Efficiency in Part-of-Speech Tagging

Naghmeh Fazeli
summer semester 2016
Supervisor: Dr. Alexis Palmer
“Learning a Part-of-Speech Tagger from Two Hours of Annotation” -2013

Dan Garrette
Department of Computer Science The University of Texas at Austin

Jason Baldridge
Department of Linguistics The University of Texas at Austin
How to Use Human Time Efficiently in a Low Resource Setting???

Labeling Full Sentences  Or  Producing a Tag Dictionary
2 Hours of POS tagging
By Two non-native Speakers
What are the Core Challenges?

• Limited labeled data (only 1-2k)
• Much noisier than a data from a typical corpus
Preview

• Basic Definitions
• Data Sources
• Time Bounded Annotation
• Main Approaches
Basic Definitions: Part of Speech Tagging

• Part-of-speech tagging (Tagging for short) is the process of assigning a part of speech to each word in an input text.

• Tagging is a **disambiguation** task; words are ambiguous—have more than one possible part-of-speech—and the goal is to find the correct tag for the situation.

Example: *book*(verb) that flight. 
hand me that *book*(noun).
Basic Definitions:
What is the difference between word type and token?

• The term "token" refers to the total number of words in a text, corpus etc, regardless of how often they are repeated.

• The term "type" refers to the number of distinct words in a text, corpus etc.

• the sentence "a good wine is a wine that you like" contains nine tokens, but only seven types, as "a" and "wine" are repeated.
Most word types (80-86%) are unambiguous; that is, they, have only a single tag. But the ambiguous words, although accounting for only 14-15% of the vocabulary, are some of the most common words of English, and hence 55-67% of word tokens in running text are ambiguous.

Some of the most ambiguous frequent words are *that*, *back*, *down*, *put* and *set*:

- earnings growth took a **back/JJ** seat
- a small building in the **back/NN**
- a clear majority of senators **back/VBP** the bill
- Dave began to **back/VB** toward the door
- enable the country to buy **back/RP** about debt
- I was twenty-one **back/RB** then
Basic Definitions: Open vs. Closed Class

- **Closed class** categories are composed of a small, fixed set of grammatical function words for a given language.

  prepositions, modals, determiners, particles, conjunctions

- **Open class** categories have large number of words and new ones are easily invented.

  Nouns(Googler, textlish), Verbs(Google), Adjectives(geeky)….
• Malagasy (MLG) is an Austronesian language spoken in Madagascar.

• Kinyarwanda (KIN) is a Niger-Congo language spoken in Rwanda.

• English (ENG) is the control language.
Data Sources

• **ENG:** Pen Tree Bank (PTB); 45 POS tags

• **KIN:** Transcripts of *testimonies by survivors of the Rwandan genocide* provided by the Kigali Genocide Memorial Center; 14 Pos Tags

• **MLG:** *Articles from the websites* Lakroa and La Gazette and Malagasy Global Voices, a citizen journalism site; 24 POS tags
The grand jury commented on a number of other topics.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
<th>Tag</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>coordin. conjunction</td>
<td><em>and, but, or</em></td>
<td>SYM</td>
<td>symbol</td>
<td>+, %, &amp;</td>
</tr>
<tr>
<td>CD</td>
<td>cardinal number</td>
<td><em>one, two</em></td>
<td>TO</td>
<td>“to”</td>
<td>to</td>
</tr>
<tr>
<td>DT</td>
<td>determiner</td>
<td><em>a, the</em></td>
<td>UH</td>
<td>interjection</td>
<td>ah, oops</td>
</tr>
<tr>
<td>EX</td>
<td>existential ‘there’</td>
<td><em>there</em></td>
<td>VB</td>
<td>verb base form</td>
<td>eat</td>
</tr>
<tr>
<td>FW</td>
<td>foreign word</td>
<td><em>mea culpa</em></td>
<td>VBD</td>
<td>verb past form</td>
<td>ate</td>
</tr>
<tr>
<td>IN</td>
<td>preposition/sub-conj</td>
<td><em>of, in, by</em></td>
<td>VBG</td>
<td>verb gerund</td>
<td>eating</td>
</tr>
<tr>
<td>JJ</td>
<td>adjective</td>
<td><em>yellow</em></td>
<td>VBN</td>
<td>verb past participle</td>
<td>eaten</td>
</tr>
<tr>
<td>JJR</td>
<td>adj., comparative</td>
<td><em>bigger</em></td>
<td>VBP</td>
<td>verb non-3sg pres</td>
<td>eat</td>
</tr>
<tr>
<td>JJS</td>
<td>adj., superlative</td>
<td><em>wildest</em></td>
<td>VBZ</td>
<td>verb 3sg pres</td>
<td>eats</td>
</tr>
<tr>
<td>LS</td>
<td>list item marker</td>
<td><em>1, 2, One</em></td>
<td>WDT</td>
<td>wh-determiner</td>
<td>which, that</td>
</tr>
<tr>
<td>MD</td>
<td>modal</td>
<td><em>can, should</em></td>
<td>WP</td>
<td>wh-pronoun</td>
<td>what, who</td>
</tr>
<tr>
<td>NN</td>
<td>noun, sing. or mass</td>
<td><em>llama</em></td>
<td>WP$</td>
<td>possessive wh-</td>
<td>whose</td>
</tr>
<tr>
<td>NNS</td>
<td>noun, plural</td>
<td><em>llamas</em></td>
<td>WRB</td>
<td>wh-adverb</td>
<td>how, where</td>
</tr>
<tr>
<td>NNP</td>
<td>proper noun, sing.</td>
<td><em>IBM</em></td>
<td>$</td>
<td>dollar sign</td>
<td>$</td>
</tr>
<tr>
<td>NNPS</td>
<td>proper noun, plural</td>
<td><em>Carolinias</em></td>
<td>#</td>
<td>pound sign</td>
<td>#</td>
</tr>
<tr>
<td>PDT</td>
<td>predeterminer</td>
<td><em>all, both</em></td>
<td>&quot;</td>
<td>left quote</td>
<td>‘ or “</td>
</tr>
<tr>
<td>POS</td>
<td>possessive ending</td>
<td>’s</td>
<td>”</td>
<td>right quote</td>
<td>” or ”</td>
</tr>
<tr>
<td>PRP</td>
<td>personal pronoun</td>
<td><em>I, you, he</em></td>
<td>(</td>
<td>left parenthesis</td>
<td>[ , ( , { , &lt;</td>
</tr>
<tr>
<td>PRPS</td>
<td>possessive pronoun</td>
<td><em>your, one’s</em></td>
<td>)</td>
<td>right parenthesis</td>
<td>] , ) , &gt;</td>
</tr>
<tr>
<td>RB</td>
<td>adverb</td>
<td><em>quickly, never</em></td>
<td>,</td>
<td>comma</td>
<td>,</td>
</tr>
<tr>
<td>RBR</td>
<td>adverb, comparative</td>
<td><em>faster</em></td>
<td>.</td>
<td>sentence-final punc</td>
<td>. ! ?</td>
</tr>
<tr>
<td>RBS</td>
<td>adverb, superlative</td>
<td><em>fastest</em></td>
<td>:</td>
<td>mid-sentence punc</td>
<td>; ... — -</td>
</tr>
<tr>
<td>RP</td>
<td>particle</td>
<td><em>up, off</em></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Annotation Tasks

• First Annotation Task: Directly produce a Dictionary of Words to their possible POS tags—> **Type-Supervised Training**

• Second Annotation Task: Annotating full sentences with POS tags—> **Token-Supervised Training**

• Annotators( A, B ) spent **two hours** on both tasks.
Advantages of Having both (type and token supervised) Sets of Annotations

- **Token-supervision** provides valuable frequency and tag context information
- **Type supervision** produces larger dictionaries
Comparing the Work of Two the Annotators

- Annotator A: Faster at annotating word types
- Annotator B: Faster at annotating full sentences

<table>
<thead>
<tr>
<th></th>
<th>sent.</th>
<th>tok.</th>
<th>dict.</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIN human sentences A</td>
<td>90</td>
<td>1537</td>
<td>750</td>
</tr>
<tr>
<td>KIN human TD A</td>
<td></td>
<td></td>
<td>1798</td>
</tr>
<tr>
<td>MLG human sentences B</td>
<td>92</td>
<td>1805</td>
<td>666</td>
</tr>
<tr>
<td>MLG human TD B</td>
<td></td>
<td></td>
<td>1067</td>
</tr>
<tr>
<td>ENG human sentences A</td>
<td>86</td>
<td>1897</td>
<td>903</td>
</tr>
<tr>
<td>ENG human TD A</td>
<td></td>
<td></td>
<td>1644</td>
</tr>
<tr>
<td>ENG human sentences B</td>
<td>107</td>
<td>2650</td>
<td>959</td>
</tr>
<tr>
<td>ENG human TD B</td>
<td></td>
<td></td>
<td>1090</td>
</tr>
</tbody>
</table>
Main approaches

• 1) Tag Dictionary Expansion
• 2) Weighted Model Minimisation
• 3) Expectation Maximization (EM) HMM Training
• 4) MaxEnt Markov Model (MEMM) Training
step1:
Tag Dictionary Expansion
Reasons for Expanding a Tag Dictionary

1. In a low-resource setting, most word types will not be found in the initial tag dictionary.

2. **limit ambiguity** —> EM-HMM

3. Small dictionaries interact poorly with Model Minimization: if there are too many unknown words, and every tag must be considered for them, then the minimal model assumes that they all have the same tag.
Expanding the Tag Dictionary with a Graph-based Technique

- Label Propagation (LP) — connect token nodes to each other via feature nodes
Advantages of LP Graph

This method uses character affix feature nodes along with sequence feature nodes in the LP graph to get distributions over unknown words.

Therefore, it can infer tag dictionary entries for words whose suffixes do not show up in the labeled data (or with enough frequency to be reliable predictors).
A dog barks.

The dog walks.

The man walks.
Benefits from Different Types of Features

**bigram**—>(the sequence is important)

**suffix**—> (inexpensive way for capturing morphological features, common types of morphology)
External Dictionary Usage in the Graph

English Wiktionary  (614k entries)
malagasyworld.org  (78k entries)
kinyarwanda.net  (3.7k entries)
From this graph, we extract a new version of the raw corpus that contains tags for each token. This provides the input for model minimization.
Seeding the Graph

token-supervision: labels for tokens are injected into the corresponding TOKEN nodes with a weight of 1.0.

type-supervision: any TYPE node that appears in the tag dictionary is injected with a uniform distribution over the tags in its tag dictionary entry.
What is the Result from Label propagation (LP)?
Extracting a Result from LP

- LP gives each token a distribution over the entire set of tags.

- Tokens with no associated tag labels after LP:
  1) Tags for the token have weights less than the threshold.
  2) No path from the token node from any seeded node.

- Lp has a filter not to add new tags to known words.

- Expansion: An unknown word type’s set of tags is the union of all tags assigned to its tokens. Additionally, full entries of word types given in the original tag dictionary are added.
Hidden Markov Model (HMM)

The goal of HMM decoding is to choose the tag sequence that is most probable given the observation sequence of $n$ words $w_1^n$

\[ \hat{t}_1^n = \arg\max_{t_1^n} P(t_1^n | w_1^n) \]

Bayes’s rule:

\[ \hat{t}_1^n = \arg\max_{t_1^n} \frac{P(w_1^n | t_1^n)P(t_1^n)}{P(w_1^n)} \]
Further Assumptions

1. The probability of a word appearing depends only on its own tag and is independent of neighbouring words and tags:

\[ P(w^n_1|t^n_1) \approx \prod_{i=1}^{n} P(w_i|t_i) \]
2. the **bigram** assumption, is that the probability of a tag is dependent only on the previous tag, rather than the entire tag sequence

\[
P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1})
\]
most probable tag sequence from a bigram tagger:

\[
\hat{t}_1^n = \arg \max_{t_1^n} P(t_1^n | w_1^n) \approx \arg \max_{t_1^n} \prod_{i=1}^{n} P(w_i | t_i) \underbrace{P(t_i | t_{i-1})}_{\text{emission transition}}
\]
Model Minimization

Model minimization is used to remove tag dictionary noise and induce tag frequency information from raw text.
Model Minimization

- **Vertex**: each vertex is a possible tag of each raw-corpus token.

- **Edge**: each edge connects two tags of adjacent tokens and is a potential tag bigram choice.
Model Minimization
Algorithm:

- first, selects tag bigrams until every token is covered by at least one bigram
- then, selects tag bigrams that fill gaps between existing edges
- continues until there is a complete bigram path for every sentence in the raw corpus.
Weighted Model
Minimization: Choosing the Weights

Max \left( \sum_i W_i (\text{for not yet covered words}) \frac{1}{1 + \text{new word/tag pairs added by } b} \right)
Stage one — > provides an expansion of the initial labeled data
Stage two — > turns that into a corpus of noisily labeled sentences.
Stage three — > uses the EM algorithm initialized by the noisy labeling and constrained by the expanded tag dictionary to produce an HMM.
Experiments

LP(ed) refers to label propagation including nodes from an external dictionary. Each result given as percentages for Total (T), Known (K), and Unknown (U).

<table>
<thead>
<tr>
<th>Human Annotations</th>
<th>0. No EM</th>
<th>1. EM only</th>
<th>2. With LP</th>
<th>3. LP+min</th>
<th>4. LP(ed)+min</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T  K  U</td>
<td>T  K  U</td>
<td>T  K  U</td>
<td>T  K  U</td>
<td>T  K  U</td>
</tr>
<tr>
<td>Initial data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIN tokens A</td>
<td>72  90  58</td>
<td>55  82  32</td>
<td>71  86  58</td>
<td>71  86  58</td>
<td>71  86  58</td>
</tr>
<tr>
<td>KIN types A</td>
<td>63  77  32</td>
<td>78  83  69</td>
<td>79  83  70</td>
<td>79  83  70</td>
<td>79  83  70</td>
</tr>
<tr>
<td>MLG tokens B</td>
<td>74  89  49</td>
<td>68  87  39</td>
<td>74  89  49</td>
<td>74  89  49</td>
<td>76  90  53</td>
</tr>
<tr>
<td>MLG types B</td>
<td>71  87  46</td>
<td>72  81  57</td>
<td>74  86  56</td>
<td>76  86  60</td>
<td>76  86  60</td>
</tr>
<tr>
<td>ENG tokens A</td>
<td>63  83  38</td>
<td>62  83  37</td>
<td>72  85  55</td>
<td>72  85  55</td>
<td>72  85  56</td>
</tr>
<tr>
<td>ENG types A</td>
<td>66  76  37</td>
<td>75  81  56</td>
<td>76  83  56</td>
<td>74  81  55</td>
<td>74  81  55</td>
</tr>
<tr>
<td>ENG tokens B</td>
<td>70  87  44</td>
<td>70  87  43</td>
<td>78  90  60</td>
<td>78  90  60</td>
<td>78  90  61</td>
</tr>
<tr>
<td>ENG types B</td>
<td>69  83  38</td>
<td>75  82  61</td>
<td>78  85  61</td>
<td>78  85  61</td>
<td>78  85  61</td>
</tr>
</tbody>
</table>
Differences between the Type and Token supervised Annotations

tag dictionary —> both cases
model minimization—> type scenario
Error Analysis

• One potential source of error—> the annotators task

Automatically remove improbable tag dictionary entries

A star indicates an entry in the human provided TD.
Conclusion:

• LP Graph—>Extracting a new version of raw corpus that contains tags for each token—>Input for Model Minimization

• Weighted Model Minimization—>set of tag paths(each path represents a valid tagging for the sentence)—>Noisily labeled corpus for initialising EM

• using EM algorithm to produce an HMM
One Open Issue

• Should Annotation task be done on **Types** or **Tokens**?
Provisional Answer

- **Type-supervision** + Expand + Minimize
  
- Identify Missing word/tag
  
- Better results comparing to token-supervision especially in Kinyarwanda case
Code

Learning POS Taggers for Truly Low-resource Languages-2015
Željko Agić, Dirk Hovy, and Anders Søgaard
Center for Language Technology
University of Copenhagen

• What does the paper present? Learning POS taggers for truly low resource languages.

• What are the data sources? 100 translations of (parts of) the Bible available as part of the Edinburgh Multi-lingual Parallel Bible Corpus.
Thank You.