Structured prediction: NER Tagging Conll2003

Ildar Nurgaliev, Innopolis university

Abstract

We address the problem of named entity tagging for English data from the popular conll2003 dataset that consist of four types of named entities: persons, locations, organizations and names of miscellaneous entities. We develop features, and report tagging results nearing 97.7% accuracy.

I. Introduction

Named Entity Recognition (NER) usually define the task of identifying an entity in a text and assign a class label to it. For example, localizing a word sequence designing a person, like [Paris Hilton], and assigning a class label (like PER for a Person or LOC for a Locality) to the words.

NER is first finding and identifying named entities, such as people, brand and company names in text and associationg them with the respective person or company. The process comprises of many steps:

1) You have to break the text into sentences and then into words within the sentence
2) Then you identify which words are which (noun, adjective, verb, title, common first name, etc.)
3) Then you come up with "candidates" these are words, normally nouns that will be "disambiguated" and identified.
4) The identification process takes place and some of the candidates are discharged, some are matched.

II. Problem Statement

We will concentrate on four types of named entities: persons, locations, organizations and names of miscellaneous entities that do not belong to the previous three groups. We will use the data (conll2003 english) for developing a named-entity recognition system that includes a machine learning component. The challenge is to find ways of incorporating this information in the system.

- **Item precision**
- **item f1-score**
- **Instance precision** = $\frac{\wedge(S)}{S}$, where S is count of sentences and $\wedge(S)$ - count of correct sentences.

Our task is to maximize the objective function of precision for both - item and instance and item's f1-score.

III. The data

We use *conll2003* that has 14041 sentences in train set, 3250 and 3453 sentences in testA and testB files. The English data is a collection of news wire articles from the Reuters Corpus. The annotation has been done by people of the University of Antwerp. Data format is **CoNLL**:

- One example per line (item)
- Tab-separated
- Reference label in the first position
- Other columns are features
- a sentence (instance) is divided from another by a blank line

The first item on each line is a syntactic chunk tag and the named entity tag , the second a word. The chunk tags and the named entity tags have the format I-TYPE which means that the word is inside a phrase of type TYPE. Only if two phrases of the same type immediately follow each other, the first word of the second phrase will have tag B-TYPE to show that it starts a new phrase. A word with tag O is not part of a phrase.

IV. Data analysis

TRAIN

NER-tag	#	Top-5 of TRAIN
I-ORG	3704	Commission, Union, Union, National, Farmers
B-MISC	3438	German, British, German, British, EU-wide
I-PER	4528	Blackburn, Zwingmann, van, der, Pas
I-LOC	1157	Strait, Heights, East, States, Kurdistan
B-LOC	7140	BRUSSELS, Germany, Britain, Britain, France
B-PER	6600	Peter, Werner, Nikolaus, Franz, Fischler
I-MISC	1155	Spongiform, Encephalopathy, n't, no, telling
O	170524	-DOCSTART-, rejects, call, to, boycott
B-ORG	6321	EU, European, European, Commission, European

TEST_A

NER-tag	#	Top-5 of TEST_B
I-ORG	751	and , County , Cricket , Board , Universities
B-MISC	922	West , Australian , ex-England , ENGLISH , English
I-PER	1307	Simmons , Caddick , Hussain , Such , Lewis
I-LOC	257	Road , Wells , Oval , 's , Oval
I-MISC	346	Indian , COUNTY , CHAMPIONSHIP , Cup , 96
B-PER	1842	Phil , Andy , Simmons , Nasser , Peter
B-LOC	1837	LONDON , Grace , England , Headingley , England
O	42975	-DOCSTART- , CRICKET , - , TAKE , OVER
B-ORG	1341	LEICESTERSHIRE , Leicestershire , Somerset , Essex , Derbyshire

TEST_B

NER-tag	#	Top-5 of TEST_B
I-ORG	835	UNION , Wednesday , 's , Moscow , Milan
B-MISC	702	Asian , Uzbek , Chinese , Soviet , Asian
I-PER	1156	Ladki , Shkvyrin , Shatskiku , Takagi , Yanagimoto
I-LOC	257	Arab , Emirates , Korea , Arab , Emirates
B-LOC	1668	JAPAN , AL-AIN , United , Japan , Syria
B-PER	1617	CHINA , Nadim , Igor , Oleg , Takuya
I-MISC	216	Cup , Cup , Games , Cup , World
O	38554	-DOCSTART- , SOCCER , - , GET , LUCKY
B-ORG	1661	FIFA , RUGBY , Plymouth , Exeter , FIFA

Sentence

- TRAIN has 14041 sentences
- TEST_A has 3250 sentences
- TEST_B has 3453 sentences

V. Features extracted

A. Standart features

A word is plenty of features that should be extracted. Below we present survey of features extracted for every word of a sentence.

word position	word position in a sentence
len	len(word)
type	digit or upper case or lower case or something unclear
lexem	pattern.en package
pos_tag	pos_tag is limited by NN — VB — JJ — RB other would have 0
is_word	check if it is real word
word_shape	get chape of a word. Ex: Test $\rightarrow$ Xxxx
previous word	same as the features above with the previous word
next word	the same as the features with the next word
double previous word	take lexem and pos-tag of the double previous wordword

B. Sentence topic feature

We decided to extend feature space by topic feature of a sentence. The topic is revealed per sentence, because mostly in every sentence we have an entity that needs to be revealed, thus a topic that characterizes a sentence could help. To do this we used a well-known topic modeling algorithm called LDA, to uncover the latent topics in the sentences. In total we specified that we wanted to try and identify 15 topics, because $14041 + 3250 + 3453 = 20744$ sentences that is quite a lot and the lower number of topics was not such descriptive as number of 15 topics.

Latent Dirichlet Allocation (LDA) is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics.

There will be explained some clusters created

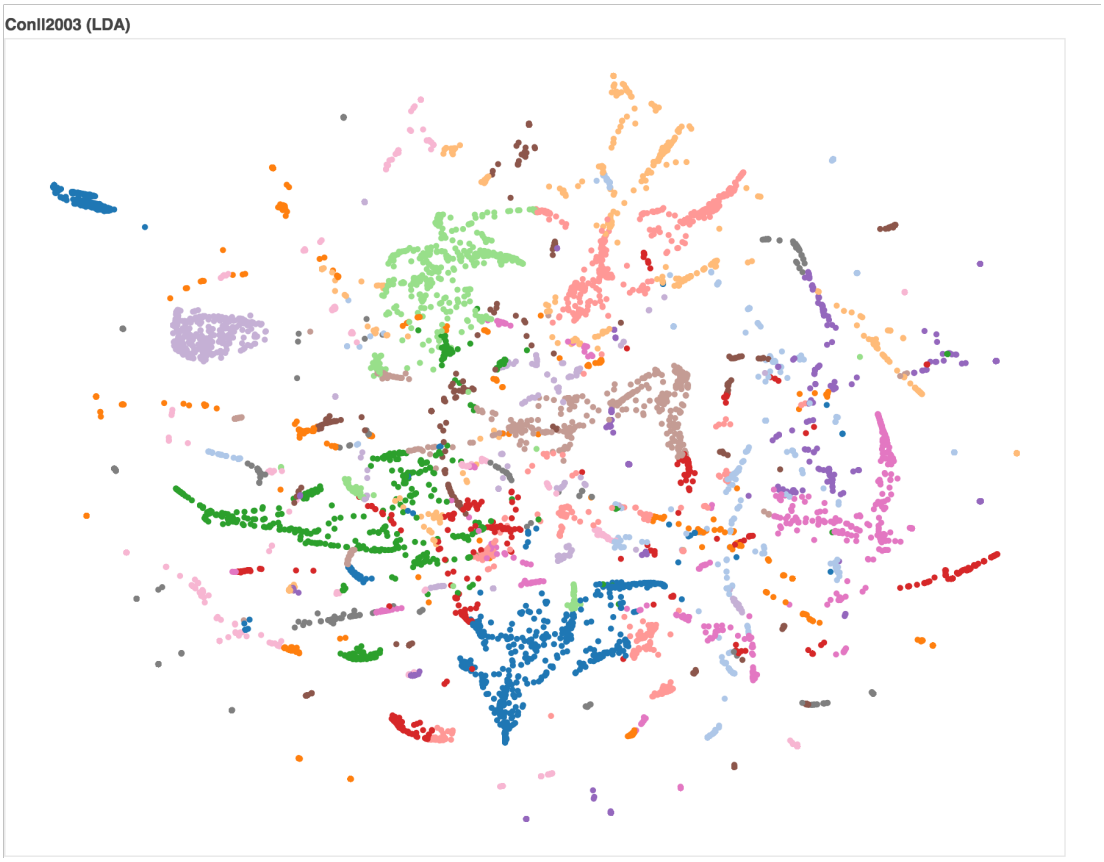

Fig. 1. clustered sentences by LDA. PCA dimension reduction is used to create the plot.

Topic 1 Concluding sentences.
Topic 2 mostly narrative sentences as ”it currently operates” ot ”I asked for a day match”
Topic 3 mostly about sport that is good for name and organization NER tag revealing
Topic 4 competition result sentences like someone (VS) someone. Mostly team names or sportsmen names.
Topic 5 mostly about disaster and war. Good for Location revealing
Topic 6 mostly about relations between countries
Topic 7 country + date, FRANKFURT 1996-08-22. Good for location revealing.
Topic 8 etc....

VI. Building model

For 4-class classification task we choosed Stochastic Gradient Descent with L1 and L2. That maximizes the logarithm of the likelihood of the training data with L2 regularization term(s) using Stochastic Gradient Descent with batch size 1. It approaches to the optimal feature weights quite rapidly.

VII. Results

Sixteen systems have participated in the CoNLL-2003 shared task. They used a wide variety of machine learning techniques and different feature sets. Here is the result table for the English test set below.

Our evaluation was designed to test the efficacy of this feature set for NER shared conll2003 task. For studying purpose we decided to test different algorithm with the same set of features. There we present best results. As result we got item accuracy = 97.7%, instance accuracy = 81% and F1-score=88%.

English	precision	recall	F
[FIJZ03]	88.99%	88.54%	88.76±0.7
[CN03]	88.12%	88.51%	88.31±0.7
[KSNM03]	85.93%	86.21%	86.07±0.8
[ZJ03]	86.13%	84.88%	85.50±0.9
[CMP03b]	84.05%	85.96%	85.00±0.8
[CC03]	84.29%	85.50%	84.89±0.9
[MMP03]	84.45%	84.90%	84.67±1.0
[CMP03a]	85.81%	82.84%	84.30±0.9
[ML03]	84.52%	83.55%	84.04±0.9
[BON03]	84.68%	83.18%	83.92±1.0
[MLP03]	80.87%	84.21%	82.50±1.0
[WNC03]*	82.02%	81.39%	81.70±0.9
[WP03]	81.60%	78.05%	79.78±1.0
[HV03]	76.33%	80.17%	78.20±1.0
[DD03]	75.84%	78.13%	76.97±1.2
[Ham03]	69.09%	53.26%	60.15±1.3
baseline	71.91%	50.90%	59.61±1.2

A. *lbfgs*

Performance by label	(#match, #model, #ref)	(precision, recall, F1)
B-ORG:	(1096, 1252, 1305)	(0.8754, 0.8398, 0.8573)
O:	(39760, 40037, 39940)	(0.9931, 0.9955, 0.9943)
B-MISC:	(751, 837, 889)	(0.8973, 0.8448, 0.8702)
B-PER:	(1622, 1786, 1775)	(0.9082, 0.9138, 0.9110)
I-PER:	(1230, 1298, 1279)	(0.9476, 0.9617, 0.9546)
B-LOC:	(1660, 1814, 1806)	(0.9151, 0.9192, 0.9171)
I-ORG:	(594, 740, 716)	(0.8027, 0.8296, 0.8159)
I-MISC:	(247, 300, 327)	(0.8233, 0.7554, 0.7879)
I-LOC:	(201, 223, 250)	(0.9013, 0.8040, 0.8499)

- Macro-average precision, recall, F1: (0.896001, 0.873747, 0.884241)
- accuracy: 47161 / 48287 (0.9777)
- Instance accuracy: 2661 / 3250 (0.8188)

VIII. Concslusion

Since 2003 year of the competition 13 years passed, now we have advanced lexer and pos tag reveling algorithms. In the work we extended the model by feature selection as POS-tag, lexeme. Moreover we used very powerful technique to get else one descriptive feature as topic of a sentence built by LDA algorithm. The most powerful feature there is lexeme + POS-tag. We got significantly higher precision and f2-score with respect to baseline.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[3] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

[4] Yoshimasa Tsuruoka, Jun'ichi Tsujii, and Sophia Ananiadou. Stochastic gradient descent training for l1-regularized log-linear models with cumulative penalty. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 477–485. Association for Computational Linguistics, 2009.