# A Multi-purpose Bayesian Model for Word-Based Morphology

Maciej Janicki

University of Leipzig

September 17, 2015

Automatische Sprachverarbeitung

# Morphology in NLP

`wahrscheinlichster`

`wahr-schein-lich-st-er`

`wahr⟨ADJ⟩-schein⟨NN⟩-lich⟨SUFF_ADJ⟩-st⟨SUP⟩-er⟨M.SG.NOM⟩`

Automatische Sprachverarbeitung

# Morphology in NLP

`wahrscheinlichster`

`wahr-schein-lich-st-er`

`wahr⟨ADJ⟩-schein⟨NN⟩-lich⟨SUFF_ADJ⟩-st⟨SUP⟩-er⟨M.SG.NOM⟩`

**provided:**

- morpheme segmentation (with or without tags)

**needed:**

- is a valid word?
- lemma, possible tags (PoS, inflectional)
- other word features

Automatische Sprachverarbeitung

# Whole Word Morphology

PHON: /kæt/

SYNT: N, SG

SEM:

# Whole Word Morphology

| | |
|---|---|
| PHON: | /kæt/ |
| SYNT: | N, SG |
| SEM: |  |

| | |
|---|---|
| PHON: | /kæts/ |
| SYNT: | N, PL |
| SEM: |  |

Automatische Sprachverarbeitung

# Whole Word Morphology

$$\begin{bmatrix} \text{PHON:} & /X/ \\ \text{SYNT:} & N,\ \textsc{sg} \\ \text{SEM:} & \spadesuit \end{bmatrix} \longleftrightarrow \begin{bmatrix} \text{PHON:} & /Xs/ \\ \text{SYNT:} & N,\ \textsc{pl} \\ \text{SEM:} & \text{many } \spadesuit \end{bmatrix}$$

- concentrates on relations between words
- no "absolute structure/analysis"
- not decomposable
- allows for non-concatenative operations

Automatische Sprachverarbeitung

# Unigram distribution

Let $u(w)$ be the unigram-based probability of the word $w$.

$$Pr(L) = Pr(|L|) \cdot |L|! \cdot \prod_{w \in L} u(w)$$

| | |
|---|---|
| ... | ... |
| sprache | $2.17 \cdot 10^{-11}$ |
| sprachen | $1.88 \cdot 10^{-12}$ |
| ... | ... |

(each word drawn independently from the unigram distribution)

Automatische Sprachverarbeitung

# Introducing rules

Let a morphological rule $r : /Xe/ \rightarrow /Xen/$ be known.
$r$ applies from left to right with probability $\pi_r = 0.53$ (*productivity*).

| | | |
|---|---|---|
| ... | ... | |
| sprache | $2.17 \cdot 10^{-11}$ | (*sprachen* derived by $r$) |
| $\rightarrow$ sprachen | 0.53 | |
| ... | ... | |

# Introducing rules

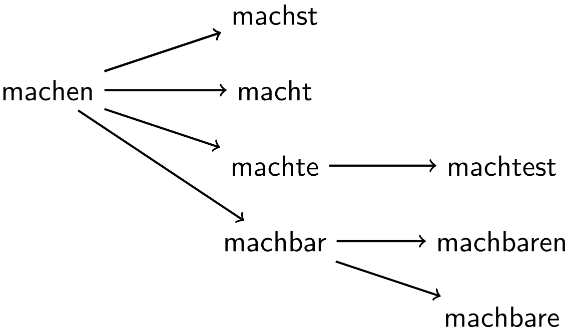
Let a morphological rule $r : /Xe/ \rightarrow /Xen/$ be known. $r$ applies from left to right with probability $\pi_r = 0.53$ (*productivity*).

| | | |
|---|---|---|
| ... | ... | |
| sprache | $2.17 \cdot 10^{-11}$ | (*sprachen* derived by $r$) |
| $\rightarrow$ sprachen | 0.53 | |
| ... | ... | |

| | | |
|---|---|---|
| ... | ... | |
| sprache | $2.17 \cdot 10^{-11} \cdot 0.47$ | (*sprachen* **not** derived by $r$) |
| sprachen | $1.88 \cdot 10^{-12}$ | |
| ... | ... | |

Automatische Sprachverarbeitung

# Lexicon as directed graph




Automatische Sprachverarbeitung

# Learning

**Model components:**

- $L$ – lexicon (graph)
- $R$ – set of rules with their productivities

defined: $P(L|R)$, $P(R)$

find:

$$\begin{aligned}\hat{R} &= \arg\max_R P(R|L) \\ &= \arg\max_R \frac{P(L|R)P(R)}{P(L)} \\ &= \arg\max_R P(L|R)P(R)\end{aligned}$$

Automatische Sprachverarbeitung

# Learning (cont.)

**Supervised learning:**

- given $L$, find $R$
- extract rules from pairs of related words
- ML estimation for rule productivities

# Learning (cont.)

**Unsupervised learning:**

- given $V(L)$, find $E(L)$ and $R$
- Find all reasonable edges.
  - find pairs of string-similar words
  - extract rules
  - choose 10k most frequent rules
  - create a “full” graph of all possible edges

Automatische Sprachverarbeitung

# Learning (cont.)

**Unsupervised learning:**

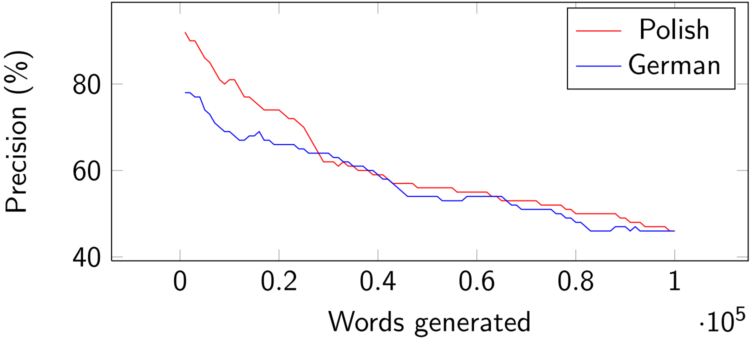
- given $V(L)$, find $E(L)$ and $R$
- Find all reasonable edges.
  - find pairs of string-similar words
  - extract rules
  - choose 10k most frequent rules
  - create a "full" graph of all possible edges
- Alternating ML estimation of $E(L)$ and $R$ ("hard EM").
  - "guess" an initial $R$
  - repeat until convergence:
  - find best $E(L)$ given $V(L)$ and $R$ (optimal branching)
  - find best $R$ given $V(L)$ and $E(L)$ (ML estimation)

ASV Automatische Sprachverarbeitung

# Lexicon expansion: task definition

- unsupervised training on 50k-wordlists (German, Polish)
- generate new words in the order of increasing cost

# Lexicon expansion: results



# Lemmatization and Tagging: task definition

- given a word, determine its lemma and PoS/inflectional tag
- training data:
  - supervised: word-lemma pairs
  - unsupervised: a set of words and a set of lemmas (without alignment)
- variants:
  - +/- **Lem**: lemmas of all unknown words included in the training data?
  - +/- **Tags**: tag of the target word given?
- baselines:
  - unsupervised: alignment based on least edit distance
  - supervised: Maximum Entropy classifier based on letter N-grams

Automatische Sprachverarbeitung

# Lemmatization and Tagging: results

**unsupervised:**

| Data | | | Results | | | Baseline | | |
|---|---|---|---|---|---|---|---|---|
| **Language** | **Lem** | **Tags** | **Lem** | **Tags** | **Lem+Tags** | **Lem** | **Tags** | **Lem+Tags** |
| German | + | + | 93% | 100% | 93% | 84% | – | – |
| | + | – | 80% | 46% | 45% | 76% | – | – |
| | – | + | 76% | 100% | 76% | 44% | – | – |
| | – | – | 61% | 34% | 28% | 43% | – | – |
| Polish | + | + | 84% | 100% | 84% | 80% | – | – |
| | + | – | 80% | 61% | 59% | 67% | – | – |
| | – | + | 80% | 100% | 80% | 41% | – | – |
| | – | – | 79% | 61% | 55% | 40% | – | – |

**supervised:**

| Data | | | Results | | | Baseline | | |
|---|---|---|---|---|---|---|---|---|
| **Language** | **Lem** | **Tags** | **Lem** | **Tags** | **Lem+Tags** | **Lem** | **Tags** | **Lem+Tags** |
| German | + | + | 97% | 100% | 97% | 89% | 97% | 89% |
| | + | – | 92% | 38% | 38% | 19% | 20% | 19% |
| | – | + | 90% | 100% | 90% | 89% | 97% | 89% |
| | – | – | 57% | 20% | 19% | 19% | 20% | 19% |
| Polish | + | + | 94% | 100% | 94% | 83% | 94% | 83% |
| | + | – | 93% | 56% | 56% | 33% | 36% | 33% |
| | – | + | 88% | 100% | 88% | 83% | 94% | 83% |
| | – | – | 68% | 40% | 38% | 33% | 36% | 33% |

# Inflection: results

**Task definition:**

- given lemma and tag, output the correct inflected form
- baseline: Maximum Entropy classifier based on letter N-grams

**Results:**

| Language | Result | Baseline |
|---|---|---|
| German | 84% | 83% |
| Polish | 86% | 84% |

ASV Automatische Sprachverarbeitung

# Conclusion

- focus on **relations** between words, rather than segmentation
- non-concatenative morphology included
- many training possibilities: unsupervised, supervised, manual editing
- one model for multiple tasks