UNIVERSITÄT PADERBORN
Die Universität der Informationsgesellschaft

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

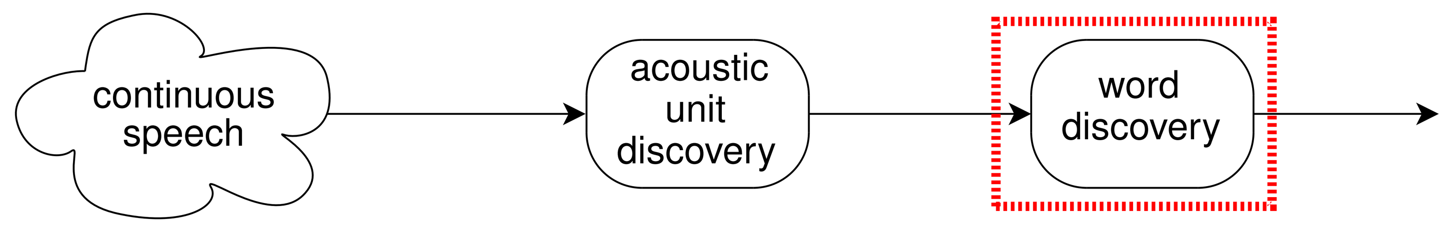
Unsupervised Word Discovery from Phonetic Input Using Nested Pitman-Yor Language Modeling

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

Sourish Chaudhuri and Bhiksha Raj
Carnegie Mellon University, USA
{sourishc,bhiksha}@cs.cmu.edu
http://mlsp.cs.cmu.edu

Introduction

- **Objective:** Unsupervised language acquisition
- "Learn a language like a child"
- Two-level hierarchical process:
  - 1. acoustic unit (phone) discovery
  - 2. lexical unit (word) discovery



Here: word discovery from character or phoneme input

Problem Formulation

- **Goal:** Segment character (or phone) sequence c_1^T

  UNCONVINCEDTHEHARRISBURGCITYCOUNCILINAPRILPASSEDANORDINANCEREQUIRINGLOCALE MPLOYERSWITHTENORMOREEMPLOYEESTOGIVEUPTOTWELVEWEEKSOFUNPAIDPARENTALLEAVE

  into words w_1^N

  UNCONVINCED THE HARRISBURG CITY COUNCIL IN APRIL PASSED AN ORDINANCE REQUIRING LOCAL EMPLOYERS WITH TEN OR MORE EMPLOYEES TO GIVE UP TO TWELVE WEEKS OF UNPAID PARENTAL LEAVE

  without a lexicon, i.e. unsupervised
- **Key idea:** Recurrent character sequences resemble words
- Learn orthography (or transcription) and language model, i.e. word probabilities, simultaneously
- Bayesian formulation:

$$\hat{w}_1^N = \underset{N, w_1^N}{\operatorname{argmax}} \Pr\left(w_1^N | c_1^T\right) \tag{1}$$

Forward Filtering/Backward Sampling Alg.

To solve (1), iterate between:

- **Forward filtering:** compute probability of candidate segmentations exploiting bigram language model
- **Backward sampling (Gibbs sampling):** draw word segmentation from above probabilities
- Use drawn word sequence to update language model

[Mochihashi, 2009]

Forward Filtering

- Input: $c_1^T = c_1 \dots c_T$: character (or phone) sequence
- $\alpha[t][k]$: probability of string c_1^t with $q_t = k$, i.e. with the last k characters, c_{t-k+1}^t being a word

$$\begin{aligned}\alpha[t][k] &= \Pr\left(c_1^t, q_t = k\right) \\ &= \sum_{j=1}^{t-k} \Pr\left(c_1^t, q_t = k, q_{t-k} = j\right) \\ &= \dots \\ &= \sum_{j=1}^{t-k} \underbrace{\Pr\left(c_{t-k+1}^t | c_{t-k-j+1}^{t-k}\right)}_{\text{Bigram probability}} \alpha[t-k][j]\end{aligned}$$

Pitman-Yor Language Model [Teh, 2006]

- Account for unknown number of words, i.e., unknown lexicon size: fall back to character (phone) language model
- Exploit prior knowledge about word frequencies (Zipf's law)
- Probability for word w in context $\mathbf{u}$ recursively calculated as

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

- In case of $\mathbf{u} = \emptyset$ use likelihood of word w_i being character (phone) sequence $c_1, \dots, c_k$ as parent probability (fall back):

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

- Probability for characters (phones) calculated as above
- Use uniform distribution as base distribution

Experiments

- **Database:** Wall Street Journal (WSJCAM0) acoustic model training data
  - 95629 word tokens (running words)
  - 10506 lexical items (unique words)
- **Input:** error free phonemic transcriptions
- Bigram language model
- Order m of phoneme language model varied from 4 to 8
- Example segmentation (character input for illustration):

$m = 4$:
UN CONVINCEDTHE HARRIS BURG CITY CO UNCIL I NAPRIL PASS ED AN ORD INANCE REQUIRINGLOCAL EMPLOYER S WITH TEN OR MORE EM PLOYEE S TO GIVE UP TO T WELVE WEEKSOFUN PAID PARENTAL LEAVE
$m = 6$:
UN CONVINCEDTHE HARRISBURG CITY COUNCIL INAPRIL PASSEDAN ORDINANCE REQUIRING LOCAL EMPLOYERS WITH TEN OR MORE EMPLOYEE S TO GIVE UP TOTWELVE WEEKSOFUNPAID PARENTAL LEAVE
$m = 8$:
UNCONVINCED THE HARRISBURG CITYCOUNCIL INAPRIL PASSED ANORDINANCE REQUIRING LOCAL EMPLOYERS WITH TEN OR MORE EMPLOYEES TO GIVE UP TO TWELVE WEEKSOFUNPAID PARENTALLEAVE

- Token discovery performance:

m	4	5	6	7	8	Ground truth
P	54.2	67.6	68.0	72.4	**73.3**	
R	49.9	51.2	52.1	56.8	**57.1**	
F	52.0	58.3	59.0	63.7	**64.2**	
Words	88070	72464	73294	74979	74471	95629

Table 1: Word token precision, recall and f-score from phoneme sequence input

- Lexical item discovery performance:

m	4	5	6	7	8	Ground truth
LP	33.0	37.1	38.7	43.7	**44.8**	
LR	56.0	65.2	64.0	65.6	**66.1**	
LF	41.5	47.3	48.3	52.5	**53.4**	
Words	17839	18466	17359	15775	15505	10506

Table 2: Word lexicon precision, recall and f-score from phoneme sequence input

Conclusions

- Unsupervised word segmentation on large vocabulary task
- 73.3% precision at 57.1% recall for word tokens
- Outlook:
  - Real data from acoustic unit discovery to be used
  - Extension to noisy data, e.g. lattice input
  - Consider variations in pronunciation

Work supported by DFG under contract no. Ha3455/9-1 within the the Priority Program SPP1527 "Autonomous Learning"