Unsupervised Multilingual Learning for POS Tagging

Benjamin Snyder, Tahira Naseem, Jacob Eisenstein and Regina Barzilay

MIT
Unsupervised Learning in NLP
Unsupervised Learning in NLP

• Has focused on monolingual settings
Unsupervised Learning in NLP

- Has focused on monolingual settings
- Performance still lags supervised learning
Unsupervised Learning in NLP

• Has focused on monolingual settings
• Performance still lags supervised learning

Question: can we improve monolingual performance when multilingual parallel data is available at training time?
Multilingual Learning for POS Tagging

Input:
untagged bilingual parallel corpus

Goal:
Induce a POS tagger for each language (test on monolingual data)

I love fish.
J’aime les poissons.
# Motivation for Multilingual Learning

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>I love fish.</td>
<td>J’aime les poissons.</td>
</tr>
<tr>
<td>I love to fish.</td>
<td>J’aime pêcher.</td>
</tr>
</tbody>
</table>
Motivation for Multilingual Learning

• Learn from differences in lexical ambiguity
  fish/poissons [N] vs. fish/pêcher [V]
Motivation for Multilingual Learning

- Learn from differences in lexical ambiguity
  
  fish/poissons [N] vs. fish/pêcher [V]

- Learn from differences in structural ambiguity
  
  1. determiner “les” signals noun
  2. “to” signals infinitival verb
Related Work

  ▸ Supervised data available in source language
  ▸ Goal: transfer annotations to target language

• Synchronous grammars for MT
  (Wu & Wong 1998, Chiang 2005)
Bilingual Graphical Models

Desiderata:

Symmetric model:
- No supervision on either side
- Information flows both ways

Minimalist approach:
- Allow language specific idiosyncrasies
  different sentence lengths, tags, tagsets etc
- Avoid over-parameterization
(1) Two Monolingual HMM's

I love fish

\[ X_1 \rightarrow X_2 \rightarrow X_3 \]

J' aime les poissons

\[ Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_4 \]
(2) Get Alignments  (using GIZA++)

I

\(x_1\)

love

\(x_2\)

\(x_3\)

\(y_1\)

\(y_2\)

\(y_3\)

\(y_4\)

\(J'\)

aime

les

poissons
(3) Form Bilingual Model

I love fish

\[ x_1/y_1 \rightarrow x_2/y_2 \rightarrow x_3/y_4 \]

\[ y_3 \rightarrow \text{les poissons} \]

\[ J' \rightarrow \text{aime} \]
Learning Task
How to Parameterize
How to Parameterize
Naive parameterization: multinomial over merged tag pair, conditioned on both languages’ previous tags.
Naive parameterization: multinomial over merged tag pair, conditioned on both languages’ previous tags.

- No parameter sharing
- For trigram tagger with 13 tags:
  28,561 unrelated multinomials ($13^4$)
  each of dimension 169 ($13^2$)
Instead, we define the generative probability of merged tag pair \((x_i, y_j)\) in terms of three factors:

\[
P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)
\]
Instead, we define the generative probability of merged tag pair \((x_i, y_j)\) in terms of three factors:

\[
P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto P(x_i | x_{i-1}, x_{i-2})P(y_j | y_{j-1}, y_{j-2})P(x_i, y_j)
\]

Transition probability in each language
Instead, we define the generative probability of merged tag pair \((x_i, y_j)\) in terms of three factors:

\[
P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)
\]

- Transition probability in each language
- “Coupling” probability: compatibility of tag pair
Instead, we define the generative probability of merged tag pair \((x_i, y_j)\) in terms of three factors:

\[
P(x_i, y_j | x_{i-1}, x_{i-2}, y_{i-1}, y_{i-2}) \propto P(x_i | x_{i-1}, x_{i-2})P(y_j | y_{j-1}, y_{j-2})P(x_i, y_j)
\]

- Transition probability in each language
- “Coupling” probability: compatibility of tag pair

Essentially, a *product of experts*. 
Bayesian Generative Story
Bayesian Generative Story

• For each language, draw:
  
  ‣ **Transition distributions** over tag space  
    (conditioned on previous two tags)
  
  ‣ **Emission distributions** over lexicon  
    (conditioned on tag)

• Draw **coupling distribution** over space of bilingual tag pairs
Bayesian Generative Story

- For each language, draw:
  - Transition distributions over tag space (conditioned on previous two tags)
  - Emission distributions over lexicon (conditioned on tag)
- Draw coupling distribution over space of bilingual tag pairs

All drawn from Dirichlet priors of appropriate dimension.
Bayesian Generative Story
(cont’d)

For each bilingual parallel sentence:
Bayesian Generative Story
(cont’d)

For each bilingual parallel sentence:
1. Draw an alignment

Alignment must be 1-1 and contain no crossing edges

Treated as observed variable (based on GIZA++ alignments)
Bayesian Generative Story
(cont’d)

For each bilingual parallel sentence:

1. Draw an *alignment*

2. Draw parallel bilingual stream of tags in sequence from left to right
   - Unaligned tags drawn according to language-specific transition parameters
     \[ P(x_i | x_{i-1}, x_{i-2}) \]
   - Aligned tag-pairs drawn jointly according to transitions and bilingual coupling parameter
     \[ \propto P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j) \]
\[ \propto \text{trans}_1(P|\#\text{START}) \cdot \text{trans}_2(P|\#\text{START}) \cdot \text{coupling}(P, P) \]
\[ \propto \text{trans}_1(V|P) \cdot \text{trans}_2(V|V) \cdot \text{coupling}(V, V) \]
$\text{trans}_2(D|V)$
\[ \propto \text{trans}_1(N|V) \cdot \text{trans}_2(N|D) \cdot \text{coupling}(N, N) \]
Bayesian Generative Story (cont’d)

For each bilingual parallel sentence:

1. Draw an alignment

2. Draw parallel bilingual stream of tags in sequence from left to right

3. Draw words according to language-specific emission parameters.
I love fish

\[ emit_1(\text{“I”} | P) \cdot emit_2(\text{“J”} | P) \cdot \ldots \]
Bayesian Inference
Bayesian Inference

- Treat words and GIZA++ alignments as observed variables: $\mathcal{X}$
Bayesian Inference

• Treat words and GIZA++ alignments as observed variables: $\mathcal{X}$

• Treat emission, transition, and coupling parameters as hidden variables: $\Theta$
Bayesian Inference

• Treat words and GIZA++ alignments as *observed variables*: $\mathcal{X}$

• Treat emission, transition, and coupling parameters as *hidden variables*: $\theta$

• Predict POS tags $y$ with highest posterior probability:

\[
\arg\max_y P(y|x) = \arg\max_y \int_\theta P(y, x|\theta) P(\theta) \, d\theta
\]
Sampling

Iteratively sample each variable conditioned on current value of others (Gibbs):

- Sample aligned tag-pairs and unaligned tags
- Sample* transition distributions
- Sample* coupling distribution
Sampling

Iteratively sample each variable conditioned on current value of others (Gibbs):

- Sample aligned tag-pairs and unaligned tags
- Sample\* transition distributions
- Sample\* coupling distribution

\*no closed form using counts, due to factored parameterization:

\[
P(x_i, y_j | ...) = \frac{P(x_i | x_{i-1}, x_{i-2}) P(y_j | y_{j-1}, y_{j-2}) P(x_i, y_j)}{Z}
\]
Sampling

Iteratively sample each variable conditioned on current value of others (Gibbs):

- Sample aligned tag-pairs and unaligned tags
- Sample transition distributions
- Sample coupling distribution

*no closed form using counts, due to factored parameterization:

\[
P(x_i, y_j|\ldots) = \frac{P(x_i|x_{i-1}, x_{i-2})P(y_j|y_{j-1}, y_{j-2})P(x_i, y_j)}{Z}
\]

So we intersperse Gibbs with a Metropolis-Hastings step
- Define tractable proposal distribution: $Q$
- Sample a new value: $z^* \sim Q$
- Accept with probability: $\text{min} \left\{ 1, \frac{P(z^*)Q(z)}{P(z)Q(z^*)} \right\}$
Metropolis-Hastings

• Define tractable proposal distribution: \( Q \)

• Sample a new value: \( z^* \sim Q \)

• Accept with probability: 

\[
min \left\{ 1, \frac{P(z^*)Q(z)}{P(z)Q(z^*)} \right\}
\]

For the coupling distribution, we use proposal:

\[
Q \equiv \text{Dir}(\text{Count}(N, N), \text{Count}(N, V), \ldots)
\]

Counts of coupled parts-of-speech according to current sampled tags
Evaluation Setup

• Evaluate on *monolingual* test-set

• Orwell’s *Nineteen Eighty Four*
  - Languages: English, Bulgarian, Serbian, Slovene
  - 94,725 tokens (English)
  - 13 coarse POS tags (Multext East corpus)

• GIZA++ alignments
  - Intersection of each direction (1-1)
  - Removal of crossing edges (< 5%)
Accuracy
(full lexicon)

Bulgarian: 89
English: 91
Serbian: 85
Slovene: 87

Learned with:
- Bayesian HMM (Goldwater 2007)
Accuracy
(full lexicon)

Learned with:
- Bulgarian
- English
- Serbian
- Slovene

Bayesian HMM (Goldwater 2007)
Accuracy
(full lexicon)

Learned with:
- Bulgarian
- English
- Serbian
- Slovene

Bayesian HMM (Goldwater 2007)
Accuracy
(full lexicon)

Learned with:

- Bayesian HMM (Goldwater 2007)
- Supervised HMM

<table>
<thead>
<tr>
<th>Language</th>
<th>89</th>
<th>94</th>
<th>92</th>
<th>91</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>97</td>
<td>97</td>
<td>97</td>
<td>97</td>
</tr>
<tr>
<td>English</td>
<td>92</td>
<td>91</td>
<td>91</td>
<td>92</td>
</tr>
<tr>
<td>Serbian</td>
<td>87</td>
<td>90</td>
<td>85</td>
<td>92</td>
</tr>
<tr>
<td>Slovene</td>
<td>88</td>
<td>89</td>
<td>95</td>
<td>87</td>
</tr>
</tbody>
</table>
Accuracy
(100 word lexicon)

Learned with:
- Bulgarian
- English
- Serbian
- Slovene

Bayesian HMM
(Goldwater 2007)
Cross-lingual Analysis

- Some language pairings much better than others (Serbian + Slovene, English + Bulgarian)
- Given gold tags, easy to predict relative performance gains using cross-lingual entropy:

\[ H \left[ P(x_i | y_j, (i, j) \in a) \right] \]
Accuracy (full lexicon)

Learned with:
- Bulgarian
- English
- Serbian
- Slovene

Bayesian HMM (Goldwater 2007)

lowest cross-lingual entropy
Open Question
Open Question

How to predict optimal pairings in *unsupervised* manner?
Open Question

How to predict optimal pairings in *unsupervised* manner?

- Family relatedness not accurate predictor
Open Question

How to predict optimal pairings in *unsupervised* manner?

- Family relatedness not accurate predictor
- Typological relatedness..?
Open Question

How to predict optimal pairings in *unsupervised* manner?

- Family relatedness not accurate predictor
- Typological relatedness..?
  - English & Bulgarian analytical, fixed word order
Open Question

How to predict optimal pairings in an *unsupervised* manner?

- Family relatedness not accurate predictor
- Typological relatedness...?
  - English & Bulgarian analytical, fixed word order
  - Serbian & Slovene inflectional, variable word order
Conclusions

• Unsupervised multilingual learning effective for POS tagging.

• Beneficial for all pairings, drastic improvement for some.

• **Unsupervised/Supervised gap:**
  - Avg over all pairings: cut by 1/3.
  - Using best pairings: cut by 1/2.

  88 91 93 97

  (full lexicon experiment)
I love fish.

J’aime les poissons.

I love to fish.

J’aime pêcher.
Accuracy (full lexicon)

Learned with:
- Bulgarian
- English
- Serbian
- Slovene

Bayesian HMM (Goldwater 2007)

average trigram entropy: $H \left[ P(x_i | x_{i-1}, x_{i-2}) \right]$
Accuracy (full lexicon)

average trigram entropy: $H \left[ P(x_i | x_{i-1}, x_{i-2}) \right]$
Tagset

- Gold Standard: Multext-East Corpus
- Tag repository: 13 categories
- Tags/Token Ratio in corpus

<table>
<thead>
<tr>
<th>Language</th>
<th>Tag/Token</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serbian</td>
<td>1.41</td>
</tr>
<tr>
<td>Slovene</td>
<td>1.40</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>1.34</td>
</tr>
<tr>
<td>English</td>
<td>2.58</td>
</tr>
</tbody>
</table>
I love fish

J’ adore les poissons
I love fish

J’ adore poissons
Blah

V

N/V

P

Foo

bar

boo