Performance Summary for CRF POS Tagging

Gabi Stanovsky

09.06.12

This document will summarize the experiments ran thus far to test the improvement of word embeddings on the POS tagging task. The experiments were conducted with the CRFSuite tool[1], on different corpora and word embeddings.

1 Annotations

These are the notations used throughout this document when describing CRF feature sets:

- $w[i]$ - The word located at the i'th location within a given sequence.
- $l[i]$ - The POS tag of $w[i]$.
- $e[i][j]$ - The j'th embedding feature of $w[i]$.

2 English POS Tagging

2.1 Corpus

The dataset used for English is the CoNLL 2000 dataset[2]. This dataset consists of approximately $220K$ words (organized in $9K$ sequences) for training, and $50K$ words (organized in $2K$ sequences) for testing. Each word is tagged for POS and BIO tags.

2.2 Experiments

Upon this dataset two experiments were run. Their results are summarized in table 1. Following is a description of each of the experiments:

[1] `http://www.chokkan.org/software/crfsuite/`

[2] `http://www.cnts.ua.ac.be/conll2000/chunking/`[test | train]

Baseline: This experiment aims at setting baseline results for the next experiment. The feature set for tag located at position t within a sequence is:

- $w[t-2]$
- $w[t-1]$
- $w[t]$
- $w[t+1]$
- $w[t+2]$
- $w[t-1]|w[t]$
- $w[t]|w[t+1]$
- $l[-1]$
- A special start | end of sequence symbol.

CW: This experiment aims at testing the improvement induced by Collobert and Weston (CW) embeddings upon the baseline results. The CRF features for word located at index t within a sequence were those of previous experiment, in addition to:

- $e[t][0] - e[t][99]$

These features were added where embeddings exists for $w[t]$

Table 1: Experimental Results for English POS Tagging

Experiment	**Acc**	**F1**
Baseline	0.9282	0.869641
CW	0.9607	0.926436

2.3 Discussion

Turian did not include this task in his article, while C&W did - although in their experiment the embeddings changed in the course of the POS training, as part of the deep learning architecture. It seems that C&W word embeddings are capable of significantly improving POS tagging task also as a CRF plug-in enhancer.

Table 2: Hebrew corpora statistics. Values show size in #tokens and #sequences

Corpus	Training	Testing
TB1	91K / 4K	18K / 750
A7	104K / 5K	21K / 1K
All	195K / 9K	40K / 2K

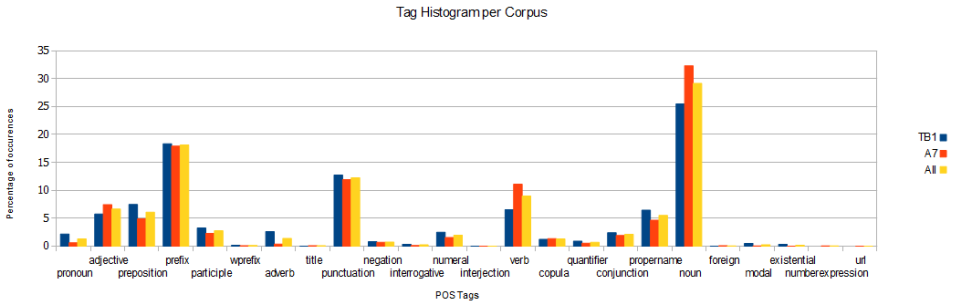

Figure 1: Tag Distribution Histogram

3 Hebrew POS Tagging

3.1 Corpora

Three corpora were used to measure the improvement of word embeddings. **TB1** and **A7** are two different corpora, while **All** is the concatenation of them. All of these were tagged for POS and segmented. The size statistics of these datasets are shown in table 2. A tag histogram is shown in figure 1.

3.1.1 Analysis of Corpora

Several experiments were conducted to measure relevant statistics with regards of above mentioned corpora. For comparison, the same statistics were measured also against the CoNLL corpus.

3.1.1.1 Unknown Tokens This experiment measures the percentage of occurrences of tokens for which there is no embedding available. Results are given in table 3.

3.1.1.2 Ambiguous Tokens This experiment measures the percentage of occurrences which appear in the corpora with several different tags. The results are grouped by the number of different possible tags for a single token. Results, in the form of an histogram per corpus, are shown in figure 2. A rougher division can part each corpus entry as either ambiguous (meaning, a token which has more than one

Table 3: Unknown Token Percentage

Corpora	Train	Testing
TB1	27.334	27.723
A7	21.604	21.473
All	24.324	24.114
CoNLL	2	2.362

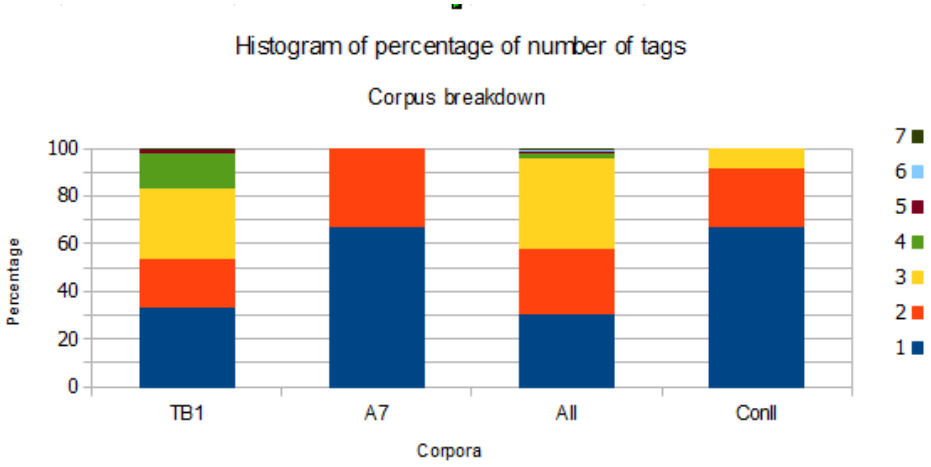

Figure 2: Ambiguous Tokens Distribution Histogram

tag available), and un-ambiguous (meaning, there is only one tag available for this token). This division statistics is available in table 4.

Table 4: Percentage of Ambiguous Tokens

TB1	A7	All	CoNLL
66.24	32.67	69.18	32.71

3.2 Experiments

Several tests were ran to achieve baseline results. These ran on each of the corpora described above. Following the baseline results, and similar to the course taken in section 2, embeddings were added as features to test their improvement.

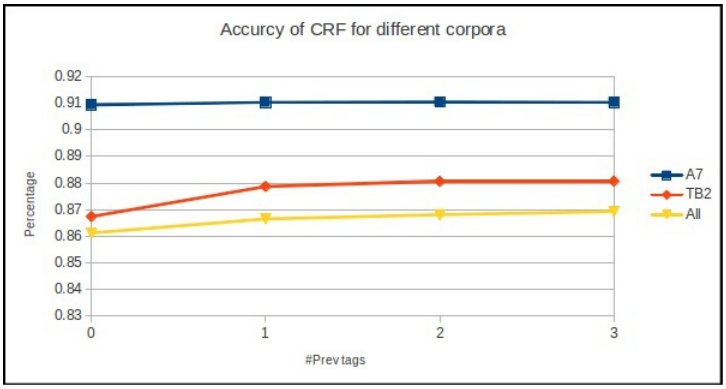

Figure 3: Hebrew Baseline Accuracy Percentage on Different Corpora and Feature Set

Baseline: In addition to using a similar feature set to that described in 2.2, another dimension of the number of previous tags was added. For example, when the number of previous tags is 2, the following features are added: $l[-2], l[-1]$. These experiments ranged from zero previous tags up to three previous tags, thus, in total, 12 experiments were ran. Their results are summarized in table 5 and graphically in figures 3, 4. The reasons for these extra experiments is explained in section 3.3.

Table 5: Hebrew Baseline Results. Values show Acc / F1

num of previous tags	**0**	**1**	**2**	**3**
TB1	0.867 / 0.725	0.879 / 0.735	0.881 / 0.74	0.881 / 0.748
A7	0.909 / 0.696	0.910 / 0.701	0.910 / 0.703	0.910 / 0.698
All	0.861 / 0.644	0.866 / 0.662	0.868 / 0.66	0.869 / 0.664

HebEmbeddings: Upon each of these corpora, a CRF training session with feature sets similar to those of section 2.2 (with one previous tag, as was done for English dataset), was conducted. The embedding features $e[t][0], ..., e[t][99]$ were taken from the Hebrew embedding experiment, after it has processed $40M$ tokens. Results are summarized in table 6

3.3 Discussion

Several observations can be made, based on the Hebrew experiments:

- In all cases, the Hebrew embeddings enhanced performance, both in accuracy measure (1% - 3%), as well as in F1 measure (6% - 12%). This is with accordance with the previously shown English

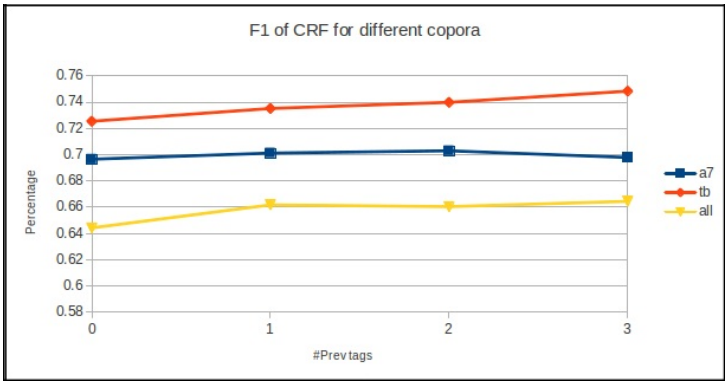

Figure 4: Hebrew Baseline F1 Percentage on Different Corpora and Feature Set

Table 6: Hebrew Embeddings Results. Values show Acc / F1.

	without embeddings	**with embeddings**
TB1	0.879 / 0.735	0.9 / 0.804
A7	0.91 / 0.701	0.94 / 0.821
All	0.866 / 0.662	0.88 / 0.723

results.

- There is less improvement found in **TB1** and **All** corpora (see table 6). This can be associated with the fact that the maximum number of unknown tokens is found in **TB1** (see table 3). In fact, there is a correlation between the percentage of known tokens and the rate of improvement gained by the addition of embeddings.

- There is a drop in performance when testing the **All** corpus, in comparison to testing its two components. The drop is mainly in F1 measure. I think this is due to different conventions and guidelines used for tagging POS of **TB1** and **A7** corpora. A supporting evidence for this can be the fact that the percentage of ambiguous tokens reaches its maximum in **All** corpus (see table 4), this is due to the fact that about $7K$ non-ambiguous tokens in **TB1** and **A7** receive different tags in both corpora - thus rendering them ambiguous in **All**.

  From brief browsing through the two corpora it seems that **A7** is noisy in comparison to **TB1**. In order to validate this assumptions the various settings were tested at section 3.2. Although there is no clear conclusion, perhaps the low F1 measure in the separate **A7** test supports this theory. A reason for the noisiness of the **A7** corpus can be the fact that is an ambiguous corpus, with no

scores for different analyses - the tag for the tests shown here was chosen based on the count of agreeing tags.