# Weakly supervised training

Peter Bell

Automatic Speech Recognition – ASR Lecture 15
9 March 2026

# What is weakly supervised training?

- Usually in ASR, we assume that each training utterance has a reliable transcription $Y$, so we can train a model to maximise $P(Y|X)$ or $(PX, Y)$
- In many practical situations this isn't the case
- We'd like to use speech data where we have imperfect knowledge of the words spoken – **weakly supervised** training
- In this lecture we'll look at two cases: *lightly supervised* and *semi-supervised* training

# Lightly supervised training

- We don't have perfect labels for each training sample, but we do have some information about what was said
- The main challenge is to find reliable $(X, Y)$ pairs
- We might sometimes have lots of text and a small amount of speech, or the other way round
- The text might differ from a verbatim transcription, sometimes in a predictable way

# Example use cases

- Audio books
- Captioned broadcast data
- Pre-prepared script material
- Cleaned non-verbatim transcriptions (eg. parliamentary transcriptions)

** Be very careful if the text has been obtained from another automatic system **

# Lightly supervised training

A standard method [Braunschweiler et al]:

1. Train an *biased* language model on the captions, interpolated with a background LM

$$p(w_i|h_i) = \lambda p_{bias}(w_i|h_i t) + (1 - \lambda) p_{bg}(w_i|h_i)$$

2. Decode the training data with a pre-existing acoustic model, and the biased LM
3. Align the captions with the ASR output
4. Select utterances where there is a good match between the captions and the automatic output

# Using broadcast captions

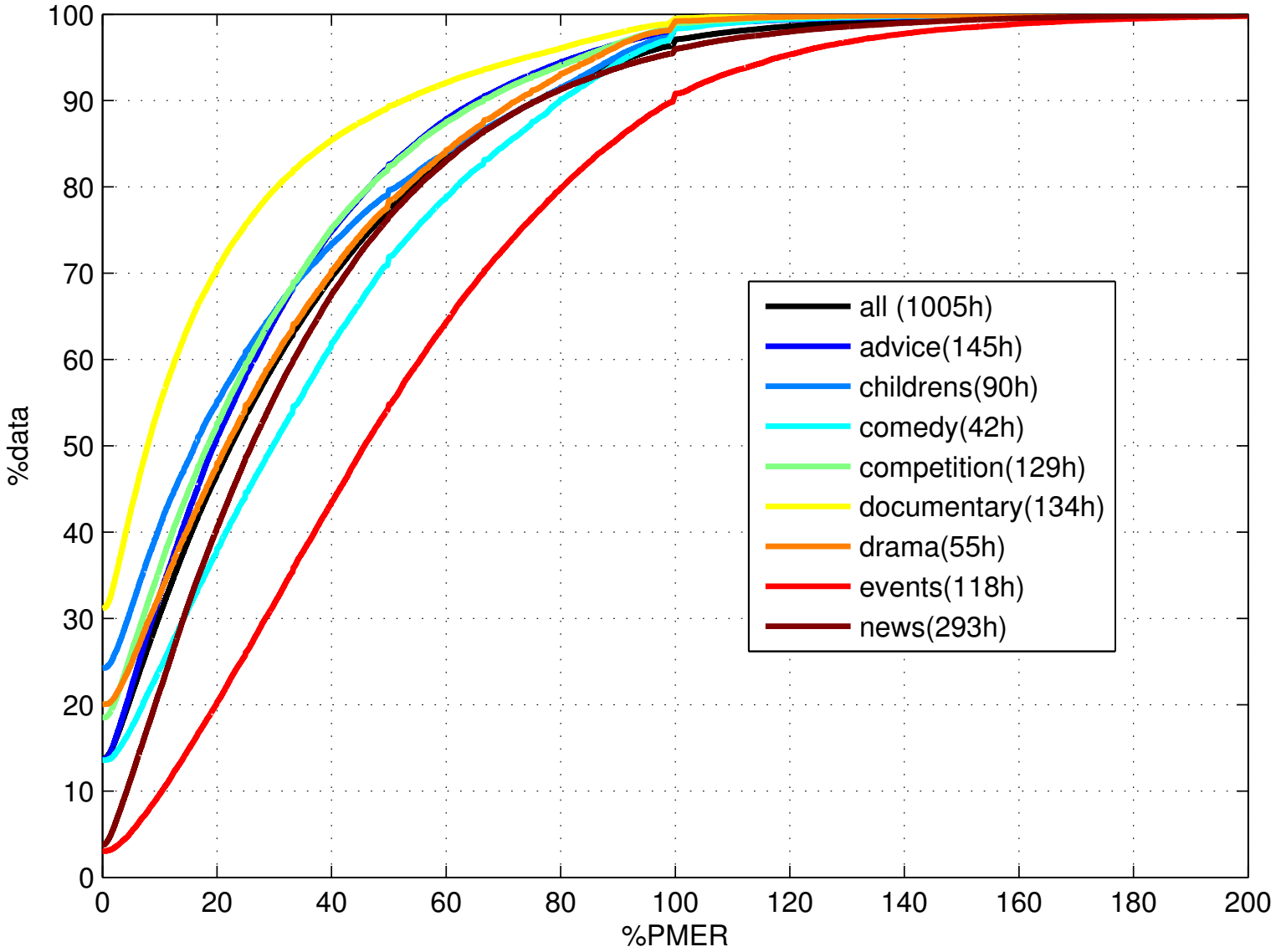
Problems with using closed captions as training data labels:

- Timings may not be accurate
- Not all words spoken are captioned
- Words may appear in the captions that were never actually spoken
- Limited speaker information is available (in the form of colour changes in the subtitles)

```
he loves your ******** ** PICTURE he thinks ****** YOU'LL do ********** well in milan
```

```
he loves your PICTURES SO MUCH    he thinks YOU'RE GONNA  do INCREDIBLY well in milan
```

# Data selection

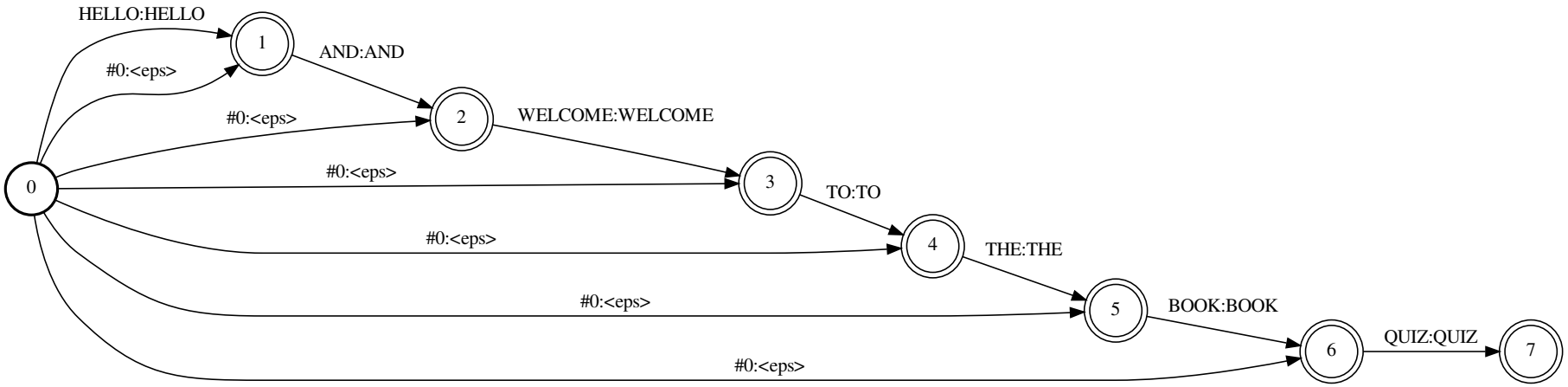


# An alternative alignment method

- The biased LM approach is quite computationally costly, and can lead to bias towards data that we can already recognise well
- We have used an alternative approach based on constructing weighted finite state transducers for each utterance
- This allows us to use much stronger constraints – based on the captions – at decoding time

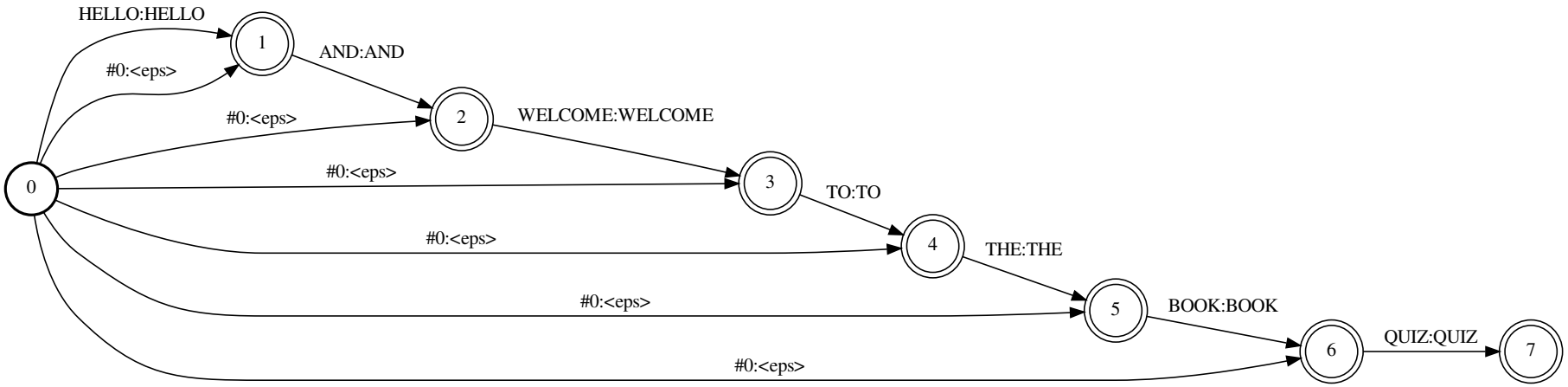
# Alignment with WFSTs

A $G$ transducer that allows any substring of the original captions – known as a *factor transducer*



# Alignment with WFSTs

A $G$ transducer that allows any substring of the original captions – known as a *factor transducer*

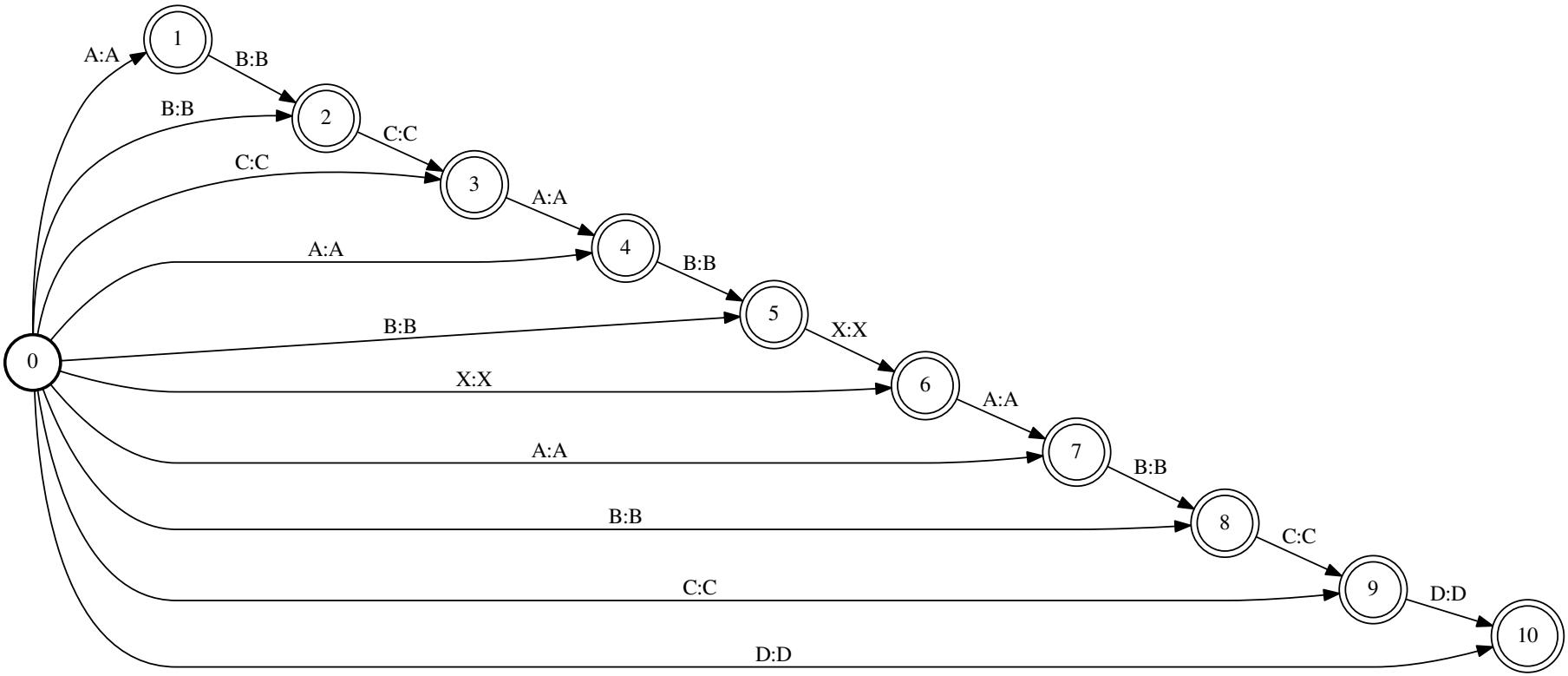



```
welcome to the ???? ???? and in the final
```

```
hello and welcome to the book quiz and in the final series in this contest
```

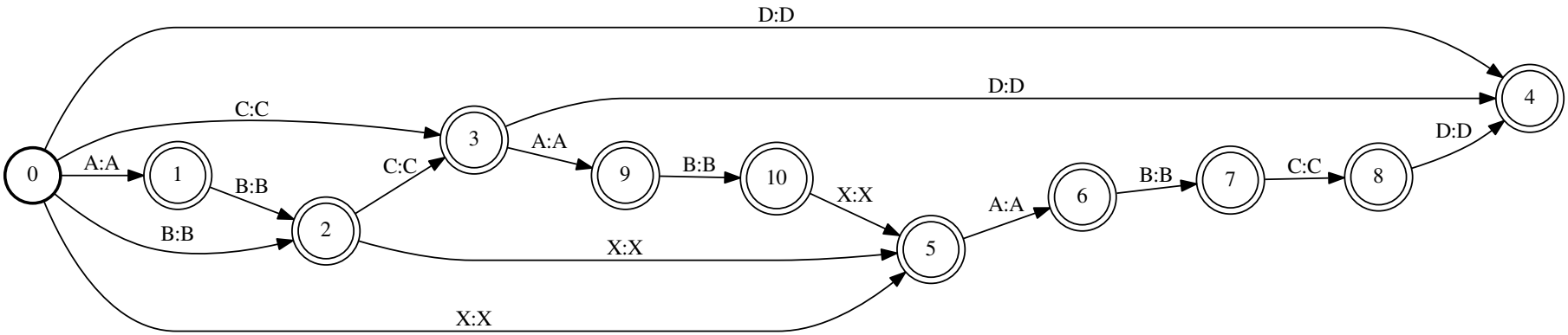
# Factor transducer

A B C A B X A B C D



# Determinised version

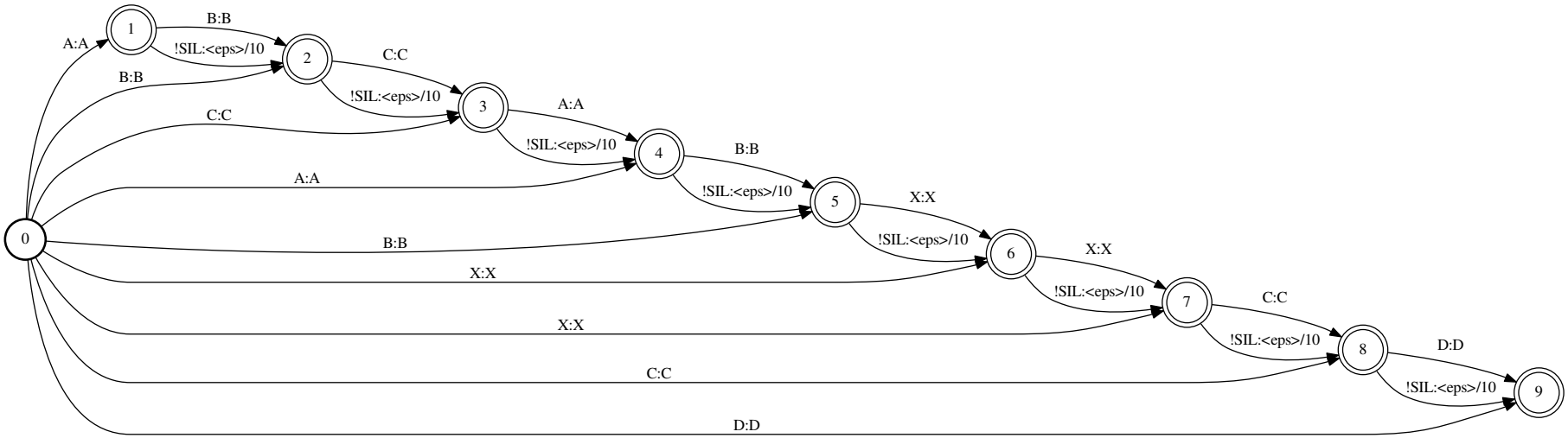
A B C A B X A B C D




The determinised version is very efficient during decoding

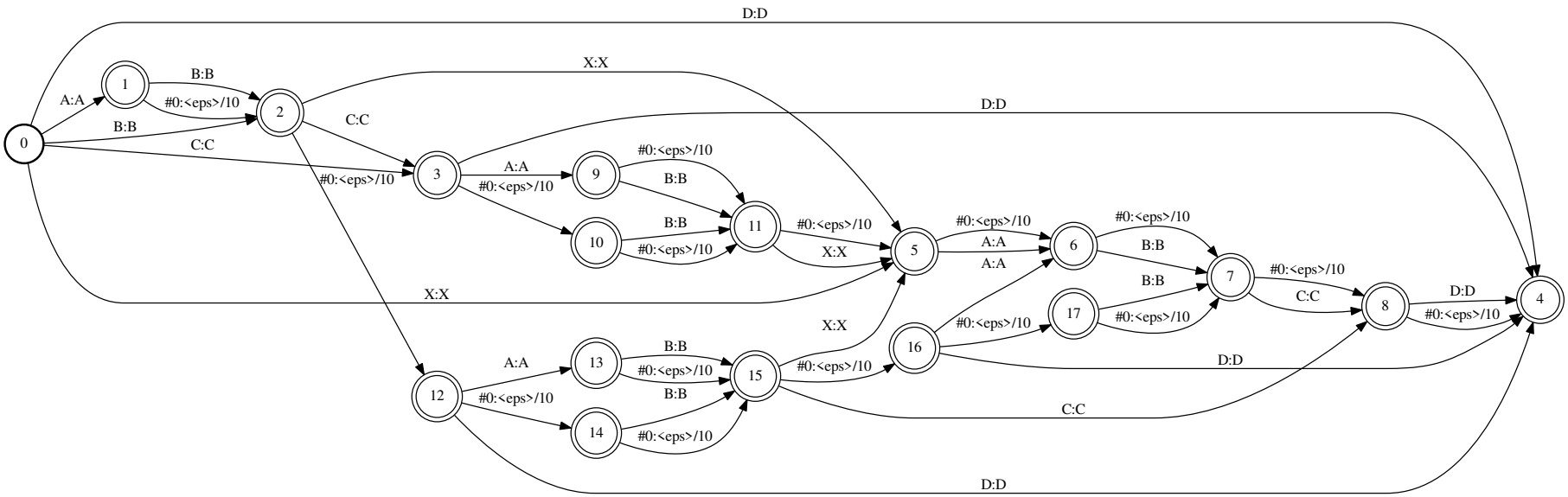
# Add word skips

A B C A B X A B C D



# Determinised version

A B C A B X A B C D



# Hallucinations

- Poor data filtering in lightly-supervised training can lead to a systematic mismatch between $X$ and $Y$ pairs
- In an HMM-system, this can often lead to a large number of deletion errors
- In end-to-end systems (especially AED models) it can lead to a high level of insertion errors ("hallucinations")
- Whisper is notably prone to this type of error. See [Frieske and Shi, 2024]

# Hallucination examples

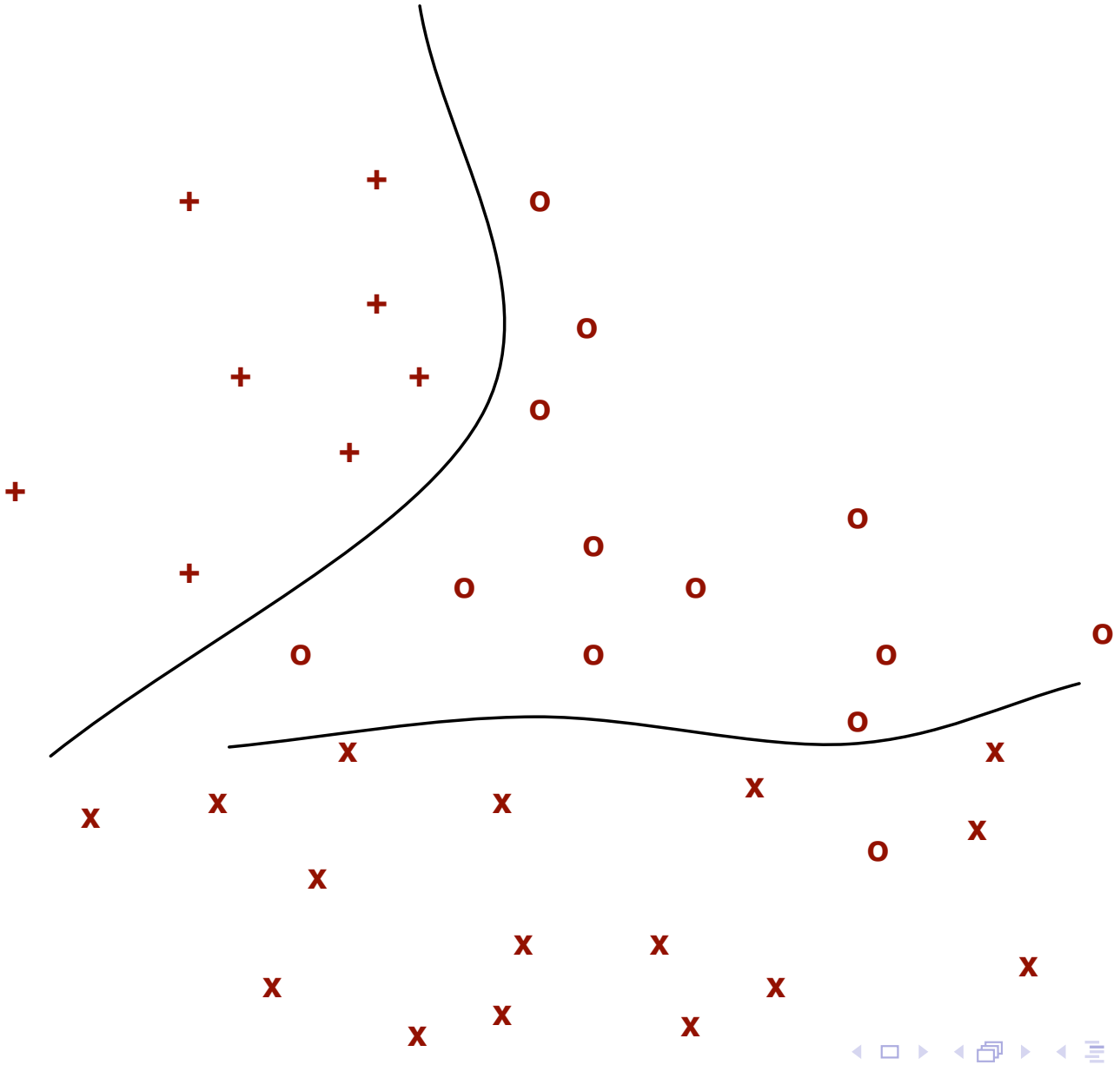
| Transcript | Language | English Translation |
|---|---|---|
| “Gracias por ver el video.” | Spanish | “Thank you for watching the video.” |
| “¿Puedo ver qué vamos hacer el video?” | Spanish | “Can I see what we’re going to do in the video?” |
| “ご視聴ありがとうございました” | Japanese | “Thank you for watching.” |

# Semi-supervised training

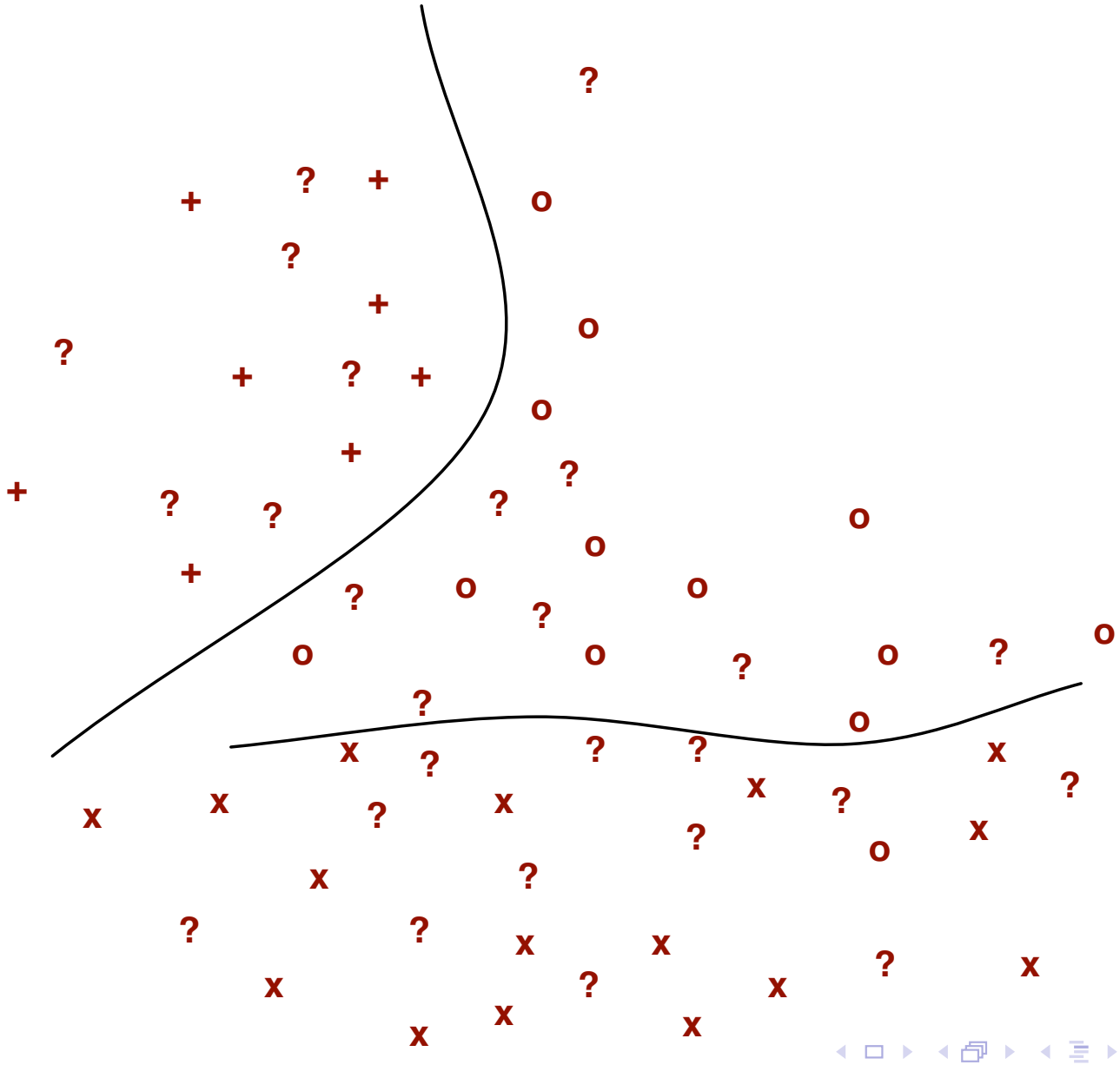
- Assume we have a weak or domain-mismatched initial model
- Use this model to generate labels for new training data
- Retrain or update the models on newly-labelled data (perhaps including the original data too)

Also sometimes called *self-training*.

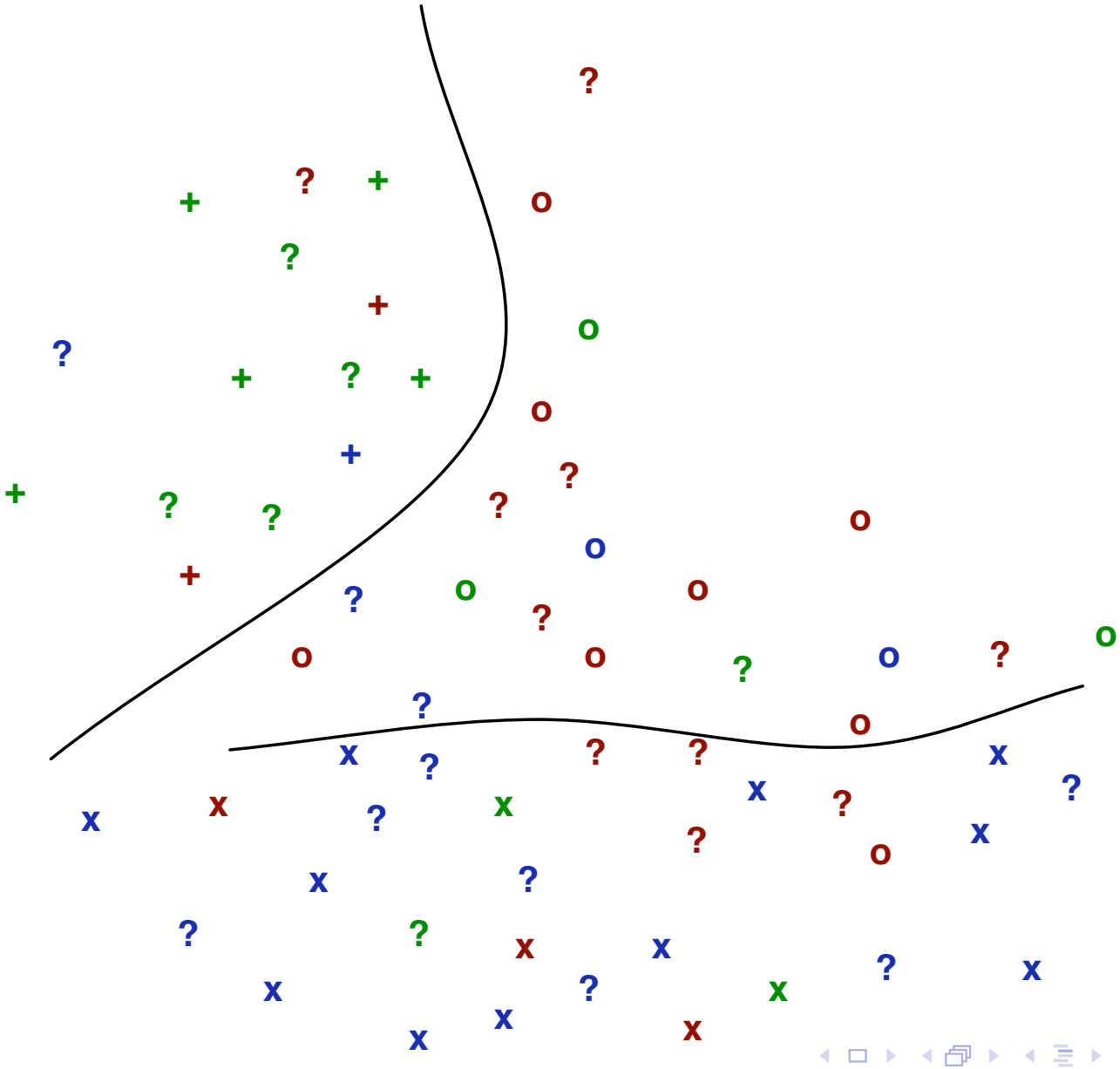
# The problem of semi-supervised training



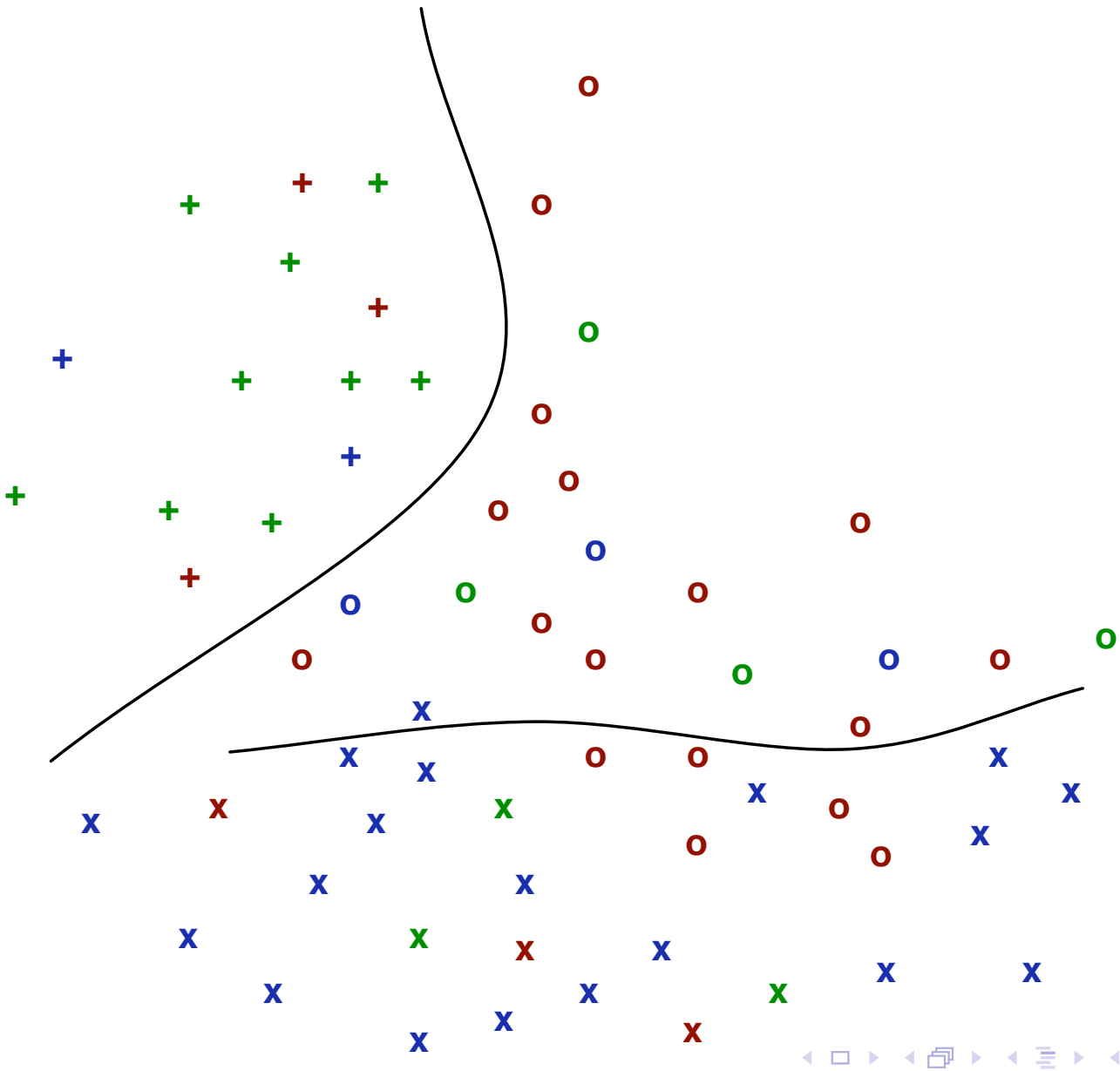
# The problem of semi-supervised training



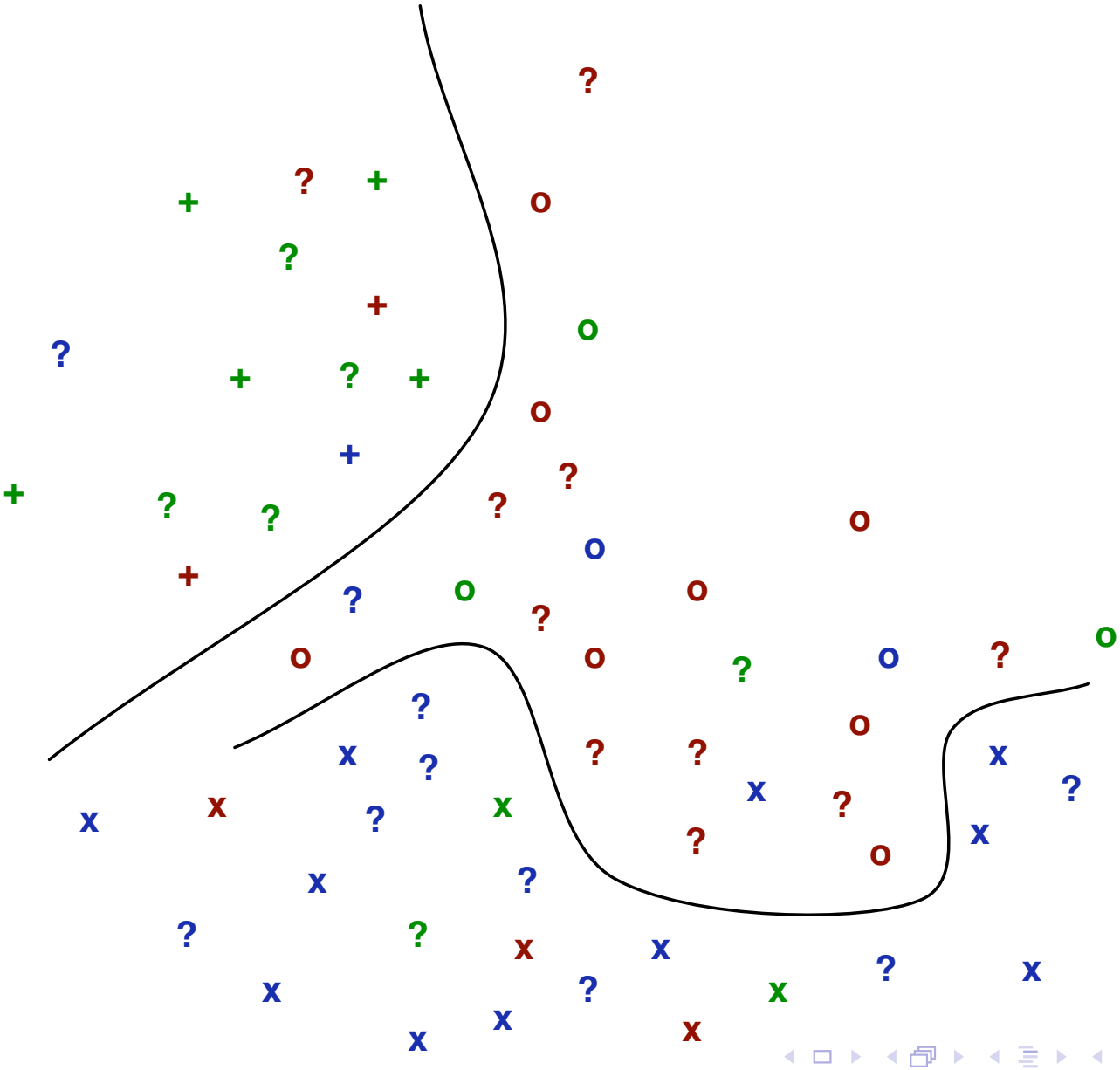
# The problem of semi-supervised training



# The problem of semi-supervised training



# The problem of semi-supervised training

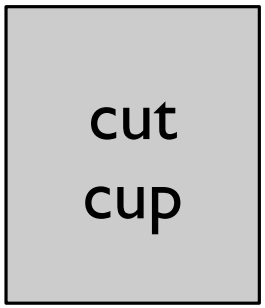


# The problem of semi-supervised training

- We don't want to train further on incorrect captions
- Traditional solution: apply data filtering based on confidence scores
- But this selection is biased towards data where the acoustic model is already confident – away from samples that will provide the most useful discriminative information
- Solution [Manohar, 2018]: use a lattice to incorporate uncertainty about the transcription, train a sequence-level criterion
- Requires a strong language model for the best performance (Wallington et al, 2021)

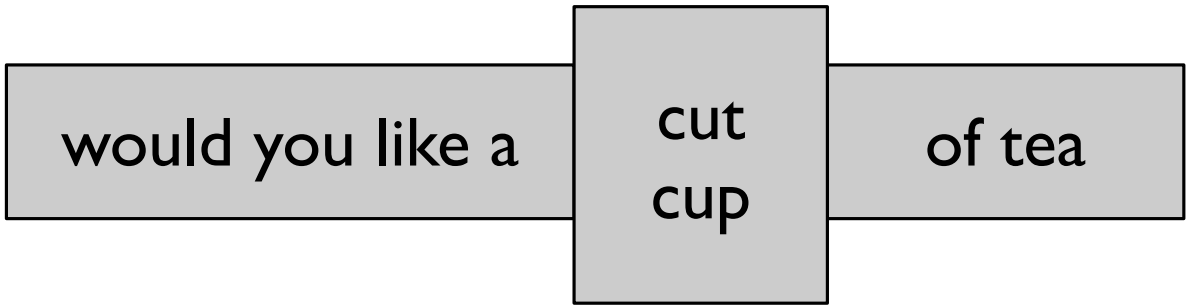
# Where does the extra information come from?

In ASR, sequence-level modelling provides the addition information, via the language model



# Where does the extra information come from?

In ASR, sequence-level modelling provides the addition information, via the language model



# Training HMMs with a semi-=supervised objective

KL objective function

$$\mathcal{F}_{\mathrm{KL}}(\theta) = D_{\mathrm{KL}}(P_W||P) = \sum_u \frac{p_W(X_u)}{p(X_u)}$$

Differentiate with respect to activations at time t

$$\frac{\partial \mathcal{F}_{\mathrm{KL}}}{\partial \log \theta_s(t)} = \sum_j \frac{\partial \mathcal{F}_{\mathrm{KL}}}{\partial \log p(x_t|j)} \frac{\partial \log p(x_t|j)}{\partial \theta_s}$$

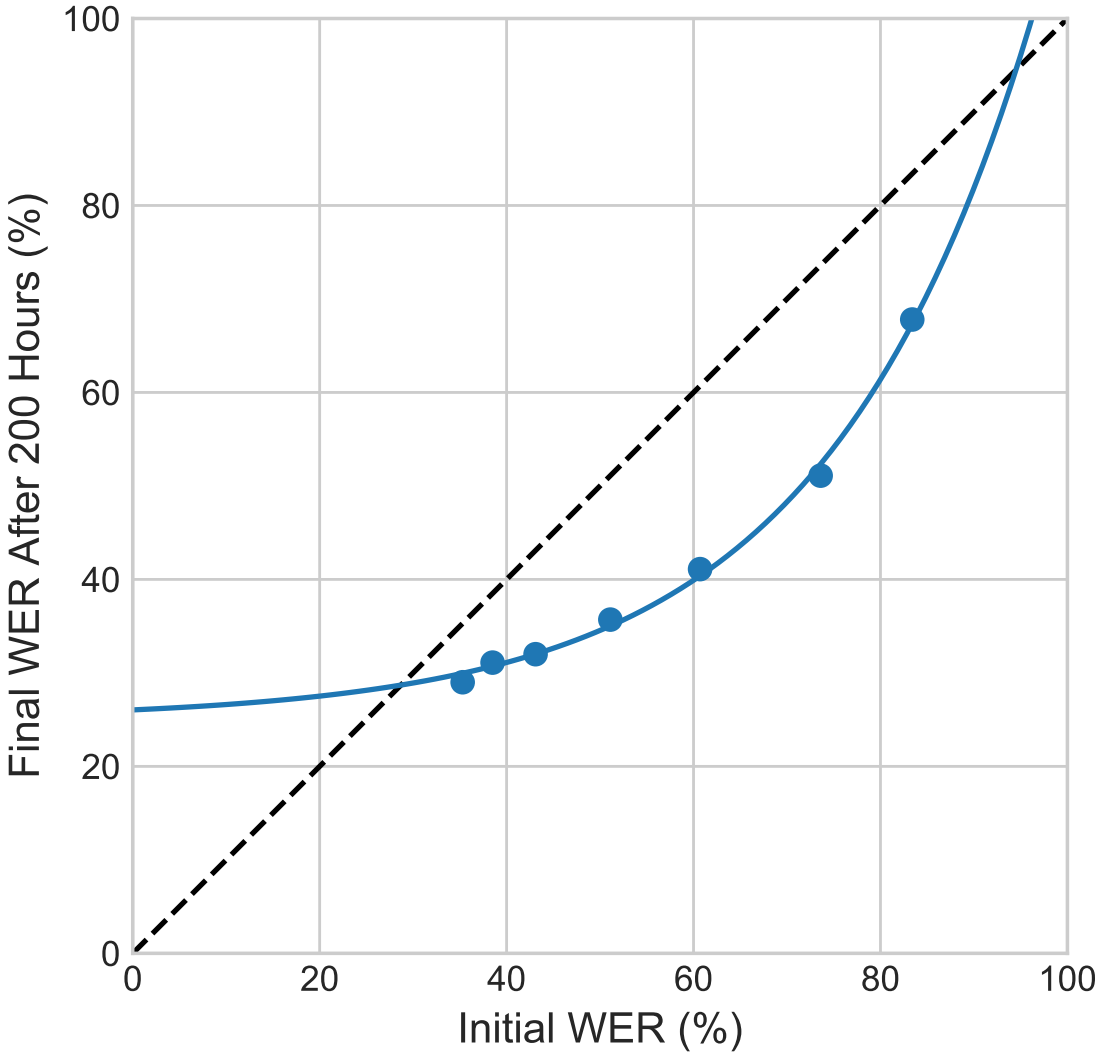
$$= \sum_j (\gamma_j^W(t) - \gamma_j(t)) \frac{\partial \log p(x_t|j)}{\partial \theta_s(t)}$$

With state occupancy probabilities given by

$$\gamma_j^W(t) = p(q_t = j|X_u, \mathcal{M}^W)$$

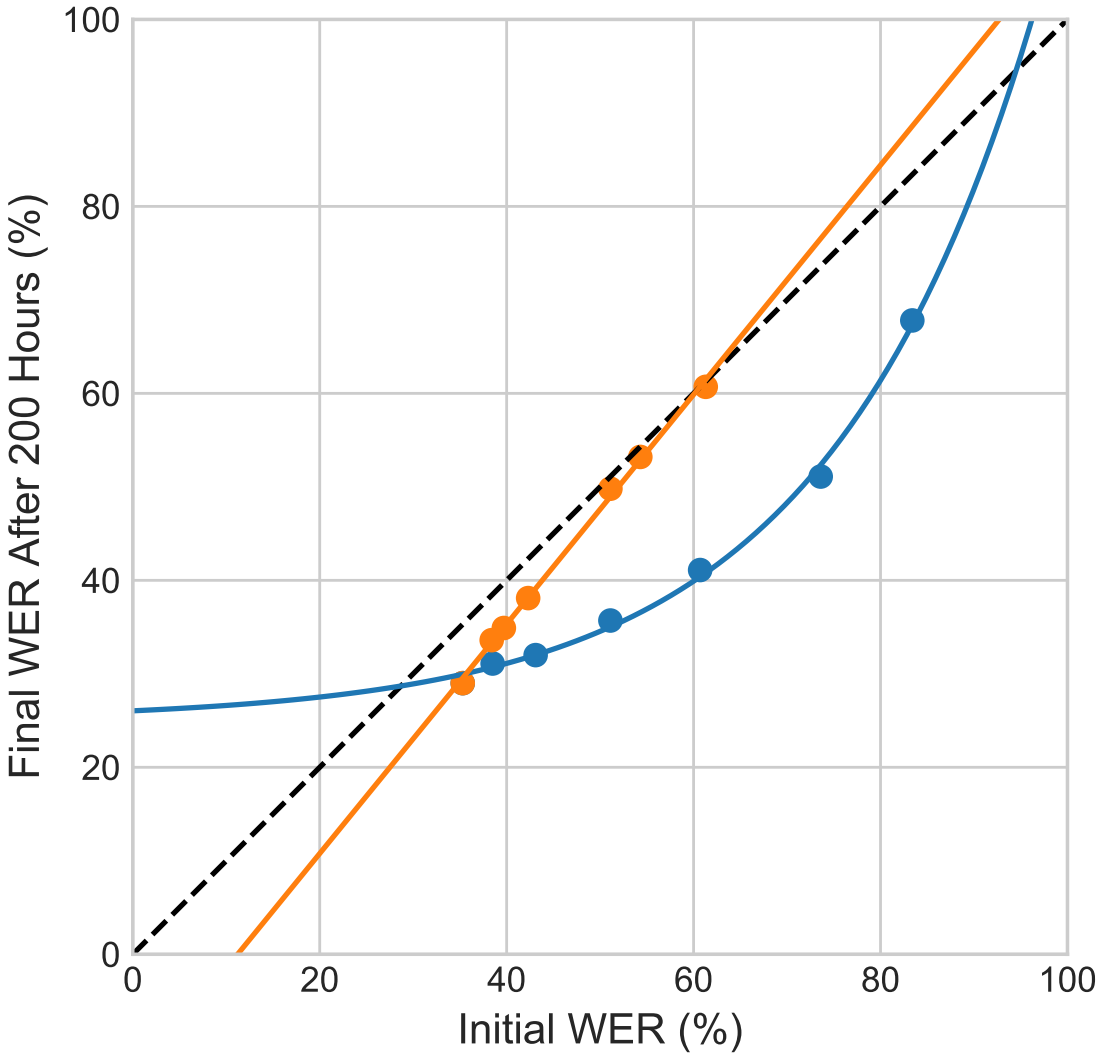
$$\gamma_j(t) = p(q_t = j|X_u, \mathcal{M})$$

# Example: Tagalog



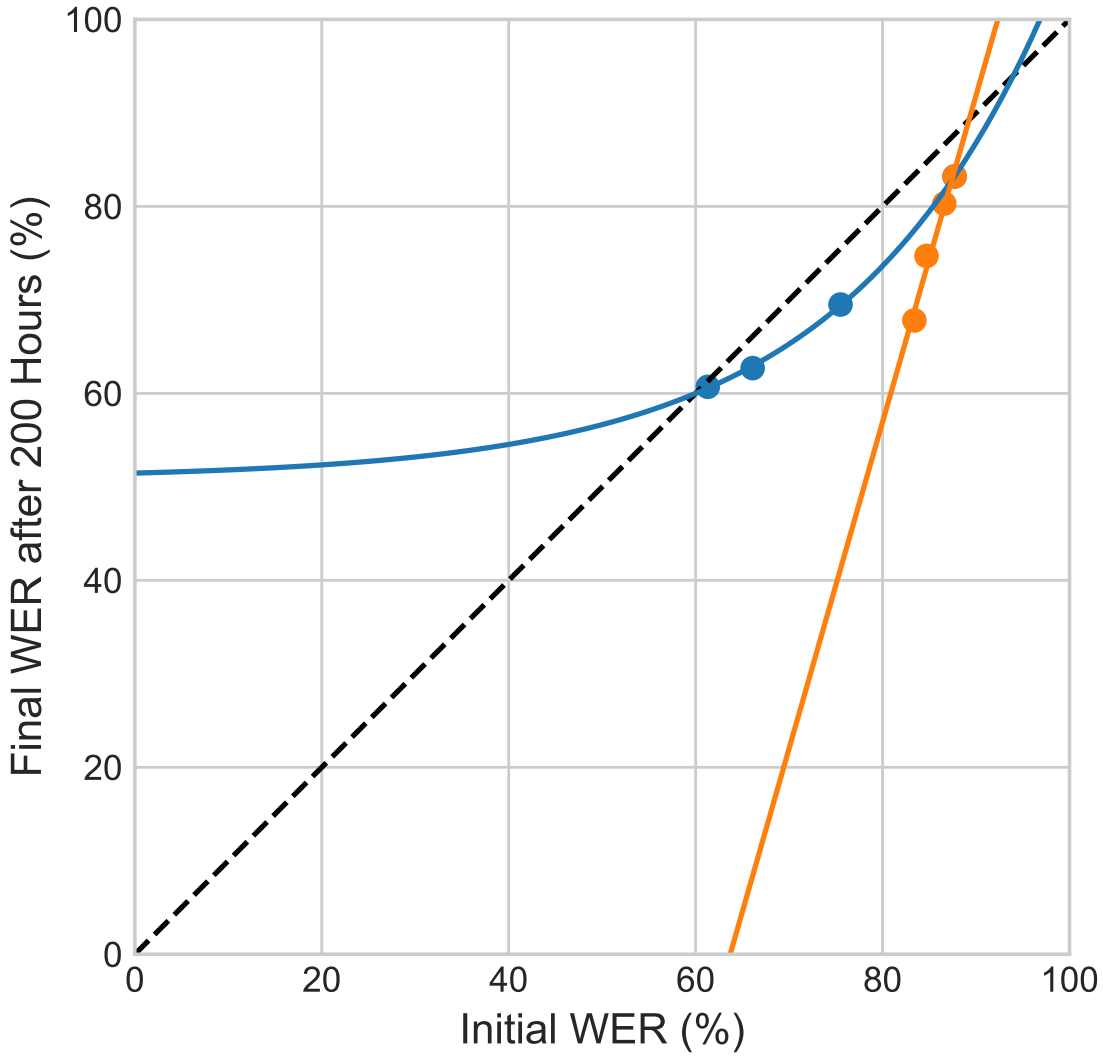

From Wallington et al (2021)

# Example: Tagalog




Blue = varying quality of AM using best LM; orange = varying quality of LM using best AM

# Example: Tagalog




Blue = varying quality of AM using worst LM; orange = varying quality of LM using worst AM

# More weakly supervised training...

- Self-supervised training (Hao, next lecture)
- Cross-lingual semi-supervised training with "decipherment" (Ondrej's guest lecture, next week)

# Reading


- Braunschweiler et al (2010), "Lightly supervised recognition for automatic alignment of large coherent speech recordings". Proc. Interspeech 2010, 2222-2225, doi: 10.21437/Interspeech.2010-611. https://www.isca-archive.org/interspeech_2010/braunschweiler10_interspeech.html
- Manohar et al (2018), "Semi-Supervised Training of Acoustic Models Using Lattice-Free MMI". Proc ICASSP, pp. 4844-4848, doi: 10.1109/ICASSP.2018.8462331. https://ieeexplore.ieee.org/document/8462331
- Wallington et al, 2021, "On the Learning Dynamics of Semi-Supervised Training for ASR". Proc. Interspeech 2021, 716-720, doi: 10.21437/Interspeech.2021-1777 https://www.isca-archive.org/interspeech_2021/wallington21_interspeech.html
- Frieske and Shi (2024), "Hallucinations in Neural Automatic Speech Recognition: Identifying Errors and Hallucinatory Models". https://arxiv.org/pdf/2401.01572