Statisical Dialogue Systems
Talk 1 – Intro, Inputs & Outputs

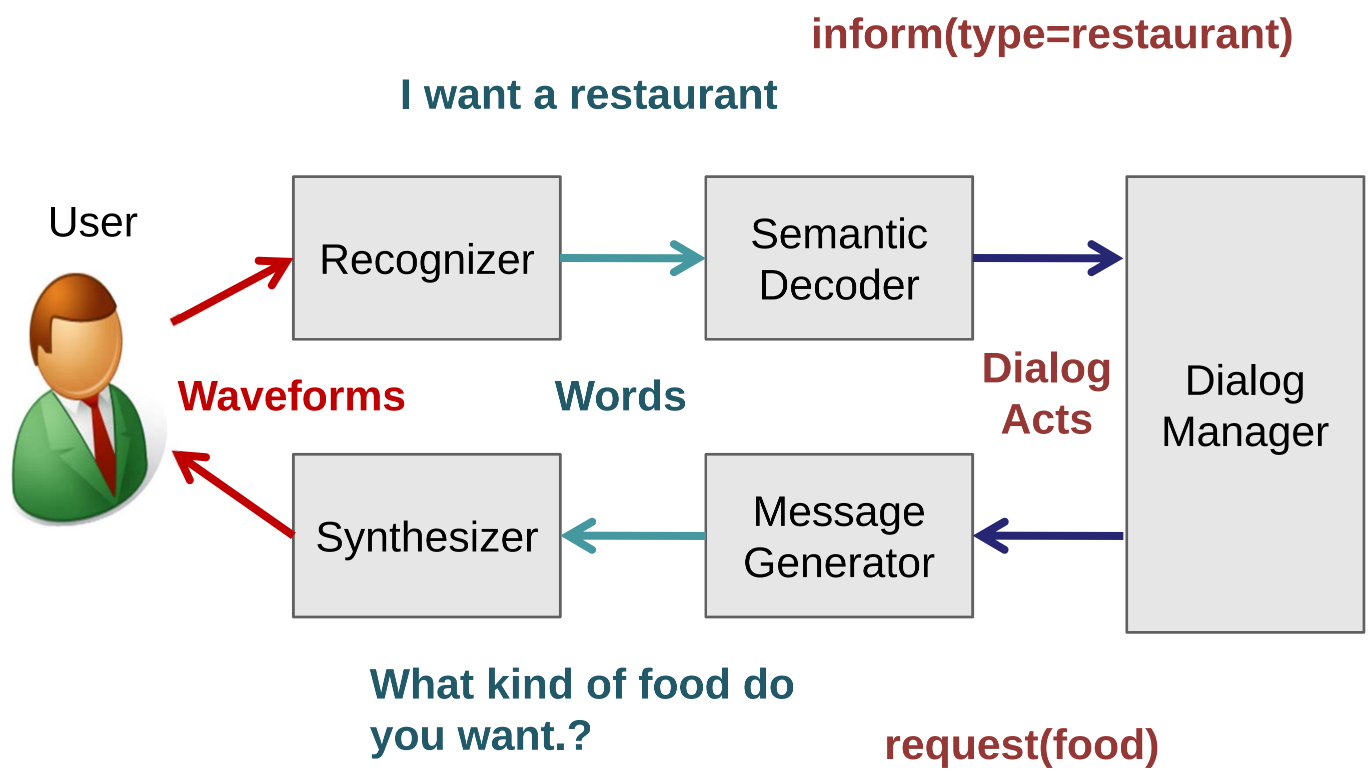
CLARA Workshop

Presented by Blaise Thomson

Cambridge University Engineering Department
brmt2@eng.cam.ac.uk
http://mi.eng.cam.ac.uk/~brmt2

Spoken Dialogue Systems – Example - Siri

Human-machine spoken dialogue

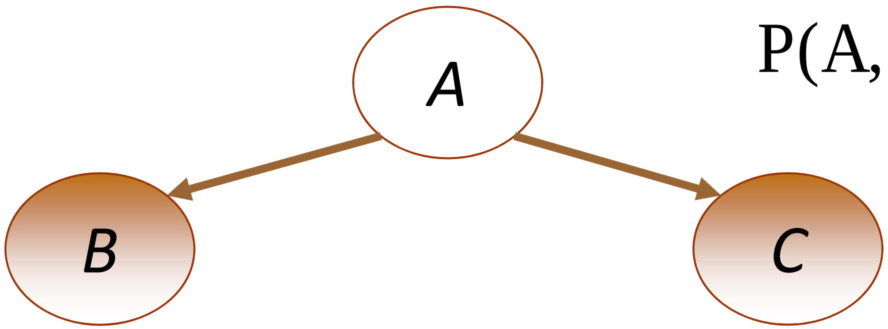


Typical structure of a spoken dialogue system

Before we start – Bayesian networks

- Bayesian networks are a graphical representation of a statistical model we will use extensively in these talks

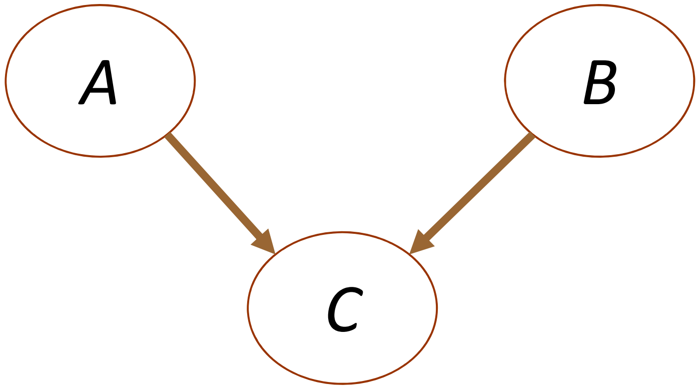
- Definition:
  - A directed acyclic graph (nodes and arrows)
  - Nodes are random variables
  - The joint distribution of all the nodes factorizes as the product of the probability of each node given its parents in the graph
  - Observed variables are coloured



$$P(A,B,C) = P(A)P(B \mid A)P(C \mid A)$$

Before we start – Bayesian networks

- These networks encode some useful independence assumptions

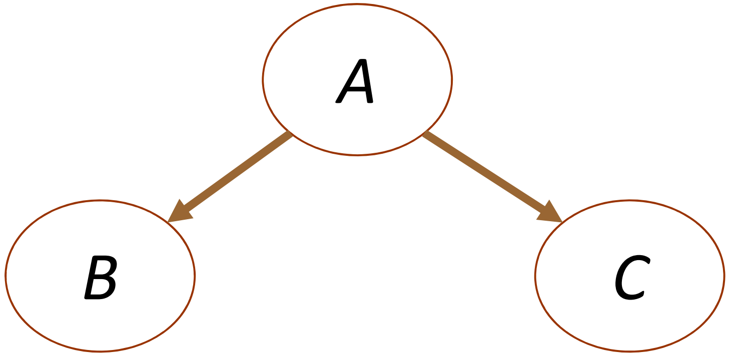


$$
\begin{aligned}
P(A,B) &= \sum_{C} P(A,B,C) \\
&= \sum_{C} P(A)P(B)P(C \mid A,B) \\
&= P(A)P(B)
\end{aligned}
$$

A&B are independent

Before we start – Bayesian networks

- These networks encode some useful independence assumptions



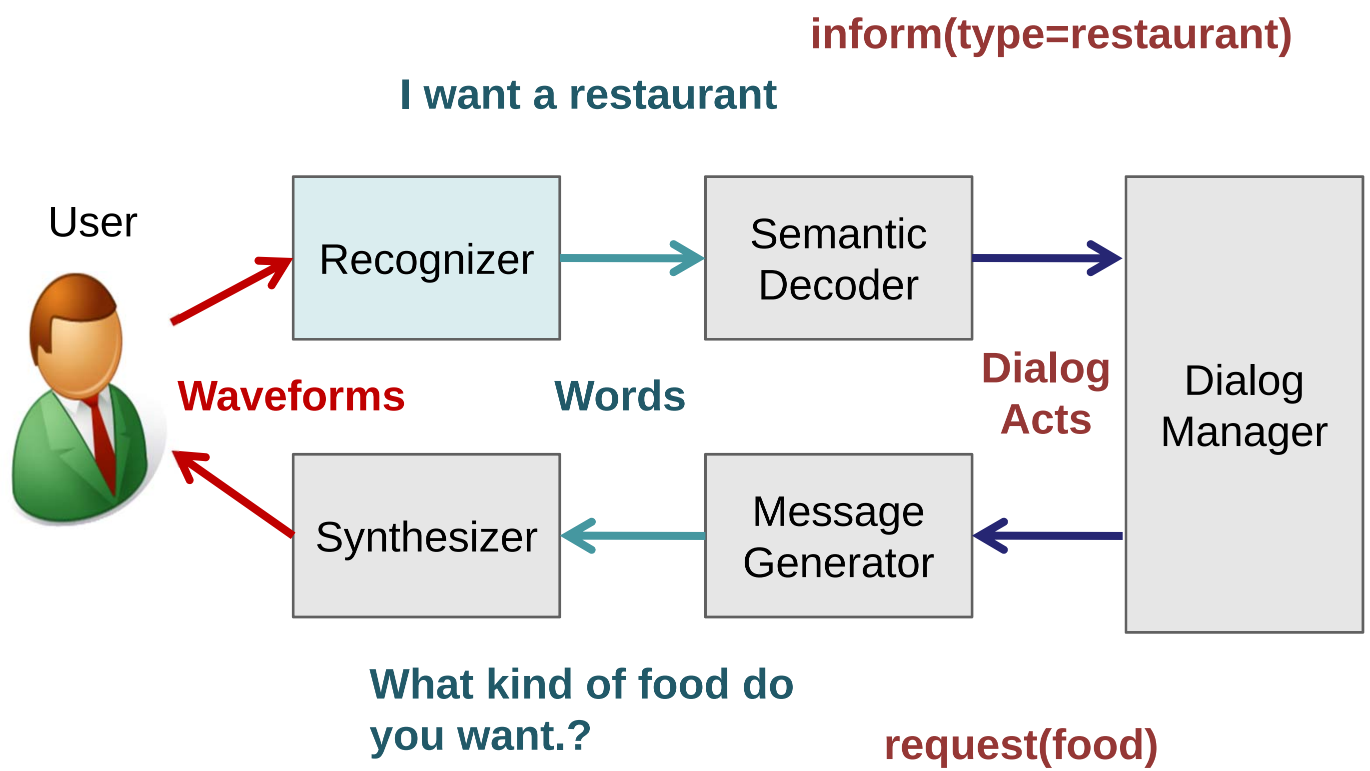
$$P(B,C \mid A) = P(A,B,C)/P(A)$$
$$= P(A)P(B \mid A)P(C \mid A)/P(A)$$
$$= P(B \mid A)P(C \mid A)$$

B&C are conditionally independent given A

Outline – Talk 1

- Speech Recognition
  - Hidden Markov Models
- Semantic Decoding
  - Phoenix
  - SVM decoders
- Dialogue management
- Output generation
  - Templates
- Text-to-speech
  - Hidden Markov Models
  - Unit selection

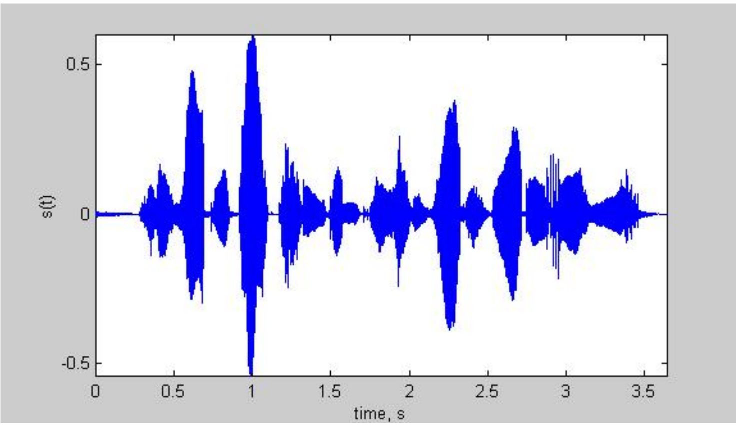
Human-machine spoken dialogue



Speech recognition – Front end

- Split audio stream into frames (about 10ms)

- For each frame do a Fourier transform and extract various features (usually Mel Frequency Cepstral Coefficients / Perceptual Linear Predictors)

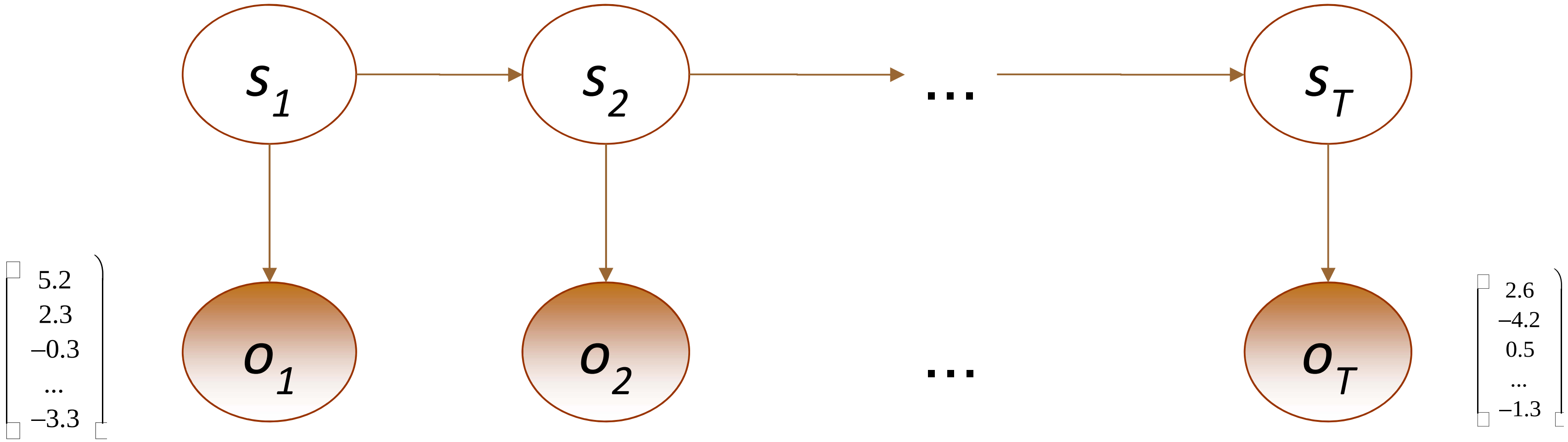


$$\begin{bmatrix} 5.2 \\ 2.3 \\ -0.3 \\ \ldots \\ -3.3 \end{bmatrix} \quad \begin{bmatrix} 2.6 \\ -4.2 \\ 0.5 \\ \ldots \\ -1.3 \end{bmatrix}$$

Speech recognition – The model

- Assume a sequence of states, s_t (model phones/sounds)
- Each state is hidden, and has an observation probability f^n
- Assume Markov property:

$$P(s_{t+1} \mid s_t, s_{t-1}, o_{t-1}, \ldots) = p(s_t \mid s_t)$$
$$P(o_t \mid s_t, s_{t-1}, o_{t-1}, \ldots) = p(o_t \mid s_t)$$



- Called a Hidden Markov Model (HMM)

Speech recognition – Bringing together

- Not all phone sequences should be allowed
  - Restrict phone transition probabilities so that they correspond to words, *w*
  - e.g. state 1 is first part of word, which must be followed by state 2, which is second part, etc.

- Transition between words is governed by a language model:

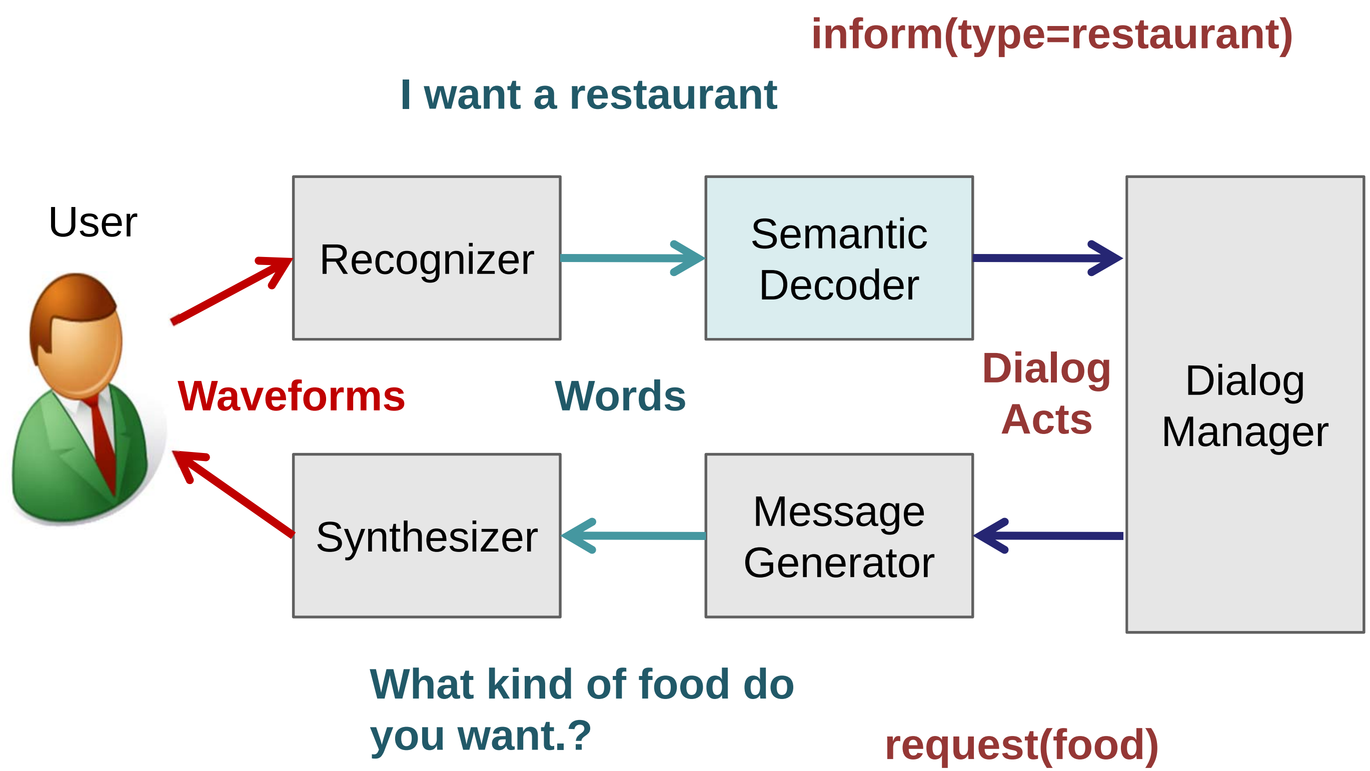
  $p(w_n \mid w_{n-1}, w_{n-2})$ [Trigram model]

- Observation model is called the acoustic model:
  - $p(o \mid s)$
  - Typically use a Gaussian Mixture Model (GMM)

Speech recognition – Inference

- Need to compute the probability of state sequences p(S | O)
- Use message passing algorithm (will discuss next time in context of dialogue systems)
- When training, the states are typically estimated using the Expectation-Maximisation algorithm
  - Fix probability models and estimate states
  - Fix state estimates and estimate probability models
  - Repeat
- Free toolkits: ATK/HTK (Cambridge, C), Sphinx (CMU, Java)
- Commercial versions: Nuance Dragon Naturally Speaking

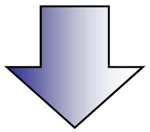
Human-machine spoken dialogue



Semantic decoding - Intro

- Lots of disfluencies in speech – grammars tend to break
- We don't care about the exact meaning
  - We just want to know what the user wants
  - Idea of speech act / dialog act (Austin / Searle / Traum)
- Our (very) simple formalism:

Is there um maybe a cheap place in the centre of town please?



inform (price = cheap, area = centre)

dialogue act type

semantics slots and values

Semantic decoding – a simple approach

Is there um maybe a cheap place in the centre of town please?

Word / Bigram	x_i
is	1
there	1
yes	0
um	1
...	
is there	1
is you	0
...	

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{pmatrix}$$

Semantic decoding – a simple approach

inform (price = cheap, area = centre)

Act type
inform

Assign type number (e.g. 3)

Item	y_i
price=cheap	1
price=exp	0
...	...
area=centre	1
area=west	0
...	...

$$\begin{bmatrix} 1 \\ 0 \\ \ldots \\ 1 \\ 0 \\ \ldots \end{bmatrix}$$

Semantic decoding – a simple approach

- When training we have lots of input vectors $\mathbf{x}_t$ and output vectors $\mathbf{y}_t$
- Use your favourite supervised learning algorithm
  - Naïve Bayes
  - Logistic regression
  - Support Vector Machines (Mairesse et al, 2009)
  - Others?
- Act type is multi-class labeling task, others are all just binary

Semantic decoding – summary so far

- Words are inputs
  - Convert them to vectors (1/0)
  - Add bigrams / trigrams
- Act type + slot values are outputs
  - Convert them to vectors (1/0)
- Run your favourite learning algorithm
- In practice – may help to post-process a bit

Semantic decoding – Further approaches

- Hidden Vector State model

  HMM structure, with hidden stack of concepts

  He & Young (2005)

- Using Markov Logic Networks

  Meza-Ruiz (2008)

- Transformation based approach

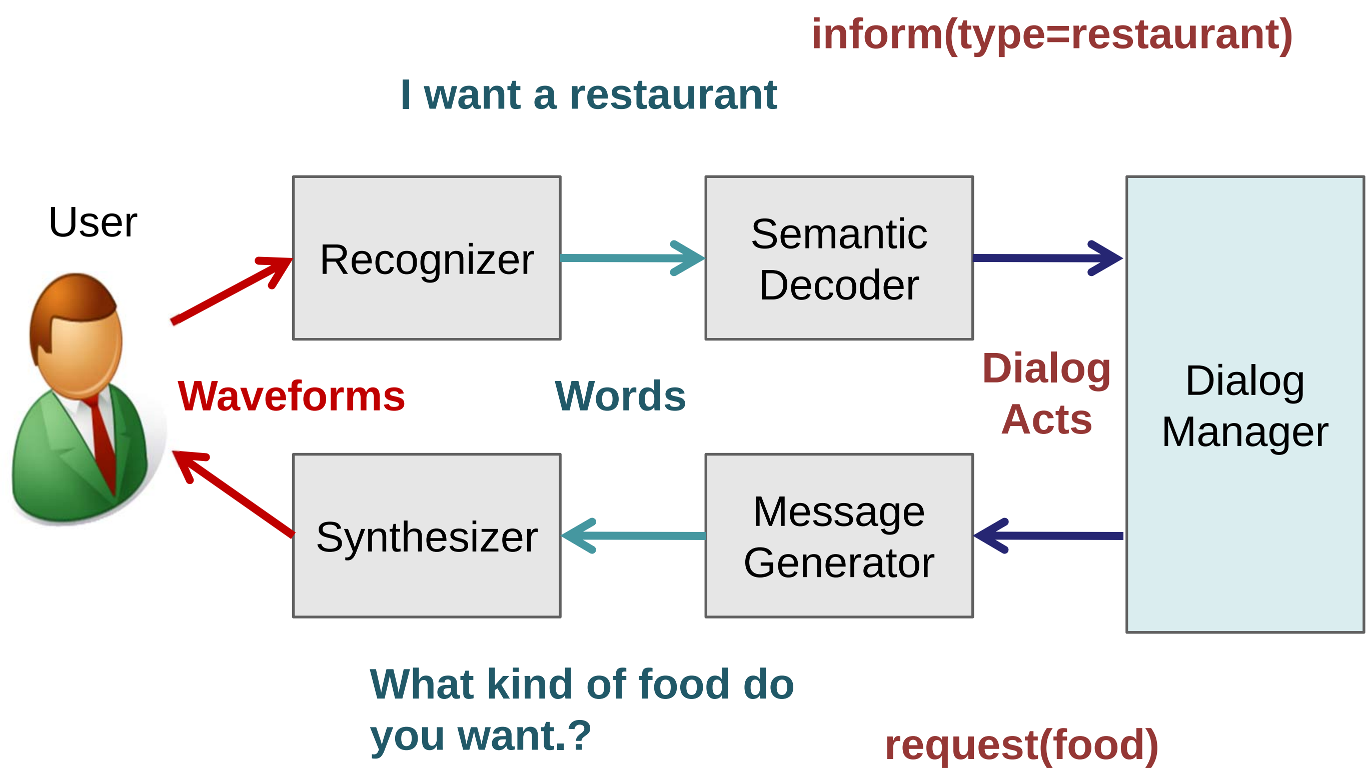
  Jurcicek et al (2009)

- Using Combinatory Categorial Grammars

  Supervised - Zettlemoyer & Collins (2009)

  Unsupervised - Artzi & Zettlemoyer (2011)

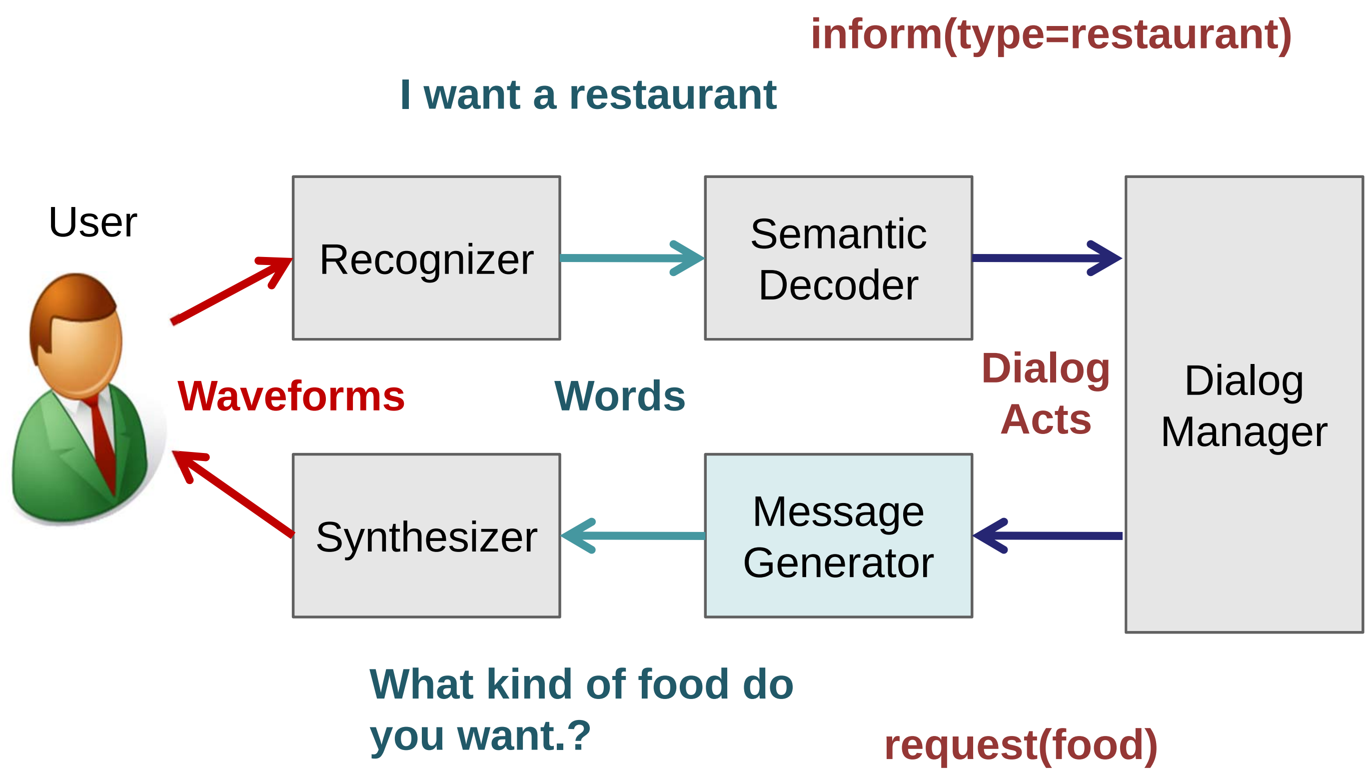
Human-machine spoken dialogue



Dialogue management

- The dialogue manager takes in what the user said and decides what to say back
- Split into two components:
  - State model (what has happened)
  - Policy (what to do)
- We will discuss these in detail in lectures 2&3

Human-machine spoken dialogue



Output generation - Templates

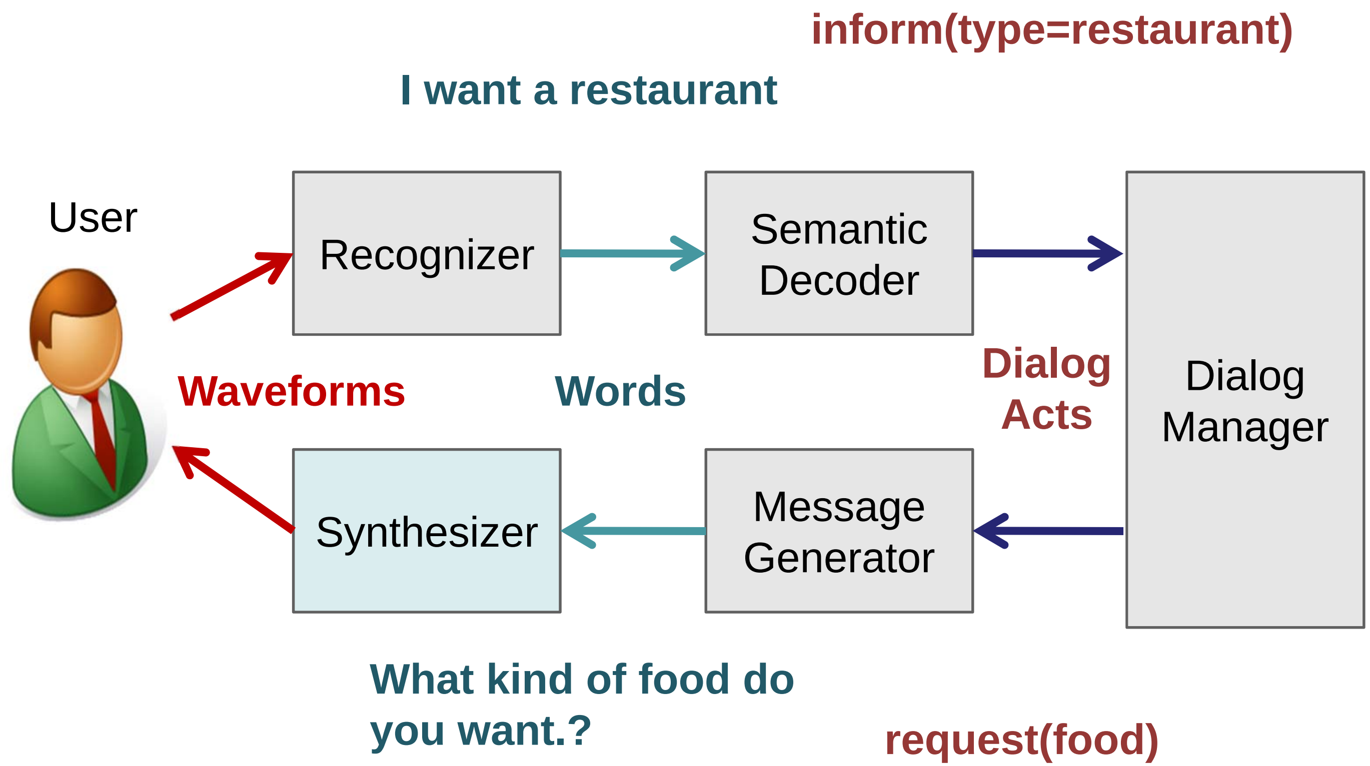
- To generate natural sentences, many systems use templates

  inform(name=$X, area=$Y) => “$X is in the $Y of town”

  inform(name=“Char Sue”, area=centre) =>
  “Char Sue is in the centre of town”

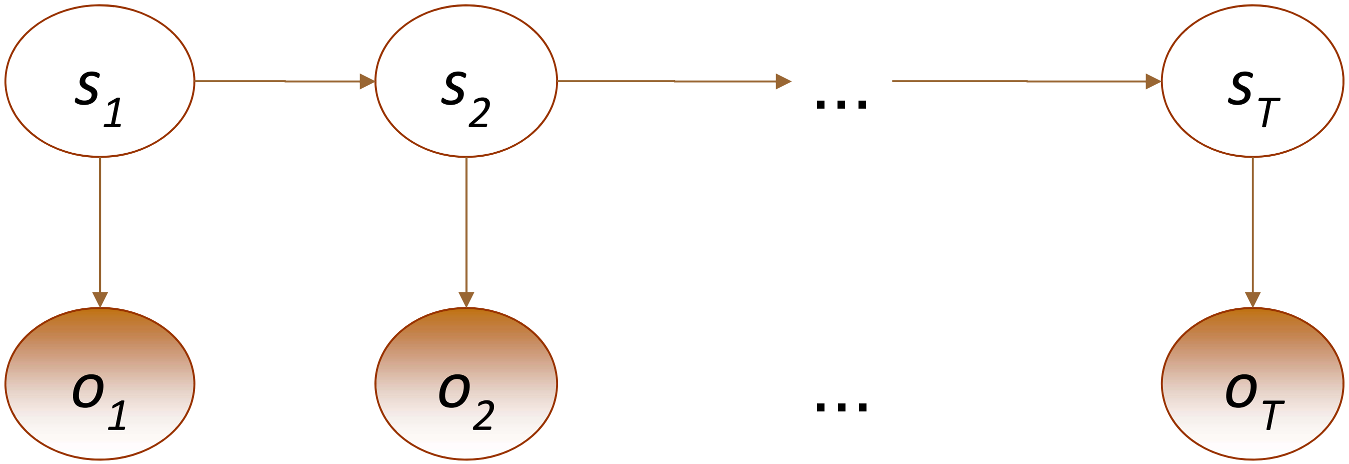
- Some work has been done on learning the generator
  - Overgenerate and rank (Langkilde & Knight 1998)
  - Bayesian Networks (Mairesse et al 2010)

Human-machine spoken dialogue



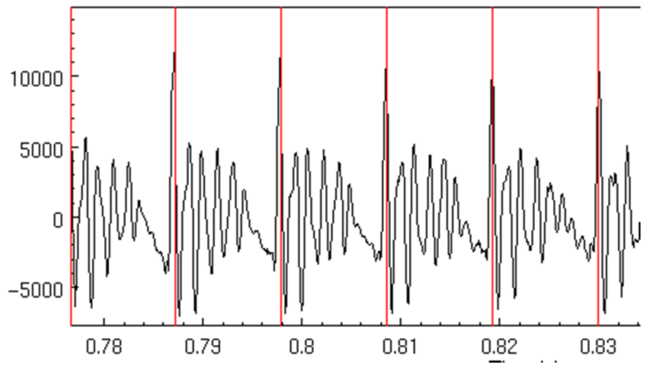
Text-to-speech – HMM approach

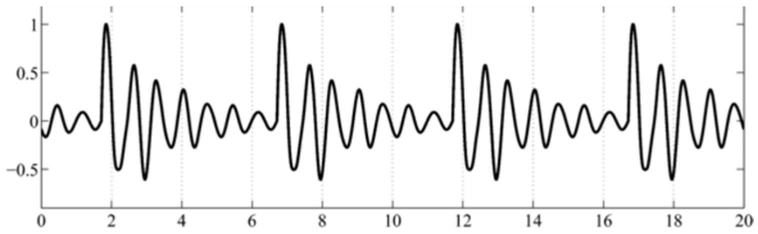
- Simply use the same Hidden Markov Model as in speech recognition
- Generate the most likely sequence
- Need to use slightly different observations so that we can recover the speech waveform



Text-to-speech – Unit selection

- Record lots of speech samples
- Stitch together segments and adjust pitch / duration
- e.g. Pitch Synchronous Overlap Add (PSOLA)





Text-to-speech: Summary

- Unit selection:
  - More natural sounding
  - Requires more data
  - Difficult to change for emotion / etc.

- HMM approach:
  - Less natural
  - Less data needed
  - Allows for change in emotion / etc

- Toolkits: Festival, Flite, HTS (for HMMs), DFKI MARY
- Commerical: Google, Microsoft SAPI, Nuance

Semantic decoding – some lab / home work

- Download a sample decoder at:
  - http://mi.eng.cam.ac.uk/~brmt2/clara.tar.gz
- Build / adapt your own decoder