Inducing Probabilistic Syllable Classes Using Multivariate Clustering

authors: Karin Müller, Bernd Möbius, Detlef Prescher, IMS, Uni Stuttgart

Kolloquium, 19 June 2000

Overview

1. Multivariate Clustering & EM-Algorithm

2. Experiments
   3- and 5-dimensional models for English and German

3. Evaluation
   - Pseudo-Disambiguation (English + German 3-dimensional model)
   - Qualitative Evaluation (English + German 5-dimensional model)
   - Evaluation with a Grapheme-Phoneme Conversion Task for German

4. Discussion & Conclusion

Multivariate Clustering

- Clustering of multivariate Data > 2 dimensions via the EM algorithm, necessary for syllables that consists at least of 3 components: onset, nucleus and coda

- Tasks of EM-based clustering:
  - Induction of a smooth probability model on given data
  - Automatic discovery of class structure in the data

EM-Algorithm

- Probabilistic Method
- Works with unannotated data (no manually added class labels)
- Soft-clustering algorithm and not hard clustering
- Aim: Maximization of the probability of the training corpus (= Maximum-Likelihood-Method)

- Probability of a n-dimensional syllable (e.g. Onset, Nucleus, Coda) $y = (y_1, ..., y_d) \in Y_1 \times .. \times Y_d, d \geq 3$ is defined as:

$$p(y) = \sum_{c \in C} p(c, y) = \sum_{c \in C} p(c)p(y|c)$$

$$= \sum_{c \in C} p(c) \underbrace{p(y_1|c)...p(y_d|c)}_{``independence''}$$

Experiments - The Input

German:	newspaper corpus of 31 million words (STZ)
English:	collection of written and spoken data of 100 million words (BNC)

Look-up of words and their syllabification in a pronounciation dictionary $\Downarrow$

syllables and their frequency of occurance

Experiment - 3 dimensional models

- Induction of 3-dimensional model (Onset, Nucleus, Coda)
  German: 9327 syllable types
  English: 15.598 types

class 0 0.212					
NOP[I]	0.282	l	0.999	NOP[I]	0.460
t	0.107			n	0.121
l	0.074			N	0.096
d	0.071			z	0.079
b	0.065			t	0.042
s	0.060			ts	0.013
				f	0.012

Figure 1: Class #0 of a three-dimensional English Model with 12 classes

Experiment - 5 dimensional models

- Induction of 5-dimensional model (Onset, Nucleus, Coda, Position, Stress)
  Values: Position (INI, MED, FIN, ONE), Stress (USTR, STR)
  German: 16.595 syllable types
  English: 24.365 types

class 46 0.007				
NOP[E] 0.630 ts 0.256 d 0.0743 n 0.001	E 0.990	nt 0.602 t 0.128 n 0.092 pt 0.010 ks 0.004	INI 0.627 FIN 0.331 MED 0.040	STR 0.596 USTR 0.403

Figure 2: Class #46 of a five-dimensional German Model with 50 classes

Evaluation - Pseudo-Disambiguation

Aim: Finding the best 3-dimensional model

1. Step: Construction of 3 Evaluation corpora of 3000 syllables (on,n,cod) from original data

2. Step: Attaching randomly chosen onsets/nuclei/codas (e.g. on' to all syllables of the evaluation corpus (whereas syl, on' $\in$ Training Corpus; syl' $\notin$ Train $\cup$ Eval)

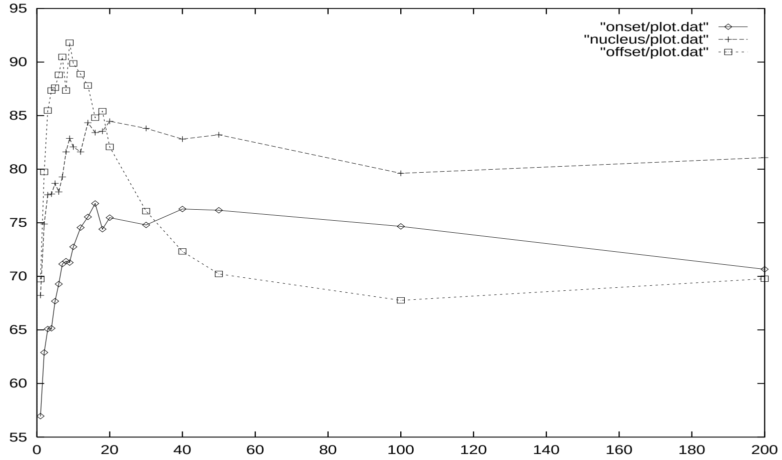
3. Step: Calculating the accuracy

4. Step: Phonological Interpretation

Evaluation - Pseudo-Disambiguation Results

Accuracy of the Onset: number of times the model decided

$$p(on, n, cod) \geq p(on', n, cod)$$



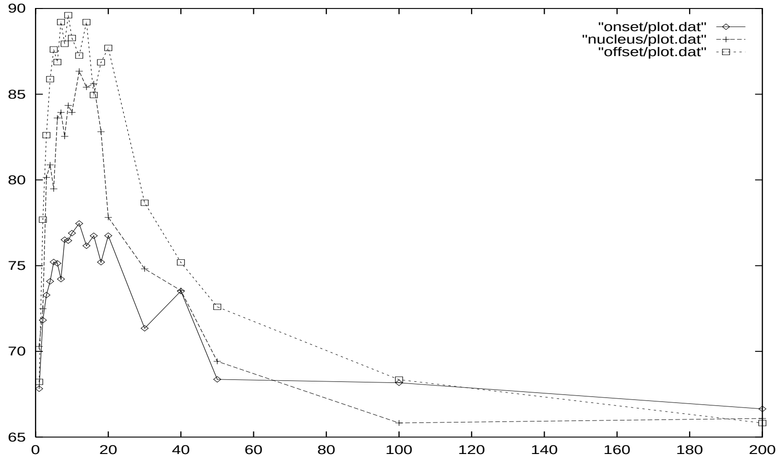


Figure: Evaluation on pseudo-disambiguation task for English (left) and German (right)

Qualitative Evaluation - English Model

- Class of High Frequency Words - Function Words

class 0 0.071				
D 0.745 NOP[@] 0.166	@ 1	NOP[@] 0.877 m 0.0792 n 0.0027	ONE 0.999	STR 1
class 17 0.023				
NOP[@] 0.973	@ 1	NOP[@] 0.325 r* 0.317	ONE 0.944 INI 0.050	STR 1

- Class of Productive word-forming suffixes

class 3 0.040				
t 0.206 s 0.106 d 0.104 NOP[I] 0.101 n 0.052	I 0.993	N 0.466 d 0.167 z 0.152 Nz 0.012	FIN 0.997	USTR 0.999

Qualitative Evaluation - English Model

class 10 0.028				
S 0.257 m 0.227 d 0.063 t 0.059 NOP[@] 0.007	@ 0.926 I 0.031 I@ 0.015 E 0.005	n 0.388 nt 0.191 nz 0.088 l 0.066 nts 0.049 ns 0.048	FIN 0.999	USTR 0.999

- Open Word class in different positions (stressed)

class 14 0.026				
m 0.116 p 0.108 k 0.090 g 0.088 t 0.080 pl 0.052 st 0.051	eI 0.426 A: 0.165 E 0.140 O: 0.110	t 0.162 s 0.131 n 0.088 d 0.079 k 0.079 nd 0.052 ts 0.037	ONE 0.696 FIN 0.276	STR 0.984

Qualitative Evaluation - German Model

- Class #7: high frequency words - function words (in, ich, ist, sind, sich, in)
- Class #26: productive prefixes (ver-, er-, zer-, vor-, her-)
- Class #34: suffixes (-lich, -ig)
- Class #4: lexical entries in stessed, word-initial syllables

Evaluation with a g2p Conversion Task

1. **Step:** Using a context-free grammar to produce all possible derivations

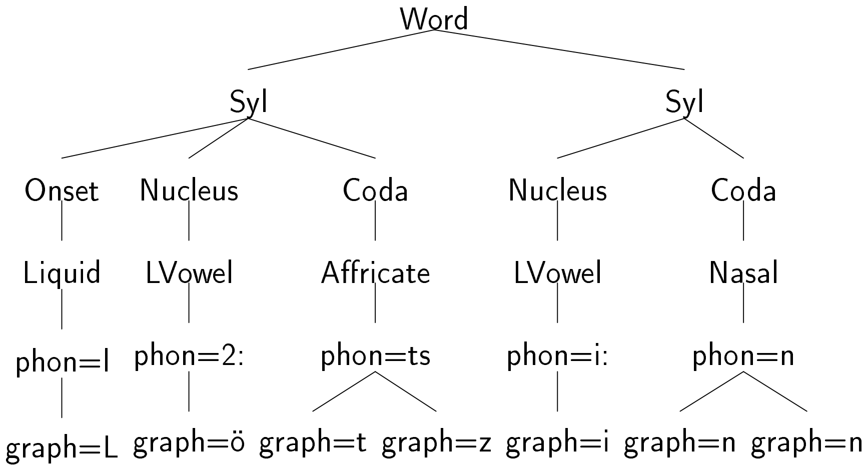
2. **Step:** Applying a probabilistic syllable model to rank the pronounciation hypotheses

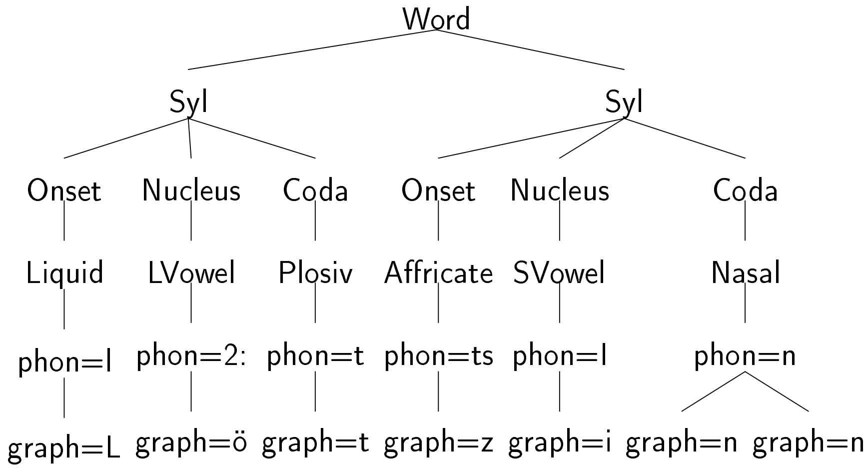
3. **Step:** Predicting pronunciation by choosing the most probable analysis

4. **Step:** Calculating the word accuracy

Evaluation with g2p

1. **Step:** extracting all possible derivations
   Example: /Lötzinn/ (tin solder) - 2 analysis out of 100





Evaluation with g2p

2. **Step:** Ranking the pronounciation hypotheses
   Key idea: product of syllable probabilities and using a hybrid model of
   (a) the empirical probability and
   (b) the model probability (empirical probability of the word is zero)

   - Probability of Wrong analysis: [l 2: ts][i: n]
     $(7.5 \cdot 10^{-7}) \cdot (3.1 \cdot 10^{-7}) = \mathbf{2.3 \cdot 10^{-13}}$
   - Probability of Correct analysis: [l 2: t][ts I n]
     $(1.5 \cdot 10^{-7}) \cdot (6.5 \cdot 10^{-6}) = \mathbf{9.8 \cdot 10^{-13}}$

Evaluation with g2p

3. **Step:** Calculating the Word accuracy
   Test set consists of 1835 unseen words.
   Out of 295.102 words not seen in the STZ, manually were eliminated:

   - Foreign words (e.g. Rugby, Sahib, Siphon, Souchong)
   - Flexional wordforms (e.g. "Astchens, Rondells)
   - Acronyms (e.g 3SAT, CD, FM, PALPlus, VIVA)
   - Proper names (e.g. Saudi, Castro, Frankfurt, Ghali, Kanther)
   - Verbs (e.g aufh"ulfst, aufluvt, felge, fiepend, kakelt)
   - Words $\geq$ 3 syllables (e.g. Ölofen, Aktenschrank, Suppenhuhn, Tageslicht)

g2p system	3-dim baseline	3-dim classes	5-dim baseline	5-dim classes
word accuracy	66.8 %	67.4 %	72.5 %	75.3 %

Discussion

- Pseudo-Disambiguation: the onset is the most important part of the syllable to disambiguate a syllable (e.g. shown for ASR by Fosler-Lussier, Greenberg, and Morgan (1999);)

- Qualitative Evaluation: in comparison with Rooth etal (1999) semantic relations were revealed between (verb/object-pairs) it cannot be expected to yield meaningful properties

- G2P: (a) Damper etal (1999) reported for the best data-driven method an accuracy of 72% (without alignment of grapheme to phonemes), however we obtained a word accuracy of 75.3%.
  (b) Bouma (2000) reported a 92.6% word accuracy for Dutch but with preprocessing and alignment with graphemes, mapping many2one graphemes.

Conclusion

- Pseudo-Disambiguation statistically supports the importance of the onset of a syllable in both languages.
- The clustering algorithm provides meaningful syllable classes.
- The clustering model increases the performance of a g2p conversion system.

Syllabaries: Frequency counts

Level W.J.M., Wheeldon L. (1994): *Do speakers have access to mental syllabary?* Phonetic encoding, consists of accessing articulatory gestural scores for each of these syllables in a "mental syllabry".

- 2 different frequencies were the base: word form frequency and lemma frequency $\Longrightarrow$ 12.000 syllable types

- Calculated: a) overall frequency of occurence
  b) frequency of occurence dependent from syllable position (first, second ...)

- Criteria:

  – low-frequency words: < 10 for word form and lemma counts

– high-frequency words:$>$ 10 for word form and lemma counts
– low-frequency syllables: $<$ 300 overall and position dependant counts
– high-frequency syllables: $>$ 300 overall and position dependant counts

VOCABULARY

• 4 groups of 16 bisyllabic Dutch words (13 nouns, 3 adjectives). They differed in the combination of word frequency and syllable frequency of their consitutent words.