Revisiting CoNNL-2003 with Classical Machine Learning

Matthew A. Hernandez,

INFO 521, Fall 2024

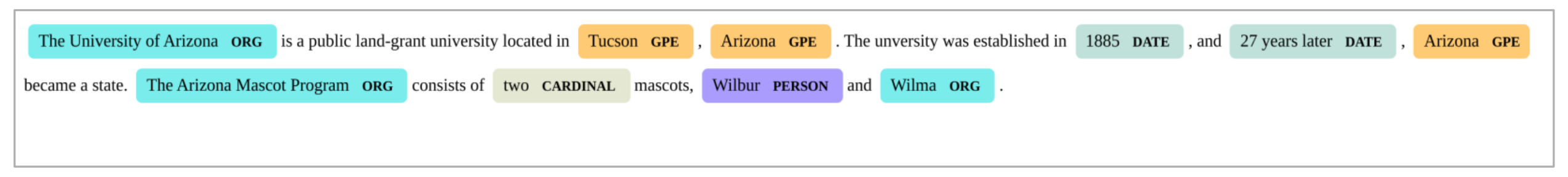
University of Arizona, Linguistics Department

Abstract

We present a brief survey on classical machine learning algorithms in the context of the CoNNL-2003/02 Shared Task. A named entity recognition (NER) system is built to recognize and classify objects in a body of text into predefined categories.

Introduction

The Named Entity Recognition and Classification (NERC) task refers to the process that identifies phrases in text into various categories. Common entity types include ORGANIZATION (ORG), PERSON (PER), LOCATION (LOC), and MISCELLANEOUS (MISC). Tagging is the process of labeling each word in a sentence with its respective named entity (NE). As an example:



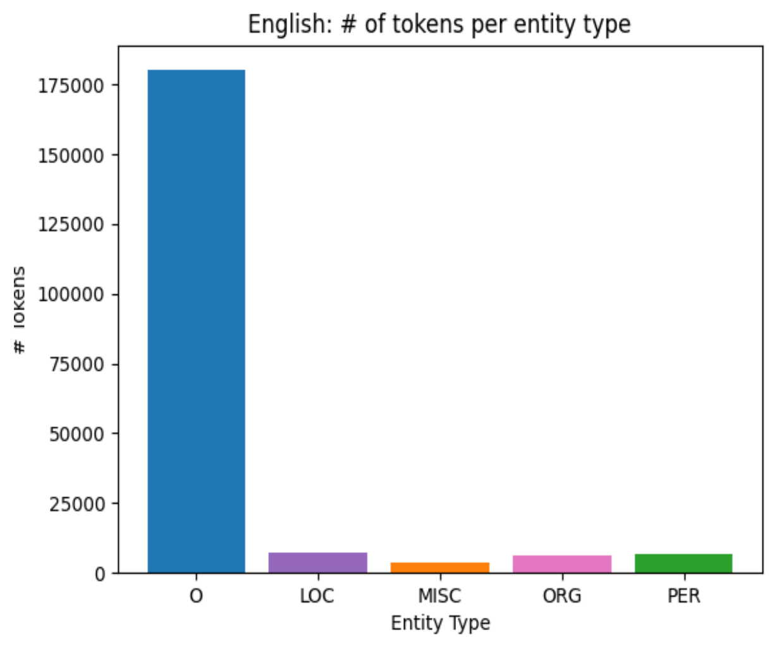
The University of Arizona ORG is a public land-grant university located in Tucson GPE , Arizona GPE . The unversity was established in 1885 DATE , and 27 years later DATE , Arizona GPE became a state. The Arizona Mascot Program ORG consists of two CARDINAL mascots, Wilbur PERSON and Wilma ORG .

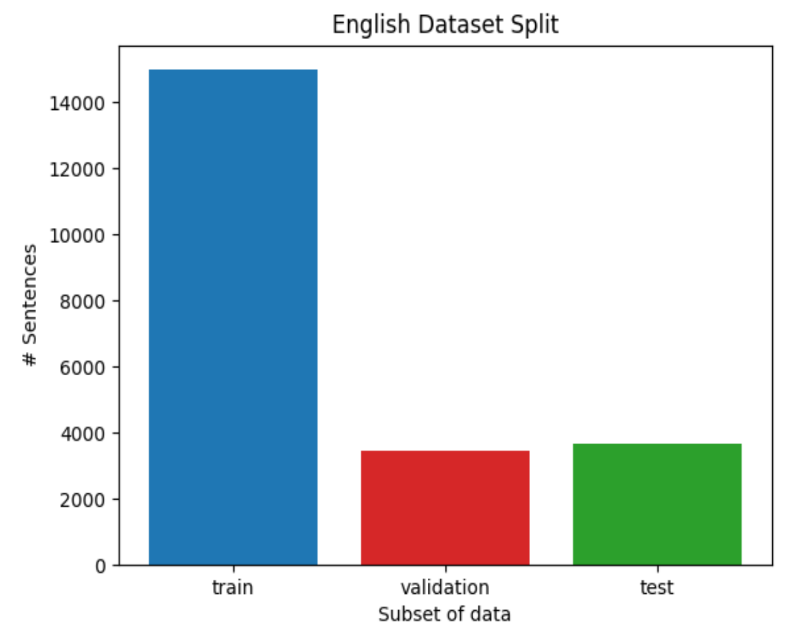
Our contributions:

- We include a principled framework that motivates the use of machine learning by creating two baseline systems.
- The analysis between generative and discriminative machine learning algorithms and recent advancement in natural language processing suggest the CoNNL-2003 dataset is subject to temporal drift.

CoNNL-2003

The English dataset was preprocessed into training, development, and test sets, Additionally, the number of tokens per entity type highlight the imbalanced nature of the CoNNL-2003 dataset and of NER.





The paper describes a baseline & benchmark methodology for developing a NER system in the context of the CoNNL-2003 Shared Task [3]. This paper will focus on the English dataset and compare the Spanish corpus from the CoNNL-2002 Shared Task [4].

Methodology

Suppose $X = (X_1, \ldots., X_n)$ is a sequence of random variables taking discrete values in finite set $T = (t_1, \ldots., t_k)$, the state space. A first-order Markov model is predicated on the Markov assumption where the probability of the next time step is conditioned only on the current state.

$$P(X_{n+1} = tag|X_1, \ldots, Xn) = P(X_{n+1} = tag|X_n)$$

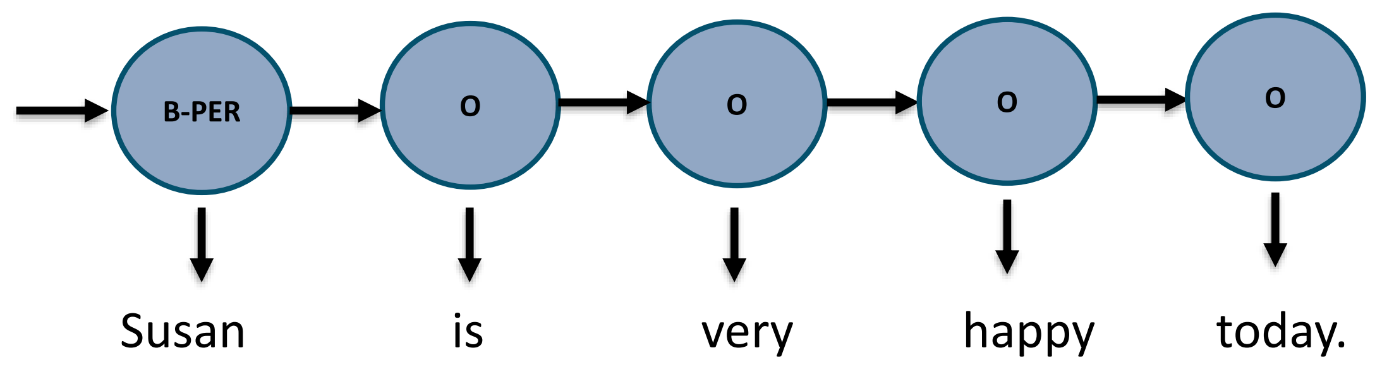
We will assume the sequence X of named entity tags is not observable. Given a sequence of words $W = (w_1, \ldots, w_n)$, we are interested in the best sequence of tags $\hat{t}_{1:n}$:

$$\hat{t}_{1:n} \approx argmax_{\hat{t}_{1:n}} P(tag_{1:n}|word_{1:n})$$

We apply Bayes' rule and simplify the equation by recognizing the denominator is constant for each sentence and calculate the best sequence $\hat{t}_{1:n}$:

$$P(X,W) \approx \hat{t}_{1:n} \approx argmax_{\hat{t}_{1:n}} \prod_{i=1}^{T} P(tag_i\,|tag_{i-1})P(word_i\,|tag_i)$$

HMM



The motivation for better modeling is clear: we want to use better feature representations at each time step. The maximum-entropy model directly computes the conditional probability.:

$$\hat{t}_{1:n} \approx argmax_{\hat{t}_{1:n}} \prod_{i=1}^{T} P(tag_i\,|tag_{i-1}, word_i)$$

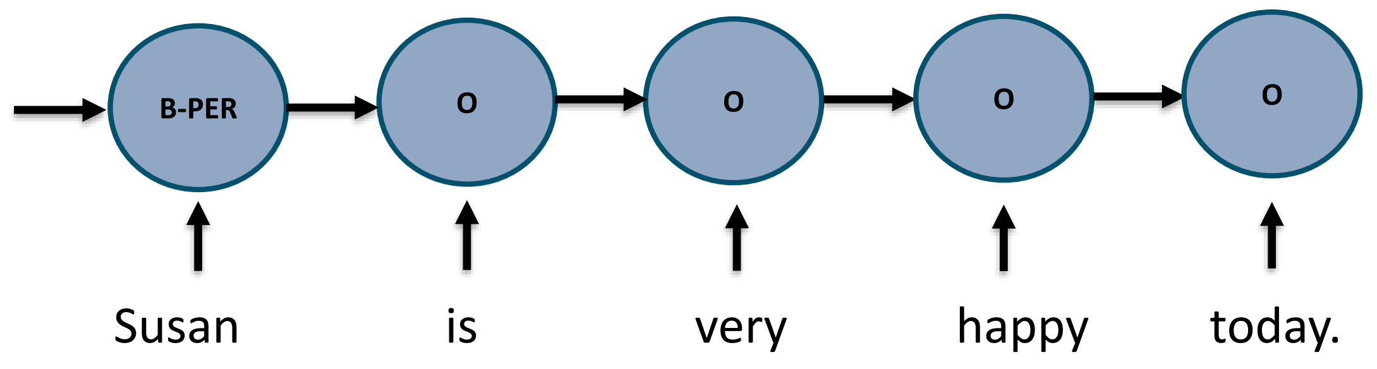
We cannot analytically compute the sequence and resort to approximation methods, namely gradient descent:

$$\hat{t}_{1:n} \approx argmax_{\hat{t}_{1:n}} \prod_{i=1}^{T} \frac{1}{Z(word_i, tag_{i-1})} \exp\left[\mathbb{W} \cdot f(tag_{i-1}, word_i)\right]$$

The equation above is multinomial logistic regression in disguise. Thus, our final equation is:

$$P(X|\,W) \approx \hat{t}_{1:n} \approx argmax_{\hat{t}_{1:n}} \prod_{i=1}^{T} \mathrm{softmax}\,(\mathbb{W}\mathbb{x}_i + \mathbb{b})$$

MEMM



Baselines

Two rule-based systems were computed for the English and Spanish corpora. The **AlwaysNonEntity** is a naive baseline that labels all tokens as 'O' or as non-entities, see Table 1.

Model	Dev		Test	
	Acc	F1	Acc	F1
AlwaysNonEntity	83.2	0.0	82.2	0.0
SingleEntity	86.2	40.0	84.8	**35.0**

Table 1: Results on English dev and test set (Accuracy and Micro F1).

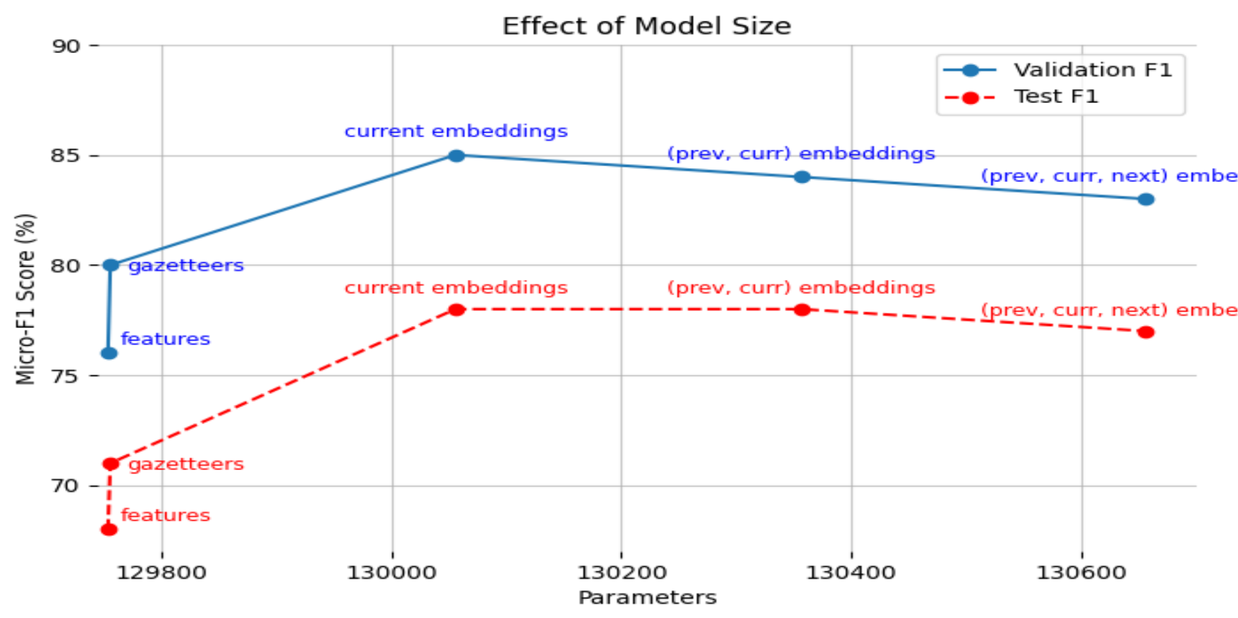
Model	Test	
	Acc	F1
AlwaysNonEntity	88.0	0.0
SingleEntity	74.4	**17.0**

Table 2: Results on Spanish test set.

The less naive baseline **SingleEntity** is based on a lookup table where only the beginning (B-) of entities are added, see Table 2. Named entity recognition is an imbalanced problem and traditionally we use the micro-F1 score for evaluation.

Effect of Model Size

The addition of features for the MEMM allows for a model to recognize more difficult entities. This increase in features, and in effect the parameters, improved the performance of the system.



Results

The performance of our algorithms and their variants is in Tables 3 & 4. The ME approach clearly shows discriminative modeling gives good results; it easily outperforms the baselines and beats an HMM as well.

Run Description	Decoding	Test	LOC	MISC	ORG	PER	Overall
MEMM-sk λ=0.1, η=0.1, epochs=15	Greedy	Val Test	89. 84.	79. 72.	75. 70.	89. 83.	**84.0** **78.0**
(Me)MM λ=0.1, η=0.1, epochs=15	Greedy	Val Test	90. 84.	79. 70.	74. 69.	88. 83.	**84.0** **78.0**
HM(Me) α=100	Viterbi	Val Test	86. 80.	82. 72.	71. 62.	76. 57.	79.0 69.0

Table 3: English evaluation on tuned models. The (Me) caption indicates the model was built from scratch and -sk indicates it was imported from scikit-learn. λ is regularization, η is the learning rate, and α is the smoothing value.

While the Spanish dataset was evaluated exclusively by an HMM, the results in Table 4 suggest a first-order model is an appropriate benchmark for the task.

Run Description	Decoding	Test	LOC	MISC	ORG	PER	Overall
1. HM(Me)	Greedy	Test	71.	37.	72.	69.	**68.0**

Table 4: Spanish evaluation on tuned (α = 100) model. The (Me) caption indicates model is built from scratch.

Conclusion

We conclude that advancements in NLP have improved our model (see Table 7) with little feature engineering and can reach competitive performance if a more complex model is chosen. However, we are not interested in performance and propose the CoNNL-2003 dataset is dated and we should instead access the generalization of these models. That is, we should start considering the upper bound of new named entities.

	Validation	Test	Time (s)	# Parameters
Features	76.%	68.%	91	129,754
Feats + Gazetteers	80.%	72.%	86	129,756
Feats + Gazetteers + $\mathbf{E}_i$	85.%	78.%	111	130,056
Feats + Gazetteers + $\mathbf{E}_i$ + $\mathbf{E}_{i-1}$	84.0%	**78.%**	159	130,356
Feats + Gazetteers + $\mathbf{E}_i$ + $\mathbf{E}_{i-1}$ + $\mathbf{E}_{i+1}$	83.%	77.%	190	130,656

Table 7: Effect of ME model size with different feature sets. 26 hand-crafted features were used.The previous and current embeddings perform marginally better than other sets.

Citations

[1] 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. 2003, pp. 142–147.

[2] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). 2002.