Unsupervised Morphological Segmentation With Log-Linear Models

Hoifung Poon
University of Washington

Joint Work with
Colin Cherry and Kristina Toutanova
Machine Learning in NLP
Machine Learning in NLP

Unsupervised Learning
Machine Learning in NLP

Unsupervised Learning  Log-Linear Models

?
Machine Learning in NLP

Unsupervised Learning

Log-Linear Models

Little work except for a couple of cases
Machine Learning in NLP

Unsupervised Learning

Global Features

Log-Linear Models
Machine Learning in NLP

Unsupervised Learning

Log-Linear Models

Global Features
We developed a method for **Unsupervised Learning** of **Log-Linear Models** with **Global Features**.
We applied it to morphological segmentation and reduced F1 error by 10%–50% compared to the state of the art.
Outline

- **Morphological segmentation**
- Our model
- Learning and inference algorithms
- Experimental results
- Conclusion
Morphological Segmentation

- Breaks words into morphemes
Morphological Segmentation

- Breaks words into **morphemes**
  governments
Morphological Segmentation

- Breaks words into **morphemes**

  governments ⇒ govern – ment – s
Morphological Segmentation

- Breaks words into **morphemes**

  governments \(\Rightarrow\) govern – ment – s

  lm$pxtm

  (according to their families)
Morphological Segmentation

- Breaks words into morphemes

  governments  ⇒  govern – ment – s

  lm$pxtm  ⇒  l – m$pm – t – m

(according to their families)
Morphological Segmentation

- Breaks words into **morphemes**
  
  governments $\Rightarrow$ govern – ment – s
  
  lm$p$xtm $\Rightarrow$ l – m$p$m – t – m

  (according to their families)

- Key component in many NLP applications

- Particularly important for morphologically-rich languages (e.g., Arabic, Hebrew, …)
Why Unsupervised Learning?

- **Text**: Unlimited supplies in any language
- **Segmentation labels?**
  - Only for a few languages
  - Expensive to acquire
Why Log-Linear Models?

Can incorporate arbitrary overlapping features
E.g., Al – rb (the lord)

- **Morpheme features:**
  - Substrings Al, rb are likely morphemes
  - Substrings Alr, lrb are **not** likely morphemes
  - Etc.

- **Context features:**
  - Substrings between Al and # are likely morphemes
  - Substrings between lr and # are **not** likely morphemes
  - Etc.
Why Global Features?

- Words can inform each other on segmentation
- E.g., Al – rb (the lord), l – Al – rb (to the lord)
State of the Art in Unsupervised Morphological Segmentation

- Use directed graphical models
- Morfessor [Creutz & Lagus 2007]
  Hidden Markov Model (HMM)
- Goldwater et al. [2006]
  Based on Pitman-Yor processes
- Snyder & Barzilay [2008a, 2008b]
  - Based on Dirichlet processes
  - Uses bilingual information to help segmentation
    - Phrasal alignment
    - Prior knowledge on phonetic correspondence
      E.g., Hebrew \( w \leftrightarrow \) Arabic \( w, f; \ldots \)
Unsupervised Learning with Log-Linear Models

Few approaches exist to this date

- Contrastive estimation [Smith & Eisner 2005]
- Sampling [Poon & Domingos 2008]
This Talk

- First log-linear model for unsupervised morphological segmentation
- Combines contrastive estimation with sampling
- Achieves state-of-the-art results
- Can apply to semi-supervised learning
Outline

- Morphological segmentation
- **Our model**
- Learning and inference algorithms
- Experimental results
- Conclusion
Log-Linear Model

- State variable $x \in X$
- Features $f_i: X \rightarrow \mathbb{R}$
- Weights $\lambda_i$
- Defines probability distribution over the states

\[
P(x) = \frac{1}{Z} \exp \left( \sum_i \lambda_i \cdot f_i(x) \right)
\]
Log-Linear Model

- State variables $x \in X$
- Features $f_i: X \rightarrow R$
- Weights $\lambda_i$
- Defines probability distribution over the states

$$P(x) = \frac{1}{Z} \exp \left( \sum_i \lambda_i \cdot f_i(x) \right)$$

$$Z = \sum_{x' \in X} \exp \left( \sum_i \lambda_i \cdot f_i(x') \right)$$
States for Unsupervised Morphological Segmentation

- Words
  \( wv lAvwn, A lrb, \ldots \)

- Segmentation
  \( w \rightarrow vlAv \rightarrow wn, A l \rightarrow rb, \ldots \)

- Induced lexicon (unique morphemes)
  \( w, vlAv, wn, Al, rb \)
Features for Unsupervised Morphological Segmentation

- Morphemes and contexts
- Exponential priors on model complexity
Morphemes and Contexts

- Count number of occurrences
- Inspired by CCM [Klein & Manning, 2001]
- E.g., $w - vlAv - wn$

$$vlAvwn$$

($$\#\#_\#\#$$)

$$w$$
($$\#\#_vl$$)

$$vlAv$$
($$\#w_wn$$)

$$wn$$
($$Av_\#\#$$)
Complexity-Based Priors

- **Lexicon prior:** \( \Theta \)
  - On lexicon length (total number of characters)
  - Favor fewer and shorter morpheme types

- **Corpus prior:** \( \Psi \)
  - On number of morphemes (normalized by word length)
  - Favor fewer morpheme tokens

- E.g., \( l - Al - rb, Al - rb \)
  - \( l, Al, rb \) \( \Rightarrow \) \( - 5 \Theta \)
  - \( l - Al - rb \) \( \Rightarrow \) \( - 3/5 \Psi \)
  - \( Al - rb \) \( \Rightarrow \) \( - 2/4 \Psi \)
Lexicon Prior Is Global Feature

- Renders words **interdependent** in segmentation
- E.g., lAlrb, Alrb
  
lAlrb \Rightarrow ?
  
Alrb \Rightarrow ?
Lexicon Prior Is Global Feature

- Renders words **interdependent** in segmentation
- E.g., lAlrb, Alrb

\[
\begin{align*}
lAlrb & \Rightarrow l - Al - rb \\
Alrb & \Rightarrow ?
\end{align*}
\]
Lexicon Prior Is Global Feature

- Renders words *interdependent* in segmentation
- E.g., lAlrb, Alrb

  lAlrb  $\Rightarrow$  l – Al – rb

  Alrb  $\Rightarrow$  Al – rb
Lexicon Prior Is Global Feature

- Renders words **interdependent** in segmentation
- E.g., lAlrb, Alrb

\[
\begin{align*}
lAlrb & \Rightarrow l - Alrb \\
Alrb & \Rightarrow ?
\end{align*}
\]
Lexicon Prior Is Global Feature

- Renders words **interdependent** in segmentation
- E.g., lAlrb, Alrb

\[
\begin{align*}
lAlrb & \Rightarrow 1 - Alrb \\
Alrb & \Rightarrow Alrb
\end{align*}
\]
Probability Distribution

For corpus $W$ and segmentation $S$

$$P(W, S) = \frac{1}{Z} \cdot \exp \left\{ \lambda_{\sigma} \cdot n_{\sigma} + \sum_{\sigma \in \text{Lex}(W, S)} \lambda_{\sigma} \cdot n_{\sigma} - \sum_{\sigma \in \text{Lex}(W, S)} \Theta \cdot \text{Len}(\sigma) - \sum_{w \in W} \Psi \cdot \text{NumMorph}(w) / \text{Len}(w) \right\}$$

- Morphemes
- Contexts
- Lexicon Prior
- Corpus Prior
Outline

- Morphological segmentation
- Our model
- Learning and inference algorithms
- Experimental results
- Conclusion
Learning with Log-Linear Models

- Maximizes likelihood of the observed data
  = Moves probability mass to the observed data
- From where? The set $X$ that $Z$ sums over

$$Z = \sum_{x' \in X} \exp \left( \sum_i \lambda_i \cdot f_i(x') \right)$$

- Normally, $X = \{ \text{all possible states} \}$
- Major challenge:
  Efficient computation (approximation) of the sum
- Particularly difficult in unsupervised learning
Contrastive Estimation

- Smith & Eisner [2005]
- $X$ = a *neighborhood* of the observed data
- Neighborhood $\Rightarrow$ Pseudo-negative examples
- Discriminate them from observed instances
Problem with Contrastive Estimation

- Objects are independent from each other
- Using global features leads to intractable inference
- In our case, could not use the lexicon prior
Sampling to the Rescue

- Similar to Poon & Domingos [2008]
- Markov chain Monte Carlo
- Estimates sufficient statistics based on samples
- Straightforward to handle global features
Our Learning Algorithm

- Combines both ideas
- Contrastive estimation ⇒ Creates an informative neighborhood
- Sampling ⇒ Enables global feature (the lexicon prior)
Learning Objective

- **Observed**: $W^*$ (words)
- **Hidden**: $S$ (segmentation)
- Maximizes log-likelihood of observing the words

$$L(W^*) = \log \sum_S P(W^*, S)$$
Neighborhood

- **TRANS1** = Transpose any pair of adjacent characters
- **Intuition**: Transposition usually leads to a non-word
- E.g.,
  - lAlrb \(\Rightarrow\) Allrb, l1Arb, ...
  - A1rb \(\Rightarrow\) lArb, Arlb, ...

\(\)
Optimization

Gradient descent

\[ \frac{\partial L (W^*, S)}{\partial \lambda_i} = E_{s \mid w^*} [f_i] - E_{w, s} [f_i] \]
Supervised Learning and Semi-Supervised Learning

- Readily applicable if there are labeled segmentations \((S^*)\)

\[
\frac{\partial L (W^*, S^*)}{\partial \lambda_i} = \mathbb{E}_{s \mid w^*, s^*} [f_i] - \mathbb{E}_{w, s} [f_i]
\]

- **Supervised**: Labels for all words
- **Semi-supervised**: Labels for some words
Inference: Expectation

- Gibbs sampling
- $E_{S|W^*}[f_i]$
  For each observed word in turn, sample next segmentation, conditioning on the rest
- $E_{W,S}[f_i]$
  For each observed word in turn, sample a word from neighborhood and next segmentation, conditioning on the rest
Inference: MAP Segmentation

Deterministic annealing

- Gibbs sampling with temperature
- Gradually lower the temperature from 10 to 0.1
Outline

- Morphological segmentation
- Our model
- Learning and inference algorithms
- **Experimental results**
- Conclusion
Dataset

- **S&B**: Snyder & Barzilay [2008a, 2008b]
  - About 7,000 parallel short phrases
  - Arabic and Hebrew with gold segmentation
- Arabic Penn Treebank (ATB): 120,000 words
Methodology

- **Development set:** 500 words from S&B
- Use trigram context in our full model
- **Evaluation:** Precision, recall, F1 on segmentation points
Experiment Objectives

- Comparison with state-of-the-art systems
  - Unsupervised
  - Supervised or semi-supervised
- Relative contributions of feature components
Experiment: S&B (Unsupervised)

- Snyder & Barzilay [2008b]
  - S&B-MONO: Uses monolingual features only
  - S&B-BEST: Uses bilingual information
- Our system: Uses monolingual features only
Results: S&B (Unsupervised)
Results: S&B (Unsupervised)

- S&B-MONO
- S&B-BEST
- Our System

F1 scores comparison: S&B-MONO has a score of 65, S&B-BEST has a score of 80, and Our System has a score of 65.
Results: S&B (Unsupervised)
Reduces F1 error by 40%
Reduces F1 error by 21%
Experiment: ATB (Unsupervised)

- Morfessor Categories-MAP [Creutz & Lagus 2007]
- Our system
Reduces F1 error by 11%
Experiment: Ablation Tests

- Conducted on the S&B dataset
- Change one feature component in each test
  - Priors
  - Context features
Results: Ablation Tests

Both priors are crucial

Corpus prior only

Lexicon prior only

No priors
Results: Ablation Tests

Overlapping context features are important

No context features
Experiment: S&B (Supervised and Semi-Supervised)

- Snyder & Barzilay [2008a]
  - S&B-MONO-S: Monolingual features and labels
  - S&B-BEST-S: Bilingual information and labels

- Our system: Monolingual features and labels
  Partial or all labels (25%, 50%, 75%, 100%)
Results: S&B (Supervised and Semi-Supervised)
Results: S&B (Supervised and Semi-Supervised)
Results: S&B (Supervised and Semi-Supervised)
Results: S&B (Supervised and Semi-Supervised)

![F1 bar chart]

- S&B MONO-S
- S&B BEST-S
- Our-S 25%
- Our-S 50%
Results: S&B (Supervised and Semi-Supervised)
Results: S&B (Supervised and Semi-Supervised)

Reduces F1 error by 46% compared to S&B-MONO-S

Reduces F1 error by 36% compared to S&B-BEST-S
Conclusion

- We developed a method for **Unsupervised Learning** of **Log-Linear Models** with **Global Features**
- Applied it to morphological segmentation
- Substantially outperforms state-of-the-art systems
- Effective for semi-supervised learning as well
- **Easy to extend with additional features**
Future Work

- Apply to other NLP tasks
- Interplay between neighborhood and features
- Morphology
  - Apply to other languages
  - Modeling internal variations of morphemes
  - Leverage multi-lingual information
  - Combine with other NLP tasks (e.g., MT)
Thanks Ben Snyder …

For his most generous help with S&B dataset