What is Computational Linguistics?

- *"the scientific study of language from a computational perspective"*[1].

  Aims can be scientific: *"trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon"*[1].

  Or technological: *"to provide a working component of a speech or natural language system"*[1].

- Knowledge/rule-based ("symbolic", "hand-crafted", "top-down"):

  From linguistics (attempt to describe human linguistic capability)

  Began in the 50s (e.g. Chomsky)

- Data-driven ("stochastic", "statistical", "empirical", "bottom-up")

  From computer science / engineering (attempt to use & describe language as observed)

  Began at about the same time (e.g. Shannon)

- Developed almost separately from each other (philosophical differences – Chomsky apparently still claims corpus linguistics "does not exist").

- Last few years have seen increasing synthesis

[1] http://www.aclweb.org

Statistical NLP – Applications

Web Search Engine

"Bag of words" approach: remove common function words from search term, rank documents based on $tf.idf$ score:

$$tf.idf = (term\ frequency).(inverse\ document\ frequency)$$

$$= \left[\frac{\Sigma_{doc}(matches)}{\Sigma_{doc}(words)}\right] / \left[\frac{\Sigma_{corpus}(matches)}{\Sigma_{corpus}(words)}\right]$$

Can use PoS-tagging to rank words in order of importance (e.g. nouns/verbs above adjectives).

Speech Recognition

/sɪkskwɪd/ → *six quid*? Or *sick squid*?

Disambiguation requires some sort of linguistic knowledge.

Individual word probabilities might rule out "squid".

Even better to use context: "I've got . . . " → *six*, "I've got a . . . " → *sick*.

Can do this using n-grams (probabilities of words based on previous $n-1$ words) allow

$$p(got\ a\ sick) \propto p(sick|got\ a)$$

$$p(got\ a\ six) \propto p(six|got\ a)$$

When is statistical NLP not enough?

Question Answering

Extension of the search engine to answer open-domain questions from a large document corpus (like the web).

Q: *"Who is the president of the USA?"*

Attempting to match based on word matches is OK for simple examples:

A1: *"George W Bush is the president of the USA."*

But what do we do with more complicated ones:

A2: *"Bill Clinton used to be the president of the USA."*

A3: *"George W Bush, despite polling fewer votes than Al Gore, is the president of the USA."*

Syntactic structure required to identify answers, choose between possibilities.

Semantic knowledge required to rule out false answers.

Machine Translation initially might seem simple - translate word-for-word via a dictionary.

But what do we do with, say, case agreement:

"I bite the dog" → *den Hund*,

"the dog bites me" → *der Hund*

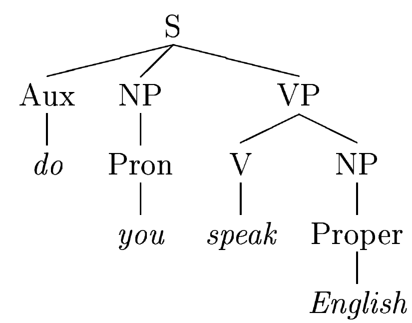
Could we use e.g. probabilities based on whether previous/next word is a verb?

"whoever I bite the dog still likes me", *"the cat next to the dog bites me"*, ...

Similar problems with coreference resolution: *"Give it to me"* → *ihn/sie/es*

Semantics: *"I'm going ..."* → *gehe/fahre/reise*

The Nature of Natural Language

Phonology (Intonation)	/ʤʉ spiˑk ɪŋglɪʃ/ H*L L*L L*LH H%	Sounds
Orthography	`do you speak English?`	Words
Syntax		Grammatical Structure
Semantics	*speak(you, English)?*	Logical Meaning
Pragmatics	*speak(* {*x1, x2,* . . .}, *English)?* *yes!* . . .	Contextualised or Implied Meaning

Syntactic Processing

Regular Grammars (Finite State Automata) (as commonly used in speech recognition - HMMs)

Valid parse is one that finishes in end state and uses exactly all words in string. Not expressive enough, and don't give (easily) useful output structure.

Context-Free Grammars (Push-Down Automata)

$$
\begin{array}{rcl@{\qquad\qquad}rcl}
S & \rightarrow & NP\ VP & Det & \rightarrow & the \\
NP & \rightarrow & Det\ N & N & \rightarrow & dog \\
VP & \rightarrow & V_{intran} & V_{intran} & \rightarrow & sleeps \\
VP & \rightarrow & V_{tran}\ NP & V_{tran} & \rightarrow & bites
\end{array}
$$

Valid parse is one that covers exactly all words in string using valid rules. Expressive power not bad, structure available, efficient parsing techniques (chart parsing), but can't express e.g. agreement easily.

Context-Sensitive Grammars (for later)

Attribute-Value Grammars

- With a CF grammar, agreement is difficult to capture:

the dog sleeps	* the dog sleep	* dog sleep
the dogs sleep	* the dogs sleeps	dogs sleep

$$\begin{array}{rcl} S & \rightarrow & NP_s\ VP_s \\ S & \rightarrow & NP_p\ VP_p \\ NP_s & \rightarrow & Det_s\ N_s \\ NP_p & \rightarrow & Det_p\ N_p \\ NP_p & \rightarrow & N_p \\ & \ldots & \end{array}$$

  Ends up with many hundreds of rules.

- We can capture generalisations with a AVG:

$$\begin{array}{rcl} S & \rightarrow & NP[num = N]\ VP[num = N] \\ NP[num = N] & \rightarrow & Det[num = N]\ N[num = N] \\ NP[num = p] & \rightarrow & N[num = p] \end{array}$$

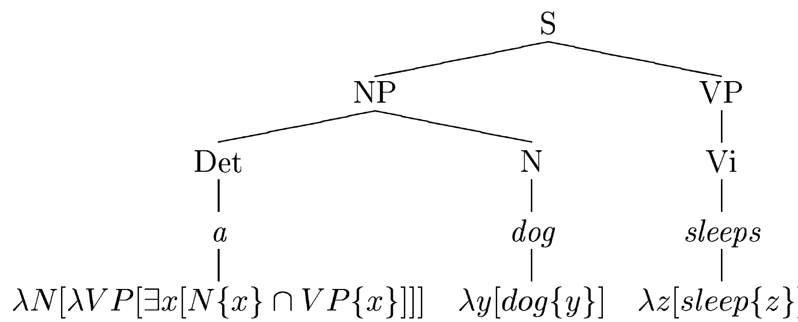
- Still fails to capture many generalisations about rules.

Semantic Representation

- Montague Semantics

  First order logic, lambda calculus - familiar.

  Compositional, easy to integrate with syntactic parser:



  Give each word semantic representation, and add a semantic application rule to the parser to build constituent representations:

$$\begin{array}{lcll} A & \rightarrow & B\ C & : B(C) \\ NP & \rightarrow & Det\ N & : \lambda VP[\exists x[dog\{x\} \cap VP\{x\}]] \\ S & \rightarrow & NP\ VP & : \exists x[dog\{x\} \cap sleep\{x\}] \end{array}$$

  Model-theoretic interpretation (possible worlds, intension) to determine truth values.

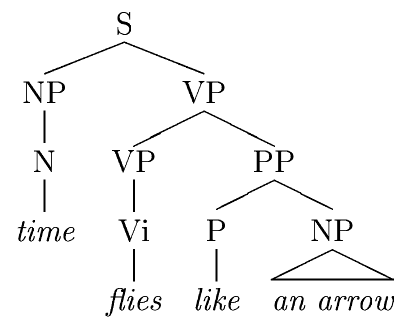
  But what to do with coreference? And what about hyperintension (logical omniscience)?

- Dynamic Predicate Logic, Discourse Representation Theory, Situation Semantics
- Shalom, Howard & Christian . . .

Statistics again - Parse Disambiguation

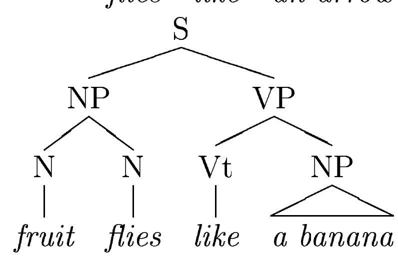
- Structural ambiguity: *"X flies like a Y"*

"Time flies like an arrow"



fly(time) & similar(time, arrow)

"Fruit flies like a banana"



like(fruit_flies, banana)

- Worse than you might think - (Martin et al. 1987) got 455 parses for:
  "List the sales of the products produced in 1973 with the products produced in 1972."
- Either use real-world knowledge (as we do?) or (much simpler) probabilistic methods.

Stochastic CFGs

- Associate each rule with a probability, e.g.:

$$\begin{array}{llll} S & \rightarrow & NP\ VP & 1.0 \\ NP & \rightarrow & N & 0.2 \\ NP & \rightarrow & Det\ N & 0.5 \\ NP & \rightarrow & N\ N & 0.3 \\ \ldots & & & \end{array}$$

  Parser calculates total probability of each parse, returns most probable structure.

- Easy to train from hand-annotated corpora using MLE (count number of occurrences of each rule)

$$p(A \rightarrow B_i) = \sum_s \frac{n_s(A \rightarrow B_i)}{\sum_x n_s(A \rightarrow B_x)}$$

- Can be trained from unmarked corpora using EM and the Inside-Outside algorithm (generalisation of speech recognition techniques)
- May need normalisation on parse length (higher number of rules → lower p)
- Problem with Zipfian nature of NL - all unseen structures assigned zero probability, and estimated probabilities for rare structures may be inaccurate. Can use smoothing and back-off techniques to help.
- But this wouldn't help with *"fruit flies . . . "*

Advanced SCFGs, Stochastic AVGs

- Make rule probabilities dependent on daughter categories (SEDCFG), or words (Lexicalised SCFG), or both e.g.:

$$\begin{array}{lll} NP & \rightarrow & N_{time}\ N_{flies} \quad 0.005 \\ NP & \rightarrow & N_{fruit}\ N_{flies} \quad 0.09 \\ \ldots & & \end{array}$$

- Training methods similar to SCFGs, but sparsity of training data becomes a serious problem, especially with lexicalised versions.
- We can extend this probabilistic approach to AVGs - can even condition rule probability on attribute values, e.g.:

$$\begin{array}{llll} NP[num = s] & \rightarrow & N[num = s] & 0.1 \\ NP[num = s] & \rightarrow & Det[num = s]\ N[num = s] & 0.9 \\ NP[num = p] & \rightarrow & N[num = p] & 0.5 \\ NP[num = p] & \rightarrow & Det[num = p]\ N[num = p] & 0.5 \\ \ldots & & & \end{array}$$

- Problem with training of AVGs - missing probability mass associated with unseen rules. Other techniques proposed (Abney 1997 - random fields & statistical sampling). Does it matter in practice?

Dialogue Systems

Wizard:	Welcome to the route planning service. How may I help you?
Caller:	Hello, um, well to plan a route really.
Wizard:	Where would you like to go?
Caller:	From Malvern
Wizard:	Mm
Caller:	to Kirbythore in Cumbria.
Wizard:	Can you spell that please?
Caller:	Yeah – K-I-R
Wizard:	Yeah
Caller:	B-Y
Wizard:	Yeah
Caller:	Uh – T-H-O-R-E
Wizard:	T-H-O-R-E?
Caller:	Yeah. It's near Penrith.

Statistical Approach Used in some applications today. Depends on detailed knowledge of domain (e.g. sentence including *"route"* → *route planning subroutine*, sentence including *"from"* → *destination value*).

Linguistic Approach required for true open-domain application.

Need *context* for anaphora, ellipsis, clarifications.

Also need interface to *phonology*, and need to identify speech acts / conversational moves.

Must also handle (amongst others) illocutionary force, grounding, hesitation.

Head-driven Phrase Structure Grammar (HPSG)

- (Pollard & Sag 1992, Ginzburg & Sag 2000) Integrates phonology, syntax, semantics and context in an attribute-value matrix or *sign*.

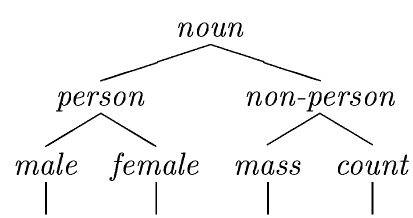
$$\begin{bmatrix} \text{PHON} & \text{sleeps} \\ \text{CAT} & \begin{bmatrix} \text{HEAD} & \begin{bmatrix} \textit{verb} \\ \text{NUM} & \boxed{1}\ \textit{sing} \end{bmatrix} \\ \text{SPR} & \left\langle \begin{bmatrix} \text{CAT}|\text{HEAD} & \begin{bmatrix} \textit{noun} \\ \text{CASE} & \textit{nom} \\ \text{NUM} & \boxed{1} \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{INDEX} & \boxed{2} \end{bmatrix} \end{bmatrix} \right\rangle \\ \text{COMPS} & \langle\ \rangle \end{bmatrix} \\ \text{CONT} & \begin{bmatrix} \text{SIT} & \boxed{3} \\ \text{SOA} & \begin{bmatrix} \text{RELN} & \textit{sleep} \\ \text{AGENT} & \boxed{2} \end{bmatrix} \end{bmatrix} \\ \text{CTXT} & \begin{bmatrix} \text{MAX-QUD} & \\ \text{SAL-UTT} & \\ \text{C-INDICES} & \langle \boxed{2}, \boxed{3}, \ldots \rangle \end{bmatrix} \end{bmatrix}$$

HPSG (2)

- Highly generalized rule *schemata* over phrasal types (instead of individual rules over phrasal categories), e.g.:

$$\begin{bmatrix} \text{CAT} & \begin{bmatrix}\text{HEAD} & \boxed{1}\\ \text{SPR} & \langle\,\rangle \\ \text{COMPS} & \boxed{2}\end{bmatrix} \\ \text{CONT} & \boxed{3}\end{bmatrix} \rightarrow \boxed{4} \begin{bmatrix} \textit{head-spr-phr} & \\ \text{CAT} & \begin{bmatrix}\text{HEAD} & \boxed{1}\\ \text{SPR} & \langle\boxed{4}\rangle \\ \text{COMPS} & \boxed{2}\langle\,\rangle\end{bmatrix} \\ \text{CONT} & \boxed{3}\end{bmatrix}$$

- Highly lexicalised. Default inheritance in the lexicon allows information to be specified in a type hierarchy.



- Allows use of information & constraints from other levels (e.g. syntax ↔ semantics), including contextual information.
- Stochastic version is (theoretically) possible (Brew 1995) – associate probabilities with rule schemata / phrasal types – but is it trainable?

Information States

- We can represent the information state for a conversational participant as an AVM (Ginzburg 1992).

  Can include dialogue context, beliefs, future plans:

$$\begin{bmatrix} \text{COM} & \{plan(route), depart(Malvern)\} \\ \text{QUD} & \langle spell(X,Y)?, destination(X)?\rangle \\ \text{LATEST-MOVE} & spell(X, `KIRBYTHORE') \\ \text{PLAN} & \langle check(Y, `KIRBYTHORE'), ask(date), \ldots\rangle \end{bmatrix}$$

- We also need some sort of update algorithm, e.g.:

```
if ( LATEST-MOVE == ask( Q? ) )
then
    push( Q?, QUD );
end;
if ( LATEST-MOVE == assert( P ) )
then
    if ( top( Q?, QUD ) and answers( P, Q? ) )
    then
        pop( Q?, QUD );
        push( P, COM );
    end;
end;
```

- This gives us a way to calculate a participant's beliefs and the questions being discussed - i.e. a way to build *context*.

Ellipsis

- Elliptical expressions are incomplete fragments which can only be fully interpreted in context. Examples include *"Yes"*, *"From Malvern"*, *"T-H-O-R-E?"*.
- Using HPSG we can use dialogue context to fully interpret these expressions, e.g.:

$$\begin{bmatrix} \text{PHON} & \text{yes} \\ \text{CONT} & \begin{bmatrix} \textit{proposition} & \\ \text{SIT} & \boxed{1} \\ \text{SOA} & \begin{bmatrix} \text{RELN} & \textit{true} \\ \text{PROP} & \boxed{2} \end{bmatrix} \end{bmatrix} \\ \text{CTXT} & \begin{bmatrix} \text{MAX-QUD} & \boxed{2} \\ \text{C-INDICES} & \langle \boxed{1} \rangle \end{bmatrix} \end{bmatrix}$$

- SHARDS is a working implementation (`http://pc320.dcs.kcl.ac.uk:8080/`).
- Other more complex types of ellipsis include bare answers, sluices, VP ellipsis, and clarification ellipsis.

Clarification Ellipsis

- Clarifications make up about 3% of dialogue, although this rises in situations where accuracy of understanding is important (e.g. Map Task more like 10%).
- Correct interpretation is important:

Caller	I'd like to go to Kirbythore
System	OK. Would you like to go via Preston or Malvern?
Caller	Malvern?

S	I'm sorry, I don't understand, please repeat
C	Malvern?
S	I think you would like to go to Malvern
C	Aaaah

S	OK. When would you like to go?
C	What? Go where?
S	I'm sorry, I don't understand, please repeat
C	Aaaah

S	Malvern is a small town in Shropshire
C	Ah, OK, via Preston then.
S	OK. When would you like to go?
	...

- But what meaning(s) can we assign to CE? And how can we disambiguate between CE and other lexically identical moves?

Clarification Ellipsis

- At least two possible meanings: a “check” type (*“Did you really say X?”*) and a “reference” type *“Who/what/where is X?”*.

$$\begin{bmatrix} \text{PHON} & \text{malvern} \\ \text{CONT} & \begin{bmatrix} \textit{question} & \\ \text{PROP} & \textit{ask}(\boxed{1}, \boxed{2}) \end{bmatrix} \\ \text{CTXT} & \begin{bmatrix} \text{SPKR} & \boxed{1} \\ \text{MOVE} & \boxed{2}\textit{go(via malvern)?} \end{bmatrix} \end{bmatrix} \qquad \begin{bmatrix} \text{PHON} & \boxed{3}\ \text{malvern} \\ \text{CONT} & \begin{bmatrix} \textit{question} & \\ \text{Q-INDEX} & \boxed{4} \\ \text{PROP} & \textit{content}(\boxed{5}, \boxed{4}) \end{bmatrix} \\ \text{CTXT} & \begin{bmatrix} \text{SAL-UTT} & \boxed{5}\begin{bmatrix} \text{PHON} & \boxed{3} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

- There are other possible meanings (e.g. “lexical gap”), but what are they?
- There are other possible forms (e.g. reprise sluices *“Where?”*), but what are they?
- Disambiguation between meanings and from other move types: could use information state (first mention of X?), intonation, relation between form and meaning.
- That’s it.