Unsupervised Part-Of-Speech Tagging with Anchor Hidden Markov Models

Karl Stratos\textsuperscript{1}

Joint work with Michael Collins\textsuperscript{2} and Daniel Hsu\textsuperscript{2}

\textsuperscript{1}Bloomberg (work done while at Columbia University)

\textsuperscript{2}Columbia University (Michael Collins is currently on leave at Google)
Unsupervised POS Tagging

- Quintessential unsupervised problem in NLP

  This/DET is/VERB unlabeled/ADJ text/NOUN

- Naively estimating an HMM with the EM algorithm
  - Terrible performance!
  - Problem 1. Model misspecification
  - Problem 2. Suboptimal learning

- Extensions
  - Better models
    - Hard-clustering HMM (Brown et al., 1992),
    - Feature-rich models (Berg-Kirkpatrick et al., 2010)
  - Better learning
    - Contrastive estimation (Smith and Eisner, 2005)
    - Sparse prior (Johnson, 2007)
This Work

- **New model**: Anchor HMM
  - Each POS tag is “anchored” at some *unambiguous* word

```
NOUN      loss
ADP       on
NUM       1
DET       the
```

- **New learning algorithm**
  - Based on non-negative matrix factorization *(Arora et al., 2012)*
  - Exact, simple, and efficient

- Competitive with state-of-the-art on universal tagset
Overview

Anchor HMM

Learning Anchor HMM
   Non-Negative Matrix Factorization (NMF)
   Parameter Estimation

Experiments
Anchor HMM: Definition

- HMM with *structural restriction* on emission probabilities

\[
p(x_1 \ldots x_N, h_1 \ldots h_N) = \pi(h_1) \times \prod_{i=1}^{N} o(x_i|h_i) \times \prod_{i=2}^{N} t(h_i|h_{i-1})
\]

- \(\pi\): initial tag probabilities
- \(o\): emission probabilities
- \(t\): transition probabilities

Restriction: each tag has at least 1 “anchor word” that belongs to that tag only.

\(o(\text{loss}|\text{NOUN}) = 0.0001\)

Reasonable assumption for POS tags
True for all 10 languages in universal treebank (with 12 tags)
Anchor HMM: Definition

- HMM with *structural restriction* on emission probabilities

\[
p(x_1 \ldots x_N, h_1 \ldots h_N) = \pi(h_1) \times \prod_{i=1}^{N} o(x_i|h_i) \times \prod_{i=2}^{N} t(h_i|h_{i-1})
\]

\(\pi\): initial tag probabilities
\(o\): emission probabilities
\(t\): transition probabilities

- **Restriction**: each tag has at least 1 “anchor word” that belongs to that tag *only*.

\[
o(\text{loss}|\text{NOUN}) = 0.0001 \quad o(\text{loss}|h \neq \text{NOUN}) = 0
\]
Anchor HMM: Definition

- HMM with *structural restriction* on emission probabilities

\[
p(x_1 \ldots x_N, h_1 \ldots h_N) = \pi(h_1) \times \prod_{i=1}^{N} o(x_i|h_i) \times \prod_{i=2}^{N} t(h_i|h_{i-1})
\]

- **\(\pi\):** initial tag probabilities
- **\(o\):** emission probabilities
- **\(t\):** transition probabilities

- **Restriction:** each tag has at least 1 “anchor word” that belongs to that tag *only.*

- **Reasonable assumption for POS tags**
  True for all 10 languages in universal treebank (with 12 tags)

\[
o(\text{loss}|\text{NOUN}) = 0.0001 \quad o(\text{loss}|h \neq \text{NOUN}) = 0
\]
Game Plan

▶ Will exploit the anchor restriction to derive an exact parameter estimation algorithm.

▶ Key step: non-negative matrix factorization (NMF) of word-context co-occurrence matrix
Overview

Anchor HMM

Learning Anchor HMM

Non-Negative Matrix Factorization (NMF)
Parameter Estimation

Experiments
Context Representation

- $X \in \{1 \ldots n\}$: word
- $H \in \{1 \ldots m\}$: POS tag of $X$
Context Representation

- \( X \in \{1 \ldots n\} \): word

- \( H \in \{1 \ldots m\} \): POS tag of \( X \)

- Pick “context” representation \( Y \in \mathbb{R}^d \) of \( X \).

- Define matrix \( \Omega \in \mathbb{R}^{n \times d} \) with rows \( \Omega_x := \mathbf{E}[Y | X = x] \).
Context Representation

- $X \in \{1 \ldots n\}$: word

- $H \in \{1 \ldots m\}$: POS tag of $X$

- Pick “context” representation $Y \in \mathbb{R}^d$ of $X$.

- Define matrix $\Omega \in \mathbb{R}^{n \times d}$ with rows $\Omega_x := \mathbf{E}[Y | X = x]$.

- **Conditions on $Y$**
  1. Conditional independence
     \[
P(Y | X, H) = P(Y | H)
     \]
  2. Non-degeneracy
     \[
     \text{rank}(\Omega) = m
     \]
Example $Y$

- Indicator vector of neighboring words $Y \in \{0, 1\}^{2n}$

```
the   dog   saw   the   cat
```

1. $p(\text{dog, the}| \text{saw, VERB}) = p(\text{dog, the}| \text{VERB})$ ✓

2. $\Omega \in \mathbb{R}^{n \times 2n}$ has rank $m$. ✓*

---

*Unless the model is degenerate.
Factorization of $\Omega$

- Under the conditions, $\Omega_x := \mathbb{E}[Y|X = x]$ factorizes:

$$\Omega_x = \sum_{h=1}^{m} p(h|x) \times \mathbb{E}[Y|h]$$
Factorization of $\Omega$

- Under the conditions, $\Omega_x := \mathbf{E}[Y|X = x]$ factorizes:

$$
\Omega_x = \sum_{h=1}^{m} p(h|x) \times \mathbf{E}[Y|h]
$$

- If $x$ is an anchor: $\Omega_x = \mathbf{E}[Y|h_x]$
Factorization of $\Omega$

- Under the conditions, $\Omega_x := \mathbb{E}[Y|X = x]$ factorizes:

$$
\Omega_x = \sum_{h=1}^{m} p(h|x) \times \mathbb{E}[Y|h]
$$

- If $x$ is an anchor: $\Omega_x = \mathbb{E}[Y|h_x]$

$\Omega_x$ form a **convex hull** with anchor words at $m$ vertices.
Finding Anchors (Arora et al., 2012)
Finding Anchors (Arora et al., 2012)
Finding Anchors (Arora et al., 2012)
Finding Anchors (Arora et al., 2012)
Finding Anchors (Arora et al., 2012)
Finding Anchors (Arora et al., 2012)
**Input.** $\Omega$ with rows $\Omega_x = \mathbb{E}[Y|X = x]$, number of anchors $m$

**Conditions.** $Y \independent X | H$, rank($\Omega$) = $m$
NMF

**Input.** $\Omega$ with rows $\Omega_x = \mathbb{E}[Y|X = x]$, number of anchors $m$

**Conditions.** $Y \perp X \mid H$, $\text{rank}(\Omega) = m$

1. Find anchor rows $\Omega_{a_1} \ldots \Omega_{a_m}$. 
NMF

**Input.** $\Omega$ with rows $\Omega_x = E[Y \mid X = x]$, number of anchors $m$

**Conditions.** $Y \independent X \mid H$, rank$(\Omega) = m$

1. Find anchor rows $\Omega_{a_1} \ldots \Omega_{a_m}$.

2. Express each row $\Omega_x$ as a convex combination of anchor rows:

$$
\Omega_x = \sum_{h=1}^{m} p(h \mid x) \times \Omega_{a_h}
$$

Can be solved with Frank-Wolfe.
**NMF**

**Input.** $\Omega$ with rows $\Omega_x = \mathbb{E}[Y | X = x]$, number of anchors $m$

**Conditions.** $Y \independent X | H$, $\text{rank}(\Omega) = m$

1. Find anchor rows $\Omega_{a_1} \ldots \Omega_{a_m}$.

2. Express each row $\Omega_x$ as a convex combination of anchor rows:

$$
\Omega_x = \sum_{h=1}^{m} p(h|x) \times \Omega_{ah}
$$

Can be solved with Frank-Wolfe.

**Output.** $p(h|x)$ for all tags $h$, words $x$
Overview

Anchor HMM

Learning Anchor HMM
  Non-Negative Matrix Factorization (NMF)
  Parameter Estimation

Experiments
Basic Idea

- $\Omega_x = \mathbb{E}[Y | X = x]$ can be estimated from unlabeled data.

- NMF of $\Omega$ gives “flipped” emission probabilities $p(h|x)$.
  - Use them to solve for model parameters.
Algorithm

1. Estimate $\hat{\Omega}$ by counting word-context cooccurrences:

$$[\hat{\Omega}_x]_i = \hat{p}(y_i|x) = \frac{\text{count}(x, y_i)}{\text{count}(x)}$$

2. Compute $\hat{p}(h|x) \leftarrow \text{NMF}(\hat{\Omega}, m)$.

3. Use Bayes’ rule to recover emission parameters

$$\hat{o}(x|h) \leftarrow \frac{\hat{p}(h|x) \times \hat{p}(x)}{\sum_{x=1}^{n} \hat{p}(h|x) \times \hat{p}(x)}$$

4. Given $\hat{o}$, recover $\hat{t}$ and $\hat{\pi}$ (easy).
Practical Issues: Dimensionality Reduction

- Context $Y \in \mathbb{R}^{2n}$ is sparse and high-dimensional.
  - Cumbersome to work with.

- Can use projection $\Pi \in \mathbb{R}^{2n \times d}$ to reduce dimension
  - Conditional independence does not break: $Y \Pi \indep X \mid H$
  - Must ensure that $\Omega \Pi$ has rank $m$.

- Various choices of $\Pi$:
  - Random projection (Arora et al., 2012)
  - Projection onto best-fit subspace (i.e., SVD)
  - Projection based on canonical correlation analysis (CCA)
  - Projection based on hard-clustering assumption
Practical Issues: Better Anchors

- **Issue.** Anchors tend to be extremely rare words
  
  **Fix.** Only consider top $K$ frequent words as anchor candidates.
Practical Issues: Better Anchors

✍️ **Issue.** Anchors tend to be extremely rare words

**Fix.** Only consider top $K$ frequent words as anchor candidates

✍️ **Issue.** No spelling information used

**Fix.** Augment $\Omega_x$ with spelling features
Overview

Anchor HMM

Learning Anchor HMM
  Non-Negative Matrix Factorization (NMF)
  Parameter Estimation

Experiments
Setting

- **Dataset.** Universal treebank *(McDonald et al., 2013)*
  - 12 POS tags for 10 languages
  - Hyperparameters tuned on English portion

- All models trained with 12 hidden states and evaluated on many-to-1 accuracy

- **Models.**
  - **EM:** HMM trained with EM
  - **BROWN:** Brown clusters *(Brown et al., 1993)*
  - **ANCHOR:** Anchor HMM
  - **ANCHOR-FEAT:** Anchor HMM + spelling features
  - **LOG-LINEAR:** Log-linear model with same features *(Berg-Kirkpatrick et al., 2010)*
Context for Learning Anchor HMM

- \( Y \in \mathbb{R}^{2n} \): previous and next words

```
the  dog    saw    the  cat
```

- Choice of dimensionality reduction

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (English)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>48.2</td>
</tr>
<tr>
<td>Best-Fit</td>
<td>53.4</td>
</tr>
<tr>
<td>CCA</td>
<td>57.0</td>
</tr>
<tr>
<td>Hard</td>
<td><strong>66.1</strong></td>
</tr>
</tbody>
</table>
Results: 12 Universal Tags

<table>
<thead>
<tr>
<th>Model</th>
<th>de</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>id</th>
<th>it</th>
<th>ja</th>
<th>ko</th>
<th>pt-br</th>
<th>sv</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>46</td>
<td>60</td>
<td>61</td>
<td>60</td>
<td>50</td>
<td>52</td>
<td>60</td>
<td>52</td>
<td>60</td>
<td>42</td>
</tr>
<tr>
<td>BROWN</td>
<td>60</td>
<td>63</td>
<td>67</td>
<td>66</td>
<td>59</td>
<td>66</td>
<td>60</td>
<td>48</td>
<td>67</td>
<td>62</td>
</tr>
<tr>
<td>ANCHOR</td>
<td>61</td>
<td>66</td>
<td>69</td>
<td>68</td>
<td>64</td>
<td>60</td>
<td>65</td>
<td>54</td>
<td>65</td>
<td>51</td>
</tr>
<tr>
<td>ANCHOR-FEAT</td>
<td>63</td>
<td>71</td>
<td>74</td>
<td>72</td>
<td>67</td>
<td>60</td>
<td>69</td>
<td>62</td>
<td>66</td>
<td>61</td>
</tr>
<tr>
<td>LOG-LINEAR</td>
<td><strong>68</strong></td>
<td>62</td>
<td>67</td>
<td>62</td>
<td>61</td>
<td>53</td>
<td><strong>78</strong></td>
<td>61</td>
<td>63</td>
<td>57</td>
</tr>
</tbody>
</table>

- Anchor HMM: generally good performance
  - Spelling features help.
Results: 45 Original Tags (English)

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM</td>
<td>62.6 (1.1)</td>
</tr>
<tr>
<td>CLUSTER</td>
<td>65.6</td>
</tr>
<tr>
<td>ANCHOR</td>
<td>67.2</td>
</tr>
<tr>
<td>ANCHOR-FEAT</td>
<td>67.7</td>
</tr>
<tr>
<td>LOG-LINEAR</td>
<td><strong>74.9 (1.5)</strong></td>
</tr>
</tbody>
</table>

- **Behind LOG-LINEAR**
- Possible reason: spelling features more important with fine-grained tags
### Discovered Anchor Words (for 12 Tags)

<table>
<thead>
<tr>
<th>German</th>
<th>English</th>
<th>Spanish</th>
<th>French</th>
<th>Italian</th>
<th>Korean</th>
</tr>
</thead>
<tbody>
<tr>
<td>empfehlen</td>
<td>loss</td>
<td>y</td>
<td>avait</td>
<td>radar</td>
<td>완전</td>
</tr>
<tr>
<td>wie</td>
<td>1</td>
<td>hizo</td>
<td>commune</td>
<td>però</td>
<td>중에</td>
</tr>
<tr>
<td>;</td>
<td>on</td>
<td>-</td>
<td>Le de</td>
<td>sulle</td>
<td>경우</td>
</tr>
<tr>
<td>Sein</td>
<td>one</td>
<td>especie</td>
<td>président qui</td>
<td>-</td>
<td>줄</td>
</tr>
<tr>
<td>Berlin</td>
<td>closed</td>
<td>Además</td>
<td>(</td>
<td>Stati</td>
<td>같아요</td>
</tr>
<tr>
<td>und</td>
<td>are</td>
<td>el</td>
<td>à</td>
<td>Lo</td>
<td>많은</td>
</tr>
<tr>
<td>,</td>
<td>take</td>
<td>países</td>
<td>États</td>
<td>legge</td>
<td>,</td>
</tr>
<tr>
<td>-</td>
<td>,</td>
<td>la</td>
<td>Unis</td>
<td>al</td>
<td>볼</td>
</tr>
<tr>
<td>der</td>
<td>vice</td>
<td>España</td>
<td>Cette</td>
<td>far-</td>
<td>자신의</td>
</tr>
<tr>
<td>im</td>
<td>to</td>
<td>en</td>
<td>quelques</td>
<td>di</td>
<td>받고</td>
</tr>
<tr>
<td>des</td>
<td>York</td>
<td>de</td>
<td></td>
<td>la</td>
<td>맞있는</td>
</tr>
<tr>
<td>Region</td>
<td>Japan</td>
<td>municipio</td>
<td></td>
<td>art.</td>
<td>위한</td>
</tr>
</tbody>
</table>

- **loss** ≈ noun  
- **1** ≈ number  
- **on** ≈ preposition  

- Not perfect, but reasonable
Summary

- New model & algorithm for unsupervised POS tagging
  - Anchor HMM: each tag “anchored” at unambiguous word
  - NMF-based learning: exact, simple, and efficient

- Automatically discovers anchor words
  - Interpretable model
    
    \[
    \begin{align*}
    h_1 & \quad \text{loss} \\
    h_2 & \quad \text{on} \\
    h_3 & \quad 1 \\
    h_4 & \quad \text{the}
    \end{align*}
    \]

- Future directions
  - Can exploit anchor assumption to learn richer model families?
  - Can we relax the anchor assumption further?
EXTRA SLIDES
Relation to Other HMM Variants

- HMM emission probabilities in matrix form $O$

$$O_{x,h} := o(x|h)$$

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>loss</td>
<td>0.4</td>
<td>0.1</td>
</tr>
<tr>
<td>set</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>hit</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>ran</td>
<td>0.1</td>
<td>0.4</td>
</tr>
</tbody>
</table>

- General HMM

- Hard-clustering HMM
  (Brown et al., 1992)

- Anchor HMM