가중 투표 기반의 앙상블 기법을 이용한 한국어 개체명 인식기

권순재O, 허윤석, 이건철, 임지수 최호정, 서정연

soon91jae@gmail.comO, yoonseok419@gmail.com, psogv0308@gmail.com,

sgcde@hotmail.co.kr, nphew@naver.com, seojy@sogang.ac.kr

A Korean Named Entity Recognizer using Weighted Voting based Ensemble Technique

Sunjae KwonO, Yoonseok Heo, Kyunchul Lee, Jisu Lim, Hojeong Choi, Jungyun Seo

요 약

본 연구에서는 개체명 인식의 성능을 향상시키기 위해, 가중 투표 방법을 이용하여 개체명 인식 모델을 앙상블 하는 방법을 제안한다. 각 모델은 Conditional Random Fields의 변형 알고리즘을 사용하여 학습하고, 모델들의 가중치는 다목적 함수 최적화 기법인 NSGA-Ⅱ알고리즘으로 학습한다. 실험 결과 제안 시스템은 $F_1 Score$ 기준으로 87.62%의 성능을 보여, 단독 모델 중 가장 높은 성능을 보인 방법보다 2.15%p 성능이 향상되었다.

주제어: 개체명 인식, 앙상블 학습, 다목적 함수 최적화

1. 서론

개체명 인식은 텍스트에서 인명, 기관명, 지명과 같이 고유한 의미를 갖는 표현을 찾아 이를 사전에 정의된 범주로 분류하는 작업이다. 최근 개체명 인식 분야에 기계학습[1,2,3,4]과 지식 베이스[5]와 방법을 적용하는 연구가 활발히 진행되고 있다. 하지만 정교하게 설계된 기계학습 기법에 좋은 자질을 선택하여 개체명인 식기를 학습하는 방법과 지식 베이스를 구축하여 개체명 인식기의 성능을 높이는 접근 방법은 많은 비용이 소요된다.

본 연구에서는 다양한 자질과 알고리즘을 사용하여 학습한 개체명 인식 모델을 가중 투표 방법으로 앙상블 하는 방법을 제안한다. 개체명 인식기는 CRFs(Conditional Random Fields)와 그 변형 알고리즘을 사용하여 학습하며[1,2,3,4], 모델 앙상블을 위한 가중치는 NSGA-Ⅱ 알고리즘을 이용한 다목적 최적화 기법을 사용하여 학습한다.

2. 관련 연구

과거부터 지금까지 개체명 인식기를 결합하여 강력한 개체명 인식기를 만드는 앙상블 기반의 개체명 인식기를 개발하는 것이 활발히 연구되어 왔다. [6]는 학습한 모델들을 검증 데이터의 성능을 기반으로 가중 투표 방법을 사용하여 결합하는 방법을 제안하였다. 그리고 [7]에서는 유전자 알고리즘 등의 전역 최적화 기법을 사용하여 가중치를 결정하는 모델을 제안하였고, 기존 검증 데이터의 성능을 기반으로 결합하는 모델보다 높은 성능을 보였다. [8]는 $F_1 Score$를 목적 함수로 가중치를 학습하는 단일 함수 최적화 방법보다, Recall과 Precision을 목적 함수로 가중치를 학습하는 다목적 함수 최적화 방법에서 더 성능이 높은 개체명 인식기를 구축 할 수 있다는 것을 보였다.

3. 가중 투표 기반의 앙상블 방법을 이용한 개체명 인식기

3.1 개체명 인식 모델 구축

본 연구에서는 기계학습 기반의 개체명 인식기를 학습하기 위해 표 1과 같은 자질을 사용하였다. 어휘, 품사 자질은 해당 형태소와 앞 뒤 두 형태소 정보를 실험적으로 적용한다. 음절 자질은 해당 형태소의 접두부와 접미부 음절 정보이다. 예를 들어, 한국정보과학회’라는 형태소에서 접두부 2 음절은 ‘한국’이고, 접미부 2음절은 ‘학회’이다. 이 접두부와 접미부의 길이 또한 실험적으로 적용한다. 형태소 길이 자질은 해당 형태소의 길이 정보를 의미한다. 개체명 사전 자질은 해당 형태소의 개체명 사전 태그 정보이다. 기분석 사전 자질은 기분석 사전 태그 정보로, 기분석 사전은 훈련 데이터에서 3회 이상 같은 태그로 분류되는 개체명을 모아놓은 사전이다. 그리고 워드 임베딩 자질은 word2vec skip-gram[9] 알고리즘으로 학습한 단어 벡터 표현 정보이다. 마지막으로 워드 임베딩 클러스터 자질은 위의 단어 벡터 표현을 DBSCAN 알고리즘[10]으로 클러스터링 한 정보이다.

- "본 연구는 미래창조과학부 및 정보통신기술진흥센터의 SW중심대학지원사업의 연구결과로 수행되었음"(R2215-16-1003)

표 1. 개체명 인식기 자질 정보

자질 정보	설명
어휘 자질	(-2 ~ 2) 위치의 형태소 어휘정보
품사 자질	(-2 ~ 2) 위치의 형태소 품사정보
음절 자질	형태소의 접두, 접미부 음절 정보
형태소 길이 자질	형태소 길이 정보
개체명 사전 자질	개체명 사전 태그 정보
기분석 사전 자질	기분석 사전 태그 정보
워드 임베딩 자질	형태소의 벡터 표현
워드 임베딩 클러스터 자질	워드 임베딩 자질의 군집 정보

기계학습 기반의 개체명 인식기는 CRFs(Conditional Random Fields)[1]와 그 변형 알고리즘 4가지를 사용하여 학습하였고, 사용한 알고리즘은 표 2.와 같다.

표 2. 개체명 인식 학습 알고리즘

사용 모델	학습 방법
CRFs	L-BFGS[1]
CRFs	Frank-Wolfe Structured Support Vector Machines[2]
CRFs	Average Structured Perceptron[3]
LSTM-CRFs	Stochastic Gradient Descent with Dropout[4]

3.2 개체명 인식 모델 결합

3.1절의 방법을 사용하여 구축한 모델들은 앙상블 기법인 가중 투표[6] 방법으로 결합하여 최종 모델을 구성한다. 이때, 최종 모델은 식 1과 같이 동작한다.

$$\underset{t \in T}{\arg max}\left(Score^{t}(x) = \sum_{i=1}^{8} W_i \times Score_i^{t}(x)\right) \quad (1)$$

식 1에서 T는 개체명 태그 집합이다. $Score^{t}(x)$는 현재 단어 x에 대해 최종 모델에서의 태그 t의 점수를 의미한다. 그리고 $Score_i^{t}(x)$는 단어 x에 대한 i번째 개체명 인식 모델에서의 태그 t의 점수이다. 마지막으로 W는 가중치 행렬을 의미한다.

본 연구에서는 식 1의 가중치 행렬 $W=\{\overrightarrow{W_1},\ldots,\overrightarrow{W_8}\}$을 학습하기 위해, NSGA-Ⅱ 알고리즘[11]을 활용한 다목적 최적화 방법을 사용한다. 이 때, 유전자 알고리즘에서 가중치 행렬 W는 그림 1과 같이 염색체 구조로 표현할 수 있다. 최적 가중치 행렬 $\hat{W}$는 Recall과 Precision을 목적함수로 최적화 한다.

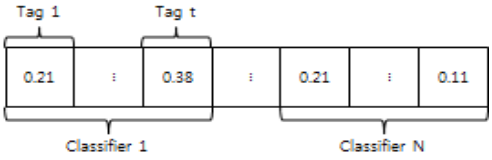

그림 1. 염색체 표현의 예시

4. 실험 및 결과

4.1 실험 환경

본 연구에서는 2016 국어 정보처리 시스템 경진대회[2] 코퍼스를 대상으로 제안하는 모델의 성능을 평가한다. 해당 코퍼스는 학습 데이터 3,555 문장, 검증 데이터 501문장, 실험 데이터 1,000문장으로 구성된다. 그리고 개체명 태그는 인명, 지명, 기관명, 날짜, 시간 5개이며, IOB2 태그 스키마[12]를 사용하여 개체명을 학습한다.

표 3. 2016 국어 정보처리 시스템 경진대회 코퍼스의 구성

	문장 수
학습 데이터	3,555 문장
검증 데이터	501 문장
실험 데이터	1,000 문장

개체명 인식기는 아래와 같은 방법으로 학습하였다. CRFs L-BFGS, CRFs Frank-Wolfe SSVM, CRFs Average Structured Perceptron 모델은 각각 pycrfsuite[3], pystruct[4], seqlearn[5]의 default 값으로 학습하였다. LSTM CRFs[6]의 drop rate는 0.5이고, 학습률은 0.005, 배치 크기는 50으로 설정하여 학습하였다. 마지막으로 가중치 행렬은 실험적으로 가장 좋은 성능을 보인 유전자 풀 크기 240, 돌연변이율 2%, 교배율 90%, 100세대에서 학습한 결과를 실험에 사용하였다.

4.2 실험 결과 및 평가

본 논문에서는 3.1절에서 소개한 자질과 알고리즘으로 구성한 8개의 모델을 3.2절에서 소개한 가중 투표 방법으로 결합하여 실험을 진행하였다. 각각의 모델들은 실험적으로 결정된 개별적인 자질 모델이 사용되었고, 각 모델들은 검증 데이터에서 실험적으로 성능이 가장 높은 모델과 다음으로 높은 모델을 사용한다. L-BFGS_1와 L-BFGS_2 모델은 CRFs를 L-BFGS 알고리즘으로 학습하였다. SSVM_1, SSVM_2 모델은 CRFs를 Frank-Wolfe SSVM 알고리즘으로 학습하였다. SP_1, SP_2 모델은 CRFs를 Structured Perceptron 알고리즘으로 학습하였다.

2) http://ithub.korean.go.kr/user/contest/contestIntroLastView.do
3) https://python-crfsuite.readthedocs.io/en/latest/
4) https://pystruct.github.io/
5) https://github.com/larsmans/seqlearn
6) https://github.com/glample/tagger

LSTM-CRFs 모델은 LSTM-CRFs 알고리즘으로 학습하였다. Voting은 위의 8개의 모델을 같은 가중치로 결합한 모델이다. Weighted Voting은 가중치 w를 학습하여 결합한 모델이다.

표 4. 실험 결과

실험 모델	Precision	Recall	F_1 Score
L-BFGS_1	86.16%	69.75%	77.09%
L-BFGS_2	85.03%	82.81%	83.91%
SSVM_1	85.37%	85.57%	85.47%
SSVM_2	84.57%	84.85%	84.71%
SP_1	86.09%	84.82%	85.45%
SP_2	84.34%	83.00$	83.66%
LSTM-CRFs_1	84.79%	82.10%	83.42%
LSTM-CRFs_2	83.47%	81.86%	82.66%
Voting	87.05%	85.84%	86.44%
Weighted Voting	**86.67%**	**88.60%**	**87.62%**

실험 결과 표 4와 같이 Weighted Voting 모델은 F_1 Score 기준으로 87.80%를 보였다. 이는 단독 성능이 가장 높은 모델인 SSVM_1보다 2.15%p, Voting 모델보다 1.18%p 높은 성능으로, 제안하는 방법을 적용하는 경우 개체명 인식 성능이 향상되는 것을 알 수 있었다. 이는 그림 2와 같이 다수의 학습 모델이 개체명을 잘못 인식하여 Voting 모델 결과가 잘못되는 경우에도, 가중치가 높은 모델이 이를 올바르게 분석한다면, Weighted Voting 모델의 결과는 정상적으로 개체명은 인식하기 때문이다.

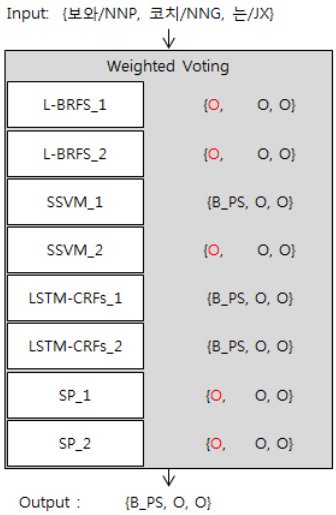

그림 2 다수의 인식기가 잘못 판단하는 경우에도 가중 투표의 결과는 정답인 예

가중 투표 기반의 모델 앙상블 방법은 전체 개체명 인식 성능은 높아지는 반면, 모델의 결합 과정에서 개체명 시퀀스가 학습이 되지 않는다는 문제가 발생하였다. 예를 들어, "psv/NNP,에인트호벤/NNP"와 같은 개체는 IOB2 태그 스키마에서 "B_OG, I_OG"라고 태그 된다. 하지만 현재 모델에서는 "O, I_OG"와 같이 IOB2 태그 스키마에서 있을 수 없는 태그가 결과로 나온다. 이는 CRFs 등의 모델에서 전이 확률 등으로 개체명 태그 시퀀스를 학습하는 방법이 적용된 반면, 본 논문에서 제안하는 방법은 태그 시퀀스를 학습에 고려하지 않기 때문에 이런 결과가 발생하였다.

5. 결론 및 향후 연구

본 연구에서는 개체명 인식에 다른 자질과 다른 알고리즘을 조합하여 구성한 모델들을 가중 투표 방법으로 결합하는 방법을 제안하였다. 제안된 시스템은 단독 모델 중 가장 높은 성능을 보인 모델보다 2.15%p, 단순 투표 방법보다 1.18%p 높은 성능을 보였다.

현재 시스템에서는 각 모델 결과를 미리 학습된 가중치에 따라 결합하는 과정에서 학습된 개체명 시퀀스가 깨지는 문제가 발생한다. 향후 연구에서는 모델 결합 과정에서 각 모델의 가중치 뿐만 아니라 전이 확률도 함께 학습하여, 개체명 시퀀스를 보존하는 방법을 연구 할 것이다.

참고문헌

[1] Lafferty, John, et al."Conditional random fields: Probabilistic models for segmenting and labeling sequence data." *ICML*, Vol.1, pp.282-289, 2001.

[2] Lacoste-Julien, et al. "Block-Coordinate Frank-Wolfe Optimization for Structural SVMs." *arXiv preprint arXiv:1207.4747*, 2012.

[3] Daume, Harold Charles. "Practical structured learning techniques for natural language processing." *ProQuest*, 2006.

[4] Lample, Guillaume, et al. "Neural architectures for named entity recognition." *arXiv preprint arXiv:1603.01360*, 2016.

[5] 이성희, 송영길 and 김학수. "원거리 감독과 능동 배깅을 이용한 개체명 인식." Journal of KIISE, Vol.43, No.2, pp.269-274, 2016.

[6] Florian, Radu, et al. "Named entity recognition through classifier combination." *Proceedings of the seventh conference on Natural language learning at HLT-NAACL*, Vol.4, 2003.

[7] Ekbal, Asif, and Sriparna Saha. "Weighted vote-based classifier ensemble for named entity recognition: a genetic algorithm-based approach.", *International Conference on Application of Natural Language to Information Systems*, pp.256-267, 2010.

[8] Saha, Sriparna, and Asif Ekbal. "Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition." *Data & Knowledge Engineering*, pp.15-39, 2013.

[9] Mikolov, T., and J. Dean. "Distributed representations of words and phrases and their compositionality." *Advances in neural information processing systems*, 2013.

[10] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

[11] Deb, Kalyanmoy, et al. "A fast and elitist multiobjective genetic algorithm: NSGA-II." *IEEE*

transactions on evolutionary computation, Vol.6, No.2, pp.182-197, 2002.

[12] Shen, Hong, and Anoop Sarkar. "Voting between multiple data representations for text chunking." *Conference of the Canadian Society for Computational Studies of Intelligence*, pp.389-400, 2005.