Case study: transcribing TED Talks

Peter Bell

Automatic Speech Recognition— ASR Lecture 16
21 March 2013

Putting it all togther

Designing a complete speech recognition system...

Real system	made by CSTR at Edinburgh and NICT in Japan
Real task	transcribing lectures from TED
Real results	compared to other research groups around the world

** This lecture is non-examinable**

TED lectures

The system is designed for automatic transcription of TED talks. (We also perform machine translation on the output).

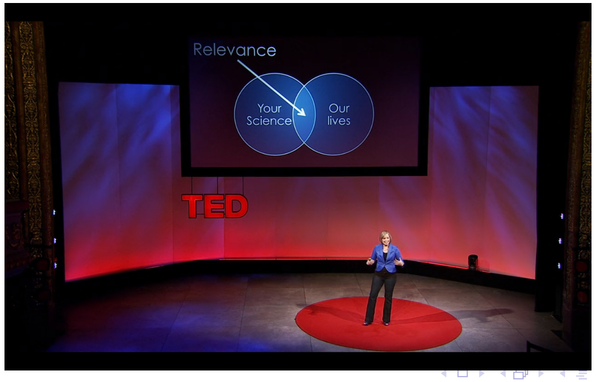
TED is a nonprofit devoted to Ideas Worth Spreading. It started out (in 1984) as a conference bringing together people from three worlds: **Technology, Entertainment, Design.** Since then its scope has become ever broader. Along with two annual conferences -- the TED Conference on the West Coast each spring, and the TEDGlobal conference in Edinburgh UK each summer -- TED includes the award-winning TED Talks video site, the Open Translation Project and TED Conversations, the inspiring TED Fellows and TEDx programs, and the annual TED Prize.

The IWSLT evaluation

- The annual IWSLT brings together researchers in speech recognition and machine translation
- In recent years, it has set a series of challenges for researchers, based on lectures from TED.
- CSTR entered this competition last year. This year, we made a combined system with NICT, who won in 2012.

The TED lecture task

The TED lecture "ASR task" is defined for the IWSLT evaluation campaign. The task consists of several development/test sets, each containing 8-11 single-speaker talks of around 10 minutes' duration. All talks are pre-segmented into utterances.



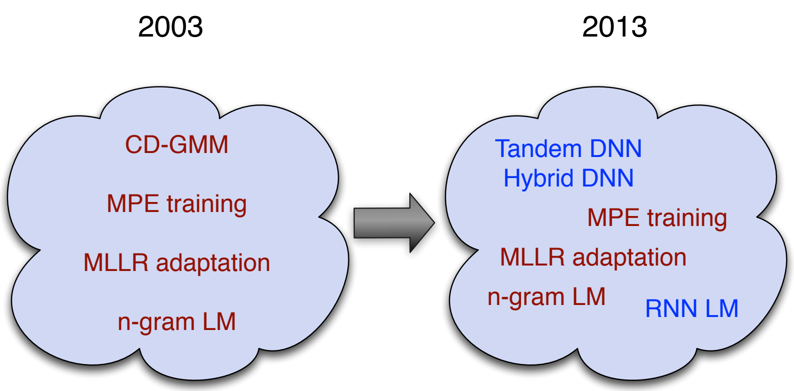
Characteristics of the task

- Generally clear, planned speech directed at an audience
- Single speaker per talk
- Training data is readily available on the web
- A wide vocabulary is used

The key components of the system

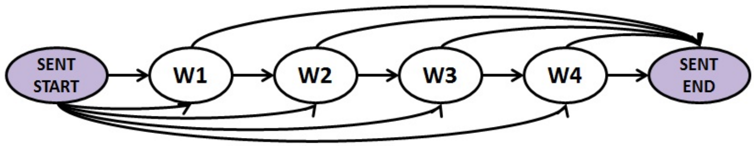
- Acoustic models based on **deep neural networks**
- Domain adaptation with **multi-level** networks
- **Recurrent neural network** language models

Advances in ASR



Preparing the training data

- 813 talks downloaded from the TED website, along with transcriptions made by the online community.
- We need to automatically align the text to the speech.
- Use a **Viterbi alignment** with the option to start and end at any word.



- After stripping out the silence, we retrieve 143 hours of speech data available for model training.

Basic acoustic models

First, need a standard cross-word triphone HMM-GMM system:

- PLP acoustic features with first, second and third deltas (52 dimensions), rojected to 39 dimensions with an HLDA transform
- The acoustic features are normalised for mean and variance across the talks
- Standard left-to-right HMM topology, with three emitting states per phone
- 10,000 tied HMM-states, modelling clustered cross-word triphones
- 190,000 Gaussians in total

Tandem vs Hybrid systems: recap

Tandem:

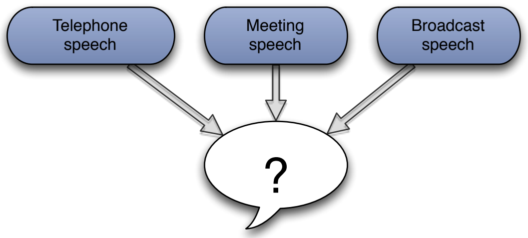
- Neural networks are used to derive features for training data, which are augmented with the standard acoustic features and used to train GMMs.
- Can use decorrelated posterior features (eg over monophones), or bottleneck features

Hybrid:

- Neural networks used to generated posterior probabilities over states, used as likelihoods in decoder, scaled by state priors.
- In modern systems, we model the probabilities of tied triphone states.

We always use **deep** neural networks with RBM pre-training.

Domain adaptation



- If we are starting out on a new speech recognition task, it may be helpful to use data that we already have from other domains.
- But speech varies a lot in style, accent, and environmental conditions – how can we use out-of-domain data without harming performance?

MAP adaptation

We can start with a model trained on another domain and adapt it to the new domain using MAP (See lecture 10):

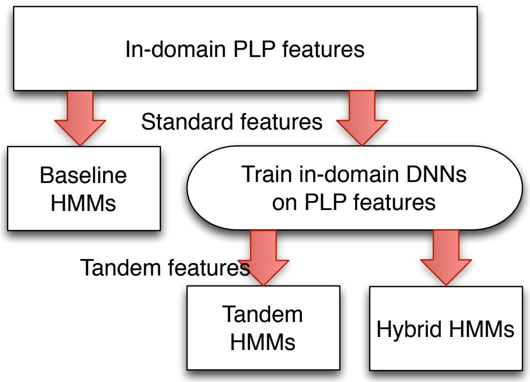
$$\hat{\mu}_j = \frac{\tau \mu_j^{\text{OD}} + \sum_t \gamma_j(t) x_t}{\tau + \sum_t \gamma_j(t)}$$

This usually works better for domain adaptation (where we are building a completely new system) than for speaker adaptation, because there is likely to be wider phonetic coverage.

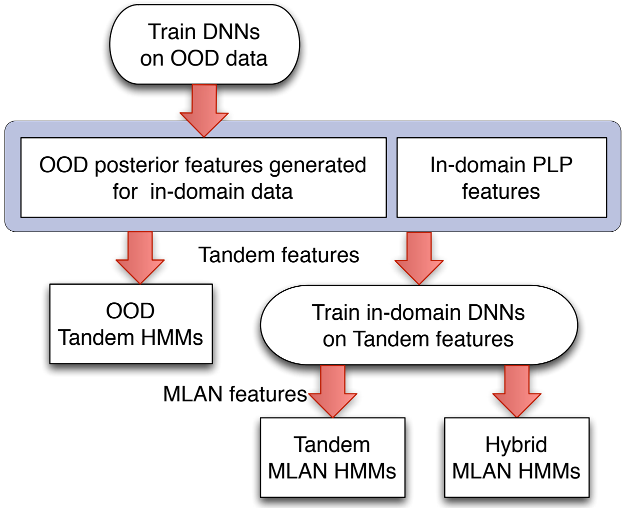
Domain adaptation with DNNs

- Features derived from neural networks are known to provide a degree of domain-independence.
- Features trained on one domain may be used add discriminative ability to another domain (perhaps one where data is limited)
  - can carry out additional training iterations...
  - ... but pre-training already provides a degree of regularisation when training data is limited
  - in tandem framework, retrain GMMs
- Multi-level adaptive networks (MLAN): use a second DNN to discriminatively select which OOD features are most effective in the new domain.

Standard scheme



MLAN scheme



Feature space adaptation

How to enable the hybrid system to benefit from speaker adaptation? (Particularly important in the when there is a lot of data for each speaker)

- The simplest method is to perform feature-space adaptation on the input acoustic features
- Estimate a single CMLLR transform for each speaker using the baseline GMMs.
- Retrain the hybrid DNNs on the speaker-normalised feature space.

Language models

In summary:

- Decoding is performed with a **trigram** model
- Lattices are rescored with a **4-gram** model
- We later score complete sentences with a **recurrent neural network** model
- Models are trained on the transcriptions of TED talks, with selected out-of-domain data

Training data

The best language model was trained at NICT in Japan, using:

	Corpus	Tokens
In-domain	TED Talks	2.4M
Out-of-domain	News Commentary v7	4.5M
	English Gigaword 5th ed.	2.7G

Language model adapation

- There is a relatively small amount of in-domain data. We need to select out-of-domain text that matches the TED domain.
- Use **cross-entropy difference** metric which biases towards sentences that are both like the in-domain corpus D_I and unlike the average of the out-of-domain:

$$D_s = \{s \mid H_I(s) - H_O(s) < \tau\}, \ s \in D_O$$

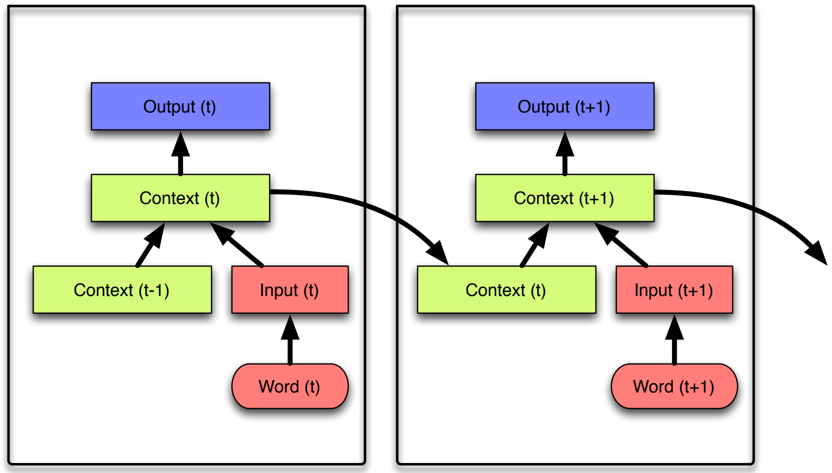
- The threshold τ is empirically set to minimize the perplexity of the development set

The final n-gram language model

- Back-off model using modified Kneser-Ney smoothing
- Some statistics:

Unigrams	**55,000**
Bigrams	**24,000,000**
Trigrams	**129,000,000**
4-grams	**43,000,000**

Recurrent neural network language models



Rescoring *n*-best lists

- We can process the output lattices of the recogniser to obtain *n*-best lists.
- With *n*-best lists, we can compute LM probabilities over whole sentences.
- This allows us to use language models which are not finite-state.

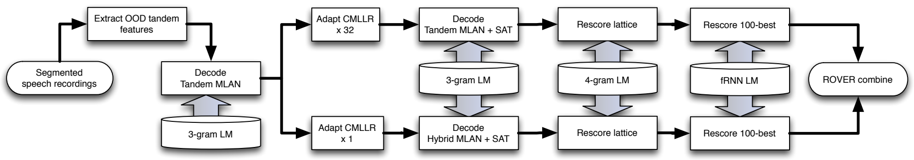
```
-6796.31 -20.6985 8 <s> A BIT TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6787.48 -20.9777 8 <s> BUT WE TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6850.2  -19.0512 8 <s> A BIT TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6841.38 -19.3304 8 <s> BUT WE TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6787.89 -21.4919 8 <s> THEY'LL BE TOO MUCH RAIN NOT ENOUGH FRAME </s>
-6841.79 -19.8447 8 <s> THEY'LL BE TOO MUCH RAIN NOT ENOUGH RAIN </s>
-6729.71 -23.5939 8 <s> A BE TOO MUCH RAIN NOT ENOUGH FRAME </s>
```


System combination

- If there are multiple systems producing different output, they may be making different mistakes.
- Try to pick the best output from all the systems...
- **ROVER**: Recogniser output voting error reduction
- Combines voting with measures of which system is most confident about each word.
- Also: confusion network decoding is a "cheat" to minimise the expected WER

The complete system

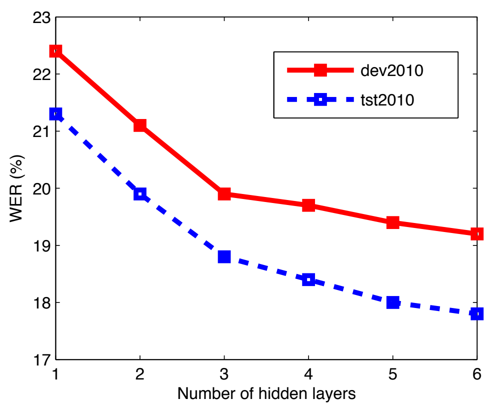
- We run a first-pass recognition which is used to estimate speaker adaptation transforms
- Then we run the complete decoding process with both Tandem and Hybrid models using MLAN features
- Finally, the systems are combined using ROVER



Tandem and hybrid adapted systems on TED

System	Dev WER (%)
PLP	31.7
+ SAT	25.3
+ MPE	20.3
Baseline tandem	23.3
+ SAT	19.7
+ MPE	17.9
Baseline hybrid	20.3
+ SAT	17.6
Tandem MLAN	20.6
+ SAT	18.1
+ MPE	**16.4**
Hybrid MLAN	17.8
+ SAT	**16.4**

Increasing the number of layers in the DNN



The effect of increasing the number of DNN layers for the hybrid MLAN systems on TED lectures (systems without SAT)

Language model experiments

System	Dev WER(%)
Tandem MLAN	14.4
+ 4gram	13.8
+ fRNN	12.8
Hybrid MLAN	14.4
+ 4gram	13.5
+ fRNN	12.7
ROVER combination	13.4
+ 4gram	12.7
+ fRNN	11.9
+ tuning	**11.7**

Results of 2012 evaluation

System	tst2011	tst2012
FBK	15.4	16.8
RWTH	13.4	13.6
UEDIN	12.4	14.4
KIT-NAIST	12.0	12.4
MITLL	11.1	12.4
NICT	10.9	12.1

NICT-UEDIN systems	tst2011	tst2012
Tandem MLAN + fRNN	10.2	11.4
Hybrid MLAN + fRNN	10.3	11.3
ROVER combination	**9.3**	**10.3**

Some conclusions

- The capabilities of speech recognition have improved a lot in the past few years
- Advances in neural networks and increases in computing power have helped a lot
- Now is an exciting time to be involved in speech recognition research!

References

- M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, “Overview of the IWSLT 2012 evaluation campaign,” in *Proc. of the 9th International Workshop on Spoken Language Translation*, 2012.
- H. Yamamoto, Y. Wu, C.-L. Huang, X. Lu, P. Dixon, S. Matsuda, C. Hori, and H. Kashioka, “The NICT ASR system for the IWSLT2012,” in *Proc. IWSLT*, 2012.
- E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, “The UEDIN systems for the IWSLT 2012 evaluation,” in *Proc. IWSLT*, 2012.
- P. Bell, P. Swietojanski, and S. Renals, “Multi-level adaptive networks in tandem and hybrid ASR systems,” in *Proc. ICASSP*, 2013, to appear.
- T. Mikolov, M. Karafiát, L. Burget, J. Černokcý, and S. Khudanpur, “Recurrent neural network based language model,” in *Proc. Interspeech*, 2010.