PROMODES: A probabilistic generative model for word decomposition

Sebastian Spiegler, Bruno Golénia, Peter Flach

University of Bristol, Department of Computer Science

September 30th, 2009
Outline

Introduction

Algorithm
  Overview
  Probabilistic Generative Model
  Parameter estimation

Experiments
  Setup
  Experiments: Morpho Challenge Competition 1

Conclusions
Introduction

Algorithm
   Overview
   Probabilistic Generative Model
   Parameter estimation

Experiments
   Setup
   Experiments: Morpho Challenge Competition 1

Conclusions
Introduction

Morphology group @ University of Bristol

- goal: online morphological analysis for a text-to-speech system
- tools: machine learning approaches with different degrees of supervision (e.g. semi-supervised)
- target languages: under-resourced indigenous languages (e.g. Zulu)
- training data: small datasets

Our objective for Morpho Challenge

- adaptation of algorithms to large-scale experiments
- application of pure machine learning approaches
- language-independent approach
- no further morpheme analysis in terms of labelling (e.g. signatures, paradigms)
Outline

Introduction

Algorithm
- Overview
- Probabilistic Generative Model
- Parameter estimation

Experiments
- Setup
- Experiments: Morpho Challenge Competition 1

Conclusions
PROMODES = Probabilistic Generative Model for Different Degrees of Supervision

Outline:
1. Probabilistic Generative Model (PGM)
2. Parameter Estimation
3. Application of PGM → experiments
**PROMODES** = **Probabilistic Generative Model for Different Degrees of Supervision**

Outline:
1. Probabilistic Generative Model (PGM)
2. Parameter Estimation
3. Application of PGM → experiments
Algorithm: Probabilistic generative model

Description

- Description of **data generation** process based on observable and hidden variables
- Observable variables: **word** \( w \)
- Hidden variables: its **segmentation** \( b \)
- Goal: forming conditional distribution \( Pr(b|w) \)
- **Decision**: \( \arg \max_{b_k} Pr(b_k|w) = \arg \max_{b_k} Pr(b_k) \cdot Pr(w|b_k) \)
- **Problem**: Evaluation of **exponential** number of segmentations

Example for PGM

| word \( w \) | segmentation \( b \) | segmentation given word | \( Pr(b|w) \) |
|-------------|---------------------|----------------------|-------------|
| unbreakable | \( \langle 0000000000 \rangle_1 \) | \( \langle \text{unbreakable} \rangle_1 \) | 0.02 |
| ... | \( \langle 0100001000 \rangle_k \) | \( \langle \text{un, break, able} \rangle_k \) | ... |
| ... | \( \langle 1111111111 \rangle_{2m} \) | \( \langle u, n, b, r, e, a, k, a, b, l, e \rangle_{2m} \) | 0.01 |
Algorithm: Probabilistic generative model

Linearization of PGM

- Segmentation perspective → position perspective
- Observable variables: letter transitions in certain position, $Pr(b_i | w_i) = Pr(x \rightarrow y)$
- Hidden variables: boundary value in certain position, $Pr(b_i), b_i \in \{0, 1\}, 1 \leq i \leq |w| - 1$
- Goal: position-wise decision whether to place a boundary or not
  \[
  \arg \max_{b_i} Pr(b_i | w) = \begin{cases} 
  1, & \text{if } Pr(b_i = 1) \cdot Pr(w_i | b_i = 1) > Pr(b_i = 0) \cdot Pr(w_i | b_i = 0) \\
  0, & \text{otherwise.}
  \end{cases}
  \]
- Advantage: linear evaluation

Example for linear PGM

unbreakable

\[
\begin{align*}
\text{un} & \quad nb & \quad br & \quad re & \quad ea & \quad ak & \quad ka & \quad ab & \quad bl & \quad le \\
un \rightarrow & & & & & & & & & \\
\uparrow & & & & & & & & & \\
\text{n} & \quad b & \quad r & \quad e & \quad a & \quad k & \quad a & \quad b & \quad l & \quad e \\
\downarrow & & & & & & & & & & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\downarrow & & & & & & & & & & \langle \text{un, break, able} \rangle
\end{align*}
\]

$u$ = abstract start symbol
Parameter estimation

Model parameters

- $X$: probability distribution over letter transitions
- $Z$: probability distribution over boundaries/non-boundaries
- $\theta = \{X, Z\}$

1) Frequency-based $\rightarrow$ Maximum likelihood estimates (MLE)

- separate pre-processing step
- all possible substrings collected in forward trie
- segmentation based on peaking successor variety $\rightarrow$ crude method

2) Probability-based $\rightarrow$ Expectation Maximization (EM)

- Initialization of model parameters $\theta$
- Alternating between calculating likelihood of parameter estimates (E) and maximization (M)
- Convergence criterion: Kullback-Leibler divergence
Example: re-estimation of transition probability \( Pr(x \rightarrow y) = p_{xy} \)

\[
Pr_{re-estimated}(x \rightarrow y) = \frac{\sum_{j=1}^{|W|} \sum_{i=1}^{m_j} \sum_{r=0}^{1} \left( P(b_i = r|w_{ji}, \theta) \sum_{y' \in A} \mu_{xy, x'y'} \right)}{\sum_{y' \in A} \sum_{j'=1}^{|W|} \sum_{i'=1}^{m_{j'}} \sum_{r'=0}^{1} \left( P(b_i' = r'|w_{j'i'}, \theta) \sum_{y'' \in A} \mu_{x'y', x''y''} \right)}
\]

\( P(b_i = r|w_{ji}, \theta) \): posterior probability of hidden variable given data

\( \mu_{xy, x'y'} \): counting function with \( \mu_{xy, x'y'} = \begin{cases} 1, & \text{if } x' = x \text{ and } y' = y \text{ in } w_j \text{ at } i\text{th position,} \\ 0, & \text{otherwise.} \end{cases} \)

→ equivalently for probability distribution over boundaries/non-boundaries
Outline

Introduction

Algorithm
  Overview
  Probabilistic Generative Model
  Parameter estimation

Experiments
  Setup
  Experiments: Morpho Challenge Competition 1

Conclusions
Experiments

Setup

**PROMODES 1 (P1):** frequency-based parameter estimation (pre-processing with trie-based alg.), single word analysis

**PROMODES 2 (P2):** probability-based parameter estimation (Expectation Maximization), initialization → random segmentation, single word analysis

**PROMODES COMMITTEE (PC):** different initializations of EM, committee decision (multiple analysis for each word)

Committee of unsupervised learners

- combination of different solutions into cumulative vector → majority vote

<table>
<thead>
<tr>
<th>word</th>
<th>committee (multiple analyses)</th>
<th>cumulative vector</th>
<th>segmentation vector</th>
<th>segmentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>unbreakable</td>
<td></td>
<td>1311114212</td>
<td>0100001000</td>
<td>un,break,able</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Results

<table>
<thead>
<tr>
<th>Language</th>
<th>Precision P1</th>
<th>Precision P2</th>
<th>Precision PC</th>
<th>Recall P1</th>
<th>Recall P2</th>
<th>Recall PC</th>
<th>F-measure P1</th>
<th>F-measure P2</th>
<th>F-measure PC</th>
<th>av. M#</th>
<th>av. WL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic (nv)</td>
<td>.8110</td>
<td>.7696</td>
<td>.7706</td>
<td>.2057</td>
<td>.3702</td>
<td>.3696</td>
<td>.3282</td>
<td>.5000</td>
<td>.4996</td>
<td>8.80</td>
<td>5.77</td>
</tr>
<tr>
<td>Arabic (vw)</td>
<td>.7485</td>
<td>.6300</td>
<td>.6832</td>
<td>.3500</td>
<td>.5907</td>
<td>.4797</td>
<td>.4770</td>
<td>.6097</td>
<td>.5636</td>
<td>8.75</td>
<td>9.90</td>
</tr>
<tr>
<td>English</td>
<td>.3620</td>
<td>.3224</td>
<td>.3224</td>
<td>.6481</td>
<td>.6110</td>
<td>.6110</td>
<td>.4646</td>
<td>.4221</td>
<td>.4221</td>
<td>2.25</td>
<td>8.70</td>
</tr>
<tr>
<td>Finnish</td>
<td>.3586</td>
<td>.3351</td>
<td>.4120</td>
<td>.5141</td>
<td>.6132</td>
<td>.4822</td>
<td>.4225</td>
<td>.4334</td>
<td>.4444</td>
<td>3.58</td>
<td>13.50</td>
</tr>
<tr>
<td>German</td>
<td>.4988</td>
<td>.3611</td>
<td>.4848</td>
<td>.3395</td>
<td>.5052</td>
<td>.3461</td>
<td>.4040</td>
<td>.4212</td>
<td>.4039</td>
<td>3.26</td>
<td>11.12</td>
</tr>
<tr>
<td>Turkish</td>
<td>.3222</td>
<td>.3536</td>
<td>.5530</td>
<td>.6642</td>
<td>.5870</td>
<td>.2835</td>
<td>.4339</td>
<td>.4414</td>
<td>.3748</td>
<td>3.63</td>
<td>10.80</td>
</tr>
</tbody>
</table>
Experiments: Analysis of results

Arabic (non-/vowelized)
- high number of morphemes per word in gold standard
- segmenting into short morphemes preferred

Other languages: English, German, Finnish, Turkish
- lower number of morphemes per word in gold standard (3-4 morphemes per word)
- PROMODES tended to over-segment
- some examples for English:
  - bluefield → blu e field
  - bluefields → blu e field s
  - cartographer → car to gra p h er
  - choreograph → chore o gra p h

PROMODES, ground truth segmentation boundary
Outline

Introduction

Algorithm
  Overview
  Probabilistic Generative Model
  Parameter estimation

Experiments
  Setup
  Experiments: Morpho Challenge Competition 1

Conclusions
Conclusions

PROMODES algorithm

- unsupervised morphological analysis based on probabilistic generative model
- different parameter estimation approaches (MLE, EM), committee of unsupervised learners
- Very good results on Arabic and Finnish, good results on other languages in competition 1

Future work

- optimization of probabilistic generative model
- investigation in behaviour of committee

Morpho Challenge in general

- workshop as discussion forum for different research groups
- valuable experiences on large datasets
- opportunity of applying our algorithms to different languages
Thank you for your attention!