UNIVERSITY OF CAMBRIDGE

Cambridge ALTA
Institute for Automated Language Teaching and Assessment

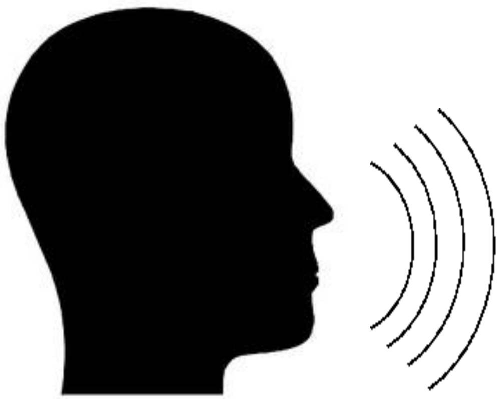
Challenges for AI in Spoken Communication

Dr Kate Knill
kate.knill@eng.cam.ac.uk

March 2017

Department of Engineering

Spoken Communication



Speaker Characteristics
Environment/Channel

Pronunciation
Prosody

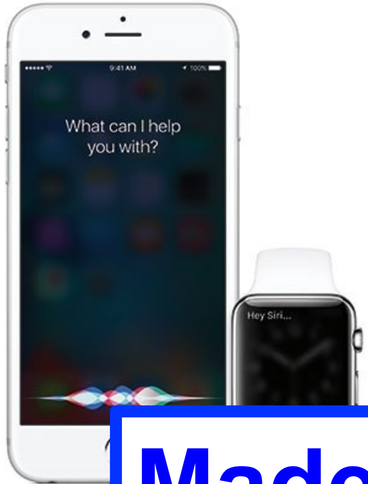
Message Construction | Message Realisation | Message Reception

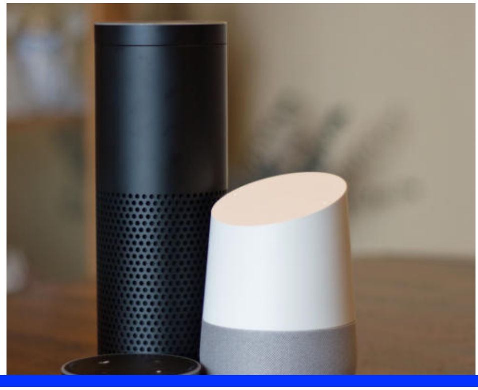
Spoken communication is a very rich communication medium

Driving factors for using speech

- Voice User Interfaces
  - Speed – e.g. dictating faster than typing text messages
  - Hands-free – e.g. driving, cooking, across the room from device
  - Intuition – everyone knows how to talk, natural replies easy to obtain
  - Empathy – conveyed through the rich medium of voice
- Data Analysis and Retrieval
  - Quantity of Data – a lot of data is in spoken form e.g. calls, radio, agents
  - Quality of Data – information about human interactions e.g. Microsoft Xiaoice

Speech is solved …





Made possible by Deep Learning

skype™ Translate

… but we' re not there yet

EXPRESSIVE SYNTHESIS

UNDERSTANDING

CONTEXT AWARE

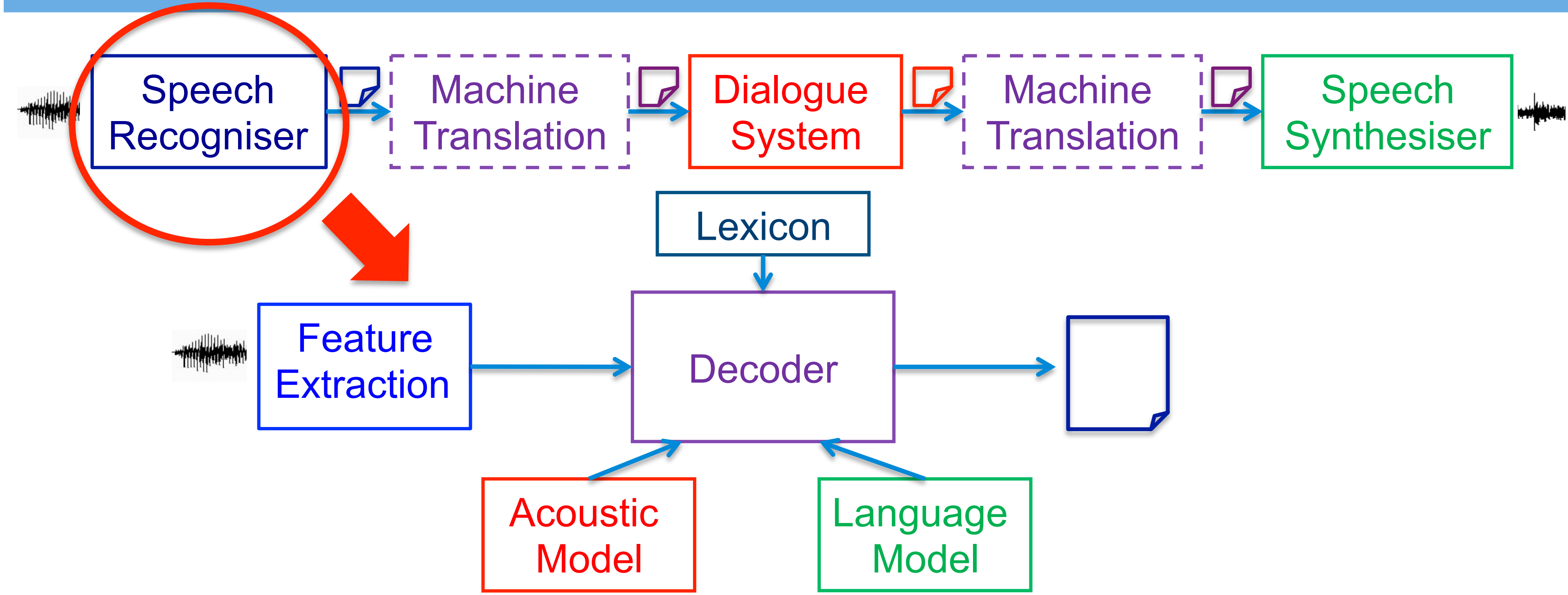
ROBUST RECOGNITION

NATURAL TRANSLATION

Unique challenges of spoken language

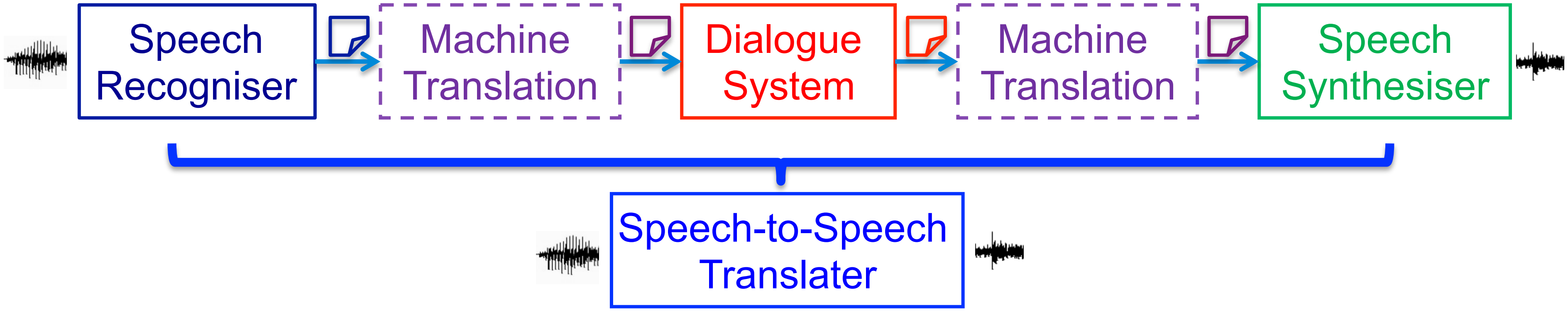
- Very rich communication medium
  - Content encoded in sound waves, words, tone, and rhythm
- Sequence-to-sequence modelling problem
  - speech synthesis: word sequence (discrete) ➔ waveform (continuous)
  - speech recognition: waveform (continuous) ➔ word sequence (discrete)
  - machine translation: word sequence (discrete) ➔ word sequence (discrete)
- The sequence lengths on either side can differ
  - waveform sampled at 5/10ms frame-rate, words, dialogue actions …

Speech-to-speech systems



- Separate modules allow flexible systems to be constructed
- Large gains achieved through applying Deep Learning to modules
- Non optimal, module errors propagated through pipeline
- Pre-define the sequences and connections between modules

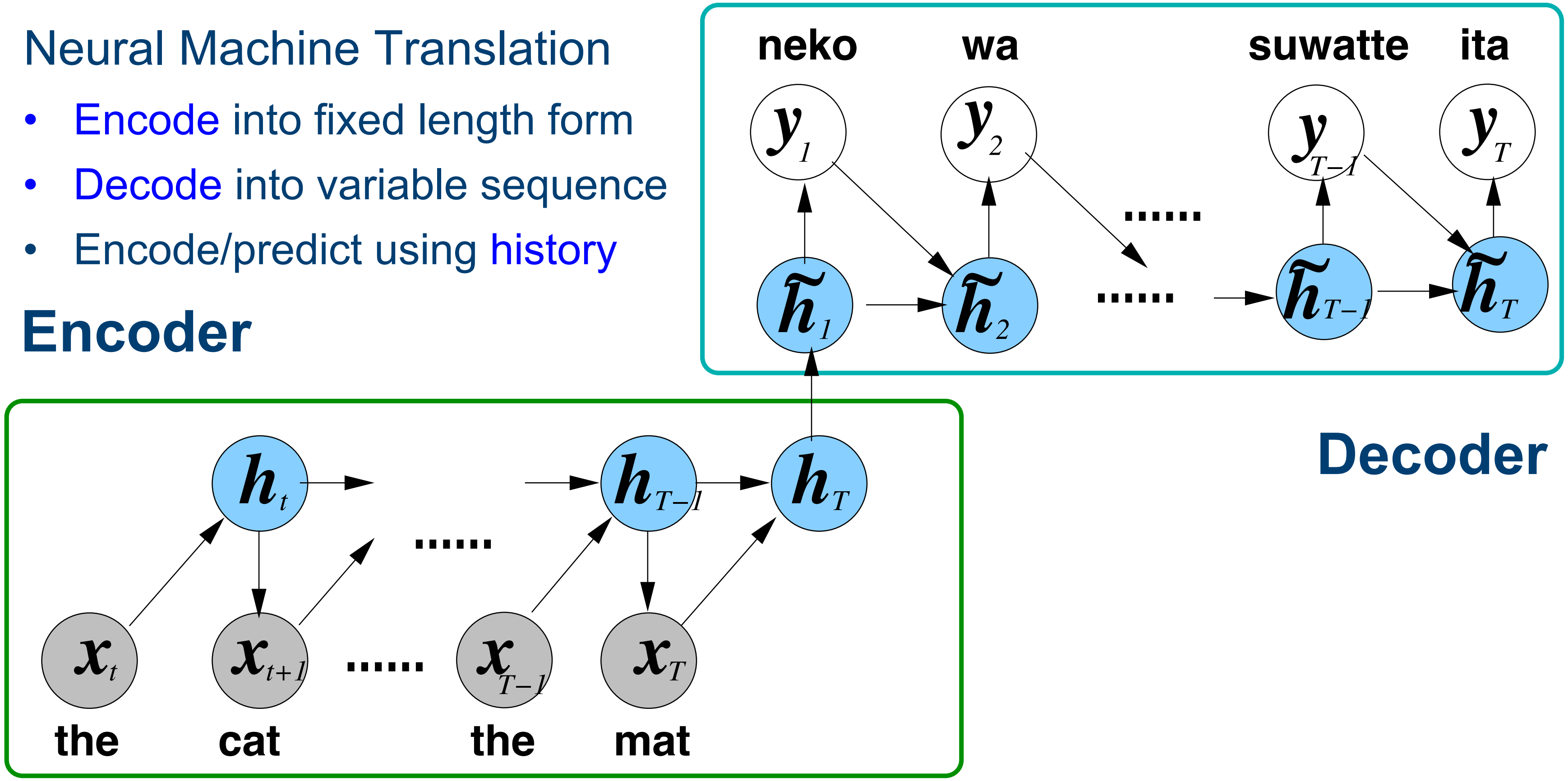
Integrated end-to-end systems



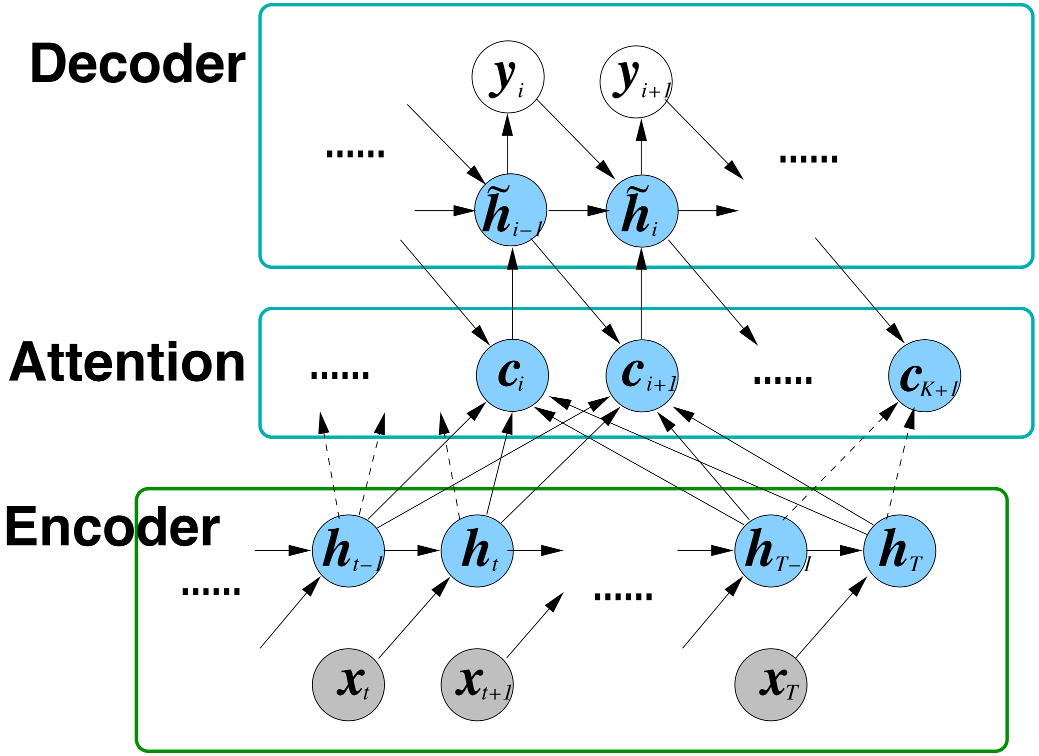
- Optimised together for full system
- Use deep learning to model sequence-to-sequence mappings
- Don’t have to predefine sequences and connections between modules

End-to-end system example

- Neural Machine Translation
  - Encode into fixed length form
  - Decode into variable sequence
  - Encode/predict using history



End-to-end systems: attention based model



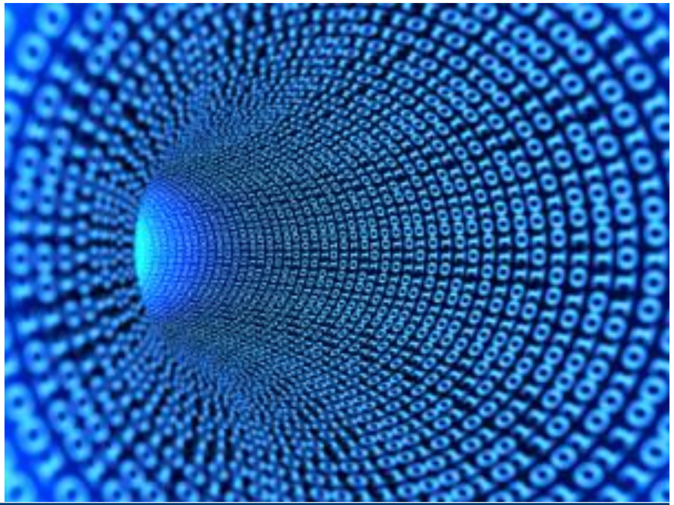
- Attention provides focus
  - Focus on most useful history
  - Emphasise key data

Need annotated training data that may not be available yet

Challenges for AI: Data Overload

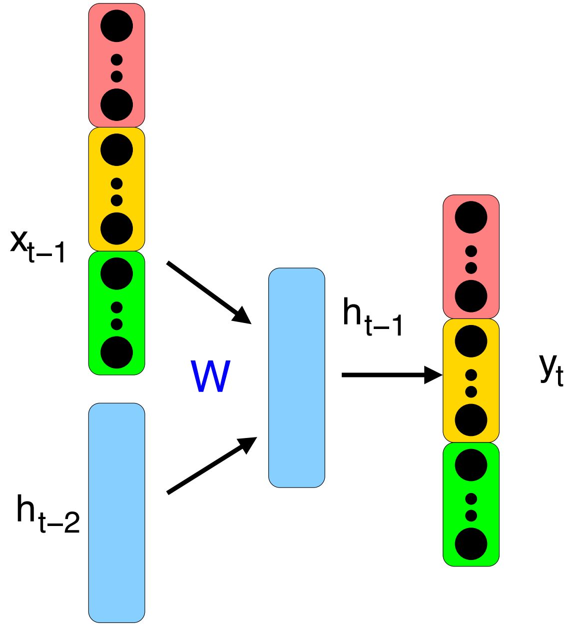
- Huge amounts of data are being collected e.g. in 2016
  - 3.7bn Google US voice searches, 2bn Siri requests, 5.2m Amazon Echo sold
- Problem:
  - Too much data to use and sample
    - which data to exploit?
    - which data to transcribe?
- Potential solution:

1. Combination of Data Mining and Active learning
   - System learns which data helps give most gains
2. Continuous Adaptation
   - Reinforce “winning” strategies



Challenges for AI: Lack of data

- For many domains and languages there is a lack of data
- Problem:
  - Insufficient data to build robust models
    - speech and/or text
- Potential solutions: exploit "other" data
  1. Multi-task training
     - Share network layers across tasks
  2. Cross-language/multilingual training
     - Share network layers across languages
     - Multilingual – language independent networks
       - e.g. IARPA Babel - audio data search in 26 languages



New applications: voice as a user interface

- Conversational speech systems
  - Infotainment in e.g. self driving cars (EPSRC Open Domain Statistical SDS)
  - Language learning and assessment (Cambridge ALTA Institute)
  - Mental health maintenance (EPSRC Natural Speech Automated Utility for Mental Health)
  - Robot support of elderly and disabled

- Speech-to-speech/text translation for any language
  - Support business in new areas e.g. Africa
    (IARPA Babel, EPSRC Improving Target Language Fluency in Statistical Machine Translation)
  - Rapid emergency response (IARPA Babel)

New applications: exploiting speech data

- Cross-language information retrieval
  - Search
  - Summarisation
  - Data Analysis

- Data analysis
  - Learn how humans converse
  - Health monitoring and early detection
  - Feedback on performance: education, agents, gaming

Cambridge University Engineering Speech Group

- Speech Group works on many aspects of spoken language processing
  - automatic speech recognition
  - statistical machine translation
  - statistical dialogue systems
  - statistical speech synthesis
- World-wide reputation for research
- Hidden Markov Model Toolkit
  - Used by R&D groups worldwide in academia and industry
  - Active development for current state-of-the-art approaches
  - Range of extensions: HMM Synthesis (HTS), RNN LMs

Conclusions

- Spoken language is a very rich communciation medium
- AI has advanced speech technology significantly in recent years
- Challenges still remain to achieve “speech communication”
  - End-to-end integrated systems
  - Data – too much, too little
- Potential for many new applications

Spoken Language Versus Written

ASR Output

okay carl uh do you exercise yeah actually um i belong to a gym down here gold's gym and uh i try to exercise five days a week um and now and then i'll i'll get it interrupted by work or just full of crazy hours you know

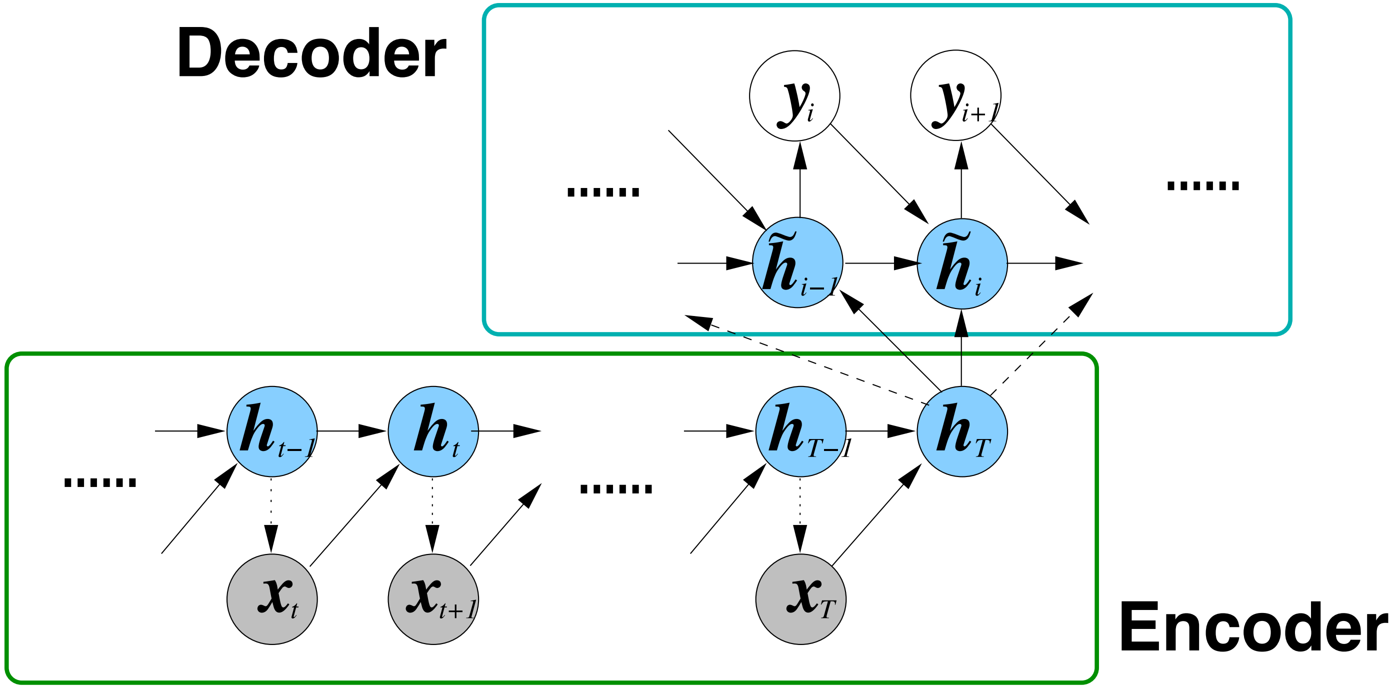
Meta-Data Extraction Markup

Speaker1: / okay carl {F uh} do you exercise /
Speaker2: / {DM yeah actually} {F um} i belong to a gym down here /
/ gold's gym / / and {F uh} i try to exercise five days a week {F um} /
/ and now and then [REP i'll + i'll] get it interrupted by work or just full of crazy hours {DM you know } /

Written Text

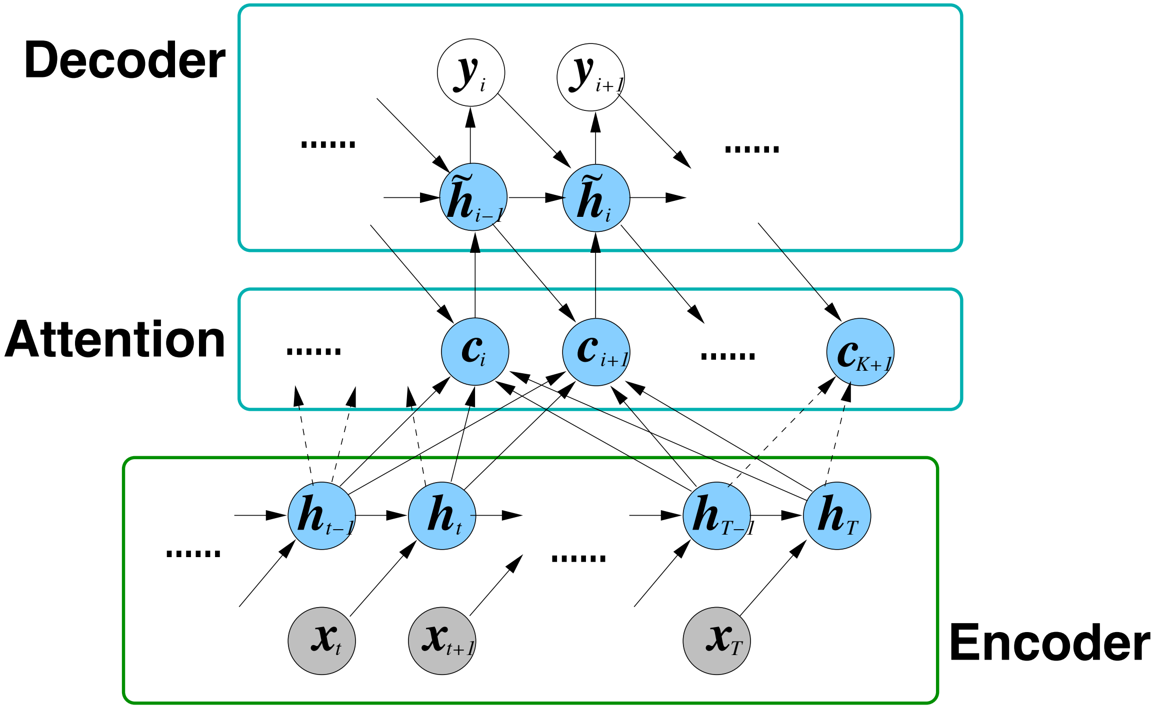
Speaker1: Okay Carl do you exercise?
Speaker2: I belong to a gym down here, Gold's Gym, and I try to exercise five days a week and now and then I'll get it interrupted by work or just full of crazy hours.

End-to-end systems: RNN encoder-decoder



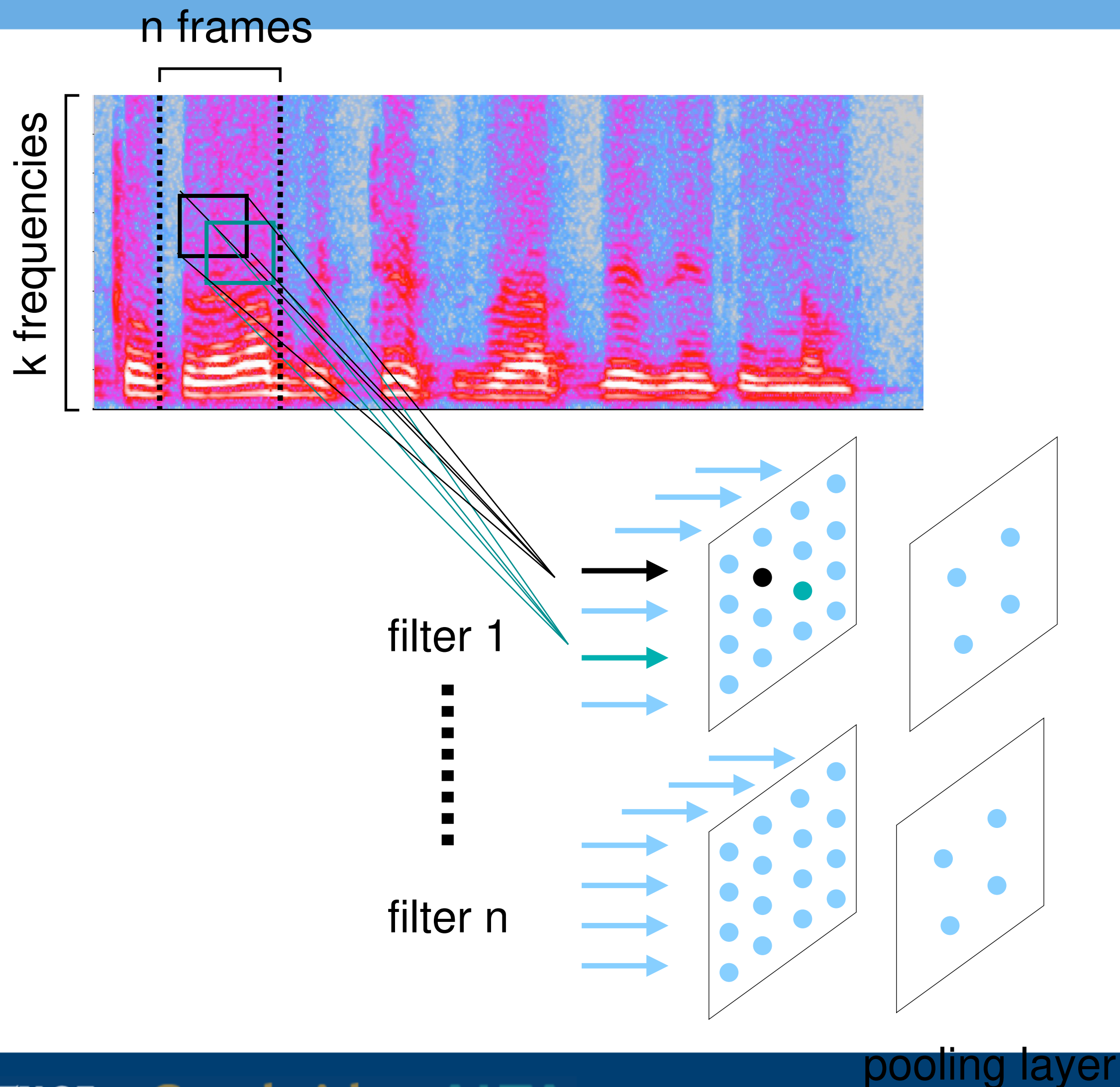
$$p(\mathbf{y}_{1:L}|\mathbf{x}_{1:T}) = \prod_{i=1}^{L} p(\mathbf{y}_i|\mathbf{y}_{i:i-1}, \mathbf{x}_{1:T})$$
$$\approx \prod_{i=1}^{L} p(\mathbf{y}_i|\mathbf{y}_{i:i-1}, \tilde{\mathbf{h}}_{i-2}, \mathbf{c})$$

End-to-end systems: attention based model

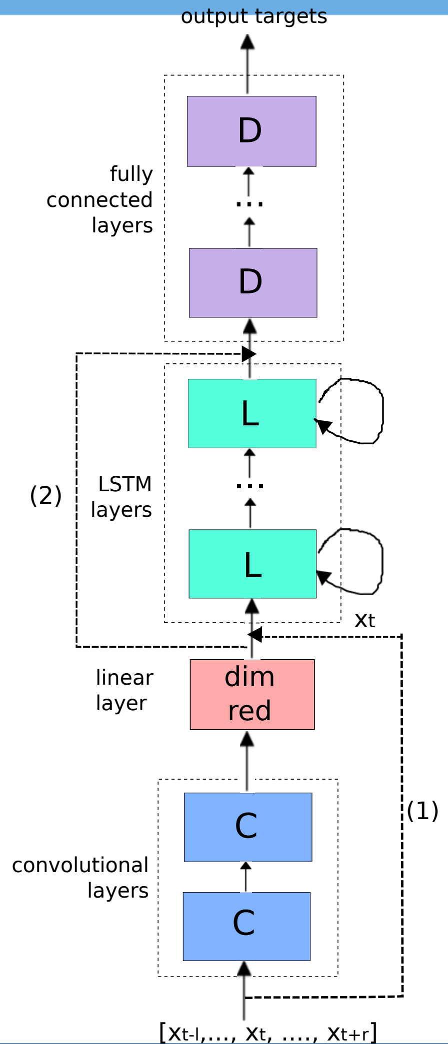


$$p(\mathbf{y}_{1:L}|\mathbf{x}_{1:T}) \approx \prod_{i=1}^{L} p(\mathbf{y}_i|\mathbf{y}_{i:i-1}, \tilde{\mathbf{h}}_{i-2}, \mathbf{c}_i) \approx \prod_{i=1}^{L} p(\mathbf{y}_i|\tilde{\mathbf{h}}_{i-1})$$

Convolutional neural network for speech



Google ASR System



Language modelling

- Model of word sequences
- Standard model n-gram

$$P(w) \quad = \quad \prod_{k=1}^{K+1} P(w_k|w_0, w_1, \ldots, w_{k-1}) \quad \approx \quad P(w_k|w_{k-1}, w_{k-2})$$

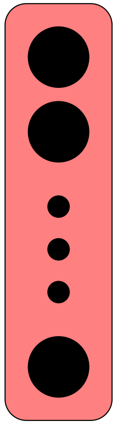
  - Very efficient
  - History limited to last 2 words

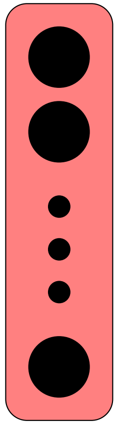
The cat sat on the ?   P(mat | on the)

猫はマットの上に?   P (座っていた |上に)

Language model neural network input and outputs

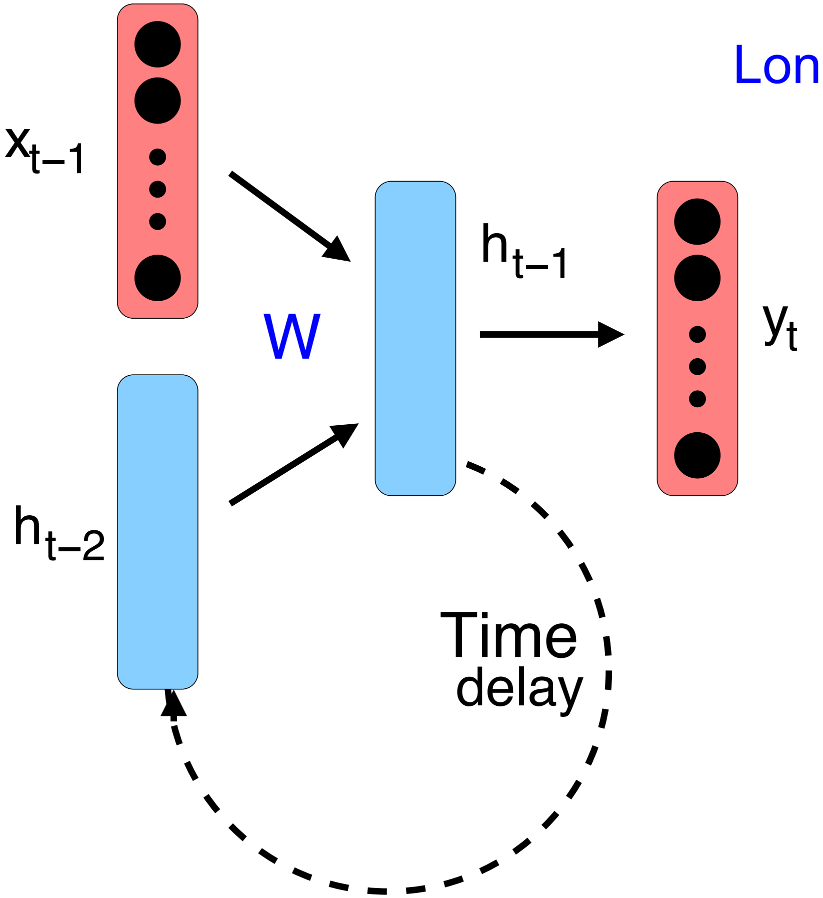
- Use neural networks to expand history

x_t  $= \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

y_t  $= \begin{bmatrix} P(\text{cat}|h) \\ P(\text{sat}|h) \\ P(\text{on}|h) \\ P(\text{the}|h) \\ P(\text{mat}|h) \end{bmatrix}$

vocabulary = {cat,sat,on,the,mat}
word at time t is "sat"
"h" is the history (preceeding words)

Recurrent neural network language models



Longer history ➔ more accurate prediction

The cat sat on the ?

P (mat | The cat sat on the)

猫はマットの上に?

P (座っていた | 猫はマットの上に)

- Improved history modelling
  - Long-short term memory
  - Bidirectional