Communication: Natural Language Processing

- Communication = action
  - INFORM: “There is a wumpus in (2,2).”
  - QUERY: “Is there a pit in (1,2)?”
  - REQUEST: “Please help me carry the gold”
  - ACKNOWLEDGE: “OK”
  - PROMISE: “I’ll shoot the wumpus; you grab the gold.”
  - REQUEST to INFORM: “Tell me if you smell a stench”

Language utterances

- Computer languages can attach semantics directly to the symbols
  - x = 23;
- Natural languages are fragments of information sufficient to allow the hearer to determine what is meant.
  - “Can you reach the salt?”
  - “Let’s vote them off the island.”

NLP is the Hardest AI Problem

- "After John proposed to Mary, they found a preacher and got married. For the honeymoon, they went to Hawaii"
  - Who got married? Who went to Hawaii?
- Jane told Sue she was going to get Mike a kite for his birthday. Sue said, "Don't! He already has one. He will make you take it back."
  - What does "it" refer to? Which kite will be taken back?

Why NLU is hard

- Language can be about all aspects of human affairs
  - love and death, hopes and fears, pride and embarrassment
  - the intricacies of social, religious and political institutions
  - times and places, real and imaginary
- Understanding natural language requires the ability to represent and reason with knowledge about all of these things

NLP Tasks (1)

- Man-machine dialogue during problem solving

  "Open the pod bay doors, HAL"

  "Make a copy of this PPT file, change it to be black on white background, make a PDF file, and post it on the course web page."

  "Show me what houses you have for sale. What is the nearest school to that one? (pointing)"

NLP Tasks (2)

- Language Translation

  Universal translator that you wear like an earring?

- Information retrieval

  "Find all papers published in the medical literature on AIDS vaccines"

  "Has anyone else experienced occasional pauses in Powerpoint under XP?"

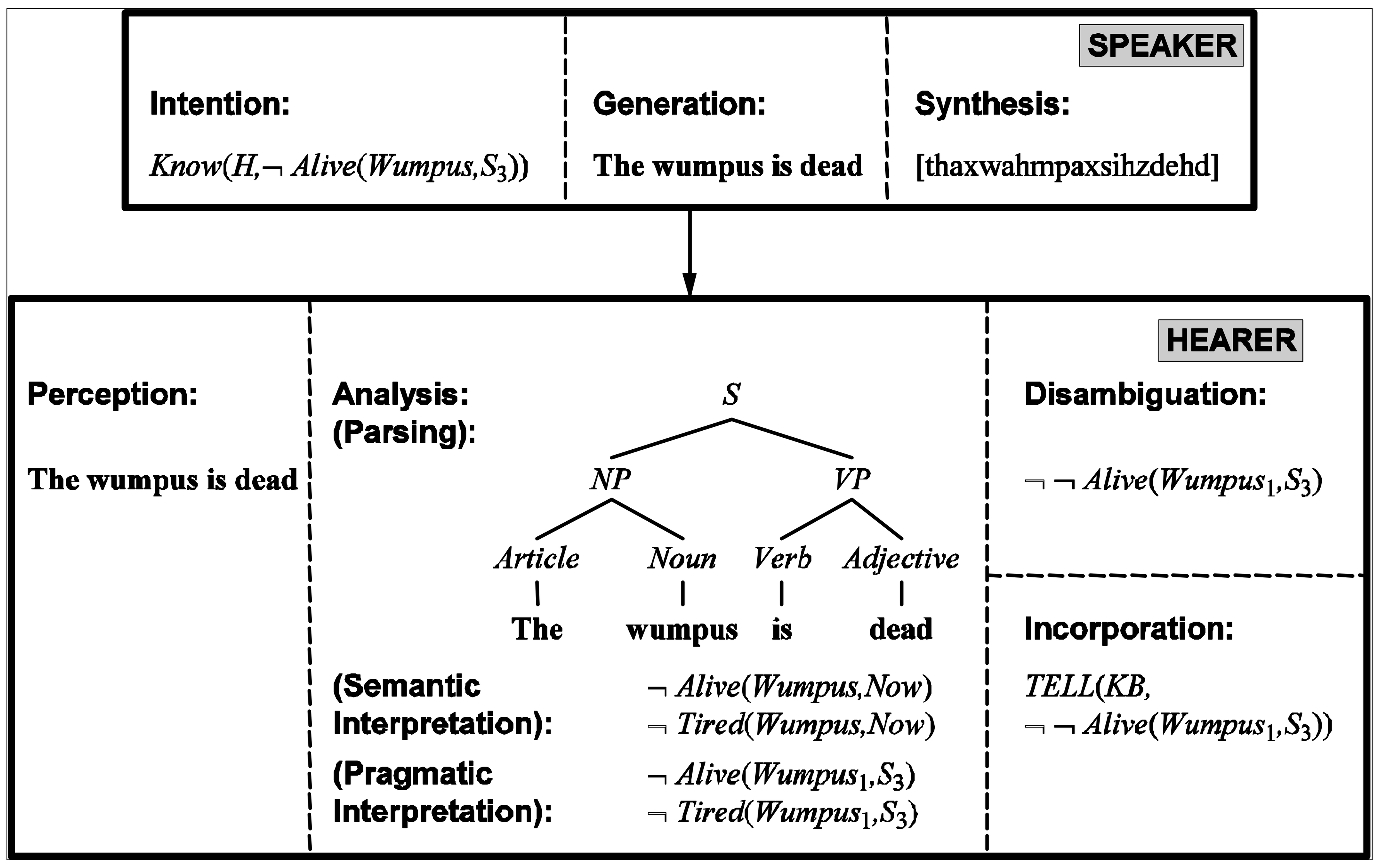
NLP Tasks (3)

- Information Extraction
  - Flipdog.com, monster.com: Spider the web and extract job ads. Build a database of all known job positions and allow searching

Phases/Levels of NLP

- Intention: Know(H, : Alive(Wumpus,t_3))
- Generation: “The wumpus is dead.”
- Synthesis: [th][ax][w][ah][m][p][ax][s][ih][z][d][eh][d]
- Perception: “The wumpus is dead”
- Analysis: set of alternative meanings
- Disambiguation: figuring out which meaning is correct
- Incorporation: believing the result

Communication



Analysis and Disambiguation

- Parsing
- Semantic interpretation
- Pragmatic interpretation
- Disambiguation
- Discourse analysis

Parsing

- Grammars
  - Context-free grammars
  - Definite Clause Grammars

Context-Free Grammar E_0

S → NP VP	I + feel a breeze
| S conjunction S	I feel a breeze + and + I smell a wumpus
NP → Pronoun	I
| Name	John
| Noun	pits
| Article Noun	the + wumpus
| Digit Digit	3 4
| NP PP	the wumpus + to the east
| NP RelClause	the wumpus + that is smelly
VP → Verb	stinks
| VP NP	feel + a breeze
| VP Adjective	is + smelly
| VP PP	turn + to the east
| VP Adverb	go + ahead
PP → Preposition NP	to + the east
RelClause → **that** VP	that + is smelly

Lexicon

Noun → stench | breeze | glitter | nothing | agent | wumpus | pit | pits | gold | east | …

Verb → is | see | smell | shoot | feel | stinks | go | grab | carry | kill | turn | …

Adjective → right | left | east | dead | back | smelly | …

Adverb → here | there | nearby | ahead | right | left | east | south | back | …

Pronoun → me | you | I | it | …

Name → John | Mary | Boston | Aristotle | …

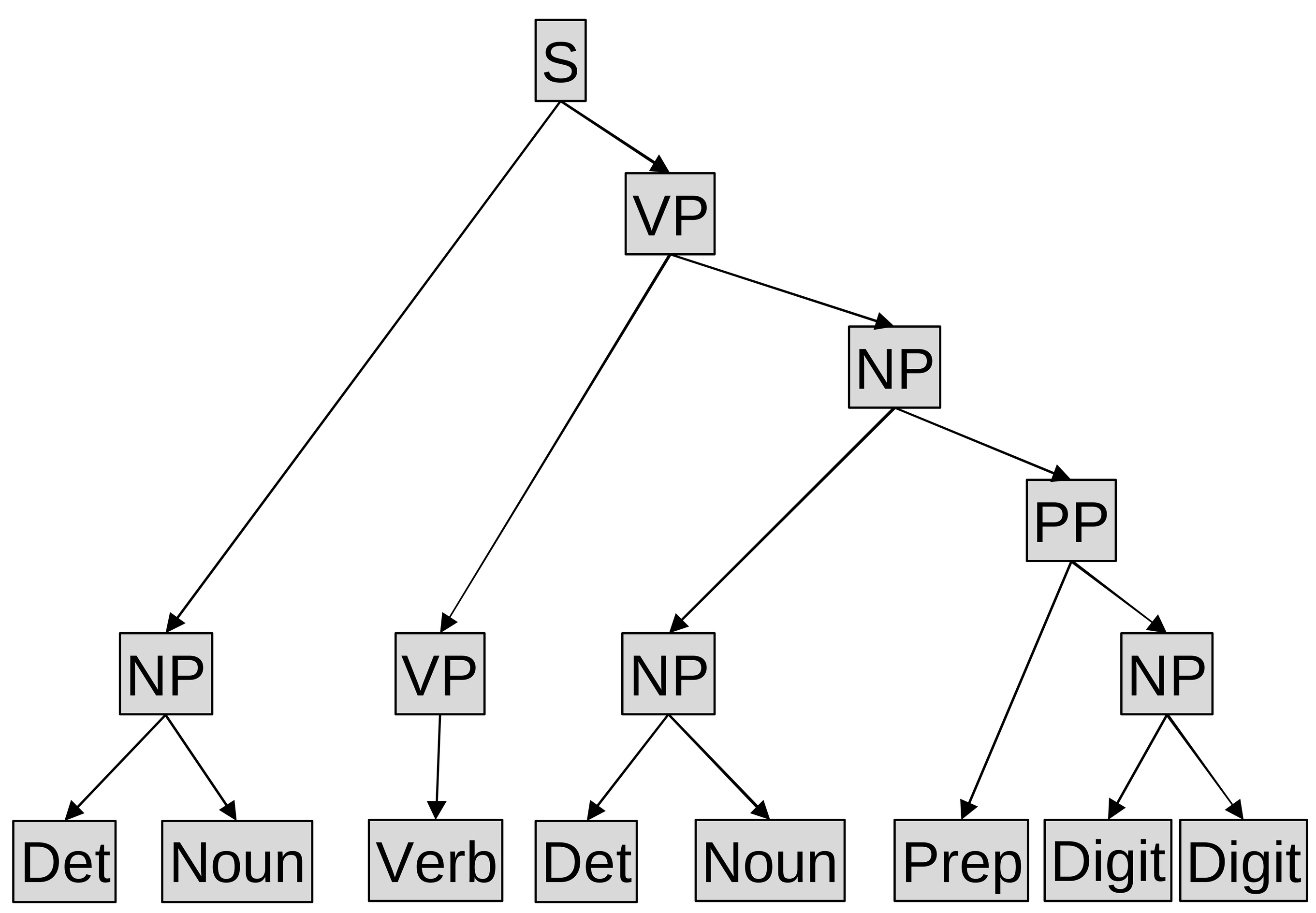
Article → a | the | an | …

Preposition → to | in | on | near | …

Conjunction → and | or | but | …

Digit → 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9

Parsing



The arrow killed the wumpus in 4 4

Parsing Natural Language

- Computer languages use restricted context-free grammars that can be parsed efficiently
  - LR(1), LL(1)
- General CFG requires $O(n^3)$ time
  - Chart parser: mixed top-down and bottom-up parsing based on dynamic programming

Problems with our grammar

- Overgeneration
  - “Me smell a wumpus”
  - “Go me the gold”
  - “Give to 1 2”
- We want some kinds of type restrictions or rules of agreement

Augmented Grammars
Add arguments to non-terminals

- Noun Cases

  Noun(subject) → I

  Noun(object) → me

  Noun(_) → arrow | wumpus | …

  S → NP(subject) VP

  VP → VP NP(object)

  NP(case) → Noun(case)

Verb Subcategories: restrictions on VP parts

Verb	Subcats	Example
give	[NP,PP] [NP,NP]	give the gold to me give me the gold
smell	[NP] [Adjective] [PP]	smell a wumpus smell awful smell like a wumpus
is	[Adjective] [PP] [NP]	is smelly is in 2 2 is a pit
died	[]	died
believe	[S]	believe the wumpus is dead

Adding subcategories to the lexicon and grammar

Verb([NP,PP]) → give | hand | …

VP(subcat) → Verb(subcat)

| VP(subcat + [NP]) NP(object)

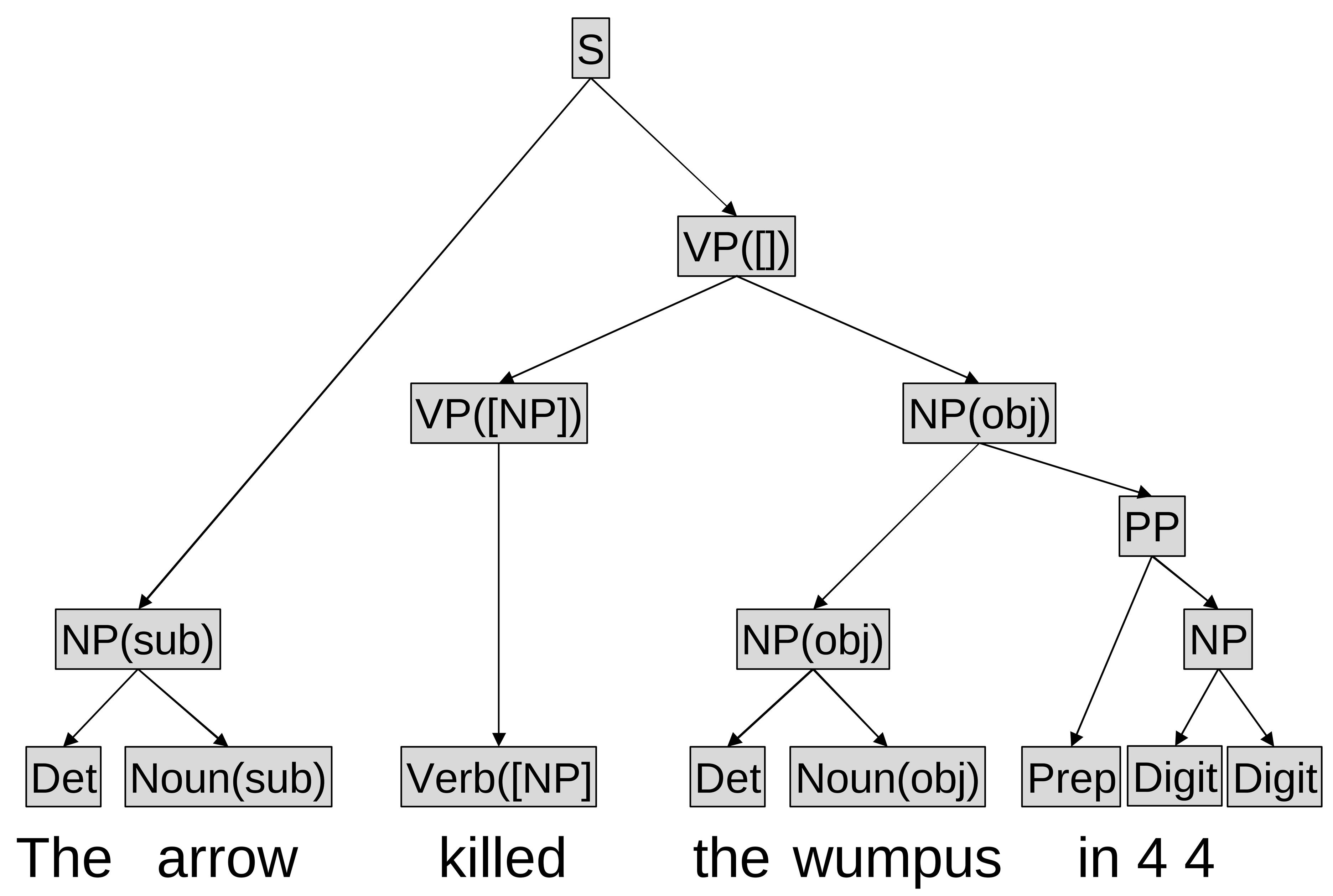
| VP(subcat + [Adjective]) Adjective

| VP(subcat + [PP]) PP

S → Noun(subject) + VP([ ])

This can all be implemented easily using Prolog! In fact, Prolog was invented for this purpose.

Revised Parse



Semantic Interpretation

- Idea: Attach quasi-logical formula to each grammar rule to represent the meaning
- Each rule *composes* the meanings of the non-terminals on the rhs to produce the meaning of the non-terminal on the lhs.

Semantic augmentations

S(rel(obj)) → NP(obj) VP(rel)

VP(rel(obj)) → Verb(rel) NP(obj)

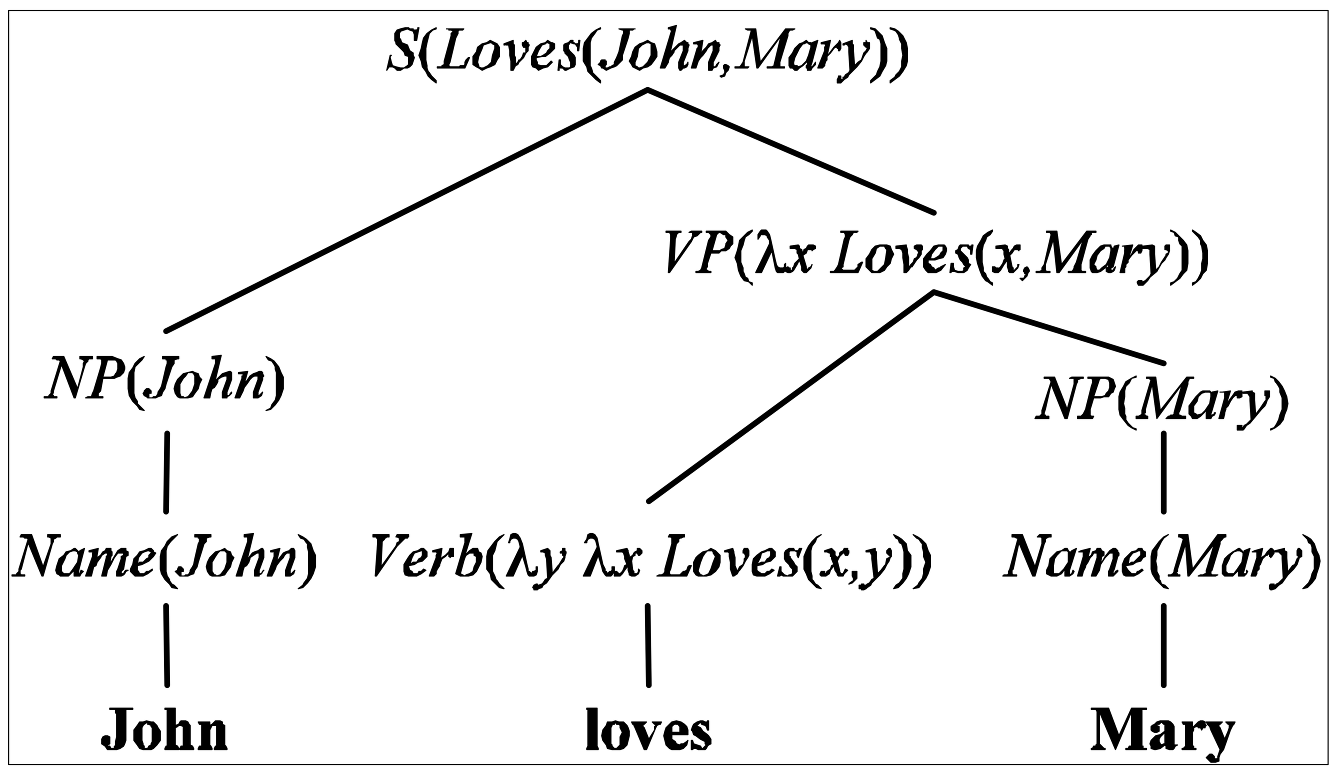
NP(obj) → Name(obj)

Name(John) → **John**

Name(Mary) → **Mary**

Verb(λx λy Loves(x,y)) → **loves**

Compositional Semantics: Use lambda application



(λy λx Loves(x,y)) Mary == λx Loves(x,Mary)

(λx Loves(x, Mary)) John == Loves(John,Mary)

Complications

- Temporal analysis
  - “John loves Mary”
  - “John loved Mary”
- Quantification
  - “Every agent smells a wumpus”
    - Is there just one wumpus?
    - 8a2Agents 9w2Wumpuses smells(a,w)
    - 9w2Wumpuses 8a2Agents smells(a,w)

More Complications

- Indexicals
  - “I” denotes the speaker
  - “today” denotes the day in which the sentence was spoken

Disambiguation

- Syntactic and Semantic analysis generally produces multiple candidate interpretations
- Disambiguation attempts to rule out incorrect interpretations and find the correct one

Ambiguities

- Squad helps dog bite victim
- Helicopter powered by human flies
- British left waffles on Falkland Islands
- Teacher strikes idle kids
- Drunk gets nine months in violin case

Almost every sentence has multiple interpretations

- "The batter hit the ball."
  - What just happened in the Mariners' game?
  - How did this ball get so sticky?
  - The mad scientist unleashed a tidal wave of cake mix towards the ballroom (!)

Syntactic Ambiguities

- Natural languages are syntactically ambiguous (one sentence can have multiple legal parses)
- "Teacher strikes idle kids"
  - [S [NP teacher][VP strikes [NP [Adj Idle][N Kids]]]]
  - [S [NP [Adj teacher][N strikes]][VP [V idle][NP [N kids]]]]

Semantic Ambiguities

- bank:
  - financial institution
  - part of a river
  - kind of hockey shot

Non-Literal Language

- Metonymy: part-for-whole
  - "Chrysler announces a new model"
    - companies can't talk
    - a company spokesman made the announcement
  - "The Red Sox need a strong arm"
    - they actually need the entire pitcher
- Metaphor
  - "The popularity of botox has jumped"
    - jump → move upwards → increase

Disambiguation = Reasoning under uncertainty

- $\text{argmax}_{\text{interp}}$ P(interp| words, situation)
- How do we compute P(interp | words…)?
  - World model: could this happen in the world? (sales don't jump; teachers are unlikely to strike students)
  - Mental model: would the speaker have meant this?
  - Semantic language model: would the speaker have chosen these words if he meant this?

Formalizing and reasoning with these models is the key bottleneck to natural language understanding

Discourse Understanding

- Understanding multiple sentences
  - provides additional constraint for disambiguation
- Sentences in a discourse are related to one another. These relationships can be identified and exploited

Resolving Pronoun References: An Example

- "Dana dropped the cup on the plate. <u>It</u> broke."
- What is the referent of "it"?

The whole discourse

“Dana was quite fond of a special blue cup. The cup had been a present from a close friend. Unfortunately, one day while setting a place at the table, Dana dropped the cup on the plate. It broke.”

- The first sentence introduces a “focus space” in which the “cup” is the main focus. The cup is mentioned again, which reinforces the focus.

Example Discourse

1. A funny thing happened yesterday
2. John went to a fancy restaurant
3. He ordered the duck
4. The bill came to $50
5. John got a shock when he realized he had no money
6. He had left his wallet at home
7. The waiter said it was all right to pay later
8. He was very embarrassed by his forgetfulness

Discourse Coherence Relations

1. A funny thing happened yesterday
   - Introduces new “focus space” and Evaluates it
2. John went to a fancy restaurant
   - Enables 3.
3. He ordered the duck
   - Causes 4.
4. The bill came to $50
   - 2-4 serve as “Ground” for the rest of the story; implies John ate the duck
5. John got a shock when he realized he had no money
6. He had left his wallet at home
   - Explains 5. 5-6 enable 7
7. The waiter said it was all right to pay later
   - 5-7 cause 8
8. He was very embarrassed by his forgetfulness

Resolving Pronoun References

1. A funny thing happened yesterday
2. John went to a fancy restaurant
3. <u>He</u> ordered the duck {John, restaurant}
4. The bill came to $50
5. John got a shock when <u>he</u> realized <u>he</u> had no money {shock, John, $50, bill, duck, …}
6. <u>He</u> had left <u>his</u> wallet at home {shock, John, …}
7. The waiter said <u>it</u> was all right to pay later
8. <u>He</u> was very embarrassed by <u>his</u> forgetfulness {waiter, home, wallet, money, shock}

Natural Language Summary

- Statements in natural language are communications actions
- Natural Language processing must exploit many constraints:
  – meanings of individual words (lexicon)
  – grammatical constraints (including case roles and verb subcategories)
  – discourse coherence constraints
  – language model
  – speaker model
  – world model
- We have reasonably good formalisms for all of these except the world model

Task-Specific Natural Language Processing

- Information Retrieval
- Information Extraction
- Language Translation

Evaluating Information Retrieval Methods

- For standard classification problems, we use false positives (FP) and false negatives (FN) to evaluate learning

Predicted Class	True Class	
	spam	nonspam
spam	TP	FP
nonspam	FN	TN

For Information Retrieval we have Precision and Recall

- Suppose we have a document collection containing R relevant documents out of N total documents.
- A particular IR system will choose M documents to retrieve and present to the user. Suppose only K of these are relevant
- Precision: K/M = fraction of retrieved documents that are relevant
- Recall: K/R = fraction of all relevant documents that are retrieved

TR=true relevant;
FR=false relevant;
FI=false irrelevant;
TI=true irrelevant;
Precision = TR/(TR+FR)
Recall = TR/(TR+FI)

Retrieved?	True Class	
	relevant	irrelevant
yes	TR	FR
no	FI	TI

Precision and Recall

- What is more important?
  - Finding one relevant document ➔ high precision
    - Google: "I'm feeling lucky"
  - Finding all relevant documents ➔ high recall
- Different users and applications have different goals

Information Extraction from the Web

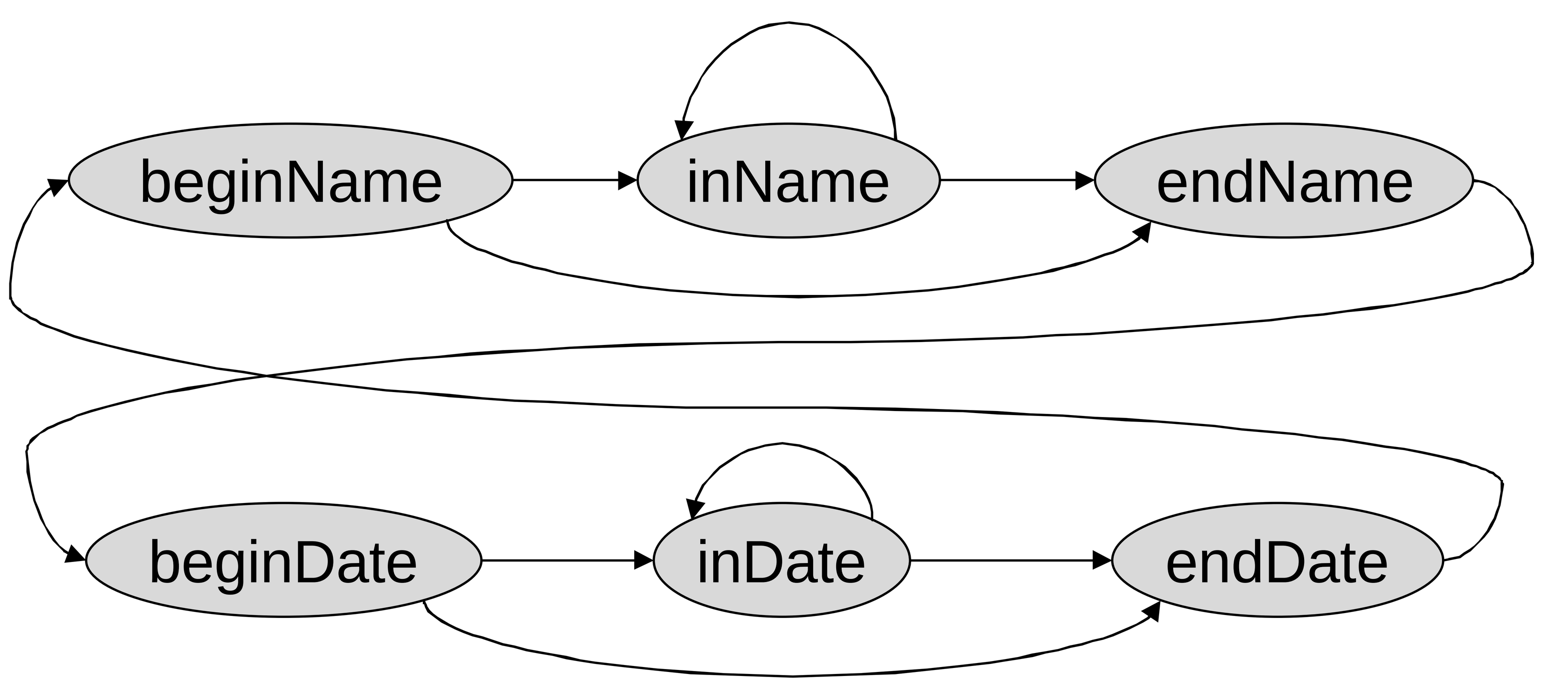
<dl><dt>Srinivasan Seshan (Carnegie Mellon University) <dt><i>Making Virtual Worlds Real</i><dt>Tuesday, June 4, 2002<dd>2:00 PM , 322 Sieg<dd>Research Seminar

* * * name name * * affiliation affiliation affiliation * * * * title title title title * * * date date date date * time time * location location * event-type event-type

name: Srinivasan Seshan
affiliation: Carnegie Mellon University
title: Making Virtual Worlds Real
date: Tuesday, June 4, 2002
time: 2:00pm
location: 322 Sieg
event-type: Research Seminar

HMM Parsing

- 8 classes
- usually modeled by 3 states for each class: {beginX, inX, endX}



HMM Parse

<dl><dt>Srinivasan Seshan (Carnegie Mellon University) <dt><i>Making Virtual Worlds Real</i><dt>Tuesday, June 4, 2002<dd>2:00 PM , 322 Sieg<dd>Research Seminar

beginX →	inX →	endX →	beginName →	endName →	beginX →	endX →	beginAffil
↓	↓	↓	↓	↓	↓	↓	↓
<dl>	<dt>		Srinivasan	Seshan		(	Carnegie

Language Translation

- Translation Tasks
  - Rough Translation
    - Get the “gist” of a passage
    - Can be ungrammatical (e.g., web surfing; emergency communications)
  - Restricted Source Translation
    - weather, travel
  - Preedited Translation
    - Original is written in restricted vocabulary and grammar so that it can be easily translated: “Caterpillar English”, Xerox manuals
  - Literary Translation
    - All nuances of text preserved.

Language Translation (2)

- Example: Systran (Altavista)
  - English → Italian → English

In chapter 22, we saw how an agent could communicate with another agent (human or software), using utterances in a common language. Complete syntactic and semantic analysis of utterances is *necessary* to extract the full meaning of the utterances, and is *possible* because the utterances are short and restricted to a limited domain

In chapter 22 we have seen as an agent could communicate with an other agent (to be human or software) that using the expressions in a language mutual come to an agreement. Complete syntactic and the semantic analysis of the expressions is necessary to extract the complete meant one of the utterances and is possible because the expressions short and are limited to a dominion

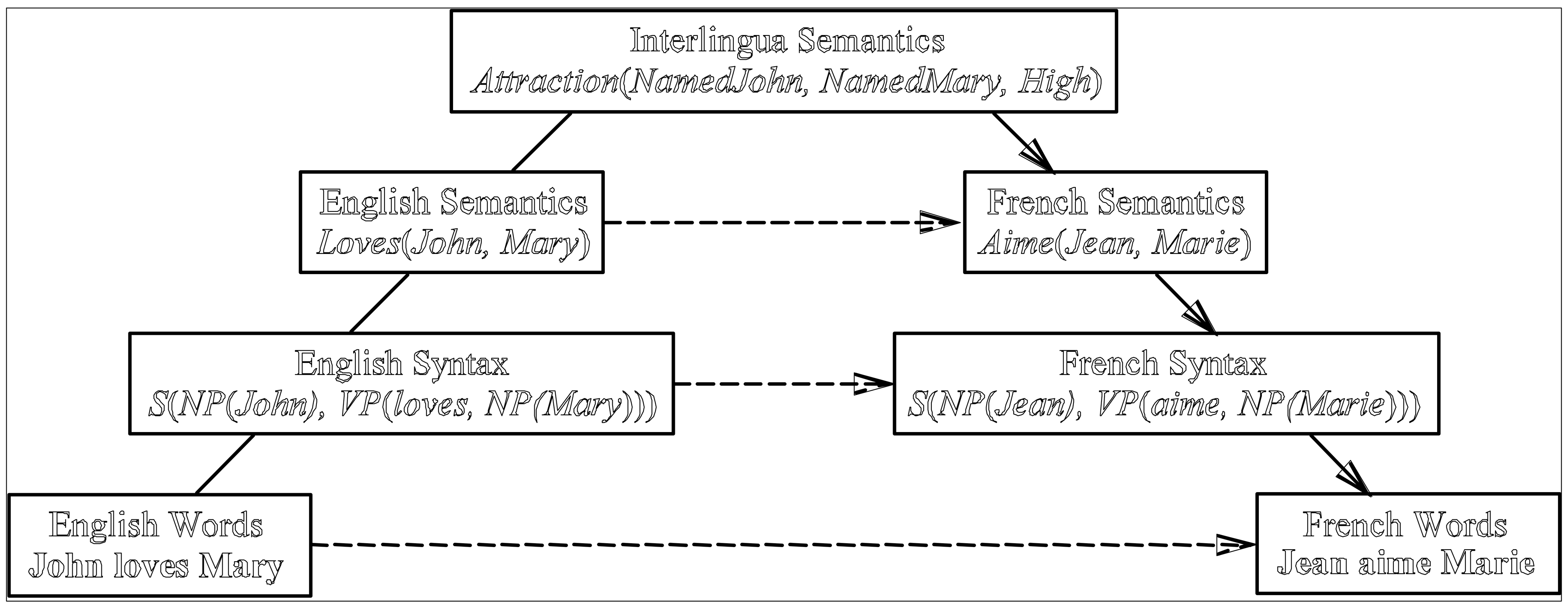
Problem 1: Grammars are Different

- English: “brown dog” → French “chien brun” (adjectives come after nouns)
- English: “I can come at 3pm” → German “ich kann um drei Uhr kommen.” (verb moves to the end)

Problems: Conceptual Categories Don’t Match

- “you” in English could be “tu” or “vous” in French
  - “tu”: for close friends and family
  - “vous”: for everyone else
- “doux” in French could mean “soft”, “sweet”, or “gentle” in English
  - English generally has more words than other languages, and therefore makes more distinctions

Four “levels” for Machine Translation



Rule-Based Translation: SYSTRAN

- Rules map sequences of English words to sequences of French words
  - Some rules can operate on single words
  - Other rules must match word sequences in English and produce word sequences in French
- Major hand-engineering effort
- Currently the most successful approach

Grammar-based translation

- Parse English sentence
- Apply rules to map from English parse tree to French parse tree
- Map the words by using English word and French parsing context

Mapping at the Semantic Level

- Parse English text and perform Semantic Analysis
- Apply rules to map English semantics to French semantics (possibly looking at English parse tree and words to help)
- Generate French sentence from French semantics

Interlingua

- Semantic representation that makes all distinctions necessary across both languages
- Generally only feasible in limited domains
- Parse English into Interlingua
- Generate French from Interlingua
- Advantage: Each language can be handled separately: $O(n)$ vs. $O(n^2)$

Statistical Machine Translation

- bilingual "corpora"
  - Hansards: record of parliamentary debate. produced in multiple languages in Canada, Hong Kong, the EU, and the UN.

A Simple Probabilistic Model (IBM 3)

- Goal: argmax_F P(F|E)
  – Mostly likely French sentence given English sentence
- argmax_f P(F|E) = argmax_F P(E|F) P(F)
  – P(F) is language model for French
  – P(E|F) is the translation model

Translation Model

- Fertility: How many destination words does this word map to?
- Offset: Where does this word move to?

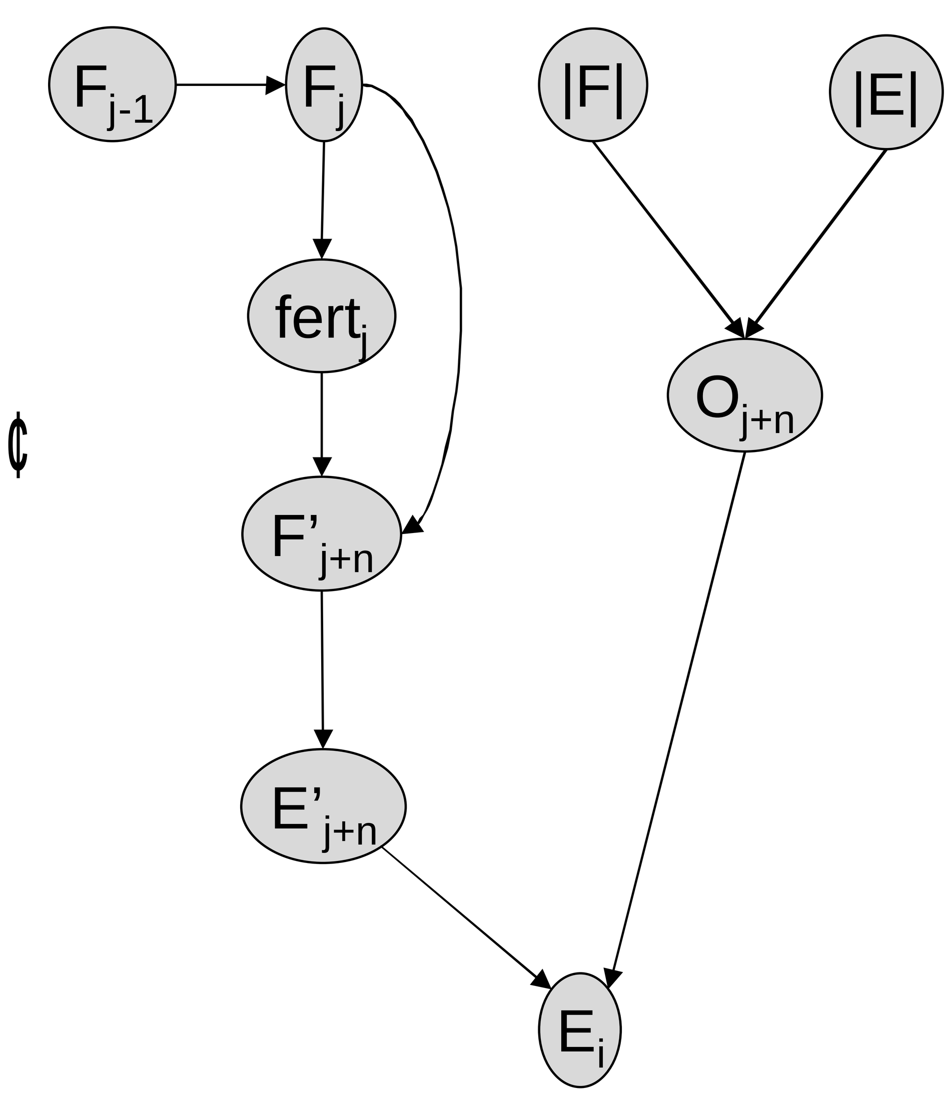
Source	Le	chien	brun	n’	est	pas	allé	à	la	maison
Fertitilty	1	1	1	1	1	0	1	0	0	1
Offset	0	+1	–1	+1	–1		0			0
English	The	dog	brown	not	did		go			home
Result	The	brown	dog	did	not		go			home

Probabilistic Model

- Fertility: $P(\text{fert} \mid F_j)$
- Word choice: $P(E_i \mid F_j)$
- Offset: $P(O_j \mid j, |E|, |F|)$
- Language model: $P(F_j \mid F_{j-1})$

Probabilistic Model

$P(E|F) = \prod_i P(E_i|O_j,E'_{j+n})$ ¢
$P(O_{j+n} \mid j, |E|, |F|)$ ¢
$P(E'_{j+n} \mid F'_{j+n})$ ¢
$P(F'_j \mid \text{fert}$ ¸ $n, F_{j-n+1})$ ¢
$P(\text{fert} \mid F_j)$ ¢
$P(F_j \mid F_{j-1})$ ¢
$P(|F|)$ ¢ $P(|E|)$



Summary

- Natural Language Understanding is one of the hardest tasks in AI
  – Large amounts of knowledge about people and the world are needed
- Many levels of processing
  – syntax
  – semantics
  – discourse
- Many language tasks
  – communication
  – information retrieval
  – information extraction
  – machine translation

