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

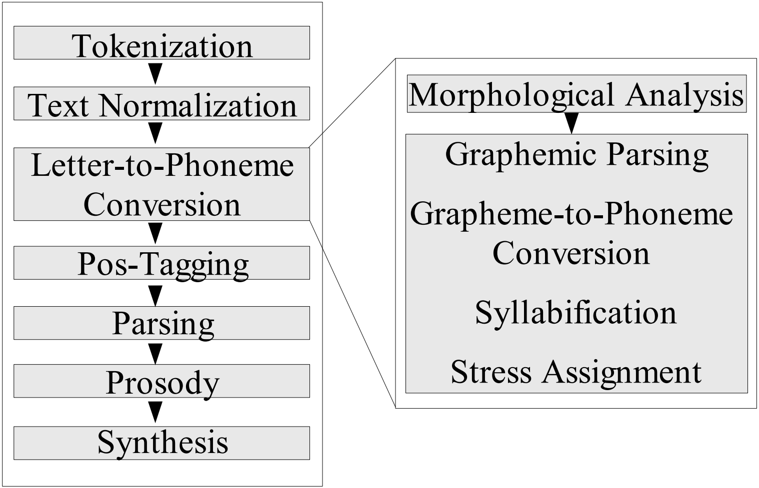
Vera Demberg

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

May 31, 2006

Overview

What part of a TTS system are we talking about?



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)

SMOR

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
  - *Kirschsaft*
  - *Adhäsionskurven*

Results for Experiments with SMOR

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

morphology	Precision	Recall	F-Meas.	PER
CELEX annotation				2.64%
ETI	0.754	**0.841**	0.795	**2.78**%
SMOR-large segments	**0.954**	0.576	0.718	3.28%
SMOR-heuristic	0.902	0.754	0.821	2.92%
SMOR-CELEX-weighted	0.949	0.639	0.764	3.22%
SMOR-newLex	0.871	0.804	**0.836**	3.00%
no morphology				3.63%

Unsupervised Morphologies

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

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

Unsupervised Morphologies

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

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

Syllabification

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

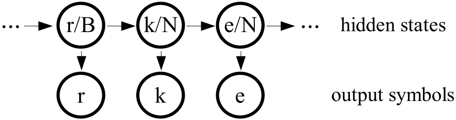
Syllabification as a Tagging Problem

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

- Definition:

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

- Model sketch:



Smoothing the Syllabification HMM

Kneser-Ney Smoothing is superior to Schmid Smoothing.

WER for k=4	schmid	kneser-ney
nomorph, proj.	3.43%	3.10%
ETI, proj.	2.95%	2.63%
CELEX, proj.	2.17%	1.91%
Phonemes	1.84%	1.53%
Phonemes (90/10)	0.18%	0.18%

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

WER	k=4
constraint	3.10%
no constraint	3.48%

- 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.

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

- Importance of Constraint:

WER with constraint	WER without constraint
9.9%	31.9%

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:

WER with constraint	WER without constraint
9.9%	31.9%

Smoothing

- Hard data sparsity problem since defined on syllable–stress pairs need to estimate probabilities from lower order n-gram models: $p(n\text{ - }gram) = backoff\text{-}factor * p(n\text{-}1\text{ - }gram)$
- typical type of error with initial Schmid Smoothing:
  - *5vér+1web2st*
  - problematic point is the backoff factor:

$$\frac{\Theta}{freq(w_{i-n+1}^{i-1}) + \Theta}$$

- Modified Kneser-Ney Smoothing (cf. Chen and Goodman 98) backoff factor:

$$\frac{D}{freq(w_{i-n+1}^{i-1})} N_{1+}(w_{i-n+1}^{i-1} \bullet)$$

  estimates n-gram probabilities from the number of *different* states a context was seen in.

Performance of the HMM

- Comparison of different smoothing methods:

context window	k=1		k=2	
smoothing alg.	schmid	kneser-ney	schmid	kneser-ney
Letters	14.2%	9.9%	19.7%	9.4%
Lett. + morph	13.2%	9.9%	18.6%	10.3%
Phonemes	12.6%	8.8%	17.3%	8.7%

- Performance of decision tree when input letters are annotated with stress tags:
  21.1% WER instead of 26.6% WER

Grapheme-to-Phoneme Conversion

Why not apply the HMM to grapheme to phoneme conversion?

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

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

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

Grapheme-to-Phoneme Conversion

Why not apply the HMM to grapheme to phoneme conversion?

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

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

- 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

	modular		one-step	
	Preproc.	Postproc.	constr.	no constr.
no morph	83.4%	84.8%	**86.3%**	78.5%
AWT no morph	78.9%			**73.4%**
ETI morph			86.4%	
AWT ETI morph				**78.2%**
CELEX morph	83.9%	85.6%	**86.7%**	74.7%
AWT CELEX morph	84.3%	84.1%		**78.4%**

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

corpus	HMM-KN	PbA	Chen	AWT
E - Nettalk		65.5%	67.9%	
E - Nettalk	64.6%		65.4%	
E - Nettalk (+syll)	70.6%	71.7%		
E - Teacher's WB	71.5%	71.8%		
E - beep	85.7%	86.7%		
E - CELEX	76.3%			68.3%
French - Brulex	88.4%			

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?

Delphine Bernhard.
Unsupervised morphological segmentation based on segment predictability and word segments alignment.
In *Proceedings of 2nd Pascal Challenges Workshop*, pages 19–24, Venice, Italy, 2006.

M. Bisani and H. Ney.
Investigations on joint multigram models for grapheme-to-phoneme conversion.
In *Proceedings of the 7th International Conference on Spoken Language Processing*, pages 105–108, 2002.

Stanley F. Chen and Joshua Goodman.
An empirical study of smoothing techniques for language modeling.
In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318, Morristown, NJ, USA, 1996. Association for Computational Linguistics.

Stanley F. Chen.
Conditional and joint models for grapheme-to-phoneme conversion.
In *Eurospeech*, 2003.

Mathias Creutz and Krista Lagus.
Unsupervised models for morpheme segmentation and morphology learning.
In *ACM Transaction on Speech and Language Processing*, 2006.

Lucian Galescu and James Allen.
Bi-directional conversion between graphemes and phonemes using a joint n-gram model.
In *Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001.

John Goldsmith.
Unsupervised learning of the morphology of a natural language.
Computational Linguistics, 27(2):153–198, 2001.

Samarth Keshava and Emily Pitler.
A simpler, intuitive approach to morpheme induction.
In *Proceedings of 2nd Pascal Challenges Workshop*, pages 31–35, Venice, Italy, 2006.

Yannick Marchand and Robert I. Damper.
Can syllabification improve pronunciation by analogy of English?
Natural Language Engineering, 2005.

Helmut Schmid, Bernd Möbius, and Julia Weidenkaff.
Tagging syllable boundaries with hidden Markov models.
IMS, unpublished, 2005.

Petra Wagner.
Improving automatic prediction of German lexical stress.
In *Proceedings of the 15th ICPhS*, pages 2069–2072, Barcelona, Spain, 2003.

Disambiguation

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 → *Ab<X>bild<X>ung*

Coping with Out-of-vocabulary words (OOV)

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