Aalto University
School of Electrical
Engineering

Class n-gram models for very large vocabulary speech recognition of Finnish and Estonian

Matti Varjokallio, Mikko Kurimo and Sami Virpioja

Department of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University, Espoo, Finland
firstname.lastname@aalto.fi

12.10.2016

Introduction

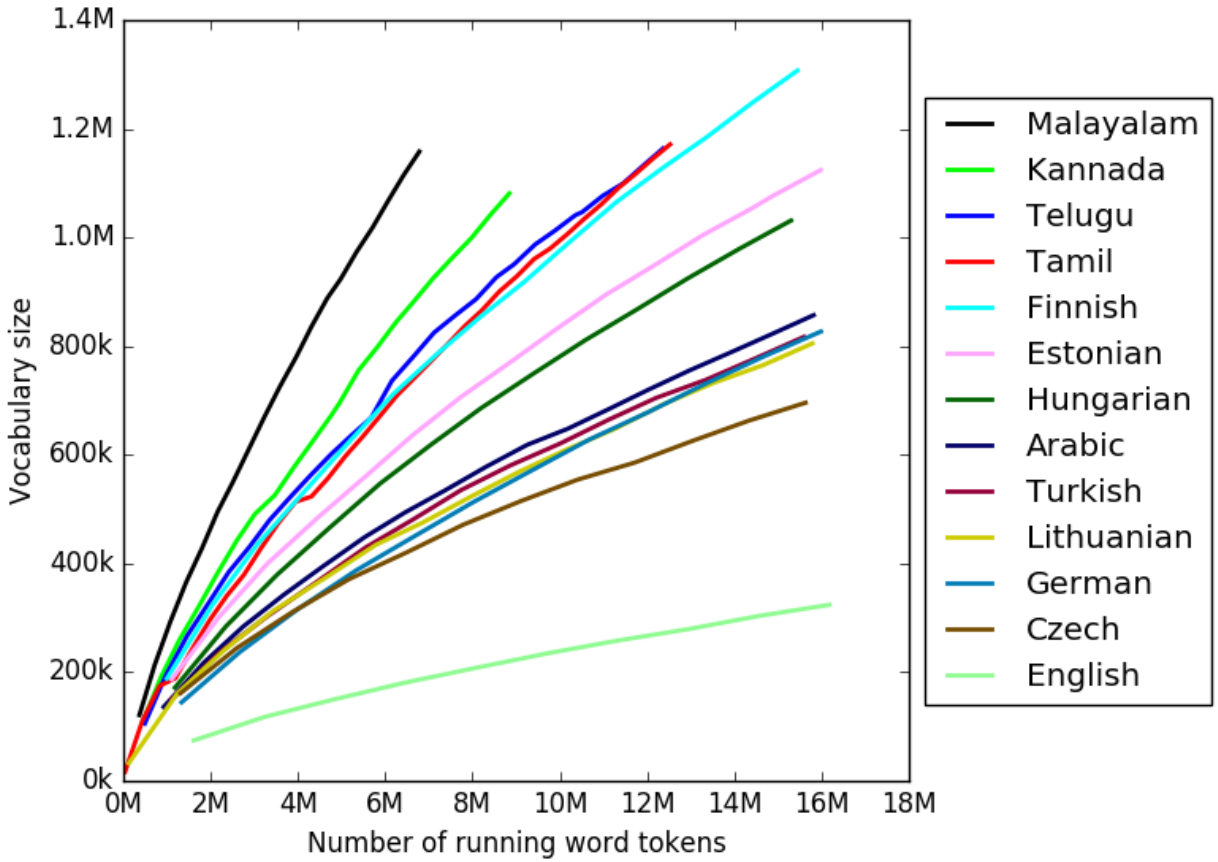
- Speech recognition of morphologically rich languages
  - N-gram models are normally used for the statistical language modelling component
  - Morphological processes like agglutination, inflection, derivation and compounding increase the vocabulary size
  - The issues with **data sparsity** and **out-of-vocabulary (OOV)** words are emphasized

Introduction

- Segmenting the training data to subwords and training the n-gram models over these units
  - Possible to opt either for unlimited vocabulary or constrained vocabulary recognition
  - Has been a common approach for languages like Finnish, Estonian, Hungarian and Turkish
- Increased type-to-token ratio likely increases the value of different word classification schemes
  - Accept some OOV rate and concentrate on improving the language model estimates
  - **This work studies class n-gram models for Finnish and Estonian very large vocabulary speech recognition**

Introduction

Vocabulary growth rates estimated from Wikipedia



Disclaimer: selected text normalization has some impact on the exact values

Research questions

- Speech recognition results compared to subword approaches
  - Perplexities are not directly comparable due to the different OOV rates
  - High requirements for the subword and word decoders
- Clustering using bigram statistics versus morphologically motivated classes
- Applying class n-gram models in the first recognition pass with a vocabulary size of several millions of words

Class n-gram models

The most common form of a class n-gram model.
Words belong only to one class.

$$P(w_i|w_{i-(n-1)}^{i-1}) = P(w_i|c_i) * P(c_i|c_{i-(n-1)}^{i-1}) \quad (1)$$

maximum likelihood estimates as given by

$$P(w|c) = \frac{f(w)}{\sum_{v \in c(w)} f(v)} \quad (2)$$

$$P(c_i|c_{i-(n-1)}^{i-1}) = \frac{f(c_{i-(n-1)}, .., c_i)}{f(c_{i-(n-1)}, .., c_{i-1})}, \quad (3)$$

Clustering via the (bigram) Exchange algorithm

(Kneser and Ney, 1991)

Algorithm 1: Exchange algorithm

```
1  compute initial class mapping
2  sum initial class based counts
3  compute initial perplexity
4  repeat
5      foreach word w of the vocabulary do
6          remove word from its class
7          foreach class k do
8              tentatively move word w to class k
9              compute perplexity for this exchange
10         move word w to class k with minimum perplexity
11 until stopping criterion is met;
```

Morphologically motivated classes (1/4)

- Omorfi, open source morphological analyzer for Finnish (Pirinen, 2015)
  - Fairly high coverage (~85%)
  - Detailed analysis
- Still challenges
  - Morphological disambiguation
  - Need to tag the remaining words not covered by the analyzer, otherwise will increase WER
  - Resulting classes are very inequal in size

```
taloitta N Abe Pl   "without houses"
autoitta N Abe Pl   "without cars"
```

Morphologically motivated classes (2/4)

Approach 1:

- FinnPOS, open source conditional random field (CRF) -based tagger (Silfverberg et al, 2015)
- High morphological tagging accuracy
- Strictly a class n-gram model
  - Separate word instances for each different analysis of a surface word

Morphologically motivated classes (3/4)

Approach 2:

- Category n-gram generalizes the class n-gram with multiple category memberships per word (Niesler and Woodland, 1999)
  - Model the morphological disambiguation with alternate categories
  - Does not increase the vocabulary size

Morphologically motivated classes (4/4)

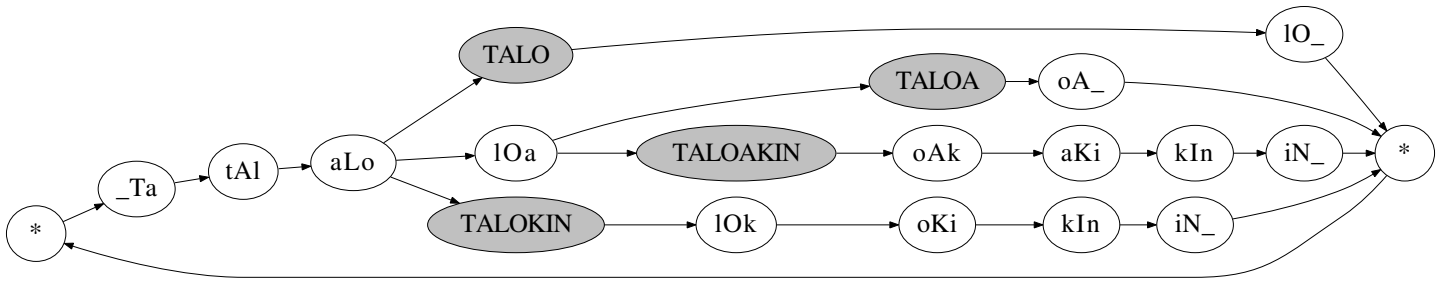
$$P(w_i|w_{i-n-1}^{i-1}) = \sum_j \left(P(w_i|c_{ji}) * \sum_s \left(P(c_{ji}|s) * \prod_{k=i-n-1}^{i-1} P(c_{sk}|w_k) \right) \right)$$

j is a possible category for the word w_i
s category sequences generated by the word history w_{i-n-1}^{i-1}

- Three types of parameters: category generation, category memberships and the n-gram term
- Expectation-maximization training
  - Tagging in the later training iterations
  - In the later iterations can limit the number of memberships to 1 per word with a very small loss in perplexity
- Further refining of classes by merging and splitting and running the exchange algorithm

Decoding with class n-grams (1/2)

- Dynamic token-passing decoding
  - Vocabulary is encoded with a lexical prefix tree
  - Tokens representing a single search hypothesis are propagated in the graph
  - Language model scores are updated dynamically to tokens
  - Search techniques: beam pruning, hypothesis recombination, language model lookahead, cross-word modelling etc.



Decoding with class n-grams (2/2)

- **Hypothesis recombination**: tokens in the same graph node and the same n-gram state may be merged
- Class n-gram model has discrete states similarly to the standard n-gram model
- Hypothesis recombination may be extended to work on n-gram state and class n-gram state tuples
  - Possible to use different context lengths

Language model perplexities

Table: Perplexities for the language models. Both word and subword model perplexities are normalized per word, but they are not directly comparable due to different OOV rates. Model combination is done via linear interpolation.

	Finnish				Estonian			
Model	Vocabulary	OOV	Model size	PPL	Vocabulary	OOV	Model size	PPL
Subword n-gram	Unlimited	-	84.2M	2071	Unlimited	-	60.7M	627
Subword n-gram	Unlimited	-	110.0M	1912	Unlimited	-	97.6M	532
Class 5-gram	2.4M	2.3%	32.3M	2194	1.6M	0.9%	34.3M	883
Word 3-gram	2.4M	2.3%	85.5M	1341	1.6M	0.9%	59.3M	285
Word + class	2.4M	2.3%	117.8M	1056	1.6M	0.9%	93.6M	249

Speech recognition results

Table: Word error rates in a broadcast news speech recognition task. Model combination is done via linear interpolation.

	Finnish				Estonian			
Model	Vocabulary	OOV	Model size	WER	Vocabulary	OOV	Model size	WER
Subword n-gram	Unlimited	-	84.2M	30.0%	Unlimited	-	60.7M	15.1%
Subword n-gram	Unlimited	-	110.0M	29.8%	Unlimited	-	97.6M	15.0%
Class 5-gram	2.4M	2.8%	32.3M	30.8%	1.6M	1.2%	34.3M	16.8%
Word 3-gram	2.4M	2.8%	85.5M	30.9%	1.6M	1.2%	59.3M	16.2%
Word + class	2.4M	2.8%	117.8M	29.2%	1.6M	1.2%	93.6M	15.0%

Conclusions

- Exchange algorithm with bigram statistics worked well for word clustering
  - Morphologically motivated classes were also evaluated but did not improve the results as much
- Perplexity reductions were 21.3% relative for Finnish and 12.8% relative for Estonian
- Word error rate improved by 5.5% relative for Finnish and 7.4% relative for Estonian compared to word n-gram baseline
- Compared to an unlimited vocabulary subword recognizer, 2.2% relative improvement for Finnish and equal result for Estonian