Understanding Social Media Texts with Minimum Human Effort on #Twitter

Tian Tian, Isabelle Tellier, Marco Dinarelli, Pedro Cardoso

Understanding Social Media Texts with Minimum Human Effort on #Twitter

Tian Tian[*1,2,3], Isabelle Tellier[†1,3], Marco Dinarelli[‡1] and Pedro Cardoso[§2]

[1]Lattice, UMR 8094, 1 rue Maurice Arnoux, 92120 Montrouge
[2]Synthesio, 8-10 rue Villedo, 75001 Paris
[3]Sorbonne Nouvelle - Paris 3

Keywords: Natural Language Processing, Machine Learning, Named Entity Recognition, Domain Adaptation, Conditional Random Fields (CRF)

1 Task

Named Entity Recognition (NER) is a traditional Natural Language Processing (NLP) task. But traditional machine learning methods are facing new problems to handle this task with Social Media data like Twitter. In this new context, the performance is often degraded like in [RCME11].

The Twitter messages have particular features. Consider the example described in Figure 1.

Today wasz Fun cusz anna Came juss for me <3: hahaha

Figure 1: A tweet example

In this example, the difficulties are manifold:

- Spelling mistakes: wasz (was), cusz (because), juss (just);
- Uppercase/lowercase inversion: Fun (fun), anna (Anna), Came (came);
- Emoticon: <3;
- Interjection: hahaha.

The alternation of uppercase/lowercase is a major problem for the NER task because the only person proper noun "anna" of our tweet begins with a lowercase instead of an uppercase, like in grammatically well-formed texts.

[*]ttian@synthesio.com
[†]isabelle.tellier@ens.fr
[‡]marco.dinarelli@ens.fr
[§]pedro@synthesio.com

2 Method and baseline

In this paper, we present our work on recognizing named entities on Twitter.

[RCME11] proposed a corpus of Twitter data annotated with named entities (Ritter corpus). We adapted this corpus to our task by annotating job titles and by merging some other types (music artist, movie and TV show into media). This corpus contains 2394 tweets (sequences), that is about 48k tokens.

First, we trained a CRF model using the Ritter corpus. We tested this model on our own reference annotated corpus on a domain named "Deezer". This reference corpus "Deezer" contains only 50 tweets (sequences), that is 850 tokens. We consider the result of this model as the baseline. The Table 1 shows numbers of each entity type for these two corpora.

Entity Type	Ritter	Deezer
facility	107	0
company	186	32
person	472	17
location	291	1
sports team	55	0
media	126	16
product	102	37
job title	87	1
other	246	1

Table 1: Named entities in Ritter corpus

3 Approach and results

Since it is easy to collect unlabeled Twitter data, our idea is to use unlabeled texts to improve the NER performance with a domain adaptation approach proposed in [GFD14]. The process consists firstly in applying the baseline model on unlabeled texts, keeping the sequences annotated with a probability greater than 0.9.

For the first step, we collected 100 million tweets in the domain "Deezer" on Twitter. Secondly, another model is trained with the selected sequences added to the Ritter corpus. Then, this new model is applied on sequences that are ignored in the previous steps and so on, in an iterative learning. In our CRF models, we used POS features predicted by the tagger of [TDTC15]. The Table 2 shows our preliminary results with only one iteration compared to the baseline. There were 10549 tweets (sequences) that reached over 0.9 probability and contain at least one entity in the first step.

We can see from this table that the iterative training method improved the NER result, even with only one iteration. We expect to get even better results with more iterations.

	Baseline			Our model		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Company	0.125	0.03	0.05	1.0	0.03	0.05
Person	0.36	0.29	0.32	0.35	0.41	0.38
Other	0.5	1.0	0.67	0.5	1.0	0.67
Product	0.5	0.05	0.10	0.5	0.03	0.05
Media	0	0	0	0.5	0.0625	0.11
Geo-Location	0	0	0	0	0	0
Job title	0	0	0	0	0	0
Micro-Average	0.28	0.09	0.11	0.34	0.1	0.12

Table 2: Results of our model

4 Conclusion and future work

In this paper, we tried to improve the NER performance on Twitter by iterative training. We plan to use normalization techniques from [LL15] and [BdMH$^+$15] to further improve our performances.

References

[BdMH$^+$15] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135, Beijing, China, July 2015. Association for Computational Linguistics.

[GFD14] Anne Garcia-Fernandez, Olivier Ferret, and Marco Dinarelli. Evaluation of different strategies for domain adaptation in opinion mining. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*, pages 3877–3880. European Language Resources Association (ELRA), 2014.

[LL15] Chen Li and Yang Liu. Improving named entity recognition in tweets via detecting non-standard words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 929–938. The Association for Computer Linguistics, 2015.

[RCME11] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language*

Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1524–1534. ACL, 2011.

[TDTC15] Tian Tian, Marco Dinarelli, Isabelle Tellier, and Pedro Cardoso. Etiquetage morpho-syntaxique de tweets avec des crf. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, pages 607–613, Caen, France, June 2015. Association pour le Traitement Automatique des Langues.