Unsupervised Part of Speech Tagging Using Unambiguous Substitutes from a Statistical Language Model

Mehmet Ali Yatbaz and Deniz Yuret
KOÇ UNIVERSITY

Introduction

• The part-of-speech (POS) tagging problem can be predicting the correct POS tag of a word in a given context using an unlabeled corpus and a dictionary with possible word–tag pairs.
• The performance of an unsupervised POS tagging system depends highly on the quality of the word–tag dictionary (Banko and Moore, 2004).
• Another issue is, unlike English, in agglutinative languages the number of theoretically possible parses can be infinite although the number of features is finite.
• We propose a dictionary filtering procedure based on likely substitutes suggested by a statistical language model.
• The procedure reduces the word–tag dictionary size and leads to significant improvement in the accuracy of the POS models.

Language models for disambiguation

• The language model is used to generate likely substitutes for the target word in the given context and these benefit the disambiguation process to the extent that the likely substitutes are unambiguous or have different ambiguities compared to the target word.
• We assume that the same hidden tag sequence that has generated a particular test sentence can also generate artificial sentences where one of the words has been replaced with a likely substitute.
• POS tags of the likely substitutes can then be used to reduce the tag set of the target word.
• The substitutes are implicitly incorporated into the disambiguation process for reducing the noise and the rare tags in the dictionary.

Sample artificial sentences generated for a test sentence from the Penn Treebank.

Currency gyrations can *whipsaw* (VB/NN) the funds .
Currency gyrations can withdraw (VB) the funds .
Currency gyrations can *restore* (VB) the funds .
Currency gyrations can modify (VB) the funds .
Currency gyrations can justify (VB) the funds .
Currency gyrations can regulate (VB) the funds .

• Table presents an example where the likely unambiguous replacements of the word target word “*whipsaw*” for a given sentence taken from the Penn Treebank are listed.
• In this example each substitute is an unambiguous verb (VB), confirming our assumption that each artificial sentence comes from the same hidden sequence.
• For all occurrences of the word “*whipsaw*”, our reduction algorithm will count the POS tags of the likely substitutes and remove the tags that have not been observed from the dictionary. In these example “NN” will be removed from the possible tags of “*whipsaw*”.

Dictionary Reduction

To reduce the dictionary with the help of the replacement words similar to a target word w, we following rules:
1. Choose the replacement word from unambiguous substitutes that are likely to appear in the target word context.
2. Substitutes must be observed in the training corpus.
3. Count the tags of the replacement for all occurrences of the target word.
4. Remove the tags that are not observed as the tag of replacements in any occurrences of the target word.

Statistical Language Modeling

• In order to estimate highly probable replacement words for a given word w in the context c, we use an n-gram language model.
• The context is defined as the 2n–1 word window W_{n+1} … W_0 … W_{n–1} and it is centered at the target word position.

\[
P(w_o \mid c) \propto P(w_{n+1} \ldots w_0 \ldots w_{n–1})
\]
\[
= P(w_{n+1})P(w_{n+2} \mid w_{n+1}) \ldots P(w_{n–1} \mid w_{n–2})
\]
\[
\propto P(w_0 \mid w_{n+1}) \ldots P(w_1 \mid w_{n+2}) \ldots P(w_{n–1} \mid w_0)
\]
where W_i represents the sequence of words w_i w_{i+1} … w_j.

• The probabilities are estimated using a 4 gram language model for all the words in the vocabulary of the unlabeled dataset D that are unambiguous and have a common tag with the target word w.
• The words with the highest \Pr(\cdot \mid c) where r \in D are selected as the replacement words of w in c.

Experimental Results

• The models in this section are trained and tested on the same unlabeled data therefore there aren’t any out-of-vocabulary words.
• We define a random and a most frequent tag (MFT) baseline on the 24K corpus. The MFT baseline simply selects the most frequent POS tag of each word from the 1M word Penn Treebank corpus.
• The 2-gram and 3-gram HMMs can be treated as the unsupervised baselines.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Baseline</td>
<td>76.98</td>
</tr>
<tr>
<td>3-gram HMM</td>
<td>77.43</td>
</tr>
<tr>
<td>2-gram HMM</td>
<td>92.22</td>
</tr>
<tr>
<td>MFT Baseline</td>
<td>96.11</td>
</tr>
</tbody>
</table>

• The success of the MFT baseline on the Noun, Adj, Pronoun and function word groups shows that tag distributions of the words in these groups are more skewed towards one of the available tags.
• The MFT baseline is significantly better on the function words. This is expected since EM tends to assign words uniformly to the available POS tags.

The abuse of the rare tags is presented. Tags that are removed by the dictionary reduction are shown in bold.

<table>
<thead>
<tr>
<th>Word</th>
<th>Tag Dictionary</th>
<th>Gold Tagging</th>
<th>EM tagging</th>
<th>Replacement POS counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>[RB,RP,IN]</td>
<td>[EN(0)</td>
<td>EN(0)]</td>
<td>[BEH</td>
</tr>
<tr>
<td>a</td>
<td>[LS,SYM,NNP]</td>
<td>[RT(0)</td>
<td>RT(0)]</td>
<td>[RT(0)</td>
</tr>
</tbody>
</table>

Reduced Dictionary

• EM can not capture the sparse structure of the word distributions therefore it tends to assign equal number of words to each POS tag.
• Together with the noisy word–tag dictionary great portion of the function words are tagged with very rare POS tags.
• The reduced dictionary (RD) removes the rare problematic POS tags of the words as a result the accuracy on the content and function words shows a drastic improvement compared to HMM models trained with the original dictionary.

Reduced Dictionary

Accuracy (%) | Noun | Verb | Adj | Adv | Pronoun | Content | Function | Total |
-------------|------|------|-----|-----|---------|---------|----------|------|
2-gram HMM   | 92.22| 83.64| 83.22| 83.96| 95.56   | 89.42   | 70.49    | 82.05|
2-gram HMM RD| 94.01| 84.90| 89.52| 85.18| 95.92   | 91.18   | 92.92    | 91.85|

• Note that with reduced dictionary by using 5 substitutes for each target word, the uniformly initialized first order HMM-EM achieves 91.85% accuracy.
• Dictionary reduction also removes some of the useful tags therefore the upper–bound (oracle score) of the 24K dataset becomes 98.15% after the dictionary reduction.
• We execute 100 random restarts of EM and select the model with the highest corpus likelihood, our model achieves 92.25% .

Conclusion

• With the help of a statistical language model, our system creates artificial replacements that are assumed to have the same POS tag as the target word and use them to reduce the size of the word–tag dictionary.
• Our method significantly improves the prediction accuracy of the unsupervised first order HMM-EM system in all of the POS groups and achieves 92.25% and 92.47% word tagging accuracy on the 24K and 48K word corpora respectively.
• We also tested our model on a coarse grained dictionary with 17 tags and achieved an accuracy of 92.8%.