인지과학, 제21권 제4호
Korean Journal of Cognitive Science
2010, Vol. 21, No. 4, 655~667.

Structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식

이 창 기† 장 명 길
한국전자통신연구원 지식마이닝연구팀

개체명 인식은 정보 추출의 한 단계로서 정보검색 분야 뿐 아니라 질의응답과 요약 분야에서 매우 유용하게 사용되고 있다. 본 논문에서는 structural Support Vector Machines(structural SVMs) 및 수정된 Pegasos 알고리즘을 이용한 한국어 개체명 인식 시스템에 대하여 기술하고 기존의 Conditional Random Fields(CRFs)를 이용한 시스템과의 성능을 비교한다. 실험결과 structural SVMs과 수정된 Pegasos 알고리즘이 기존의 CRFs 보다 높은 성능을 보였고(신뢰도 99%에서 통계적으로 유의함), structural SVMs과 수정된 Pegasos 알고리즘의 성능은 큰 차이가 없음(통계적으로 유의하지 않음)을 알 수 있었다. 특히 본 논문에서 제안하는 수정된 Pegasos 알고리즘을 이용한 경우 CRFs를 이용한 시스템보다 높은 성능(TV 도메인 F1=85.43, 스포츠 도메인 F1=86.79)을 유지하면서 학습 시간은 4%로 줄일 수 있었다.

주제어 : 한국어 개체명 인식, Structural SVMs, Pegasos 알고리즘

---

† 교신저자: 한국전자통신연구원 지식마이닝연구팀
E-mail: leeck@etri.re.kr, mgjang@etri.re.kr

서 론

개체명(Named Entity)이란 문서에서 나타나는 고유한 의미를 가지는 명사나 숫자 표현으로 인명(Person), 지명(Location), 기관명(Organization), 날짜, 시간, 화폐, 퍼센트 등을 말한다. 개체명 인식(Named Entity Recognition)은 문서에서 이러한 개체명을 추출하고 추출된 개체명의 종류를 결정하는 작업을 말한다. 개체명은 고유 명사이거나 미등록어인 경우가 많으며, 항상 새롭게 만들어지고, 또한 같은 단어라도 사용되는 문맥에 따라 상이한 의미의 변화를 보이기 때문에 단순한 사전 구축으로는 이러한 개체명을 인식하는 것은 어렵다.

1990년대에 정보추출(Information Extraction)의 목적으로 개최되었던 Message Understanding Conference (MUC)에서 개체명 인식을 정보추출의 일환으로 본격적으로 연구되기 시작하였으며[1], MUC 이후 개체명에 대한 연구가 꾸준히 진행되었으며 Conference on Computational Natural Language Learning 2002(CoNLL 2002)와 CoNLL 2003을 통해서 더욱 많은 발전이 있었다[2]. 이 대회에서 최고 성적을 낸 IBM의 시스템은 여러 기계학습 방법을 voting한 결과를 사용하였으며 영어의 경우에 약 F1=89의 성능을 보였다.

최근에 개체명 인식에 주로 이용되는 방법은 통계기반의 기계학습 방법이며, 대표적인 방법으로는 Hidden Markov Model(HMM), Maximum Entropy Model(MEM), Support Vector Machines(SVMs), Conditional Random Fields(CRFs) 등이 있다[1][3][4][5][6][7].

개체명 인식 연구가 시작된 90년대 말에는 주로 영어만을 대상으로 연구가 이루어졌으나, 최근에는 일본어, 중국어, 독일어, 한국어, 등 다양한 언어에 대해서 개체명 인식 시스템이 개발되었다[2][6][7].

Structural Support Vector Machines(structural SVMs)은 기존의 SVMs을 확장한 기계학습 알고리즘으로, 기존의 SVMs이 바이너리 분류, 멀티클래스 분류 등을 지원하는 반면에, Structural SVMs은 더욱 일반적인 구조의 문제(예를 들어, sequence labeling, 구문 분석 등)를 지원한다[8]. Structural SVMs의 학습에는 cutting-plane algorithm이 사용되어 $O(1/e^2)$의 반복(iteration)에 학습이 완료될 수 있다[8]. 최근에 1-slack 형태의 structural SVMs이 제안되었으며 기존의 structural SVMs 보다 빠른 O(1/e)의 반복

(iteration)에 학습이 완료될 수 있음이 증명되었다[9]. [10]와 [11]에서는 수정된 Fixed-Threshold Sequential Minimal Optimization(FSMO)을 structural SVMs과 1-slack structural SVMs에 적용하여 문서 분류 및 청킹 등에서 기존의 structural SVMs과 1-slack structural SVMs 보다 빠른 학습 속도를 보였다.

학습데이터의 규모가 큰 문제에서는 기존의 SVMs 등의 학습 시간이 매우 커져서 문제가 되며, 이를 해결하기 위해서 Stochastic Gradient Decent(SGD) 방법이 사용되곤 하였다[12]. SGD 방법론 중에서 Pegasos는 바이너리 분류 SVMs에 적용되어 높은 성능과 빠른 학습 속도를 보였다[13]. [14]에서는 Pegasos 알고리즘을 structural SVMs으로 확장하여 이를 의존 구문 분석기(dependency parser)에 적용하여 높은 성능과 빠른 학습 속도를 보였다.

본 논문에서는 기존의 CRFs를 이용한 한국어 개체명 인식 시스템[6][7]에 최근 각광받고 있는 structural Support Vector Machines(structural SVMs)을 적용하여 성능을 개선 시키고, Pegasos 알고리즘을 적용하여 높은 성능을 유지하면서 학습 시간을 기존 CRFs의 학습시간의 4%로 줄일 수 있음을 보인다. 2장에서는 structural SVMs의 간단한 설명과 이를 이용한 개체명 인식 시스템을 서술하고, 3장에서는 Pegasos 알고리즘을 structural SVMs으로 확장하여 개체명 인식 시스템의 학습 시간을 획기적으로 줄일 수 있음을 보인다. 4장에서는 2, 3장에서 제안된 개체명 인식 시스템을 기존의 CRFs를 이용한 개체명 인식 시스템과 비교 실험을 통하여 그 유용성을 살펴본다. 5장에서는 향후 연구과제에 대하여 생각해 보고 결론을 맺고자 한다.

Structural SVMs 기반 개체명 인식

한국어 개체명 인식에서, 입력 문장의 형태소 열을 $\mathbf{x}=<x_1, x_2, ..., x_T>$라 하고, 이에 대응되는 개체명 클래스(Person, Location, Organization 등)와 개체명 경계 정보(B: 개체명의 시작, I: 개체명의 연속, O: 개체명이 아님)가 합쳐진 BIO 태그(예를 들어, B-Person, I-Person, O 등) 열을 $\mathbf{y}=<y_1, y_2, ..., y_T>$라 하면, 개체명 인식을 위한 structural SVMs은 다음과 같이 정의된다[8].

$$\min_{w,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i}^{n}\xi_i \quad , s.t.\ \forall i, \xi_i \geq 0$$
$$\forall i, \forall \mathbf{y} \in Y \setminus \mathbf{y}_i : \mathbf{w}^T \delta\Psi_i(\mathbf{x}_i, \mathbf{y}) \geq L(\mathbf{y}_i, \mathbf{y}) - \xi_i \tag{1}$$

위 식에서, $(\mathbf{x}_i,\mathbf{y}_i)$는 학습데이터의 한 문장에 대한 형태소 열과 그에 대응되는 개체명 태그 열을 나타내고, $L(\mathbf{y}_i,\mathbf{y})$는 정답 태그 열 $\mathbf{y}_i$와 결과 태그 열 $\mathbf{y}$의 loss 함수를 나타내며, $\delta\Psi(\mathbf{x}_i,\mathbf{y}) = \Psi(\mathbf{x}_i,\mathbf{y}_i) - \Psi(\mathbf{x}_i,\mathbf{y})$로 정의하며, $\Psi(\mathbf{x},\mathbf{y})$는 자질(feature) 벡터 함수를 나타낸다. $\Psi(\mathbf{x},\mathbf{y})$는 문제의 성질에 따라 달라지는데, 그림 1은 개체명 인식 문제에서 $\Psi(\mathbf{x},\mathbf{y})$의 예를 나타낸다.

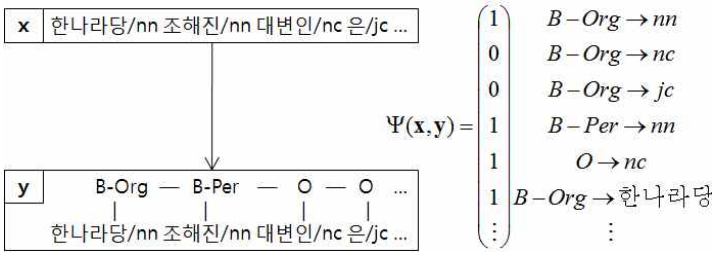

그림 1. 개체명 인식의 $\Psi(x,y)$ 예제

바이너리 분류 및 멀티클래스 분류를 지원하는 기존의 SVMs과 달리 structural SVMs는 더욱 일반적인 구조의 문제(예를 들어, 형태소 태깅, 청킹, 개체명 인식, 구문 분석 등)를 지원한다.

Structural SVMs의 학습에는 cutting-plane algorithm이 사용되어 $O(1/e^2)$의 iteration에 학습이 완료될 수 있다[8]. 최근에 1-slack 형태의 structural SVMs이 제안되었으며 기존의 structural SVMs 보다 빠른$O(1/e)$의 iteration에 학습이 완료될 수 있음이 증명되었다[9]. 개체명 인식을 위한 1-slack structural SVMs은 다음과 같이 정의된다.

$$\min_{w,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi \ , \ s.t. \ \forall i, \xi \geq 0$$
$$\forall(\hat{\mathbf{y}}_1,...,\hat{\mathbf{y}}_n) \in Y^n : \frac{1}{n}\mathbf{w}^T \sum_{i=1}^{n} \partial\Psi_i(\mathbf{x}_i, \hat{\mathbf{y}}_i) \geq \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{y}_i, \hat{\mathbf{y}}_i) - \xi. \quad (2)$$

1-slack structural SVMs의 학습을 위해서 1-slack cutting-plain 알고리즘이 사용되며, 그림 2는 이 알고리즘이다. 본 논문에서는 개체명 인식을 위한 1-slack structural SVMs의 학습을 위해 그림 2의 1-slack cutting 알고리즘과 학습속도를 좀더 개선시킨 [11]의 수정된 FSMO 알고리즘을 사용하였다.

1-Slack Cutting Plane Algorithm

1: Input: $(\mathbf{x}_1,\mathbf{y}_1), \ldots, (\mathbf{x}_n,\mathbf{y}_n), C, e$
2: $S \leftarrow \emptyset$
3: **repeat**
$(\mathbf{w}, \xi) \leftarrow \arg\min_{\mathbf{w},\xi>0} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\xi$
4: $s.t. \forall(\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_n) \in S :$
$\frac{1}{n}\mathbf{w}^T\sum_{i=1}^{n}\partial\Psi_i(\mathbf{x}_i,\hat{\mathbf{y}}_i) \geq \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{y}_i,\hat{\mathbf{y}}_i) - \xi$
5: **for** i=1, ..., n **do**
6: $\hat{\mathbf{y}}_i \leftarrow \max_{\mathbf{y}\in Y}\{L(\mathbf{y}_i,\mathbf{y}) + \mathbf{w}^T\Psi(\mathbf{x}_i,\mathbf{y})\}$
7: **end for**
8: $S \leftarrow S \cup \{(\hat{\mathbf{y}}_1, \cdots, \hat{\mathbf{y}}_n)\}$
9: **until** $\frac{1}{n}\sum_{i=1}^{n} L(\mathbf{y}_i,\hat{\mathbf{y}}_i) - \frac{1}{n}\mathbf{w}^T\sum_{i=1}^{n}\partial\Psi_i(\mathbf{x}_i,\hat{\mathbf{y}}_i) \leq \xi + e$
10: return $(\mathbf{w}, \xi)$

그림 2. 1-slack cutting-plane 알고리즘

한국어 개체명 인식을 위한 자질 벡터 Ψ(**x**,**y**)는 다음과 같은 자질로 이루어진다 (CRFs 및 수정된 Pegasos 알고리즘에 공통으로 쓰임).

- 어휘 자질: (-2,-1,0,1,2) 위치에 해당하는 형태소 어휘 정보.
- 접미사(suffix) 자질: (-2,-1,0,1,2) 위치에 해당하는 형태소 어휘의 접미사(suffix) 정보.
- 형태소 태그 자질: (-2,-1,0,1,2)위치에 해당하는 형태소의 POS tag 정보.
- 형태소 태그 + 길이 자질: 형태소의 태그 정보와 형태소의 길이 정보 조합
- 형태소의 어절 내 위치: 형태소가 어절의 시작, 중간, 끝 위치에 있는 지에 대한 정보
- 개체명 사전 + 길이 자질: 개체명 사전에 존재하는 지에 대한 정보와 형태소의 길이 정보 조합
- 개체명 사전 자질 + 형태소 길이 정보
- 15개의 정규 표현식: [A-Z]*, [0-9]*, [0-9][0-9], [0-9][0-9][0-9][0-9], [A-Za-z0-9]*, … .

Pegasos 알고리즘 기반 개체명 인식

Pegasos 알고리즘은 SGD 방법의 하나로 바이너리 분류에서 높은 성능과 짧은 학습 시간을 보였다[13]. 본 논문에서는 Pegasos 알고리즘을 structural SVMs으로 확장하여 이를 개체명 인식에 적용하였다. 먼저 Pegasos 알고리즘의 object 함수를 structural SVMs의 object 함수의 근삿값으로 교체하면 다음과 같다.

$$f(\mathbf{w};A_t)=\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}+\frac{1}{k}\sum_{(\mathbf{x}_i,\mathbf{y}_i)\in A_t} l(\mathbf{w};(\mathbf{x}_i,\mathbf{y}_i))$$
$$\text{where } l(\mathbf{w};(\mathbf{x}_i,\mathbf{y}_i))=\max_{\mathbf{y}}\left\{L(\mathbf{y}_i,\mathbf{y})-\mathbf{w}^T\,\mathrm{f}(\mathbf{x}_i,\mathbf{y}_i)+\mathbf{w}^T\,\mathrm{f}(\mathbf{x}_i,\mathbf{y})\right\} \quad (3)$$

위 식에서 k는 sub-gradient를 계산하기 위한 학습데이터의 수이고, A_t는 전체 학습데이터 $S=\{(\mathbf{x}_i,\mathbf{y}_i): i=1,2,...,n\}$의 부분 집합(subset)이며, λ는 튜닝을 위한 상수 값으로 structural SVMs의 C값과는 $\lambda = 1/C$ 관계를 갖는다. 수식 (3)에서 $f(\mathbf{w};A_t)$의

sub-gradient를 구하면 다음과 같다.

$$\nabla f(\mathbf{w}; A_t) = \lambda \mathbf{w} - \frac{1}{|A_t|} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in A_t} \{\mathrm{f}(\mathbf{x}_i, \mathbf{y}_i) - \mathrm{f}(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}$$
$$\text{where } \hat{\mathbf{y}}_i = \arg\max_{\mathbf{y}} \{L(\mathbf{y}_i, \mathbf{y}) - \mathbf{w}^T \mathrm{f}(\mathbf{x}_i, \mathbf{y}_i) + \mathbf{w}^T \mathrm{f}(\mathbf{x}_i, \mathbf{y})\}. \quad (4)$$

만약 A_t=S(즉, k=n)인 경우, Pegasos 알고리즘은 sub-gradient projection method가 되며, k=1인 경우 Pegasos 알고리즘은 stochastic gradient method의 하나의 변형이 된다.

그림 3은 개체명 인식을 위해서 수정된 Pegasos 알고리즘이다. 이 알고리즘은 입력으로 알고리즘의 반복(iteration) 회수 T와 sub-gradient를 계산하기 위한 학습데이

A modified Pegasos algorithm for NER

1: Input: S, λ, T, k
2: Initialize: Choose $\mathbf{w}_1$ s.t. $\|\mathbf{w}_1\| \le 1/\sqrt{\lambda}$, $\mathbf{v} = 0$
3: For $t = 1,2,\ldots,T$
4: Choose $A_t \subseteq S$, where $|A_t| = k$
5: $\forall (\mathbf{x}_i, \mathbf{y}_i) \in A_t : \hat{\mathbf{y}}_i = \arg\max_{\mathbf{y}} \{L(\mathbf{y}_i, \mathbf{y}) + \mathbf{w}^T \mathrm{f}(\mathbf{x}_i, \mathbf{y})\}$
6: $\eta_t = 1/\lambda t$
7: $\mathbf{w}_{t+1/2} = (1 - \eta_t \lambda)\mathbf{w}_t + \frac{\eta_t}{k} \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in A_t} \{\mathrm{f}(\mathbf{x}_i, \mathbf{y}_i) - \mathrm{f}(\mathbf{x}_i, \hat{\mathbf{y}}_i)\}$
8: $\mathbf{w}_{t+1} = \min\left\{1, \frac{1/\sqrt{\lambda}}{\|\mathbf{w}_{t+1/2}\|}\right\} \mathbf{w}_{t+1/2}$
9: $\mathbf{v} = \mathbf{v} + \mathbf{w}_{t+1}$
10: $\mathbf{w}_{\text{averaged}} = \mathbf{v}/T$
11: Output: $\mathbf{w}_{T+1}$ and $\mathbf{w}_{\text{averaged}}$

그림 3. 수정된 Pegasos 알고리즘

터의 수 k를 받는다. 초기에 벡터 $\mathbf{w}_1$을 크기가 $1/\sqrt{\lambda}$보다 작은 임의의 벡터로 설정한다. 반복(iteration) 회수가 t인 경우에, 먼저 크기가 k인 A_t를 전체 학습데이터로부터 임의로 고르고(4행), A_t 안에 있는 학습데이터로부터 가장 잘못된(most violated) 개체명 태그 열을 구한 후(5행), 학습 비율(learning rate)을 설정한 후(6행), $\mathbf{w}_{t+1/2}$를 구한 후(7행), $\mathbf{w}_{t+1/2}$를 $\{w:|w| \leq 1/\sqrt{\lambda}\}$ 집합으로 투영(projection)한 벡터를 $\mathbf{w}_{t+1}$로 설정한다(8행). 알고리즘의 결과는 최종 $\mathbf{w}_{T+1}$ 벡터와 평균화된(averaged) 벡터 $\mathbf{w}_{averaged}$가 된다(행11). 실험 결과 $\mathbf{w}_{T+1}$과 $\mathbf{w}_{averaged}$는 성능상 큰 차이가 없었으며, $\mathbf{w}_{averaged}$의 경우 학습 시간이 좀더 필요하지만 성능의 편차가 적어졌다. 본 논문의 실험에서는 $\mathbf{w}_{T+1}$만을 사용하였다.

실험 및 결과

본 논문에서 제안된 structural SVMs 및 Pegasos 알고리즘을 이용한 한국어 개체명 인식 시스템의 성능을 측정하기 위해서 ETRI의 한국어 개체명 학습데이터 셋 중에서 TV 도메인 및 스포츠 도메인을 사용한다. TV 도메인에서는 2,900문서(105,265문장)를 학습데이터로, 나머지 100문서(3,719문장)를 테스트 데이터로 사용하고, 스포츠 도메인에서는 3,500문서(81,829문장)를 학습데이터로 사용하고, 나머지 100문서(2,735문장)를 테스트 데이터로 사용한다.

ETRI의 개체명 학습데이터는 180여개의 세부분류 개체명 태그 셋으로 이루어져 있는데, 본 실험에서는 15개 대분류 태그 셋(인물, 지역, 기관, 학술분야, 이론, 인공물, 문명/문화 관련 명칭, 날짜, 시간, 수량 표현, 이벤트, 동물, 식물, 물질, 용어)을 이용한다.

수정된 FSMO 알고리즘을 사용하는 1-slack structural SVMs 및 개체명 인식을 위해 확장된 Pegasos 알고리즘은 C++로 구현하였으며, 비교 실험을 위해서 공개된 L-BFGS 알고리즘 모듈을 사용하여 C++로 CRFs를 구현하였다. CRFs의 성능 최적화를 위해서 가우시안 확률(Gaussian prior)을 {0.01, 0.1, 1, 10, 100} 값으로 바꾸어 가면서 최적의 성능치를 구했으며 500번 반복(iteration)하였다. 1-slack structural SVMs

및 Pegasos 알고리즘의 성능 최적화를 위해서 C값을 {1000, 2000, 5000, 10000, 20000}로 바꾸어 가면서 최적의 성능치를 구했으며, 1-slack structural SVMs의 경우 종료 조건으로 e=0.1을 사용하고, Pegasos 알고리즘은 k=100을 사용하고 $|1-f(\mathbf{w}_{t+1})/f(\mathbf{w}_t)|<0.001$ 이면 종료하였다. 모든 알고리즘은 똑 같은 자질을 사용하였고, 모든 실험은 인텔 코어 i7 CPU 3.33GHz와 12GB의 RAM으로 구성된 PC에서 수행되었다.

표 1은 TV 도메인에서의 한국어 개체명 인식 성능이다. CRFs 시스템의 성능을 기준으로 했을 때, 1-slack structural SVMs 및 수정된 Pegasos 알고리즘의 정확도(accuracy)가 각각 0.14, 0.16 향상되었고, F1 값은 각각 0.15, 0.44가 향상되었다. 학습 시간은 CRFs가 가장 오래 걸렸으며, 수정된 Pegasos 알고리즘은 CRFs의 학습 시간의 약 4% 밖에 걸리지 않았다. 각 알고리즘의 성능 차이가 통계적으로 유의한지 알아보기 위해서 스튜던트 t-검정을 수행하였으며 그 결과로 CRFs와 1-slack structural SVMs의 성능차이는 유의수준 0.01(신뢰도 99%)에서 통계적으로 유의하며(p값=0.0037), CRFs와 수정된 Pegasos 알고리즘의 성능차이는 유의수준 0.01(신뢰도 99%)에서 통계적으로 유의하고(p값=0.0018), 1-slack structural SVMs과 수정된 Pegasos 알고리즘의 성능차이는 통계적으로 유의하지 않았다(p값=0.69).

표 1. TV 도메인 개체명 인식 성능

기계학습 알고리즘	학습시간 (초)	정확도 (accuracy)	F1
CRFs (baseline)	16738	96.78	84.99
1-slack structural SVMs	11239	96.92 (+0.14)	85.14 (+0.15)
modified Pegasos	**649**	**96.94 (+0.16)**	**85.43 (+0.44)**

표 2는 스포츠 도메인에서의 한국어 개체명 인식 성능이다. CRFs 시스템의 성능을 기준으로 했을 때, 1-slack structural SVMs 및 수정된 Pegasos 알고리즘의 정확도(accuracy)가 각각 0.24, 0.23 향상되었고, F1 값은 각각 0.22, 0.15가 향상되었다. 학습 시간은 CRFs가 가장 오래 걸렸으며, 수정된 Pegasos 알고리즘은 CRFs의 학습 시

간의 약 4% 밖에 걸리지 않았다. 스튜던트 t-검정결과 각 알고리즘의 성능 차이는, CRFs와 1-slack structural SVMs의 성능차이는 유의수준 0.001(신뢰도 99.9%)에서 통계적으로 유의하며(p값=1.9E-07), CRFs와 수정된 Pegasos 알고리즘의 성능차이 역시 유의수준 0.001(신뢰도 99.9%)에서 통계적으로 유의하고(p값=2.6E-07), 1-slack structural SVMs과 수정된 Pegasos 알고리즘의 성능차이는 통계적으로 유의하지 않았다(p값=0.72).

표 2. 스포츠 도메인 개체명 인식 성능

기계학습 알고리즘	학습시간 (초)	정확도 (accuracy)	F1
CRFs (baseline)	14362	95.58	86.64
1-slack structural SVMs	5991	**95.82 (+0.24)**	**86.86 (+0.22)**
modified Pegasos	**610**	95.81 (+0.23)	86.79 (+0.15)

위의 실험들로부터 한국어 개체명 인식에서 structural SVMs 및 수정된 Pegasos 알고리즘이 기존의 CRFs 보다 높은 성능을 보이고(신뢰도 99%에서 통계적으로 유의함), structural SVMs 및 수정된 Pegasos 알고리즘의 성능 차이는 없으며(통계적으로 유의하지 않음), 수정된 Pegasos 알고리즘의 학습 속도가 가장 빠름(CRFs의 학습시간의 4% 소요됨)을 알 수 있다.

결 론

본 논문에서는 structural SVMs 및 수정된 Pegasos 알고리즘을 이용한 한국어 개체명 인식 시스템에 대하여 기술하고 기존의 CRFs를 이용한 시스템과의 성능을 비교하였다. 실험결과 structural SVMs 및 수정된 Pegasos 알고리즘이 기존의 CRFs 보다 높은 성능을 보이고(신뢰도 99%에서 통계적으로 유의함), structural SVMs 및 수정된 Pegasos 알고리즘의 성능 차이는 없음(통계적으로 유의하지 않음)을 알 수 있었다.

특히 수정된 Pegasos 알고리즘을 이용한 경우 CRFs를 이용한 시스템보다 높은 성능(TV 도메인 F1=85.43, 스포츠 도메인 F1=86.79)을 유지하면서 학습 시간은 CRFs의 학습 시간의 4%로 줄일 수 있었다.

앞으로 좀더 높은 성능의 한국어 개체명 인식 시스템을 개발하기 위해서 다양한 자질의 발굴이 필요할 것이다. 특히 단어 군집(word clustering) 자질 같은 대용량의 말뭉치를 사용한 자질이나, 워드넷, 위키피디아 등과 같은 외부자원을 이용한 자질을 추가하면 좀더 높은 성능 향상이 있을 것으로 기대된다.

참고문헌

[1] Borthwick, A., Sterling, J., Agichtein, E., Grishman, R., "NYU: Description of the MENE named entity system as used in MUC-7," MUC-7, 1998.

[2] Kim Sang, E. F. T., de Meulder, F., "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," CoNLL, 2003.

[3] Ratnaparkhi, A, "A Simple Introduction to Maximum Entropy Models for Natural Language Processing," University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-97-08, 1997.

[4] Lafferty, J., McCallum, A.,Pereira, F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," ICML, 282 – 2289, 2001.

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

[6] Changki Lee, et al., "Fine-Grained Named Entity Recognition Using Conditional Random Fields for Question Answering," Lecture notes in computer science (AIRS2006), 2006.

[7] Changki Lee, Yi-Gyu Hwang, Myung-Gil Jang, "Fine-grained named entity recognition and relation extraction for question answering," SIGIR 799-800, 2007.

[8] I. Tsochantaridis, et al., "Support Vector Machine Learning for Interdependent and

Structured Output Spaces," Proc. ICML, 2004.

[9] T. Joachims, T. Finley, C.N. Yu, "Cutting-Plane Training of Structural SVMs," MLJ, 2008.

[10] Changki Lee, Myung-Gil Jang, "Fast Training of Structured SVM Using Fixed-Threshold Sequential Minimal Optimization," ETRI Journal, vol.31, no.2, 121-128, 2009.

[11] Changki Lee, Myung-Gil Jang, "A Modified Fixed-threshold SMO for 1-Slack Structural SVM," ETRI Journal, vol.32, no.1, 120-128, 2010.

[12] L. Bottou and O. Bousquet, "The Tradeoffs of Large Scale learning," NIPS 20, 2008

[13] S. Shalev-Shwartz, et al., "Pegasos: Primal Estimated sub-GrAdient SOlver for SVM," Proc. ICML, 2007.

[14] Changki Lee, Soojong Lim, Myung-Gil Jang, "Large-Margin Training of Dependency Parsers Using Pegasos Algorithm," ETRI Journal, vol.32, no.3, 490-492, 2010.

1 차원고접수 : 2010. 10. 15
2 차원고접수 : 2010. 12. 7
최종게재승인 : 2010. 12. 8

(*Abstract*)

Named Entity Recognition with Structural SVMs and Pegasos algorithm

Changki Lee Myungil Jang

ETRI, Knowledge Mining Lab.

The named entity recognition task is one of the most important subtasks in Information Extraction. In this paper, we describe a Korean named entity recognition using structural Support Vector Machines (structural SVMs) and modified Pegasos algorithm. Using the proposed approach, we could achieve an 85.43% F1 and an 86.79% F1 for 15 named entity types on TV domain and sports domain, respectively. Moreover, we reduced the training time to 4% without loss of performance compared to Conditional Random Fields (CRFs).

Key words : Korean Named Entity Recognition, Structural SVMs, modified Pegasos algorithm