Name Entity Recognition Report

Diep (Emma) Vu

Abstract

In this report I will present my results of adding several word features and training the model with the viterbi algorithm on the Spanish dataset of CoNLL2002

1 Part 1: Additional Word Features

Before adding any new features, I ran the model and the F-1 score was **54.00**. Hence, in the first part of the project, besides the whole word itself as a feature, I included several word features to improve the model. Here are the added features:

- prefix: prefix of a word with the length ≤ 4, which can often provide clues about its meaning and entity type.
- suffix: suffix of a word with the length ≤ 4, which can often provide clues about its meaning and entity type.
- capitalize: check if the word is capitalized with the first character, which can help to identify proper nouns.
- word shape: represent the abstract letter pattern of the word by mapping lower-case letters to 'x', upper-case to 'X', numbers to 'd', and retaining punctuation, which can help the model to recognize words that have similar shapes but different spellings
- short word shape: represent the abstract letter pattern of the word using shorter word shape features by removing duplicates, which can help the model to learn faster
- part of speech tagging (POS): since the dataset already has POS, I utilized that feature to improve the model, which can help the model learn to recognize common patterns of word types that are associated with different entity types.

Initially I included features like *lower, upper, digits* to check if the word contains lowercase, uppercase or digits but I decided to remove them because the word shape feature already helped identify whether the character is *lower (x), upper (X) or digit (d)*. I didn't remove capitalize feature because proper nouns are often capitalized so it would be useful for the recognizer to learn that pattern.

So besides word, I added features: *capitalize, word shape, short word shape, prefix, suffix* and ran on the development set and F-1 score increased from **54.00** to **68.37**. At first I also added the feature *hyphen* (checking if the word contains hyphen) but interestingly, the feature only made the F-1 score on development set to **decrease** (from **68.37** to **68.23**) so I didn't include that feature. Perhaps the one of the reasons leading to this decrease is because there are not many words that have hyphen in the training dataset but there might be words with hyphen in the test set, so adding the hyphen feature would just make the model perform poorly.

From that list of new added features, I tried to remove some combinations of features but they only made the F-1 score lower, with the most significant is suffix (from **68.37** to **66.36**). This means that *suffix* feature was important in improving the model on the Spanish dataset. We can see the detailed F-1 score in the ***Table 1*** below when I tried to remove different features to decide which ones give the best performance on development set.

Besides adding features, I also increased window size around the token (consider 1, 2, or 3 neighboring words before and after) and found out the best window size is **3**.

Here and ***Table 2*** below are the final results on the test set with the features: *capitalize, word shape, short word shape, prefix, suffix*:

Processed 51533 tokens with 3558 phrases; found: 3857 phrases; correct: 2677.

- accuracy: 97.12%
- precision: 69.41%
- recall: 75.24%
- F-1 score: 72.20

After adding features listed above, the F-1 score significantly increased to **72.20**. Also looking at the result on the test set, we can see that the model predicts the tag *MISC* the poorest (only around **40%** in *Precision, Recall and F-1 score*. It might be because miscellaneous words don't have pattern as recognizable as Person, Location and Organization.

2 Part 2: Viterbi Algorithm

I used numpy to implement viterbi algorithm with Maximum Entropy Markov Model (MEMM) to train the model. Even though it took longer to run on the testing data (**4742.224006891251** seconds compared to **305.84402775764465** seconds in the first part), the results were so much higher. The reason why might be the viterbi algorithms knows about previous tag of the last word which gives important information and thus make the model better at learning sequence pattern.

Before adding any new features, with the viterbi algorithm on MEMM, the F-1 score on development set increased to **66.40** compared to **54.00** in the first part. After adding the new features like *prefix, suffix, capitalize, word shape, short word shape, POS* like the first part, the F-1 score on development set was **75.23** compared to **68.37** in the first part.

I also tried to remove other combinations of features but they only made the F-1 score lower and nothing made the F-1 score drop as significant as suffix, just like in Part 1 (from **75.23** to **73.88**). So even with the viterbi algorithm, suffix is still an important feature in improving the model. Interestingly, when I add the feature *hyphen*, unlike the first part, this time the feature made no difference in F-1 score (still **75.23**). Perhaps the reason why with the viterbi algorithm, the classifer's performance wasn't affected because there's information about most likely sequence of state, which is strong enough to counteract bad influence of *hyphen* feature.

Surprisingly, when I removed the *word shape* feature, the F-1 score increased the most significant (from **75.23** to **77.23**). This means that with the viterbi algorithm, the *word shape* feature becomes redundant and affects the model's performance. It might be because there's not a strong pattern that allows the classifier to distinguish. Also there might be a lot of words have same word shape, and if there are words that have different word shape which make test set not representative, the classifier can learn the pattern wrong and perform poorly.

However, this is not the case for short word shape because when I removed the feature, the F-1 score only changed slightly (**75.23** to **74.98**). Hence I decided to remove the word shape feature on this part with the viterbi algorithm. So the features that I keep besides the word itself are: *short shape, capitalize, prefix, suffix, POS*. ***Table 3*** below shows the detailed F-1 score on different combinations of features.

I still keep the window size of **3** like the first part and I ran the viterbi algorithm with these above features on the test set. ***Table 4*** below here are the results:

Processed 51533 tokens with 3558 phrases; found: 3512 phrases; correct: 2788.

- accuracy: 97.29%
- precision: 79.38%
- recall: 78.36%
- F-1 score: 78.87

Looking at the results above and in ***Table 4***, we can see that all the metrics have increased. The *Accurcy* of the model with the viterbi algorithm slight changed but still above **97%**, whereas *Precision, Recall and F-1 score* of the model with the viterbi algorithm increased significantly to all above **78%**. Specifically, in terms of F-1 score, it increased from **72.20** to **78.87**. Precision also witnessed a significant increase from **69.41%** in part 1 to **79.38%**. Thus, of all the tags, *MISC* is still the most unpredicted tag and *PER* is still the tag the model predicts best on since there might be a stronger pattern in recognizing a person than other miscellaneous things in Spanish.

Honor Code Statement

I affirm that I have carried out my academic endeavors with full academic honesty.

Features	**F-1 score**
word	**54.00**
Add features: *word, capitalize, word shape, short word shape, prefix, suffix*	68.37
Add *hyphen*	**68.23**
Remove *word shape*	67.79
Remove *POS*	68.18
Remove *short word shape*	67.78
Remove *capitalize*	67.77
Remove *prefix*	67.42
Remove *suffix*	**66.36**

Table 1: (Part 1) F-1 score on different combinations of features

Tag	**Precision**	**Recall**	**F-1 score**	**Phrases**
LOC	74.56%	74.08%	74.32	1077
MISC	40.17%	42.18%	41.15	356
ORG	67.59%	76.71%	71.86	1589
PER	78.68%	89.39%	83.69	835

Table 2: (Part 1) detailed results on test set

Features	**F-1 score**
word	**66.40**
Add features: *word, capitalize, word shape, short word shape, prefix, suffix*	75.23
Add *hyphen*	**75.23**
Remove *word shape*	**77.23**
Remove *POS*	74.84
Remove *short word shape*	74.98
Remove *capitalize*	74.95
Remove *prefix*	74.62
Remove *suffix*	**73.88**

Table 3: (Part 2) F-1 score on different combinations of features with viterbi algorithm

Tag	**Precision**	**Recall**	**F-1 score**	**Phrases**
LOC	80.47%	76.01%	78.18	1024
MISC	66.14%	48.97%	56.27	251
ORG	78.15%	81.00%	79.55	1451
PER	84.48%	90.34%	87.31	786

Table 4: (Part 2) detailed results on test set with viterbi algorithm