Letter-to-Phoneme Conversion for a German Text-to-Speech System

Vera Demberg

Institut für Maschinelle Sprachverarbeitung (IMS)
Universität Stuttgart
und
IBM Deutschland Entwicklung GmbH
Böblingen

May 31, 2006
Overview

1. Introduction
2. Morphology
   - SMOR
   - Unsupervised Morphologies
3. Syllabification
   - Hidden Markov Model for Syllabification
4. Word Stress
   - German Word Stress
   - A Rule-based System
   - HMM for Stress Assignment
5. Grapheme-to-Phoneme Conversion
6. Summary
What part of a TTS system are we talking about?

- Tokenization
- Text Normalization
- Letter-to-Phoneme Conversion
- Pos-Tagging
- Parsing
- Prosody
- Synthesis
- Morphological Analysis
  - Graphemic Parsing
  - Grapheme-to-Phoneme Conversion
  - Syllabification
  - Stress Assignment
Why use morphological information?

Pronunciation of German words is sensitive to morphological boundaries

- *Granatapfel, Sternanisöl* (compounds)
- *Röschen* (derivational suffixes)
- *vertikal vs. vertickern* (affixes)
- *Weihungen vs. Gen* (inflectional suffixes)
Problems with SMOR

- Ambiguity
  - Akt+ent+asch+en
  - Akten+tasche+n
  - Akt+en+tasche+n

- Complex Lexicon Entries
  - Ab+bild+ung+en
  - Abbildung+en

- Insufficient Coverage
  - Kirschschaft
  - Adhäsionskurven
Higher F-measure does not always correspond directly to better performance on the grapheme-to-phoneme conversion task.

<table>
<thead>
<tr>
<th>morphology</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Meas.</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX annotation</td>
<td></td>
<td></td>
<td></td>
<td>2.64%</td>
</tr>
<tr>
<td>ETI</td>
<td>0.754</td>
<td>0.841</td>
<td>0.795</td>
<td>2.78%</td>
</tr>
<tr>
<td>SMOR-large segments</td>
<td>0.954</td>
<td>0.576</td>
<td>0.718</td>
<td>3.28%</td>
</tr>
<tr>
<td>SMOR-heuristic</td>
<td>0.902</td>
<td>0.754</td>
<td>0.821</td>
<td>2.92%</td>
</tr>
<tr>
<td>SMOR-CELEX-weighted</td>
<td>0.949</td>
<td>0.639</td>
<td>0.764</td>
<td>3.22%</td>
</tr>
<tr>
<td>SMOR-newLex</td>
<td>0.871</td>
<td>0.804</td>
<td><strong>0.836</strong></td>
<td>3.00%</td>
</tr>
<tr>
<td>no morphology</td>
<td></td>
<td></td>
<td></td>
<td>3.63%</td>
</tr>
</tbody>
</table>
Unsupervised Morphologies

- Unsupervised approaches require raw text only
- they are language-independent (ideally)
- segmentation quality of unsupervised systems not sufficient

<table>
<thead>
<tr>
<th>morphology</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Meas.</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bordag</td>
<td>0.665</td>
<td>0.619</td>
<td>0.641</td>
<td>4.38%</td>
</tr>
<tr>
<td>Morfessor</td>
<td>0.709</td>
<td>0.418</td>
<td>0.526</td>
<td>4.10%</td>
</tr>
<tr>
<td>Bernhard</td>
<td>0.649</td>
<td>0.621</td>
<td>0.635</td>
<td>3.88%</td>
</tr>
<tr>
<td>RePortS</td>
<td>0.711</td>
<td>0.507</td>
<td>0.592</td>
<td>3.83%</td>
</tr>
<tr>
<td>no morphology</td>
<td></td>
<td></td>
<td></td>
<td>3.63%</td>
</tr>
<tr>
<td>SMOR+newLex</td>
<td>0.871</td>
<td>0.804</td>
<td>0.836</td>
<td>3.00%</td>
</tr>
<tr>
<td>ETI</td>
<td>0.754</td>
<td>0.841</td>
<td>0.795</td>
<td>2.78%</td>
</tr>
<tr>
<td>CELEX</td>
<td></td>
<td></td>
<td></td>
<td>2.64%</td>
</tr>
</tbody>
</table>
Unsupervised approaches require raw text only
- they are language-independent (ideally)
- segmentation quality of unsupervised systems not sufficient

<table>
<thead>
<tr>
<th>morphology</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Meas.</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bordag</td>
<td>0.665</td>
<td>0.619</td>
<td>0.641</td>
<td>4.38%</td>
</tr>
<tr>
<td>Morfessor</td>
<td>0.709</td>
<td>0.418</td>
<td>0.526</td>
<td>4.10%</td>
</tr>
<tr>
<td>Bernhard</td>
<td>0.649</td>
<td>0.621</td>
<td>0.635</td>
<td>3.88%</td>
</tr>
<tr>
<td>RePortS</td>
<td>0.711</td>
<td>0.507</td>
<td>0.592</td>
<td>3.83%</td>
</tr>
<tr>
<td>no morphology</td>
<td></td>
<td></td>
<td></td>
<td>3.63%</td>
</tr>
<tr>
<td>SMOR+newLex</td>
<td>0.871</td>
<td>0.804</td>
<td>0.836</td>
<td>3.00%</td>
</tr>
<tr>
<td>ETI</td>
<td>0.754</td>
<td>0.841</td>
<td>0.795</td>
<td>2.78%</td>
</tr>
<tr>
<td>CELEX</td>
<td></td>
<td></td>
<td></td>
<td>2.64%</td>
</tr>
</tbody>
</table>

Vera Demberg (IMS / IBM)  G2P for German TTS  May 31, 2006  7 / 25
Why a separate module for Syllabification?

- Improve g2p conversion quality (cf. Marchand and Damper 2005)
- Prevent phonologically impossible syllables
  \[.1 \ ? \ A \ L \ . \ T \ . \ B \ U \ N \ . \ D \ E\# \ S \ . \ P \ R \ AE \ . \ Z \ I: \ . \ D \ AE \ N \ . \ T \ E\# \ N/\]
  \[.1 \ K \ U: \ R\# \ . \ V \ E\# \ N \ . \ L \ I: \ N \ E: \ .1 \ A: \ L \ S/\]
- Basis for a separate stress module
Syllabication as a Tagging Problem

Using a Hidden Markov Model for Syllable Boundary Labelling
(Schmid, Möbius and Weidenkaff, 2005)

Definition:

\[
\hat{s}_1^n = \arg \max_{s_1^n} \prod_{i=1}^{n+1} P(\langle l; s \rangle_i | \langle l; s \rangle_{i-k}^{i-1})
\]

Model sketch:

![Diagram with states and symbols representing syllabication]

... → r/B → k/N → e/N → ...  hidden states

... → r → k → e → ...  output symbols
Kneser-Ney Smoothing is superior to Schmid Smoothing.

<table>
<thead>
<tr>
<th></th>
<th>schmid</th>
<th>kneser-ney</th>
</tr>
</thead>
<tbody>
<tr>
<td>nomorph, proj.</td>
<td>3.43%</td>
<td>3.10%</td>
</tr>
<tr>
<td>ETI, proj.</td>
<td>2.95%</td>
<td>2.63%</td>
</tr>
<tr>
<td>CELEX, proj.</td>
<td>2.17%</td>
<td>1.91%</td>
</tr>
<tr>
<td>Phonemes</td>
<td>1.84%</td>
<td>1.53%</td>
</tr>
<tr>
<td>Phonemes (90/10)</td>
<td>0.18%</td>
<td>0.18%</td>
</tr>
</tbody>
</table>
Syllabification – Summary

Were the goals achieved?

- Improved g2p conversion quality
  - preprocessing for AWT: WER decreased from 26.6% to 25.6%
    (significant at $p = 0.015$ according to a two-tailed binomial test)

- Used constraints to prevent ungrammatical syllables

<table>
<thead>
<tr>
<th></th>
<th>k=4</th>
</tr>
</thead>
<tbody>
<tr>
<td>constraint</td>
<td>3.10%</td>
</tr>
<tr>
<td>no constraint</td>
<td>3.48%</td>
</tr>
</tbody>
</table>

- Basis for a stress module
German Word Stress

Why a separate Word Stress Component?

- 14.5% of words in list are assigned incorrect stress (21.15% overall WER)
  - more than one primary stress: 5.3%
  - no primary stress: 4%
  - wrong position of stress: 5.2%
- decision tree model cannot capture wide enough context to decide stress
- many wrong stress annotations in CELEX
German Word Stress

Describing German Word Stress:
  - compounds
    - right-branching: [[Lébens+mittel]+punkt]
    - left-branching: [Lebens+[mittel+punkt]]
      a) [Háupt+[bahn+hof]] because *Bahnhof* is lexicalized
      b) [Bundes+[kriminál+amt]] because fully compositional
  - affixes
    - always stressed: *ein-*, *auf-*, *-ieren*...
    - never stressed: *ver-*, *-heit*, *-ung*...
    - sometimes stressed: *um-*, *voll-*... (e.g. *úmfahren* vs. *umfáhren*)
    - some influence stress: *Musík* vs. *Músiker*, *Áutor* vs. *Autóren*
  - stems
    - syllable weight
    - syllable position
A rule-based approach

Word stress rules by Petra Wagner, based on Jessen

- claims to cover 95% of German words
- just 5 rules, full affix lists publicly accessible
- overcome problem of low quality training data

But real life is not that easy

- syllable weight defined on phonemes
- perfect morphology is needed: little above 50% without compounding information
- achieved only 84% of words correct with CELEX morphology
- real text contains many foreign words which the rules get wrong
A rule-based approach

Word stress rules by Petra Wagner, based on Jessen
- claims to cover 95% of German words
- just 5 rules, full affix lists publicly accessible
- overcome problem of low quality training data

But real life is not that easy
- syllable weight defined on phonemes
- perfect morphology is needed: little above 50% without compounding information
- achieved only 84% of words correct with CELEX morphology
- real text contains many foreign words which the rules get wrong
Adapting the HMM to word stress assignment

The basic units of the model are syllable–stress-tag pairs.

\[ \hat{str}_1^n = \arg \max_{str_1^n} \prod_{i=1}^{n+1} P(\langle syl; str \rangle_i | \langle syl; str \rangle_{i-1}^{i-1}) \]

Importance of Constraint:

<table>
<thead>
<tr>
<th>WER with constraint</th>
<th>WER without constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.9%</td>
<td>31.9%</td>
</tr>
</tbody>
</table>
Adapting the HMM to word stress assignment

The basic units of the model are syllable–stress-tag pairs.

\[
\hat{str}_1^n = \arg \max_{str_1^n} \prod_{i=1}^{n+1} P(\langle syl; str \rangle_i | \langle syl; str \rangle_{i-k}^{i-1})
\]

Importance of Constraint:

<table>
<thead>
<tr>
<th>WER with constraint</th>
<th>WER without constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.9%</td>
<td>31.9%</td>
</tr>
</tbody>
</table>
Smoothing

- Hard data sparsity problem since defined on syllable–stress pairs need to estimate probabilities from lower order n-gram models:
  \[ p(n\text{-gram}) = \text{backoff-factor} \times p(n-1\text{-gram}) \]

- Typical type of error with initial Schmid Smoothing:
  - 5véř+1web2st
  - Problematic point is the backoff factor:
    \[ \Theta = \frac{freq(w_{i-n+1}^{i-1}) + \Theta}{\Theta} \]

- Modified Kneser-Ney Smoothing (cf. Chen and Goodman 98)
  - Backoff factor:
    \[ \frac{D}{freq(w_{i-n+1}^{i-1})} N_{1+}(w_{i-n+1}^{i-1} \cdot) \]

  Estimates n-gram probabilities from the number of different states a context was seen in.
Performance of the HMM

- Comparison of different smoothing methods:

<table>
<thead>
<tr>
<th>context window smoothing alg.</th>
<th>k=1</th>
<th>k=2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>schmid</td>
<td>kneser-ney</td>
</tr>
<tr>
<td>Letters</td>
<td>14.2%</td>
<td>9.9%</td>
</tr>
<tr>
<td>Lett. + morph</td>
<td>13.2%</td>
<td>9.9%</td>
</tr>
<tr>
<td>Phonemes</td>
<td>12.6%</td>
<td>8.8%</td>
</tr>
</tbody>
</table>

- Performance of decision tree when input letters are annotated with stress tags:
  21.1% WER instead of 26.6% WER
Why not apply the HMM to grapheme to phoneme conversion?

- this time defined on letter–phoneme-sequence pairs (“graphones”, e.g. \( a-\_1\_?\_A:\) )

\[
\hat{p}_1^n = \arg \max_{p_1^n} \prod_{i=1}^{n+1} P(\langle l; p \rangle_i | \langle l; p \rangle_{i-1, k})
\]

- related work :-(
  - Bisani and Ney, 2002
  - Galescu and Allen, 2001
  - Chen, 2003
Why not apply the HMM to grapheme to phoneme conversion?

- this time defined on letter–phoneme-sequence pairs ("graphones", e.g. a-1_?_A):

\[
\hat{p}_1^n = \arg \max_{p_1^n} \prod_{i=1}^{n+1} P(\langle l; p \rangle_i | \langle l; p \rangle_{i-1}^{i-1})
\]

- related work :
  - Bisani and Ney, 2002
  - Galescu and Allen, 2001
  - Chen, 2003
Issues

- **Alignment**
  
  An aligned corpus is needed as an input for the algorithm.

- **Pruning**
  
  The full graph is immense: each letter can on avg. map to 12 different phoneme-sequences.
  
  Even when Viterbi algorithm is used, approx. 8 min / word.
  
  Pruning Strategy: only ever remember the best 15 paths.

- **Smoothing**
  
  Again, Kneser-Ney Smoothing worked significantly better than Schmid Smoothing.
Integration of Constraints

Finally, I integrated the **phonological syllable constraints** and the **word stress constraint** directly into the g2p- model.

<table>
<thead>
<tr>
<th></th>
<th>modular Preproc.</th>
<th>modular Postproc.</th>
<th>one-step constr.</th>
<th>no constr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>no morph</td>
<td>83.4%</td>
<td>84.8%</td>
<td>86.3%</td>
<td>78.5%</td>
</tr>
<tr>
<td>AWT no morph</td>
<td>78.9%</td>
<td></td>
<td></td>
<td>73.4%</td>
</tr>
<tr>
<td>ETI morph</td>
<td></td>
<td></td>
<td>86.4%</td>
<td></td>
</tr>
<tr>
<td>AWT ETI morph</td>
<td></td>
<td></td>
<td></td>
<td>78.2%</td>
</tr>
<tr>
<td>CELEX morph</td>
<td>83.9%</td>
<td>85.6%</td>
<td>86.7%</td>
<td>74.7%</td>
</tr>
<tr>
<td>AWT CELEX morph</td>
<td>84.3%</td>
<td>84.1%</td>
<td></td>
<td>78.4%</td>
</tr>
</tbody>
</table>

Why is the HMM so much better than the decision tree?

- it integrates phonological constraints
- the model compresses the data much less
Performance on other Languages

Comparison to state-of-the-art models

<table>
<thead>
<tr>
<th>corpus</th>
<th>HMM-KN</th>
<th>PbA</th>
<th>Chen</th>
<th>AWT</th>
</tr>
</thead>
<tbody>
<tr>
<td>E - Nettalk</td>
<td>64.6%</td>
<td>65.5%</td>
<td>67.9%</td>
<td>65.4%</td>
</tr>
<tr>
<td>E - Nettalk (+syll)</td>
<td>70.6%</td>
<td>71.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E - Teacher’s WB</td>
<td>71.5%</td>
<td>71.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E - beep</td>
<td>85.7%</td>
<td>86.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E - CELEX</td>
<td>76.3%</td>
<td></td>
<td></td>
<td>68.3%</td>
</tr>
<tr>
<td>French - Brulex</td>
<td>88.4%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

- **Morphology**
  - SMOR lacks some information that is relevant for G2P
  - Unsupervised approaches are not yet good enough
- **Syllable boundary and stress annotation** improves conversion quality
- **The choice of a smoothing method** matters a lot
- **Joint n-gram models** are very good for grapheme-to-phoneme conversion
  - Reduction of word error rate by up to 50% wrt. a decision tree
  - A morphological preprocessing component is less important because the model captures morphemes well
- **Models that do several strongly inter-dependent steps in just one step** are superior to a pipeline architecture
- **Postprocessing of syllabification and stress** yields better results than preprocessing
Questions?


Summary

John Goldsmith. 
Unsupervised learning of the morphology of a natural language. 

Samarth Keshava and Emily Pitler. 
A simpler, intuitive approach to morpheme induction. 

Yannick Marchand and Robert I. Damper. 
Can syllabication improve pronunciation by analogy of English? 

Helmut Schmid, Bernd Möbius, and Julia Weidenkaff. 
Tagging syllable boundaries with hidden Markov models. 

Petra Wagner. 
Improving automatic prediction of German lexical stress. 
Alternative Strategies for Disambiguation

- always choose the analysis with the smallest number of morphemes
  \( Ab+fal+leim+er \ vs. \ Abfall+eimer \)

- use frequencies from taz for disambiguation
  \( Topf+es \ vs. \ top+Fes \)

- learn a weighted FST after disambiguating with manually annotated analyses from CELEX
Complex Lexicon Entries and Insufficient Coverage

Improving Recall

- heuristic: always choose the analysis with the largest number of morphemes, if this analysis has at least one common boundary with the analysis made of the smallest number of morphemes
  - Ab+bild+ung+en instead of Abbildung+en
  - not Akt+ent+asch+en instead of Akten+tasche+n

- insert morphological boundaries into the lexicon
  \[ Abbildung \rightarrow Ab<X>bild<X>ung \]

Coping with Out-of-vocabulary words (OOV)

- use the SMOR list of affixes and peal off anything you can