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

Outline

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

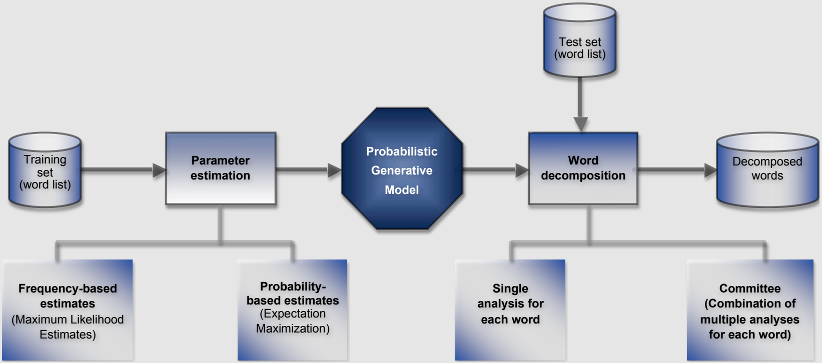
Our objective for Morpho Challenge

- 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

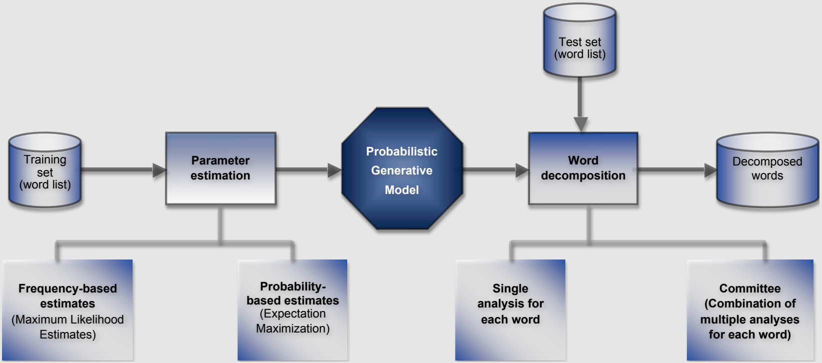
Algorithm: Overview

PROMODES = Probabilistic Generative Model for Different Degrees of Supervision



Algorithm: Overview

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 $\rightarrow$	$\langle 0000000000 \rangle_1$	$\rightarrow$	$\langle unbreakable \rangle_1$	$\rightarrow$	0.02
	...		...		...
	$\langle 0100001000 \rangle_k$	$\rightarrow$	$\langle un, break, able \rangle_k$	$\rightarrow$	0.50
	...		...		...
	$\langle 1111111111 \rangle_{2m}$	$\rightarrow$	$\langle u, n, b, r, e, a, k, a, b, l, e \rangle_{2m}$	$\rightarrow$	0.01

Algorithm: Probabilistic generative model

Linearization of PGM

- Segmentation perspective $\rightarrow$ 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

$$\begin{array}{cccccccccccc}
 & & & & & & \text{unbreakable} & & & & \\
 & & & & & & \downarrow & & & & \\
 & un & nb & br & re & ea & ak & ka & ab & bl & le \\
\$u \rightarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow \\
 & \$n & \$b & \$r & \$e & \$a & \$k & \$a & \$b & \$l & \$e \\
 & & & & & & \downarrow & & & & \\
 & \langle 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \rangle \\
 & & & & & & \downarrow & & & & \\
 & & & & & & \langle un, break, able \rangle & & & &
\end{array}$$

\$ = 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 θ
- Alternating between calculating likelihood of parameter estimates (E) and maximization (M)
- Convergence criterion: **Kullback-Leibler divergence**

Parameter estimation: Expectation Maximization

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 $\mu_{xy,x'y'} = \begin{cases} 1, \text{ if } x' = x \text{ and } y' = y \text{ in } w_j \text{ at } i\text{th position}, \\ 0, \text{ otherwise}. \end{cases}$

$\rightarrow$ 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

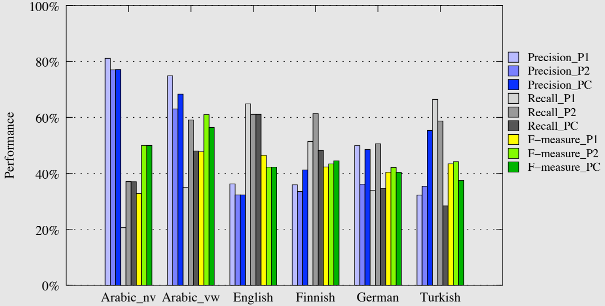
PROMODES 1 (P1):	frequency-based parameter estimation (pre-processing with trie-based alg.), single word analysis
PROMODES 2 (P2):	probability-based parameter estimation (Expectation Maximization), initialization $\rightarrow$ random segmentation, single word analysis
PROMODES COMMITTEE (PC):	different initializations of EM, committee decision (multiple analysis for each word)

Committee of unsupervised learners

- combination of different solutions into cumulative vector $\rightarrow$ majority vote

word	committee (multiple analyses)	cumulative vector		segmentation vector	segmentation
	$\langle 0101000101 \rangle \rightarrow$				
	$\langle 1000011000 \rangle \rightarrow$				
unbreakable $\rightarrow$	$\langle 0110101110 \rangle \rightarrow$	$\langle 1311114212 \rangle$	$\xrightarrow{\text{majority vote}}$	$\langle 0100001000 \rangle \rightarrow$	$\langle$un,break,able$\rangle$
	$\langle 0100001000 \rangle \rightarrow$				
	$\langle 0000001001 \rangle \rightarrow$				

Results



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

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

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!