Multilingual and Low-Resource Speech Recognition

Peter Bell

Automatic Speech Recognition – ASR Lecture 15
8 March 2021

Languages of the World

- Over 6,000 languages globally....
- In Europe alone
  - 24 official languages and 5 “semi-official” languages
  - Over 100 further regional/minority languages
  - If we rank the 50 most used languages in Europe, then there are over 50 million speakers of languages 26-50 (Finnish – Montenegrin)
- 3,000 of the world’s languages are endangered
- Google cloud speech API covers over 98 languages and more than 300 accents/dialects of those languages; Apple Siri covers over 21 languages; Google assistant has over 30

Under-resourced languages

Under-resourced (or low-resourced) languages have some or all of the following characteristics

- limited web presence
- lack of linguistic expertise
- lack of digital resources: acoustic and text corpora, pronunciation lexica, ...

Under-resourced languages thus provide a challenge for speech technology

See Besaciera et al (2014) for more

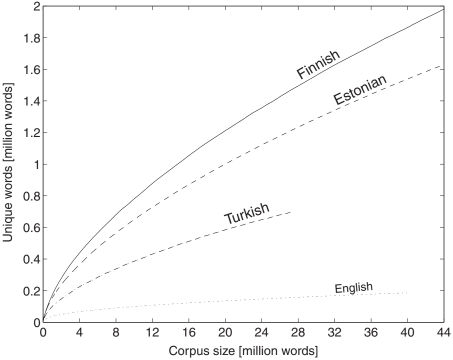
Speech recognition of under-resourced languages

- Training acoustic and language models with limited training data
- Transferring knowledge between languages
- Constructing pronunciation lexica
- Dealing with language specific characteristics (e.g. morphology)

Morphology

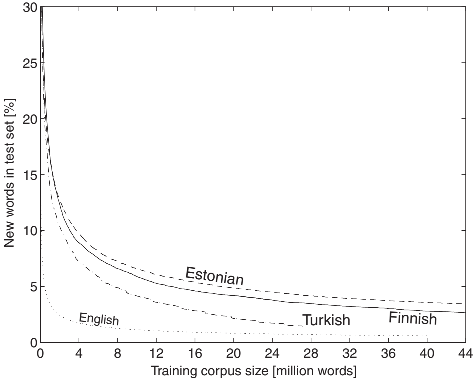
- Many languages are morphologically richer than English: this has a major effect of vocabulary construction and language modelling
- Compounding (eg German): decompose compund words into constituent parts, and carry out pronunciation and language modelling on the decomposed parts
- Highly inflected languages (eg Arabic, Slavic languages): specific components for modelling inflection (eg factored language models)
- Inflecting and compounding languages (eg Finnish)
- All approaches aim to reduce ASR errors by reducing the OOV rate through modelling at the morph level; also addresses data sparsity

Vocabulary size for different languages



Creutz et al (2007)

OOV Rate for different languages



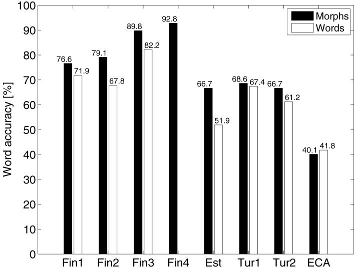
Creutz et al (2007)

Segmenting into morphs

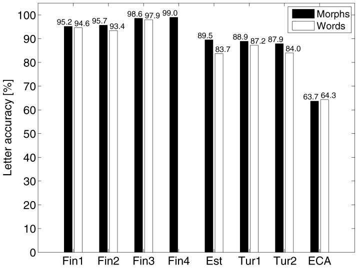
- Linguistic rule-based approaches – require a lot of work for an under-resourced language!
- Automatic approaches – use automatically segment and cluster words into their constitutent morphs
- Morfessor (http://www.cis.hut.fi/projects/morpho/)
    - "Morfessor is an unsupervised data-driven method for the segmentation of words into morpheme-like units."
    - Aims to identify frequently occurring substrings of letters within either a word list (type-based) or a corpus of text (token-based)
    - Uses a probabilistic framework to balance between few, short morphs and many, longer morphs
- Morph-based language modelling uses morphs instead of words – may require longer context (since multiple morphs correspond to one word)

Morph-based vs Word-based ASR

Speech recognition accuracies on Finnish (Fin1-Fin4), Estonian (Est), Turkish (Tur), and Egyptian Arabic (ECA), using morph- and word-based n-gram language models.



word accuracies



letter accuracies

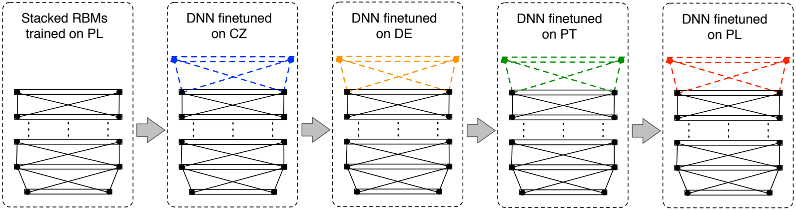
Creutz et al (2007)

Multilingual and cross-lingual acoustic models

How to share information from acoustic models in different languages?

- General principle – use neural network hidden layers to learn a multilingual representation
- Hidden layers are multilingual – shared between languages
- Output layer are often monolingual language specific
- **Multi-lingual phone sets** use a network with multilingual hidden representations directly in a hybrid DNN/HMM systems
- **Hat-swap/multi-task** train a network with an output layer for each language, but shared hidden layers
- **Multilingual bottleneck** use a bottleneck hidden layer (trained in a multilingual) way as features for either a GMM- or NN-based system

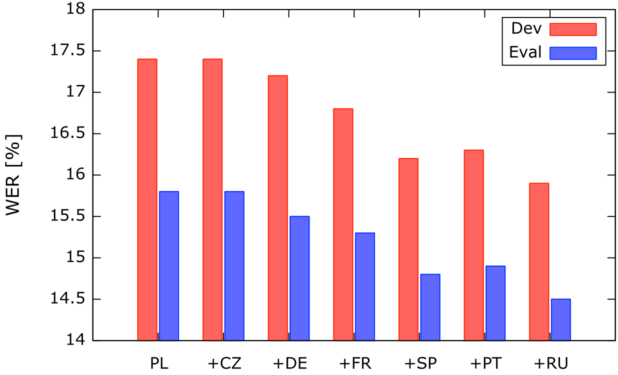
Hat Swap – architecture



Ghoshal et al, 2013

Hat Swap – experiment

Recognition of GlobalPhone Polish

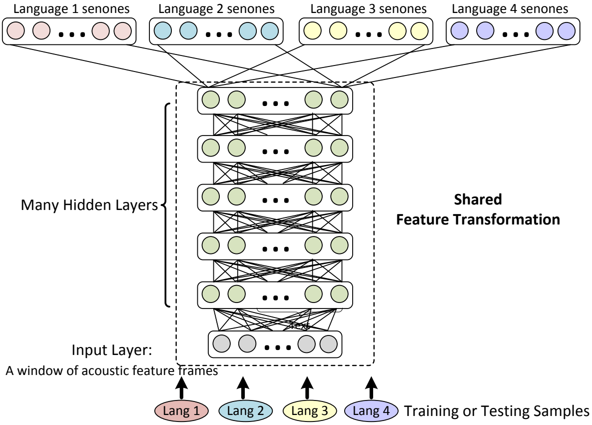


Ghoshal et al, 2013

Multi-lingual networks ("block softmax")

- Train one network for all languages:
  - separate output layer for each language
  - shared hidden layers
- Each training input is propagated forward to the output layer of the corresponding language – only that output layer is used to compute the error used to train the network for that input
- Since the hidden layers are shared, they must learn features relevant to all the output layers (languages)
- Can view this as a parallel version of hat swap

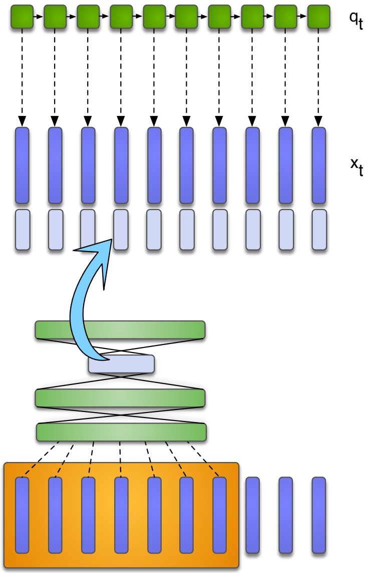
Multi-lingual networks – architecture



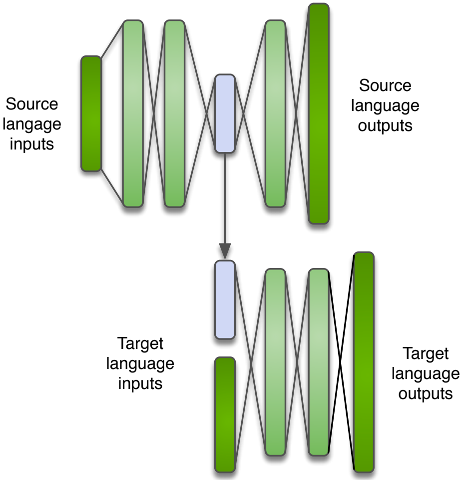
Huang et al, 2013

NB: A senone is a context-dependent tied state

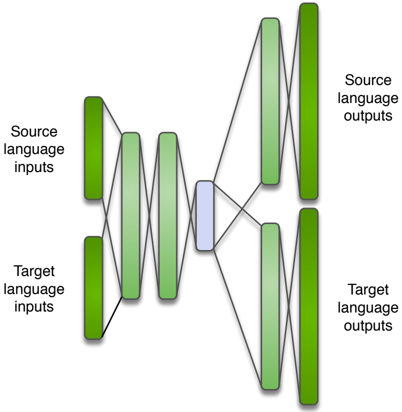
Bottleneck features



Cross-lingual bottleneck features



Multi-lingual bottleneck network



Multilingual bottleneck features – experiments

GMM-based acoustic models

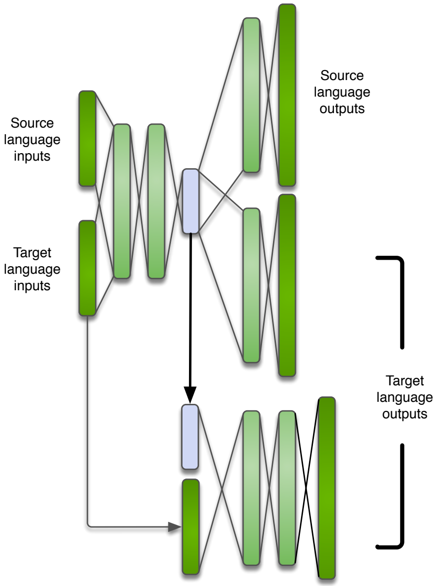
WER [%]		MFCC	MFCC+BN Bottleneck trained on GER	ENU	FRA
Test language	GER	29.97	*27.50* (8.2)	29.63 (1.1)	30.38 (-1.4)
	ENU	21.69	21.31 (1.8)	*18.85* (13.1)	22.63 (-4.3)
	FRA	37.78	37.76 (0.1)	38.72 (-2.5)	*33.95* (10.1)

WER [%]		MFCC	MFCC+BN BN trained on			
Test language	GER	34.58	GER 33.39 (3.4)	**ENU** 34.07 (1.5)	**ENU** + **FRA** 32.74 (5.3)	GER + **ENU+FRA** 31.72 (8.3)
	ENU	26.14	ENU 23.54 (9.9)	**GER** 24.81 (5.1)	**GER** + **FRA** 23.68 (9.4)	**GER** + ENU+**FRA** 21.79 (16.6)
	FRA	43.52	FRA 40.51 (6.9)	**GER** 43.65 (-0.3)	**GER** + **ENU** 41.96 (3.6)	**GER** + **ENU**+FRA 39.98 (8.1)

(Mismatched acoustic environment)

Tüske et al, 2013

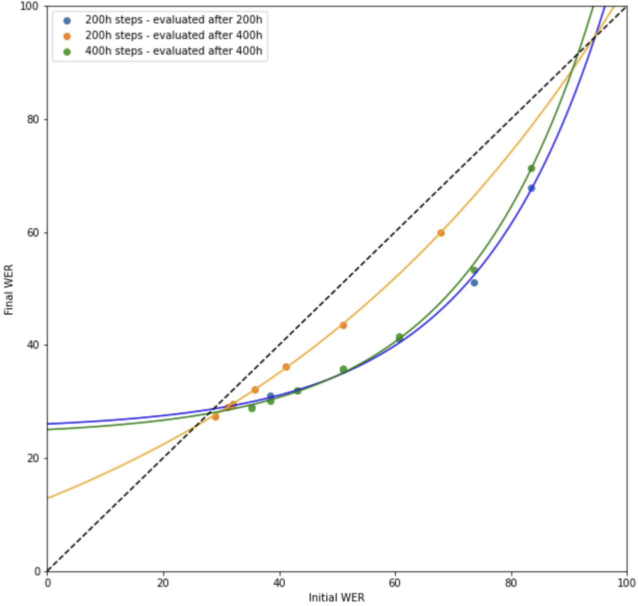
Use of BN features in HMM/DNN systems



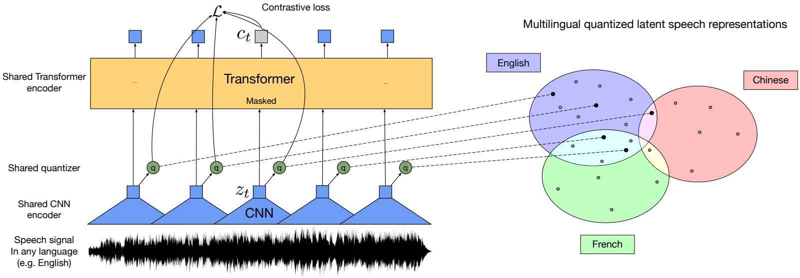
Semi-supervised training

- Assume we only have a only a small amount of data is transcribed, but much more untranscribed data $\rightarrow$ train a seed model and use it to transcribe more data
- But don't want to train further on incorrect captions
- Traditional solution: apply data filtering based on confidence scores
- This can select out the harder data that is most useful for refining the system
- Solution (Manohar, 2018): use a lattice to incorporate uncertainty about the transcription, train with LF-MMI criterion

Example: Tagalog



Self-supervised approaches



Conneau et al, 2020

Graphemes and phonemes

- Can represent pronunciations as a sequence of graphemes (letters) rather than a sequence of phones
- Advantages of grapheme-based pronunciations
  - No need to construct/generate phone-based pronunciations
  - Can use unicode attributes to assist in decision tree construction
- Disadvantages: not always direct link between graphemes and sounds (eg. in English)

Grapheme-based ASR results for 6 low-resource languages

Language	ID	System	WER (%) tg	WER (%) +cn	WER (%) cnc
Kurmanji Kurdish	205	Phonetic	67.6	65.8	64.1
		Graphemic	67.0	65.3	
Tok Pisin	207	Phonetic	41.8	40.6	39.4
		Graphemic	42.1	41.1	
Cebuano	301	Phonetic	55.5	54.0	52.6
		Graphemic	55.5	54.2	
Kazakh	302	Phonetic	54.9	53.5	51.5
		Graphemic	54.0	52.7	
Telugu	303	Phonetic	70.6	69.1	67.5
		Graphemic	70.9	69.5	
Lithuanian	304	Phonetic	51.5	50.2	48.3
		Graphemic	50.9	49.5	

IARPA Babel, 40h acoustic training data per language, monolingual training; cnc is confusion network combination, combining the grapheme- and phone-based systems
Gales et al (2015)

Speech recognition systems for low-resource languages

- Morph-based language modeling
- Transferring data between acoustic models based on multilingual hidden representations
- Grapheme-based pronunciation lexica

Speech recognition systems for low-resource languages

- Morph-based language modeling
- Transferring data between acoustic models based on multilingual hidden representations
- Grapheme-based pronunciation lexica

In the future:

- "Zero-resource" ASR (no transcribed data at all)
- Languages without written forms
- Much active research in this area (including at Edinburgh)

Reading (1)

- L Besaciera et al (2014). "Automatic speech recognition for under-resourced languages: A survey", Speech Communication, 56:85–100. http://www.sciencedirect.com/science/article/pii/S0167639313000988
- Z Tüske et al (2013). "Investigation on cross- and multilingual MLP features under matched and mismatched acoustical conditions", ICASSP. http://ieeexplore.ieee.org/abstract/document/6639090/
- A Ghoshal et al (2013). "Multilingual training of deep neural networks", ICASSP-2013. http://ieeexplore.ieee.org/abstract/document/6639084/
- J-T Huang et al (2013). "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers", ICASSP. http://ieeexplore.ieee.org/abstract/document/6639081/.
- M Gales et al (2015). "Unicode-based graphemic systems for limited resource languages", ICASSP. http://ieeexplore.ieee.org/document/7178960/

Reading (2)

- M Creutz et al (2007). "Morph-based speech recognition and modeling OOV words across languages", *ACM Trans Speech and Language Processing*, 5(1). http://doi.acm.org/10.1145/1322391.1322394 A. Conneau, et al (2020). "Unsupervised cross-lingual representation learning for speech recognition", arXiv:2006.13979. https://arxiv.org/abs/2006.13979
- V. Manohar, et al. (2018) "Semi-supervised training of acoustic models using lattice-free MMI". In Proc. IECC ICASSP (pp. 4844-4848). https://ieeexplore.ieee.org/abstract/document/8462331