서 295 class가 만들어지고 이에 대해 classification을 수행한 것이다. 2단계 개체명은 본 논문에서 사용하는 방법으로 개체명 경계 인식과 개체명 클래스 분류의 2단계로 개체명 인식이 진행된다. 실험 결과 1단계 개체명 인식인 경우 CRFs는 클래스 수가 많아지면 속도가 느려 적용할 수 없음을 알 수 있고, ME인 경우 역시 2단계 개체명 인식에 비해 학습 시간이 오래 걸리는 것을 알 수 있다. 성능도 1단계 ME 보다 2단계의 CRFs + ME가 성능이 더 좋은 것을 알 수 있다.

표 3. 세부 분류 개체명 인식 (일반 1 단계 개체명 인식 vs. 2 단계 개체명 인식)

Task	1단계 개체명 인식	1단계 개체명 인식	2단계 개체명 인식 (경계인식 + 클래스 분류)
Model	ME	CRFs	CRFs + ME
# Class	295 classes	295 classes	3 + 147 classes
학습시간 (초/iter)	154.97	12248.33	42.50
Pre.(%)	83.2	-	85.8
Rec.(%)	81.5	-	81.1
F1	82.3	-	83.4

세부 분류 개체명 인식이 질의응답 시스템에 미치는 영향을 알아보기 위해 질의응답 시스템 실험을 수행하였다. 테스트 셋은 ETRI QA TEST SET(402개의 질문 및 정답으로 구성)을 이용하였다 [5]. 각 질문에 대해 스코어는 첫 번째로 맞은 정답에 대한 reciprocal answer rank(RAR)로 계산하였다.

본 논문의 세부 분류 개체명 인식기를 이용하여 각 문장으로부터 147 종류의 answer type (세부 분류 개체명 클래스)을 추출하여 Answer Index Unit(AIU)을 생성하여 인덱싱 하였다 [5].

Table 4는 질의응답 시스템의 성능을 나타낸다. 세부 분류 개체명 인식기를 이용한 것이 이용하지 않은 것보다 성능이 26% 올라간 것을 알 수 있다.

Table 4. 질의응답 시스템 성능

Task	MRAR
QA with passage retrieval	0.525
QA with passage retrieval + AIU (세부 분류 개체명 이용)	0.662

4. 결론

본 논문에서는 질의응답 시스템을 위해서, Conditional Random Fields를 이용한 세부 분류 개체명 인식 방법을 제안하였다. CRFs가 클래스의 수가 증가하면 학습 속도가 급속히 느려지기 때문에 세부 분류 개체명 인식을 개체명 경계 인식과

개체명 클래스 분류 문제로 나누고 개체명 경계 인식에 CRFs를 이용하고 개체명 클래스 분류에 ME를 이용하였다. 실험결과 147개의 세부분류 개체명 인식에 대해서 정확도 85.8%, 재현률 81.1%, F1=83.4의 성능을 얻었고, baseline model인 1단계 ME 모델 보다 학습 시간이 27%로 줄고 성능은 증가하였다. 또한 제안된 세부 분류 개체명 인식기를 이용하여 질의응답 시스템에 적용한 결과 26%의 성능향상을 보였다.

감사의 글

본 논문은 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참고문헌

[1] 1. M. Fleischman and E. Hovy. *Fine grained classification of named entities*. COLING, 2002.

[2] G. Mann, *Fine-Grained Proper Noun Ontologies for Question Answering*, SemaNet'02: Building and Using Semantic Networks, 2002

[3] A. McCallum and W. Li. *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*, CoNLL, 2003.

[4] S. Fei, F. Pereira. *Shallow Parsing with Conditional Random Fields*, HLT & NAACL, 2003.

[5] H. Kim, J. Wang, C. Lee, C. Lee, M. Jang. *A LF based Answer Indexing Method for Encyclopedia Question Answering System*, AIRS, 2005.

[6] K. Han, H. Chung, S. Kim, Y. Song, J. Lee, H. Rim. *Korea University Question Answering System at TREC 2004*, TREC, 2004.

Appendix: Fine-Grained Named Entity Tag Set (147 classes)

1. Person	PS_NAME PS_MYTH
2. Study Field	FD_OTHERS FD_SCIENCE FD_SOCIAL_SCIENCE FD_MEDICINE FD_ART FD_PHILOSOPHY
3. Theory	TR_OTHERS TR_SCIENCE TR_TECHNOLOGY TR_SOCIAL_SCIENCE TR_ART TR_PHILOSOPHY TR_MEDICINE
4. Artifacts	AF_CULTURAL_ASSET AF_BUILDING AF_MUSICAL_INSTRUMENT AF_ROAD AF_WEAPON AF_TRANSPORT AF_WORKS AFW_GEOGRAPHY AFW_MEDICAL_SCIENCE AFW_RELIGION AFW_PHILOSOPHY AFW_ART AFWA_DANCE AFWA_MOVIE AFWA_LITERATURE AFWA_ART_CRAFT AFWA_THEATRICALS AFWA_MUSIC
5. Organization	OG_OTHERS OGG_ECONOMY OGG_EDUCATION OGG_MILITARY OGG_MEDIA OGG_SPORTS OGG_ART OGG_SOCIETY OGG_MEDICINE OGG_RELIGION OGG_SCIENCE OGG_BUSINESS OGG_LIBRARY OGG_LAW OGG_POLITICS
6. Location	LC_OTHERS LCP_COUNTRY LCP_PROVINCE LCP_COUNTY LCP_CITY LCP_CAPITALCITY LCG_RIVER LCG_OCEAN LCG_BAY LCG_MOUNTAIN LCG_ISLAND LCG_TOPOGRAPHY LCG_CONTINENT LC_TOUR LC_SPACE
7. Civilization	CV_NAME CV_TRIBE CV_SPORTS CV_SPORTS_INST CV_POLICY CV_TAX CV_FUNDS CV_LANGUAGE CV_BUILDING_TYPE CV_FOOD CV_DRINK CV_CLOTHING CV_POSITION CV_RELATION CV_OCCUPATION CV_CURRENCY CV_PRIZE CV_LAW, CVL_RIGHT,

	CVL_CRIME, CVL_PENALTY,
8. Date	DT_OTHERS DT_DURATION DT_DAY DT_MONTH DT_YEAR DT_SEASON DT_GEOAGE DT_DYNASTY
9. Time	TI_OTHERS TI_DURATION TI_HOUR TI_MINUTE TI_SECOND
10. Quantity	QT_OTHERS QT_AGE QT_SIZE QT_LENGTH QT_COUNT QT_MAN_COUNT QT_WEIGHT QT_PERCENTAGE QT_SPEED QT_TEMPERATURE QT_VOLUME QT_ORDER QT_PRICE QT_PHONE
11. Event	EV_OTHERS EV_ACTIVITY EV_WAR_REVOLUTION EV_SPORTS EV_FESTIVAL
12. Animal	AM_OTHERS AM_INSECT AM_BIRD AM_FISH AM_MAMMALIA AM_AMPHIBIA AM_REPTILIA AM_TYPE AM_PART
13. Plant	PT_OTHERS PT_FRUIT PT_FLOWER PT_TREE PT_GRASS PT_TYPE PT_PART
14. Material	MT_ELEMENT MT_METAL MT_ROCK MT_CHEMICAL MTC_LIQUID MTC_GAS
15. Term	TM_COLOR TM_DIRECTION TM_CLIMATE TM_SHAPE TM_CELL_TISSUE TMM_DISEASE TMM_DRUG TMI_HW TMI_SW