UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

Computer Science, Electrical Engineering and Mathematics
NT Communications Engineering
Prof. Dr.-Ing. Reinhold Häb-Umbach

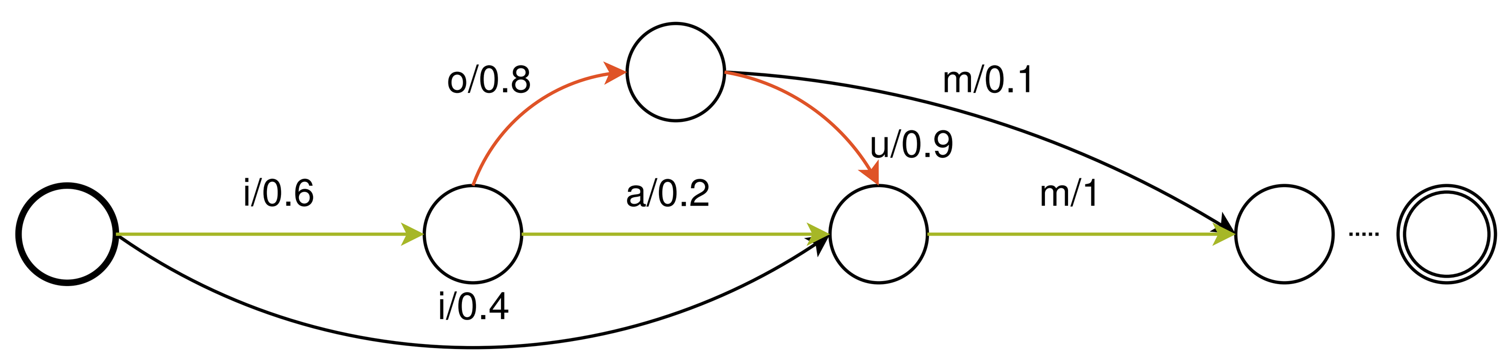
UNSUPERVISED WORD SEGMENTATION FROM NOISY INPUT

Jahn Heymann, Oliver Walter and Reinhold Haeb-Umbach
University of Paderborn, Germany
{walter,haeb}@nt.uni-paderborn.de
http://nt.uni-paderborn.de

Bhiksha Raj
Carnegie Mellon University, USA
bhiksha@cs.cmu.edu
http://mlsp.cs.cmu.edu

Introduction

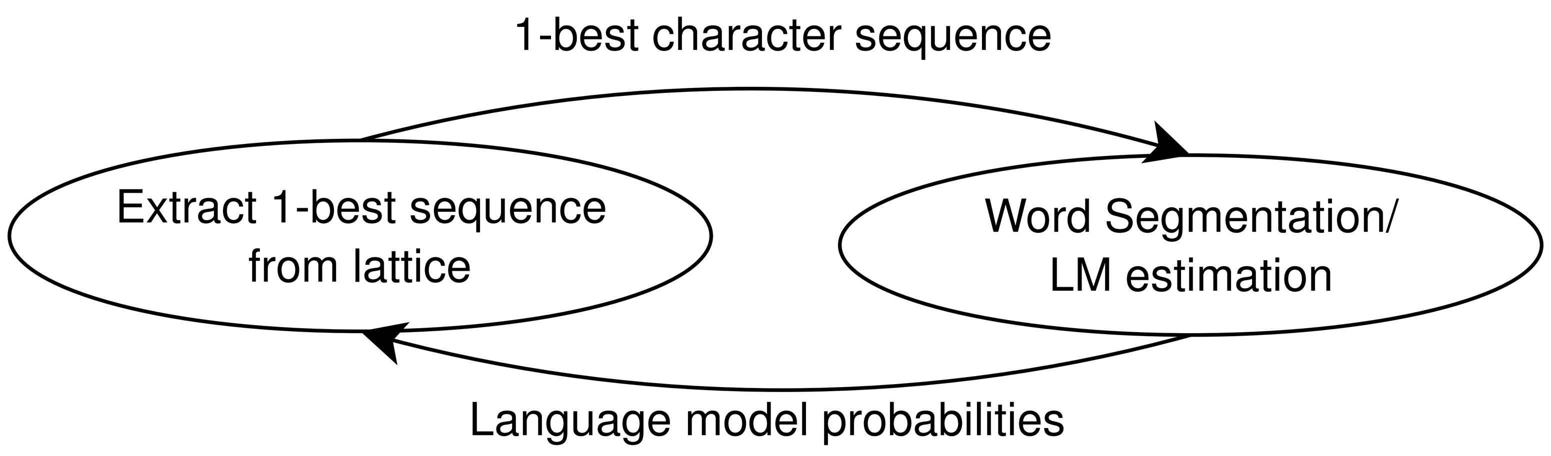
- Segmentation of character sequence into words using Bayesian nonparametric approach [Mochihashi09]
  - Example: iamatestsequence → i am a test sequence
- **Here:** Noisy character lattice with erroneous 1-best sequence (using zerogram character language model)
  - Example: ioumotasdcekunce → i am a test sequence



  - **Assumption:** correct character string present in lattice
- **Outlook:** Unsupervised language acquisition from speech

Iterative 2-step Algorithm

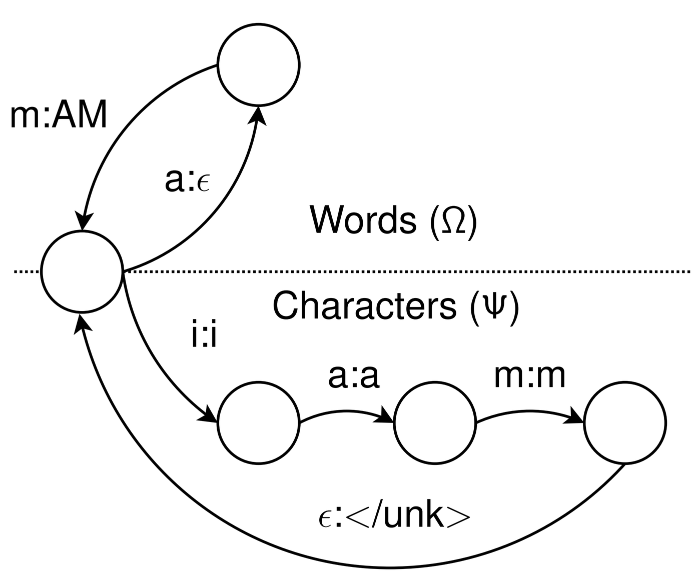
- **Iterate:** 1-best sequence extraction and word segmentation



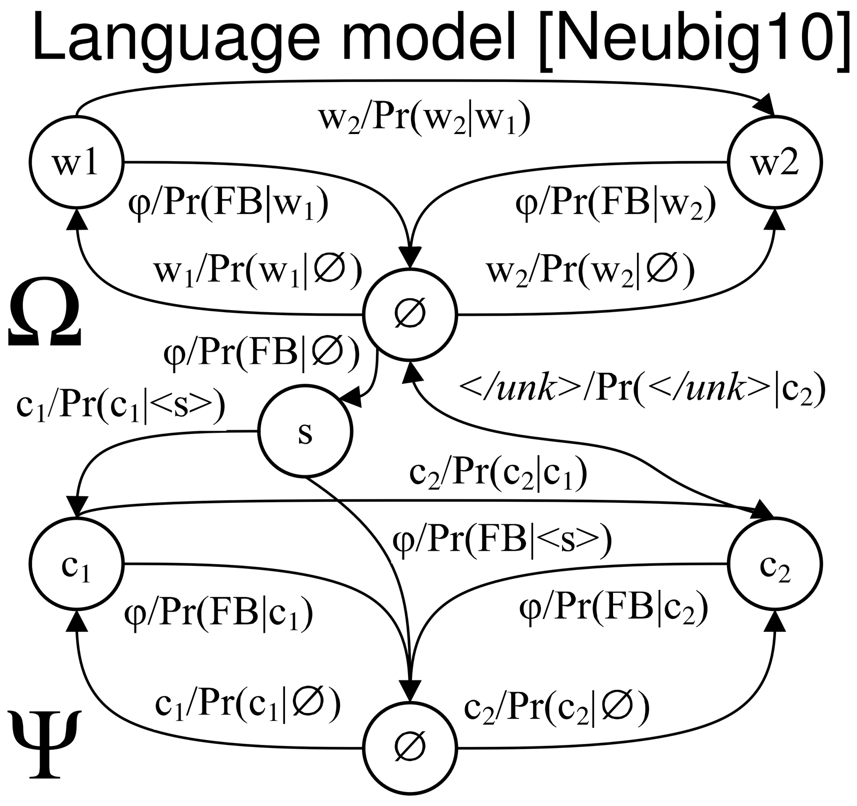
- Simultaneous error correction and word segmentation!
  - Exploiting consistency of character sequence within words

WFST based implementation

- WFST used to determine possible segmentations and their sequence probability ⇒ Sample most likely segmentation
- Lexicon WFST consists of known words and unknown character sequences



- Contains all possible subsequences for a string
- Language model WFSA resembles Pitman-Yor Language model [Neubig10]



- Building of lexicon WFST computationally feasible for a single character sequence only ⇒ Two step algorithm

References

[Mochihashi09] Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling
D. Mochihashi, T. Yamada, and N. Ueda, ACL 2009
[Neubig10] Learning a Language Model from Continuous Speech
G. Neubig, M. Mimura, S. Mori, T. Kawahara, InterSpeech 2010
[Teh06] A hierarchical Bayesian language model based on Pitman-Yor processes
YW. Teh, ACL 2006

Pitman-Yor Language Model [Teh06]

- Non-parametric i.e. unknown number of words
- Bayesian approach with power law prior (Zipf's law)
- Probability for word w in context $\mathbf{u}$ recursively calculated as

$$\Pr(w|\mathbf{u}, S, \Theta) = \frac{c_{\mathbf{u}w\cdot} - d_{|\mathbf{u}|} t_{\mathbf{u}w}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} + \frac{\theta_{|\mathbf{u}|} + d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{\theta_{|\mathbf{u}|} + c_{\mathbf{u}\cdot\cdot}} \Pr(w|\pi(\mathbf{u}), S, \Theta)$$

- Nesting: For $\mathbf{u} = \emptyset$ use likelihood of word w_i being character (phone) sequence $c_1, \ldots, c_k$ as base probability (fall back):

$$\Pr(w_i) \approx \prod_{i=1}^{k} \Pr\left(c_i | c_{i-n+1}, \ldots, c_{i-1}, S, \Theta\right)$$

- Probability for characters (phones) calculated as above

Experimental Setup

- Artificially generated lattices
- **Database:** Text prompts of WSJCAM0 training data
- White spaces between words removed
- String expanded to lattice by artificially including errors to X percent of the characters:
  - Draw action from [insert|delete|substitute] uniformly
  - Draw character uniformly in case of substitution or insertion
  - Draw probability p of correct character uniformly from $[0 \ldots 1]$
  - Add alternative character with weight $1 - p$

Experimental results

- Bigram word/8-gram character language model
- Example segmentation:

1-best sequence (zerogram at first iteration):
POWEYFSNKNZIMLCIAFBNCNUINLSVIVAHEVGGVCEOQRPHBOISSRXTHHIZBUQSIENASELDBYPMUMRTORYIRTTION RTJAGNFAATDNTRZTPBATEKTLDINGOMIAJV
after 25 iterations:
POWER FINANCIAL IS A FINANCIALSERVICES CONCERN THAT IS SIXTY NINE PERCENT HELD BY POWER CORPORATION OF CANADA A MONTREAL BASED HOLDING COMPANY

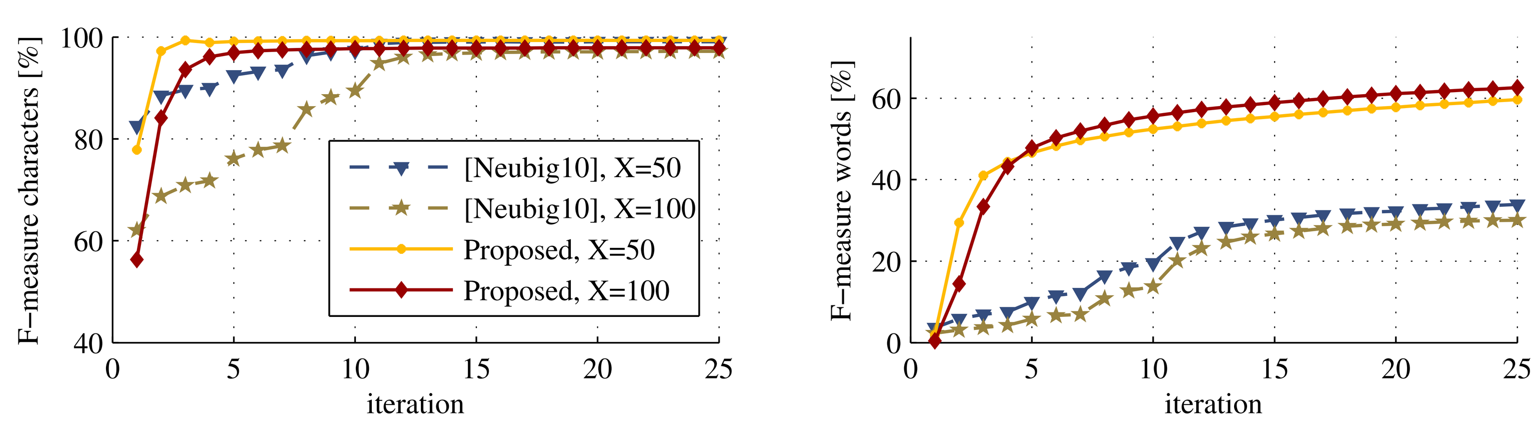


Figure 1: F-measure for characters and words over iterations at different strength of noise

- F-measure on clean error free sequence:
  - Bigram word LM: 36% [Neubig10], 65% (proposed)
  - Unigram word LM: 52% [Neubig10], 57% (proposed)

Conclusions

- Unsupervised vocabulary discovery from noisy input
- Iterative 2-step algorithm for simultaneous character error correction and word segmentation
- Significantly outperforms earlier algorithm
- Outlook: Replace input character lattice by phoneme lattice produced by ASR decoder

⇒ Unsupervised (zero-resource) speech recognition

Work supported by DFG contract no. Ha3455/9-1, Priority Programme SPP1527, NSF grant 1017256