Language Technology: Applications and Techniques

Robert Dale
Language Technology Pty Ltd
and
Macquarie University
rdale@language-technology.com

 1

Managing Expectations

From a real e-mail message posted to a mailing list:

> I need to read a file and parse it and convert it to first order logic. I would thus need some kind of natural language parser/processor and since my ultimate aim is far more ambitious I would like to use an existing (but good) NLP.
>
> I would truly appreciate any pointers to free LISP code that implements a natural language processor. I do know the basics of NLP but can't write the grammar now.

 2

Aims of This Tutorial

- To provide a broad awareness of actual and potential Language Technology applications
- To provide a framework for thinking about LT applications in terms of the linguistic resources they need
- To provide an understanding of what's involved in building LT applications

Outcomes

- By the end of this tutorial you should have:
  - an understanding of what LT is
  - an appreciation of the range of applications that LT enables
  - an insight into the technologies used in LT applications
  - an ability to assess claims about the capabilities of LT applications
  - an awareness of the major vendors and suppliers in LT technologies

Tutorial Structure

- Part 1: Applications [1 hr 15 mins]
- Break [15 mins]
- Part 2: Techniques [1 hr 15 mins]

Part 1: Applications

A Definition

- Language Technology involves the application of knowledge about human language in computer-based solutions

Two Drivers for Language Technology

- The need for intelligent, habitable, natural interfaces:
  - Telephony-based apps need voice capabilities
  - Nobody wants a keyboard on their intelligent microwave
- The problem of information overload
  - There's too much stuff on the web
  - There's too much stuff in the filing cabinet
  - Nobody has time to read all their email

Related Terms

- Natural Language Processing
- Computational Linguistics
- Speech Technology
- Language Engineering
- Intelligent Text Processing
- Document Processing
- Artificial Intelligence
- Cognitive Science

Two Dimensions

- Speech versus Text
- Input versus Output

Principal Components in a Language Technology Application

- Language input
  - recognizing
- Language processing
  - reasoning
- Language output
  - rendering

Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Machine Translation
- Text Summarisation
- Search and Information Retrieval
- Question-answering Systems

Applications of Language Technology: Language Output

- Text-to-Speech
- Tailored Document Generation
- Dynamic Web Pages

Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Speech Recognition: Key Focus and Applications

- Key Focus of the Technology
  - Deriving a textual representation of a spoken utterance
- Applications
  - Desktop command and control
  - Dictation
  - Telephony-based transaction and information services

Speech Recognition: Fundamental Issues

- Isolated word vs continuous speech
- Vocabulary size
- Speaker dependence vs speaker independence

Speech Recognition: Current State of the Art

- Cheap PC desktop software available: virtually a commodity
- 60–90% accuracy depending on circumstances
- A number of major players in telephony-based systems

Speech Recognition: Current State of the Art

- Accuracy rates good enough for general dictation and simple transactions, but depends on speaker—your mileage may vary
- Ease of handling errors is important
- Recognition is not understanding!

Speech Recognition: Fielded Products

- Desktop
  - Philips FreeSpeech (www.speech.philips.com)
  - IBM ViaVoice (www.ibm.com/viavoice)
  - Lernout & Hauspie's DragonDictate (www.lhs.com)

Speech Recognition: Fielded Applications

- Telephony-based
  - Nuance (www.nuance.com)
  - SpeechWorks (www.speechworks.com)
  - Philips (www.speech.philips.com)

Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Optical Character Recognition: Key Focus and Applications

- Key Focus of the Technology
  - Deriving a computer-readable representation of printed material
- Applications
  - Scanning documents into ASCII form for electronic archival
  - Business card readers
  - Web site construction from printed documents
  - Menu-translating pens!

Optical Character Recognition: Fundamental Issues

- Two issues: character segmentation and character recognition
- Problems: unclean data, ambiguity, and new typefaces
- Special fonts aid accuracy (look at your cheque book)
- Many OCR systems use linguistic knowledge to correct recognition errors:
  - N-grams for word choice during processing
  - Spelling correction for post-processing

Optical Character Recognition: Current State of the Art

- 90% accuracy or better on clean text
- 100–200 characters per second … as opposed to 3–4 characters per second for typing
- Market development depends on recognising not only characters, but also larger structural elements of documents
- Current apps include 'read-back' for proofreading
- US Postal Service research focuses on assigning ZIP Codes to letter images which may not contain any ZIP Code

Optical Character Recognition: Fielded Products

- Caere's OmniPage (www.caere.com)
- Xerox's TextBridge (www.scansoft.com)
- ExperVision's TypeReader (www.expervision.com)

Applications of Language Technology: Language Input

- Speech Recognition
- Optical Character Recognition
- Handwriting Recognition

Handwriting Recognition: Key Focus and Applications

- Key Focus of the Technology
  – Deriving a computer-readable representation of human handwriting
- Applications
  – Forms processing
  – Mail routing
  – PDAs

Handwriting Recognition: Fundamental Issues

- Everyone writes differently!
- Isolated letters vs cursive script
- Better to train the user than to train the system?
  - Apple Newton vs Palm's Graffiti
- Many people can type faster than they can write
  - So, handwriting appropriate where keyboards are not
- Need to integrate elaborate language models and writing style models

Handwriting Recognition: Current State of the Art

- Generally based on neural network technology
- 5–6% error rate typical for isolated letters
- Good typists tolerate up to 1% error rate on keyboards that generate random errors
- Human subjects make 4–8% errors in isolated character reading, and 1.5% errors given context

Handwriting Recognition: Fielded Products

- Isolated letters
  - Palm's Graffiti (www.palm.com)
  - Computer Intelligence Corporation's Jot (www.cic.com)
- Cursive Script
  - Motorola's Lexicus (www.motorola.com/MIMS/lexicus)
  - ParaGraph's CalliGrapher (www.paragraph.com)

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Machine Translation
- Text Summarisation
- Search and Information Retrieval
- Question-answering Systems

Spoken Language Dialog Systems: Key Focus and Applications

- Key Focus of the Technology
  - Natural voice interactive dialogs with computer-based systems
  - Spoken dialogue systems communicate with users via automatic speech recognition and text-to-speech interfaces, and mediate the user's access to a back-end database
- Applications
  - Information services: stock quotes, timetables
  - Transaction services: banking, betting, flight reservations

Spoken Language Dialog Systems: Fundamental Issues

- Telephony-based systems cannot afford a training period
- Making a conversation too realistic falsely raises expectations
- Error handling is a significant issue
- Giving dialog initiative to the user increases difficulty

Spoken Language Dialog Systems: Current State of the Art

- Commercial systems operational for limited transaction and information services
  - QTAB betting service
  - American Airlines flight information system
  - Charles Schwab's stock broking system
- Very limited, finite-state notion of dialog
- Limited natural language understanding

Spoken Language Dialog Systems: Fielded Applications

- Nuance (www.nuance.com)
- SpeechWorks (www.speechworks.com)
- Philips (www.speech.philips.com)

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Machine Translation
- Text Summarisation
- Search and Information Retrieval
- Question-answering Systems

Machine Translation: Key Focus and Applications

- Key Focus of the Technology
  - Deriving a version of a document in a language other than that used in the original document
- Applications
  - Desktop and web-based translation services
  - Spoken language translation services

Machine Translation: L&H's iTranslator

- Source English:
  - With worldwide translation and documentation services, L&H can help you do business in any market around the globe. Our technology and top-line people enable you to communicate with employees and customers anywhere.
- Translated into French …:
  - Avec traduction mondiale et la documentation entretient, L&H peut vous aider à travailler dans tout marché autour du globe. Notre technologie et gens de la sommet - ligne vous permettent de communiquer avec les employés et les clients n'importe où.
- … and back into English:
  - With world translation and the documentation maintains, L&H can help you to work in all market around the globe. Our technology and people of the summit - lign offs permit you to communicate with the employees and the customers where.

 39

Machine Translation: Systran's Web-Based Translator

- Source English:
  - With worldwide translation and documentation services, L&H can help you do business in any market around the globe. Our technology and top-line people enable you to communicate with employees and customers anywhere.
- Translated into French …:
  - Avec des services mondiaux de traduction et de documentation, L&h peut vous aider fait des affaires sur n'importe quel marché autour du globe. Nos personnes de technologie et de dessus-ligne vous permettent de communiquer avec des employés et des clients n'importe où.
- … and back into English:
  - With world services of translation and documentation, L&h can help you made deals on any market around the sphere. Our people of technology and top-line allow you to communicate with employees and customers anywhere.

 40

Machine Translation: Fundamental Issues

- The broad coverage required by mainstream translation technologies exacerbates ambiguity problems
- Effectively limited to literal language use
- Main approaches:
  - Transfer
  - Interlingua
  - Example-based
- Real systems often Machine-<u>Assisted</u> Translation

 41

Machine Translation: Current State of the Art

- Broad coverage systems already available via the Web
- Fast turnaround, acceptable error rate for gisting
- Higher accuracy can be achieved by carefully domain-targetted systems
- Controlled languages such as Caterpillar English maximise likelihood of accurate translation

 42

Machine Translation: Fielded Products

- Lernout & Hauspie's iTranslator (www.lhs.com)
- Systran—used by AltaVista (www.systransoft.com)

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Machine Translation
- Text Summarisation
- Search and Information Retrieval
- Question-answering Systems

Text Summarisation: Key Focus and Applications

- Key Focus of the Technology
  - Producing a version of a document that is shorter than the original document
- Applications
  - Information browsing
  - Voice delivery of web pages and email

Text Summarisation: Fundamental Issues

- There are different kinds of summaries:
  - Informative vs indicative
- Real summarisation requires real understanding
- Quality of 'knowledge-free' summarisation relies on aspects of the document other than content

Text Summarisation: Current State of the Art

- Commercial systems work on a ‘sentence-extraction’ model
- Sentences extracted on basis of
  - location
  - linguistic cues
  - statistical information
- Relatively knowledge-free but broad coverage as a result

Text Summarisation: Fielded Applications

- British Telecom’s ProSum
- InXight (www.inxight.com)
- Tetranet’s Extractor (http://extractor.iit.nrc.ca)
- MS Word’s Summarisation Tool

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Machine Translation
- Text Summarisation
- Search and Information Retrieval
- Question-answering Systems

Search and Information Retrieval: Key Focus and Applications

- Key Focus of the Technology
  - Concept-based search: moving beyond documents as bags of words
- Applications
  - Intelligent web search
  - Improved document retrieval

Search and Information Retrieval: Fundamental Issues

- Major failure in IR systems: vocabulary mismatch
- Information need is described using words other than those used in relevant documents
- Solved by automatic expansion of the query

Search and Information Retrieval: Current State of the Art

- Thesaurus-based vocabulary expansion
- Limited linguistic analysis to determine phrases rather than words

Search and Information Retrieval: Fielded Applications

- WebTop's concept based search (www.webtop.com)
- Lexeme's eQualia (www.lexeme.com)
- Ultraseek's Natural Language (http://software.infoseek.com)

Applications of Language Technology: Language Processing

- Spoken Language Dialog Systems
- Machine Translation
- Text Summarisation
- Search and Information Retrieval
- Question-answering Systems

Question-Answering Systems: Key Focus and Applications

- Key Focus of the Technology
  - Given a natural language query, produce an appropriate response
- Applications
  - Web-based information services
  - Desktop help systems

Question-Answering Systems: Fundamental Issues

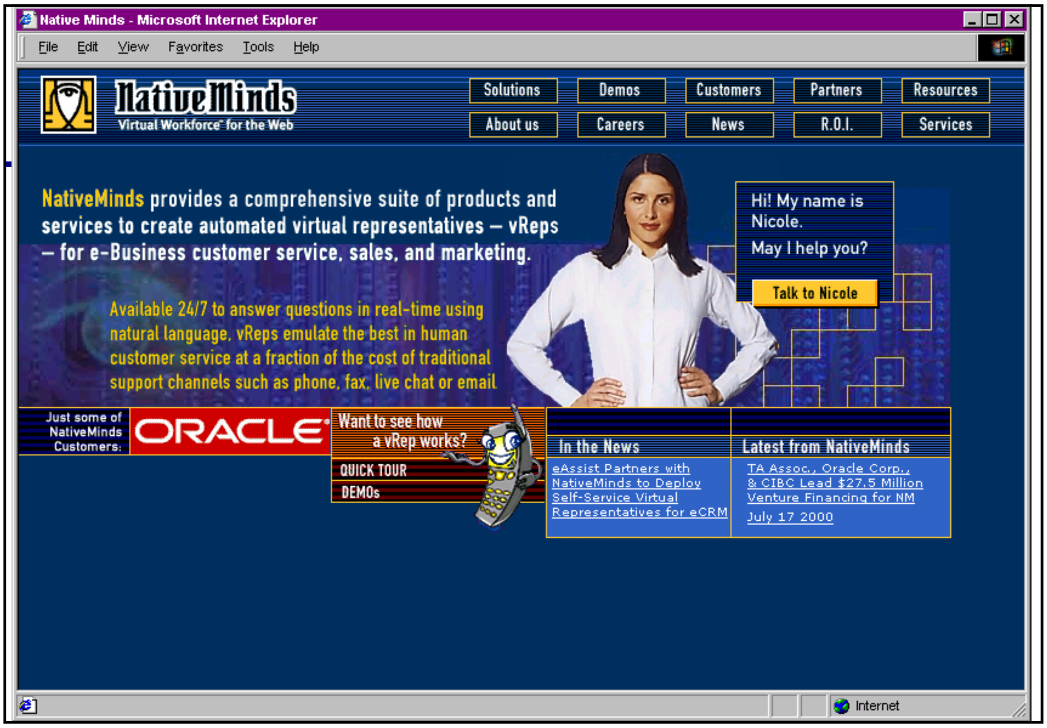
- Limiting coverage to short questions provides some restriction on syntactic structure but leaves open vocabulary issues
- Real questions often contain presuppositions and contextual assumptions
  - Where can I find my class timetable?

Question-Answering Systems: Current State of the Art

- Limited question analysis to determine query type and central queried concept
- IR techniques to return appropriate documents
- Data analysis to support construction of custom answers for common questions
- Current technology claimed capable of reducing call center expenses from $75 a call to 18c a call

Question-Answering Systems: Fielded Applications

- Ask Jeeves (www.askjeeves.com)
- Artificial Life's ALife Sales Rep (www.artificiallife.com)
- iPhrase Technologies (www.iphrase.com)
- Native Minds' vReps (www.nativeminds.com)
- Soliloquy (www.soliloquy.com)



Applications of Language Technology: Language Output

- Text-to-Speech
- Tailored Document Generation

Text-to-Speech: Key Focus and Applications

- Key Focus of the Technology
  - Production of natural sounding speech from a textual input
- Applications
  - Spoken rendering of email via desktop and telephone
  - Document proofreading
  - Voice portals

Text-to-Speech: Issues and State of the Art

- TTS in a vacuum requires reverse engineering of linguistic information
  - Appropriate use of intonation and phrasing
  - Handling homophones
- High quality diphone concatenation is readily available:
  - Short digital-audio segments are concatenated, and intersegment smoothing performed to produce a continuous sound
  - Very appropriate where audio prerecording not usable

Text-to-Speech: Fielded Applications

- Lernout & Hauspie's RealSpeak (www.lhs.com)
- British Telecom's Laureate

Applications of Language Technology: Language Output

- Text-to-Speech
- Tailored Document Generation

Tailored Document Generation: Key Focus and Applications

- Key Focus of the Technology
  - Production of individually-tailored documents based on parameter values
- Applications
  - Individual, personalised advice-giving
  - Customised personnel and policy manuals
  - Web-delivered dynamic documents

Tailored Document Generation: Issues and State of the Art

- Mail-merge is the bottom-end of this technology
- Tailored composition of document components and associated template filling can produce wide variations in output
- Going beyond mail-merge requires underlying knowledge source rich enough to drive sophisticated linguistic abilities
- Applications with complex underlying models such as project management software or CAD software can provide appropriate input

Tailored Document Generation: Fielded Applications

- KnowledgePoint (www.knowledgepoint.com)
  - Tailored job descriptions and personnel policies
  - Automated performance review systems
- CoGenTex (www.cogentex.com)
  - Automatic generation of project status reports

Summary So Far

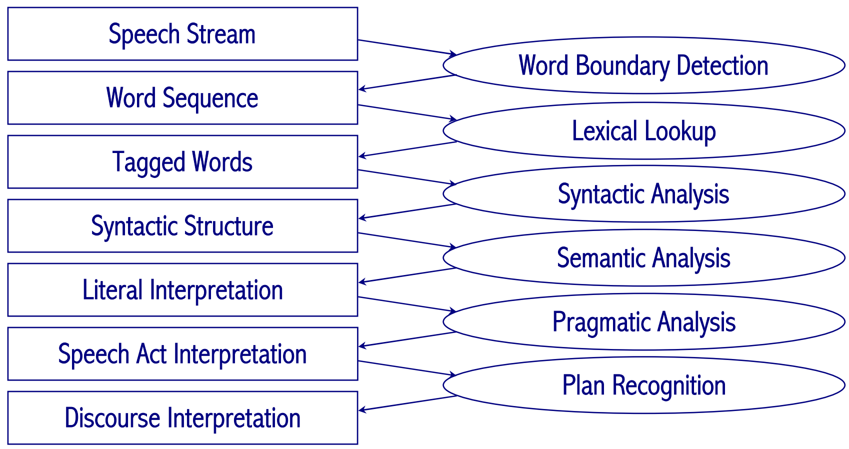
- Input technologies can achieve in excess of 90% accuracy
- Broad coverage applications have to rely on limited linguistic knowledge
- Targetted applications can use more sophisticated linguistic knowledge
- Output technologies not yet a major focus

Part 2: Techniques

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
- Conclusions and Further Information

Stages in Processing Language



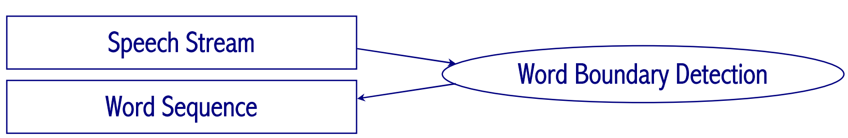
Word Boundary Detection

- recognise speech
- wreck a nice peach

Word Boundary Detection

- A speech recognition system needs to recognise the phonemes that were spoken and then assemble these into valid sequences of words
- Different people pronounce phonemes in different ways: an acoustic model captures a representation of the possible renderings of phonemes that can be matched against
- A language model indicates what sequences of words are possible

Stages in Processing Language



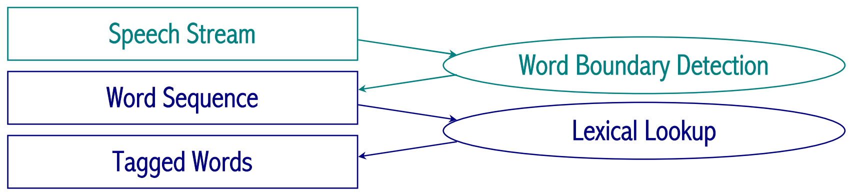
Lexical Ambiguity

- The astronomer saw the star.
- The astronomer married the star.
- King Kong sat on the bank.

Lexical Ambiguity

- Early methods were rule-based and relied on at least a partial understanding of the context
- Selectional restrictions in the lexicon:
  - marry[agent=animate, object=animate]
  - $star_1$[+animate] % famous or celebrated-person
  - $star_2$[–animate] % celestial object
- Modern techniques rely on statistical evidence derived from large bodies of text

Stages in Processing Language

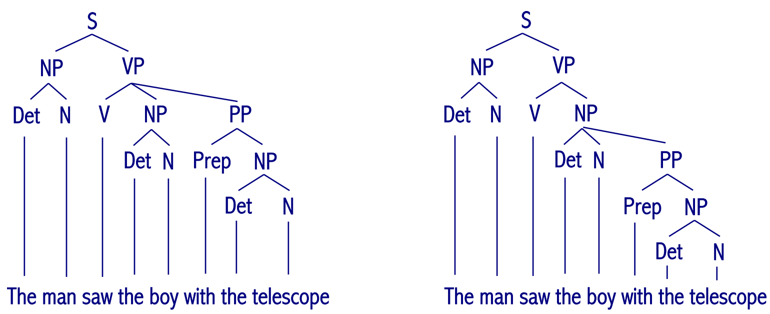


Structural Ambiguity

- The astronomer saw the star with a telescope.
- The astronomer married the star with a history.
- Visiting uncles can be a nuisance.
- I forgot how good beer tastes.

Structural Ambiguity

- The man saw the boy with the telescope



Structural Ambiguity

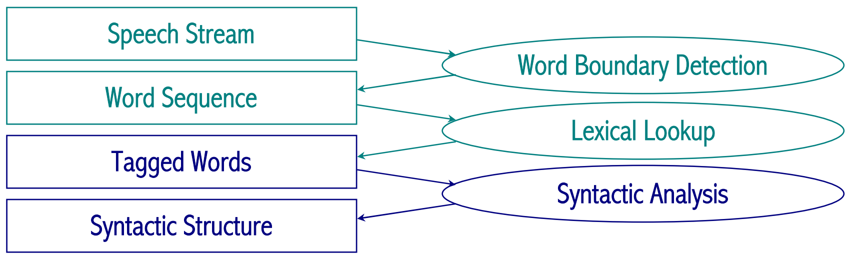
- A grammar inventorises the possible syntactic structures in a language by means of a fine set of rules
- These rules dictate how symbols in the language can be combined to create well-formed sentences

  S → NP VP

  NP → Det N

  VP → V NP
- A parser uses a set of grammar rules to attribute a syntactic structure to a well-formed string

Stages in Processing Language



Anaphora Resolution

- The councillors refused the women a permit because they feared revolution.
- The councillors refused the women a permit because they advocated revolution.

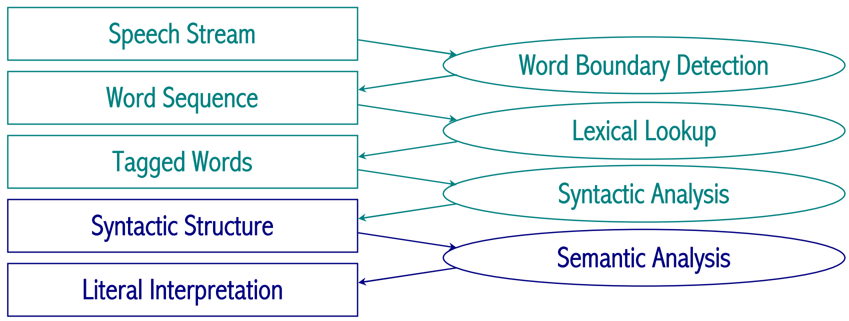
Anaphora Resolution

- The councillors refused the women a permit because they feared revolution.
  - $refuse(e_1) \wedge agent(e_1,c_1) \wedge benefactor(e_1,w_1) \wedge object(e_1,p_1) \wedge fear(e_2) \wedge agent(e_2, c_1) \wedge object(e_2, r_1) \wedge cause(e_2, e_1)$
- The councillors refused the women a permit because they advocated revolution.
  - $refuse(e_1) \wedge agent(e_1,c_1) \wedge benefactor(e_1,w_1) \wedge object(e_1,p_1) \wedge advocate(e_2) \wedge agent(e_2,w_1) \wedge object(e_2, r_1) \wedge cause(e_2, e_1)$

Anaphora Resolution

- Anaphora resolution is just one of a range of problems in semantic interpretation
- Anaphora resolution involves all kinds of linguistic knowledge: intonational, syntactic, semantic and pragmatic:
  - Maisy swore at Sabine then she insulted her.
  - Jim hurt him.
  - Andy put the cake on the table and ate it.
  - Sue went to Mary’s house and she cooked her dinner.

Stages in Processing Language



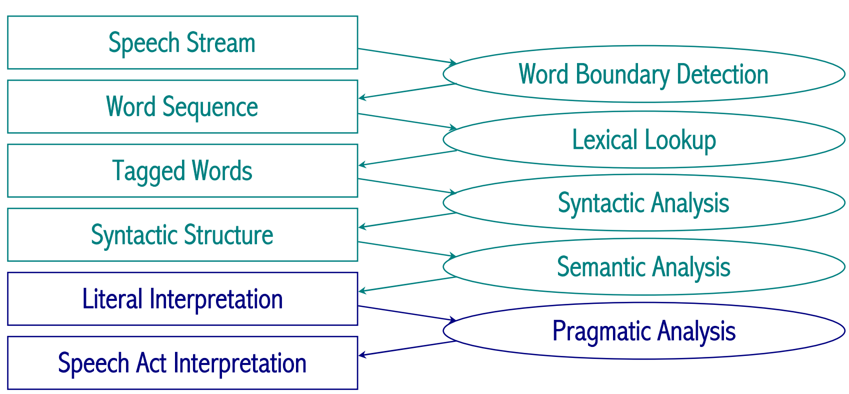
Non-literal Meaning

- Can you pass the salt?
- You’re standing on my foot.
- His handwriting is very good.

Non-literal Meaning

- We always understand language in a context
- Our rich store of world knowledge allows us to draw the appropriate inferences to construct an appropriate interpretation
- Access to a similar store of world knowledge is a significant problem for computers
- As a result, successful applications of NLP lie in areas where we can closely constrain the context and therefore the range of possible interpretations

Stages in Processing Language



Plan Recognition

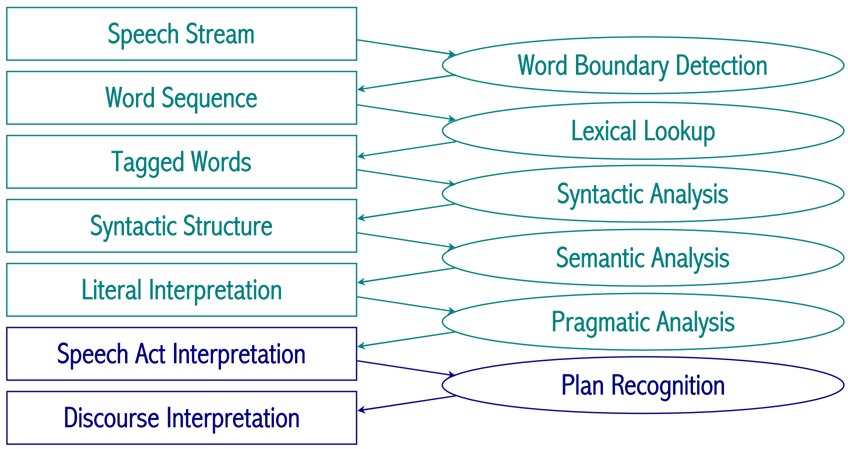
Plan inference and co-operative response:

User:	*Which students got an F in Comp248 in 1993?*
System:	None.
User:	*Did anyone fail Comp248 in 1993?*
System:	No.
User:	*How many people passed Comp248 in 1993?*
System:	Zero.
User:	*Was Comp248 given in 1993?*
System:	No.

Plan Recognition

- When we take part in dialog, we are constantly making predictions as to what the other party in the dialog wants
- Research systems use complex inferences over assumed user beliefs and intentions
- Truly intelligent systems need to do the same thing
- Meaning results from the text and the context in combination

Stages in Processing Language



Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Getting Language into the Machine

- **Speech Stream:** segment into words, represent as a stream or lattice of space-separated word tokens
- **Handwriting Recognition:** recognise characters in cursive script, represent as space-separated word tokens
- **Optical Character Recognition:** recognise characters within page layout, combine into space-separated word tokens
- **Existing Electronically Encoded Documents:** strip out formatting commands and control characters, represent as space-separated word tokens

Getting Language into the Machine

- Tokenisation:
  - the process of breaking up a sequence of characters in a text by locating the word boundaries
  - the words thus identified are tokens
  - in languages where no word boundaries are explicitly marked in the writing system, also known as word segmentation

Getting Language into the Machine

- Sentence Segmentation
  - the process of identifying sentence boundaries
  - involves sentence boundary detection, disambiguation or recognition

Tokenisation and Sentence Segmentation

- The two tasks are not independent:
  - Maria finished her Ph.D. yesterday.
  - Yesterday Maria finished her Ph.D.
- Real sentence boundary recognition is hard!
  - Two high-ranking positions were filled Friday by Penn St. University President Graham Spencer.
  - Two high-ranking positions were filled Friday by Penn St. University President Graham Spencer announced the appointments.

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Word Lists

- The minimal linguistic resource required for many applications: a list of the words in the language
  - Generally required for spell checking and correction
  - Can reduce error rates in OCR and handwriting recognition
- Spell checking can also be carried out using lists of valid character bigrams or trigrams—but this isn't enough for correction

Word Lists

- Existing IR and Text Summarisation systems can perform without word lists:
  - In simple IR, words are just strings of characters
  - In simple Text Summarisation, sentences are just sequences of words, which are strings of characters
- Benefit: absolutely broad coverage
- Cost: zero leverage of linguistic information

Word Lists

- A typical desk dictionary contains around 50000–150000 entries
- In 44 million words of Associated Press newswire text collected over 10 months, there were 300000 different tokens
- How do you build a lexicon big enough to deal with real language?
- One possibility: make use of machine readable dictionaries
- A popular MRD: Longman's Dictionary of Contemporary English

Word Lists

- How many words do you need? It has been suggested that by age 17 we know 80000 words.
- But: it has been estimated that 8000 base forms of words (morphemes) is sufficient to handle 95% of texts
- Typically, 15 most frequent words account for 25% of tokens
- 100 most frequent words account for 60% of tokens

Word Frequencies

Rank	Spoken English	Written English	French	German
1	the	the	de	der
2	and	of	le	die
3	I	to	la	und
4	to	in	et	in
5	of	and	les	des
6	a	a	des	den
7	you	for	est	zu
8	that	was	un	das
9	in	is	ure	von
10	it	that	du	fur

Dictionaries

- A dictionary (or lexicon) is a collection of words with associated information:
  - A mapping to phonetic transcriptions is required for speech recognition
  - A mapping to parts of speech is required for almost all language technology applications that do anything with the words once recognised

Dictionaries: Phonetic

- The Roman alphabet has 26 characters, but English has around 44 distinct phonemes
- Phonetic transcription traditionally notated using IPA, the International Phonetic Alphabet, but more recent encodings are computer-readable

ði ɪntəˈnæʃənəl fəˈnɛtɪk əsoʊsiˈeɪʃn

Dictionaries: Part of Speech

- Every word has a Syntactic Category or Part of Speech
- Parts of speech are important because they constrain how sentences can be put together
- Two broad types: Open Class words vs Closed Class words
- This information is needed for syntactic analysis
- Problem: dealing with unknown words

Dictionaries: Part of Speech

- Nouns
  - projector, money, infidelity, amazement, antidisestablishmentarianism ...
- Verbs
  - run, fly, walk, procrastinate, believe ...
- Adjectives
  - crazy, green, hungry, unbelievable, amazed, smart ...
- Adverbs
  - slowly, hungrily, unbelievably ...

Dictionaries: Part of Speech

- Determiners
  - a, the, this, that, these, those …
- Conjunctions
  - and, but, therefore, because …
- Prepositions
  - in, on, under, between, to, from …

Morphology and the Dictionary

- Listing information on every word in the language separately fails to observe that there are systematic relationships between words
- We can save space by recognising the morphological structure of words, and constructing them from their component parts by rule
- Morphological processing can help in providing Part of Speech information for unknown words

Inflectional Morphology

- Root Form + Affix; affix can be a Prefix, Infix or Suffix
- Part of speech remains constant; same basic meaning
- Examples:
  - deliver + s = delivers [third person singular present tense]
  - deliver + ing = delivering [present participle]
  - deliver + ed = delivered [past tense]
- Root form also known as the Base, Stem, or Lemma
- Root forms are Free Morphemes
- Affixes are usually Bound Morphemes

Derivational Morphology

- A word of one category is used to derive a word of another category
- friend [noun] + ly [suffix] = friendly [adjective]
- friendly [adjective] + ness [suffix] = friendliness [noun]

Stemming

- Many IR systems use a linguistically under-motivated but simpler process called stemming to conflate words with a common base

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Building Syntactic Representations

- A significant proportion of the work in traditional NLP has focused on syntactic analysis
  - sophisticated linguistic formalisms for capturing generalisations
  - efficient parsing techniques for broad coverage syntactic analysis

Applications of Syntactic Analysis

- Rich analysis generally required for
  - Grammar checking
  - Transfer-based and Interlingua-based Machine Translation

Applications of Syntactic Analysis

- Limited syntactic coverage required for:
  - Spoken-language dialog systems
  - Question-answering systems

Applications of Syntactic Analysis

- Shallower techniques based on finite state grammars sufficient for
  - Concept-based information retrieval
  - Information extraction technologies

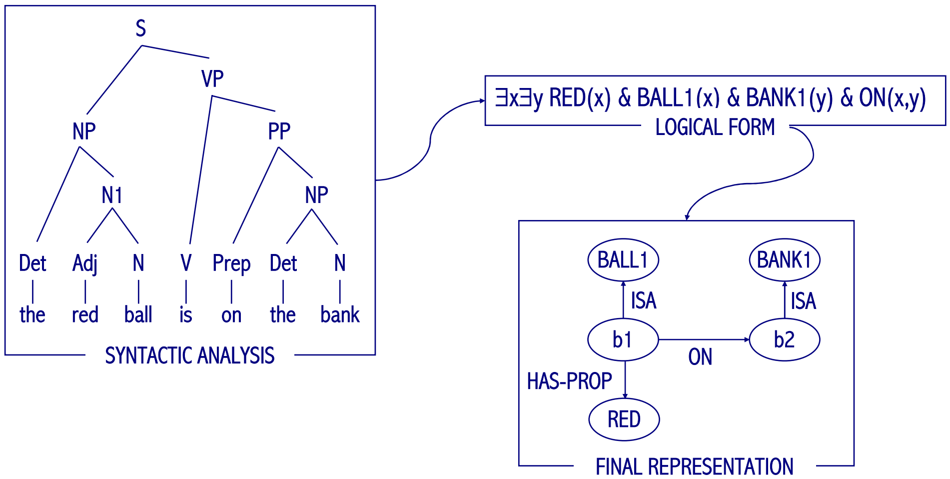
Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
  - Getting Language Into the Machine
  - Lexical Knowledge
  - Syntactic Knowledge
  - Semantic and Pragmatic Knowledge
- Conclusions and Further Information

Semantics as Logical Form

- Typically expressed using First Order Predicate Calculus:
  - variables
  - predicates
  - logical connectives
  - quantifiers
- Other forms of logic required to express possibility, necessity, temporal phenomena …

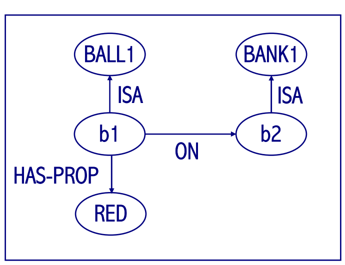
From Syntax to Semantics



From Syntax to Semantics

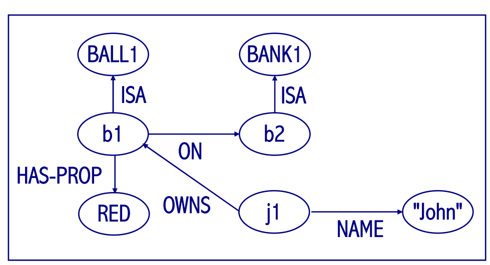
The red ball is on the bank.

$\exists x \exists y$ RED(x) & BALL1(x) & BANK1(y) & ON(x,y)



It belongs to John.

$\exists x \exists y$ NAME(x, John) & OWNS(x,y)



How Do We Get From Syntax to Semantics?

- Meaning is compositional: the meaning of a constituent is derived solely from the meanings of its subconstituents and their means of combination
- An elegant approach: the lambda calculus
- Each lexical entry expresses the meaning of the word as a lambda expression; the rules of the grammar indicate how these expressions are to be combined
- The lack of a language-wide analysis in these terms makes the approach currently impractical

Case Frames

- If we ignore quantificational phenomena, most significant aspect of meaning is 'who did what to whom'
- Semantically, each verb carries a set of case roles that specify the semantic relationships corresponding to the different participants in the event described:
  - AGENT
  - PATIENT
  - INSTRUMENT
  - SOURCE
  - DESTINATION
  - ...

Case Roles and Case Frames

We can introduce events as logical variables:

- The astronomer saw the star with a telescope
  - ∃e∃x∃y∃z SEE(e) & PAST(e) & ASTRONOMER(x) & STAR1(y) & TELESCOPE(z) & AGENT(e,x) & PATIENT(e,y) & INSTRUMENT(e,z)
- The astronomer married the star with a birthmark
  - ∃e∃x∃y∃z MARRY(e) & PAST(e) & ASTRONOMER(x) & STAR2(y) & BIRTHMARK(z) & AGENT(e,x) & PATIENT(e,y) & POSSESS(y,z)

A Feature Structure Representation

```
index:  e1
sem:    pred:  see
        time:  < now
        args:  agent:        index:  a1
                             sem:    ASTRONOMER
               patient:      index:  s1
                             sem:    STAR1
               instrument:   index:  t1
                             sem:    TELESCOPE
```

Semantic and Pragmatic Knowledge

- From a theoretical perspective, semantics and pragmatics are distinct
- In practical systems, pragmatic issues are often ‘compiled-down’ into semantics, or even into the syntax
- For practical applications this is valid because of the limited coverage required

A Grammar Rule in a Dialog System

- Semantics compiled into syntax:

balance-request →
    ([what is | what’s | my | the |] balance [please]) |
    ([tell me the | check my] balance [please])
    <request=balance>

Interlingua Mappings in Machine Translation

- Representations similar to case frames serve as interlingua: a level of representation that embodies the basic concepts in a language-independent form
- Pragmatics? Some options
  - Pragmatics compiled into semantics
  - Pragmatics as a free lunch
  - Treat special cases separately

Overview

- Traditional NLP Issues and Techniques
- How The Techniques Map to Applications
- Conclusions and Further Information

Technology Map: Spoken Language Dialog Systems

- Limited grammatical coverage: simple syntax, effectively represented by means of semantic grammars
- Rich phonetically-annotated lexica for speech recognition and synthesis
- Hard-wired, implicit pragmatics

Technology Map: Machine Translation

- Large lexica
- Rich syntactic analysis
- For transfer-based systems, structural and lexical mapping rules; limited semantic constraints
- For interlingua-based systems, some level of semantic analysis

Technology Map: Text Summarisation

- Current commercial systems use virtually no knowledge of language, other than extraction rules based on specific linguistic cues
- Interesting research direction: combination of information extraction technology with natural language generation

Technology Map: Query Systems

- Existing systems use combination of linguistic knowledge of question forms + finite state grammars
- Answers found by information retrieval with some minimal NLP
- Quality results come from string matching to hand-crafted answers for frequent questions

Finding Out More: Comprehensive Texts

- R Dale, H Moisl and H Somers (eds) [2000] Handbook of Natural Language Processing. Marcel Dekker Inc.
- D Jurafsky and J Martin [2000] Speech and Language Processing. Prentice-Hall.
- R Cole, A Zaenen and A Zampolli (eds) [1998] Survey of the State of the Art in Human Language Technology. Cambridge University Press.

Finding Out More: Industry Magazines

- Speech Technology (www.speechtechmag.com)
- PC AI (www.primenet.com/pcai)
- Multilingual Computing (www.multilingual.com)
- Language International (www.language-international.com)

Finding Out More: Research Journals

- Computational Linguistics
- Natural Language Engineering
- Machine Translation
- Speech Communication
- Computer Speech and Language

Finding Out More: Professional Associations

- Association for Computational Linguistics
  - www.aclweb.org
- European Association for Machine Translation
  - www.eamt.org
- Association for Machine Translation in the Americas
  - www.isi.edu/natural-language/organizations/AMTA.html
- European Speech Communication Association
  - www.esca-speech.org/home.html

Finding Out More: Research Conferences

- Association for Computational Linguistics
- COLING: International Conference on Computational Linguistics
- International Conference on Spoken Language Processing
- EuroSpeech
- MT Summit

Finding Out More: Mailing Lists

- Corpora (www.hd.uib.no/corpora)
- MT-List (www.eamt.org/mt-list)
- The Linguist List (http://linguistlist.org)
- Cmp-Lg [research archive] (http://arxiv.org/archive/cs/intro.html)

Follow-up Comments and Questions

- Please email rdale@language-technology.com
- Thanks for coming!