Assignment 1: Named Entity Recognition

Wen Xie
University of Houston
wxie5@uh.edu

Abstract

Named Entity Recognition (NER) is a hot topic in Natural Language Processing (NLP) community due to its fundamental necessity. In this study, we try to increase the prediction accuracy by adopting various features such as lexical features,shape features, and grammatical and syntactic features. The primary algorithm used in this paper is the typical Conditional Random Field (CRF). The data-set include 7154 sentences in total. After training the CRF model with 3-fold cross validation and 50 iterations, the best model performs well on the validation data-set with F-1 score 29.56 percent and accuracy 94.053 percent.

1 Introduction

With the widely use of social media platforms, sentences become more and more informal and irregular. Named Entity Recognition (NER) is a challenging work because of the extremely wide scope of entities needed to be recognized. First of all, there are a lot of entities such as person, company, and location. Inside the person category, there are many sub-entities such as writer, singer, dancer. However, very few labeled data-sets exist. Secondly, due to the basic characteristics of entities, the data-set is doomed to be unbalanced. The number of each category is severely unequal. Most of our focused entities only account for a small proportion of the whole data-set, of which most data are useless.

According to the survey paper (Nadeau and Sekine, 2007), Previous researchers try to increase the prediction accuracy based on three main approaches. First, researchers try to investigate different kinds of textual genre (journalistic, scientific, informal, etc.) or domain factor (gardening, sports, business, etc.) because different genres or domains have their unique terminologies. Second, people focus on the learning models. Basically, learning models can be divided into three directions. Supervised learning (SL) is the first category. SL techniques include Hidden Markov Models (HMM)(Bikel et al., 1998), Decision Trees(Sekine, 1998), Maximum Entropy Models (ME)(Borthwick et al., 1998), Support Vector Machines (SVM)(Asahara and Matsumoto, 2003), and Conditional Random Fields (CRF)(McCallum and Li, 2003). The above-mentioned models are typical ones and can be used to combine with other models for further improvements. Apart from SL, semi-supervised learning is another one. According to Nadeau and Sekine (2007), the main technique for SSL is called "bootstrapping" and involves a small degree of supervision, such as a set of seeds, for starting the learning process. At last, it's unsupervised learning models.

Third, researchers pay much attention to feature space(Nadeau and Sekine, 2007). Word-level features include different cases (Starts with a capital letter;Word is all uppercased), punctuation (ends with period; has internal period), digit (digit pattern), character (Possessive mark; first person pronoun), morphology (Prefix; suffix; singular version; stem), part-of-speech (proper name; verb; noun; foreign word), and function (alpha; non-alpha; n-gram). Besides word-level features, list lookup features are also informative because of the prior knowledge. Lists are the privileged features in NER. The terms "gazetteer", "lexicon" and "dictionary" are often used interchangeably with the term "list". Three kinds of features are included in gazetteer features that are general lists (Capitalized nouns;Stop words), list of entities (organization; government; airline), and list of entity cues (typical words in organization). Except the aforementioned features, document and corpus level features are also important. Meta information and corpus frequency are usually considered.

In this study, the CRF model is utilized to finish the assignment. The structure of this report is as

label	tokens	label	tokens
B-company	144	I-company	31
B-group	85	I-group	63
B-location	316	I-location	124
B-other	177	I-other	230
B-person	361	I-person	168
B-product	74	I-product	61
B-title	51	I-title	55
O	35679		

Table 1: train data-set labels.

follows. In section one, I talk about introduction. Methodology is documented in section two. After that, experiment results and interpretations are shown in section three. At last, it's conclusion.

2 Methodology

In this part, I will introduce my methodology from two aspects. The first part is about the features extraction. I will talk about my methods used to extract features for NER. Then I will talk about the CRF model.

2.1 Features Extraction

In the NER task, the labeled data-set include 14 types of labels. Table 1 summarize the overall information of the fourteen labels. It's clear that the data-set is extremely unbalanced. The label O account for the majority of the data-set.

According to previous researches, we know word-level features are of vital importance. So in this paper, I have selected a variety of features including alpha-case, punctuation-case, stop-case, lemma-case, uppercase, title-case, camel-case and lowercase. Besides the basic word shape features, part-of-speech tags are considered too. Three methods are used. I trained a comprehensive tagger with bigramtagger and unigramtagger based on brown corpus and treebank corpus respectively. Besides these two kinds of tags, pos-tag labels is realized based on nltk package as well. Morphology features include prefix and suffix features, which I consider the first two, three letters and the last two, three letters as additional features. Lastly, I consider the space information between adjacent words. The features of the former and latter one word are also considered as the features of the current word itself. Table 2 illustrates all features I use in the NER task.

case	morphology	function	pos
upper	word[:2]	alpha	brown
title	word[:3]	lower	treebank
digit	word[-2:]	punc	pos
camel	word[-3:]	length	
lemma			

Table 2: whole features.

2.2 Conditional Random Field

Conditional Random Field (CRF) model is asked to be used to finish the task. After traininig the CRF model, it's used to predict IOB labels of the test data-set. CRF model combines the advantages of Hidden Markov Model (HMM), which can learn useful information from the conditional states. For example, the features in section 2.1 can be used as the conditional states. Based on these features or states, the model can learn certain criteria and make predictions. In this study, I use the sklearn-crfsuite package in python.

3 Experiment and Result

In this section, we talk about the data processing, training CRF model, and analysis of the results.

3.1 Data Pre-processing

There are three data-sets including training data-set, validation data-set, and test data-set. Training and validation data-sets are from WNUT data-set. Test data-set is from anonymous social media platform. First, sentences are recovered from the raw data-set based on the space line. The training data-set includes 1969 sentences. The validation data-set includes 915 sentences while the test data-set includes 4270 sentences. Then, bigramtagger and unigramtagger are trained based on the brown corpus and treebank corpus respectively. After that, the part-of-speech tags are added in the data-set which is processed by pandas package in python. At last, I convert the data-set into dict version which is the format of the input of the CRF model. Related features are grouped based on the sentence number. Then, the data-sets are available for training, validation and testing.

3.2 Training CRF model

CRF model is imported in python from the sklearn-crfsuite package. The hyparameters are selected as below. The learning algorithm is LBFGS which is

name	result
c1	0.0589
c2	0.0498
CV score	0.4256

Table 3: fine-tune results.

label	prec	recall	f1	supp
B-company	0.64	0.21	0.31	34
I-company	0.00	0.00	0.00	8
B-group	0.42	0.05	0.09	100
I-group	0.10	0.02	0.04	43
B-location	0.60	0.47	0.53	137
I-location	0.44	0.25	0.32	72
B-other	0.27	0.11	0.15	121
I-other	0.10	0.20	0.13	75
B-person	0.60	0.50	0.55	147
I-person	0.65	0.64	0.64	80
B-product	0.50	0.06	0.11	32
I-product	0.40	0.02	0.04	88
B-title	0.00	0.00	0.00	10
I-title	0.00	0.00	0.00	13
micro avg	0.42	0.26	0.32	960
macro avg	0.34	0.18	0.21	960
weighted avg	0.43	0.26	0.30	960

metrics	overall score
accuracy score	0.9405
F1 score	0.2956

Table 4: validation results.

a limited-memory quasi-Newton code for unconstrained optimization. The maximum iterations of the single learning model is set to 200. The L1 and L2 regularization coefficients c1 and c2 are fine-tuned based on 3-fold cross validation and 50 iterations. So in total, we fine-tuned the model and tried 150 pairs of (c1,c2) based on random search cross validation algorithm. And the best result is shown in table 3. The overall F1 score is up to 29.56 percent excluding the O label predictions.

With the best parameters, I validate its performs on the dev data-set. The validation result shows in table 4.

3.3 Prediction results

With the best model, I predict the IOB labels of each token in the testing data-set. Totally, there are 4270 sentences and 42413 predictions.

4 Conclusion

From the table 4, we can find that some entities can be predicted with a good result. However, the title and product entity is hard to predict. One reason is due to the small proportion of training samples. The other reasons may include the feature extraction failure or useless to these two kinds of entities. Except these two entities, others seem to be predicted more accurately especially the person and location entities with f1 score 64 percent and 53 percent respectively.

In the future, more useful features should be considered such as chunking features and parsing features.

References

Masayuki Asahara and Yuji Matsumoto. 2003. Japanese named entity extraction with redundant morphological analysis. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 8–15. Association for Computational Linguistics.

Daniel M Bikel, Scott Miller, Richard Schwartz, and Ralph Weischedel. 1998. Nymble: a high-performance learning name-finder. *arXiv preprint cmp-lg/9803003*.

Andrew Borthwick, John Sterling, Eugene Agichtein, and Ralph Grishman. 1998. Description of the mene named entity system as used in muc-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, April 29-May 1, 1998*.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Satoshi Sekine. 1998. Description of the japanese ne system used for met-2. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.