Lecture 16
Processing Human Language by Computer: Advances and Challenges (Part II)

Rob Gaizauskas

Outline

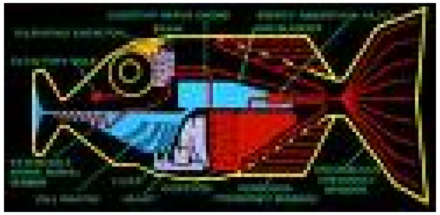
- The Significance of Language
- Turing and the Imitation Game
- Why Language Processing is so hard for a Computer
- Two Frameworks for the Computational Analysis of Language
  - Symbolic
  - Statistical
- Applying Partial Knowledge: Text Processing Applications
- Summing Up and The Road Ahead

Text Processing Applications

- Spelling correction
- Word prediction in text messaging
- Machine translation
- Document retrieval (mono/cross-lingual)
- Information Extraction/Text Mining
- Question Answering
- Plagiarism/Reuse Detection
- Summarisation
- Categorisation
- Text generation
- Dialogue Systems
- …

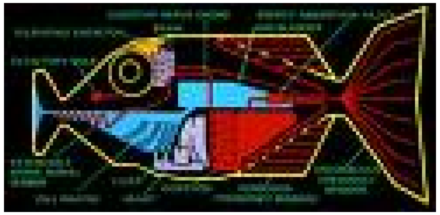
Text Processing Applications

- Spelling correction
- Word prediction in text messaging
- Machine translation
- Document retrieval (mono/cross-lingual)
- Information Extraction/Text Mining
- Question Answering
- Plagiarism/Reuse Detection
- Summarisation
- Categorisation
- Text generation
- Dialogue Systems
- …



Machine Translation

- Machine translation (MT) a focal interest of NLP researchers since the 1950’s
  - Inspired by cold war – military/strategic intelligence gathering remains significant motivation
  - EU also a prime driver
    - 23 official languages
    - €1 billion/annum on translation and interpreting – around 1% of the EU budget (interview with EU translation chief, Feb 2008)
- History
  - Initial interest in pure statistical approaches but inadequate computing power in 1950’s
  - Knowledge engineering approaches grounded in the symbolic framework dominated in 1960’s-90’s (“classical MT”)
  - Resurgence of interest in statistical approaches since 1990’s



Machine Translation

Google translate

English

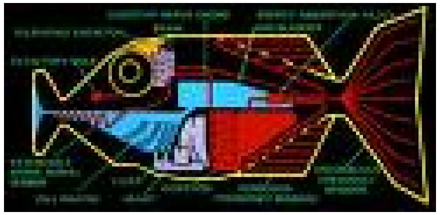
In his landmark 1950 paper "Computing Machinery and Intelligence", Alan Turing proposed conversation as the ultimate test of machine intelligence: if we cannot distinguish a computer from a human in conversation then we may deem the computer intelligent. Turing also speculated that computers would achieve this capability by the year 2000. 2000 has now come and gone, and despite the phenomenal advances in computing since 1950, the "Turing Test" has not yet been passed. Why not? And what have computer scientists, computational linguists and artificial intelligence researchers been up to in the interim in order to get computers to "understand" human language?

French

Dans son point de repère 1950 papier "Computing Machinery and Intelligence», Alan Turing a proposé la conversation comme le test ultime de la machine Intelligence: si nous ne pouvons pas distinguer un ordinateur d'un homme en conversation, alors nous mai jugent l'ordinateur intelligent. Turing spéculé que les ordinateurs permettrait d'atteindre cette capacité d'ici l'an 2000. 2000 est venu et parti, et en dépit des progrès phénoménaux dans le calcul depuis 1950, le «test de Turing" n'a pas encore été passé. Pourquoi pas? Et qu'est-ce que les scientifiques ont d'ordinateur, informatique les linguistes et les chercheurs en intelligence artificielle été à la hauteur dans le provisoires en vue d'obtenir des ordinateurs de "comprendre" le langage humain?

English

In his landmark 1950 paper "Computing Machinery and Intelligence" Alan Turing proposed conversation as the ultimate test of the machine Intelligence: if we can not distinguish a computer from a man conversation, then we may consider the computer intelligent. Turing speculated that computers would reach that capacity by the year 2000. 2000 has come and gone, and despite the phenomenal progress in computing since 1950, the "Turing test" has not yet been past. Why not? And what the computer scientists, computer linguists and researchers in artificial intelligence was at the height in the provisional to get computers to "understand" human language?



Machine Translation

Google translate

English

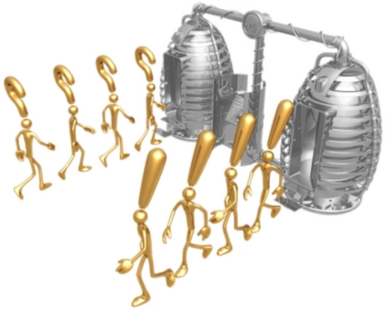
In his landmark 1950 paper "Computing Machinery and Intelligence", Alan Turing proposed conversation as the ultimate test of machine intelligence: if we cannot distinguish a computer from a human in conversation then we may deem the computer intelligent. Turing also speculated that computers would achieve this capability by the year 2000. 2000 has now come and gone, and despite the phenomenal advances in computing since 1950, the "Turing Test" has not yet been passed. Why not? And what have computer scientists, computational linguists and artificial intelligence researchers been up to in the interim in order to get computers to "understand" human language?

Chinese

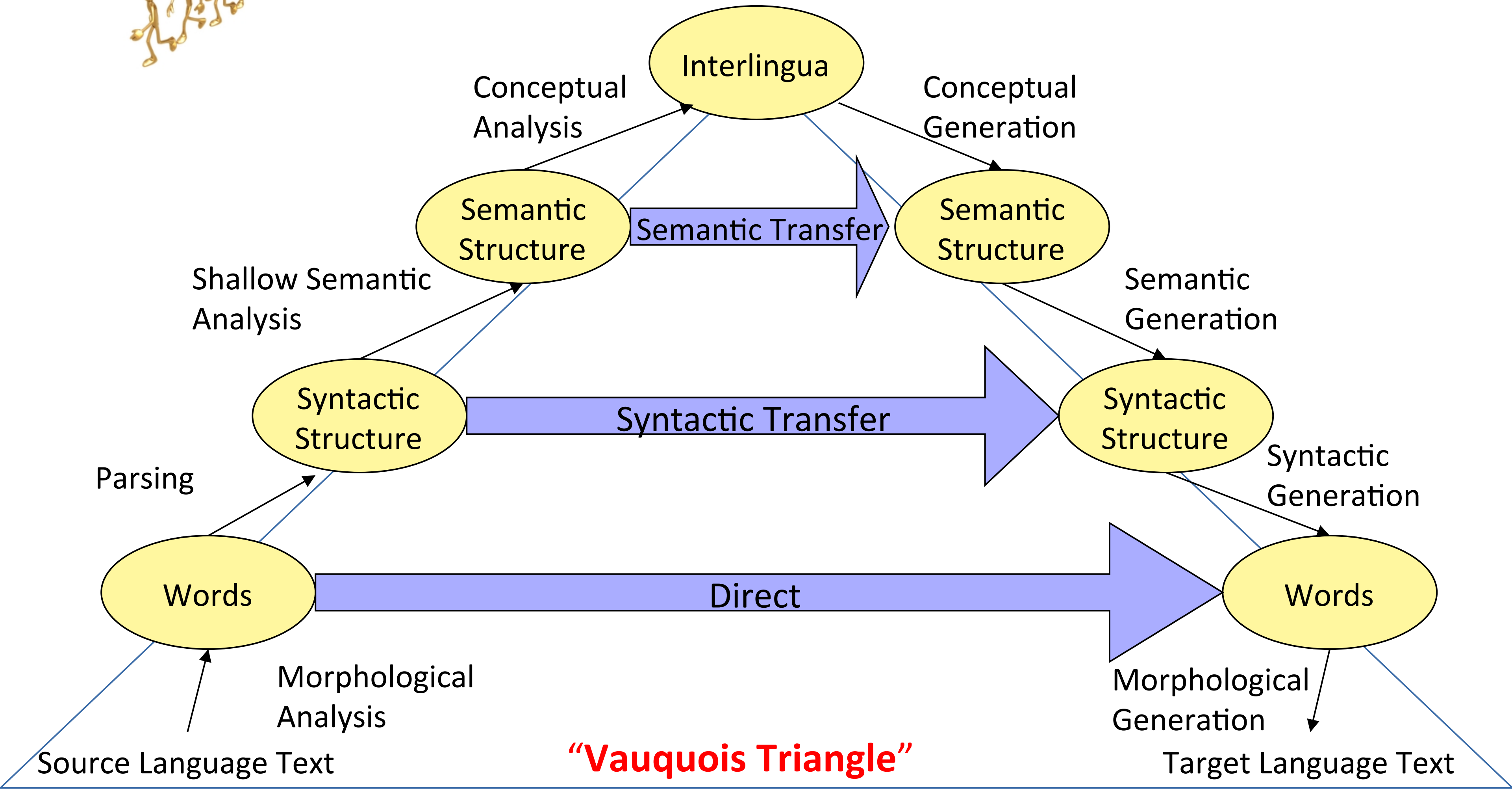
1950年在他的具有里程碑意義的論文“計算機器與智能”，圖靈提出談話的最終測試機 情報：如果我們不能區分一台計算機從一個人在 談話那麼，我們可以認為計算機智能化。圖靈還 推測，計算機將實現這一能力的一年 2000。2000年已經來了又去，儘管顯著進步 在計算1950年以來，“圖靈測試”尚未 通過。為什麼不呢？什麼樣的計算機科家，計算 語言學家和人工智能的研究人員一直在做的，在 臨時為了讓電腦“理解”人類的語言？

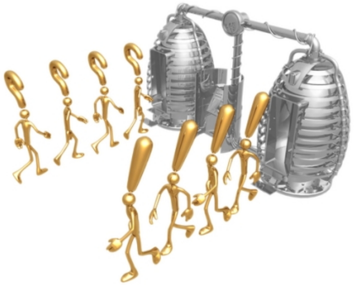
English

In his 1950 landmark paper "Computing Machinery and Intelligence", Turing proposed that the final test machine conversation Intelligence: if we can not distinguish a computer from a person Talk then we can assume that the computer intelligent. Turing was also Speculated that the computer will be a year to achieve this capability 2000. 2000 has come and gone, despite the significant progress In the calculation since 1950, "Turing test" has not yet 通过. Why not? What kind of computer scientists to calculate Linguists and artificial intelligence researchers have been doing, and in Temporary order for computer "understanding" of human language?



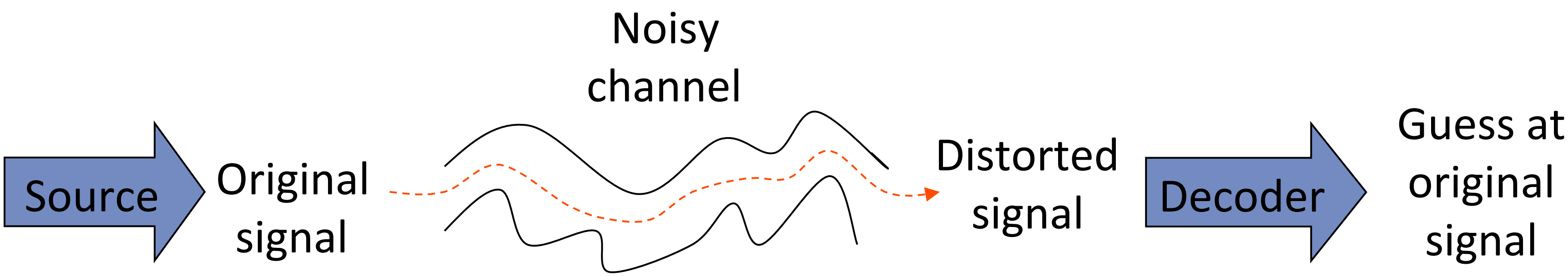
“Classical MT”

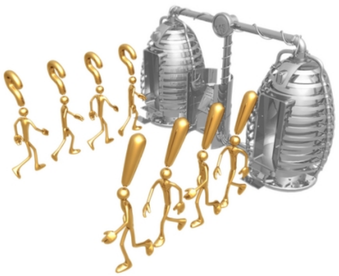




Statistical MT

- Builds on the noisy channel model





Statistical MT

- Builds on the noisy channel model

Noisy channel

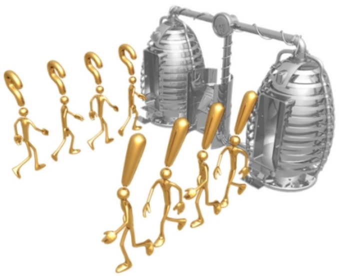
Source → Original signal →  → Distorted signal → Decoder → Guess at original signal

"My hovercraft is full of eels"

"Mon aéroglisseur est plein d'anguilles"

"My hovercraft is full of eels"

- To determine hidden source, need a model of the channel AND a model of the target language



Statistical MT

- Treat as type of Bayesian inference
- Best approximation of English translation given by

$$\hat{E} = \operatorname*{argmax}_{E \in English} P(English \mid French)$$

$$= \operatorname*{argmax}_{E \in English} \underbrace{P(French \mid English)}_{\text{Translation Model}} \underbrace{P(English)}_{\text{Language Model}}$$

- Get language model from any set of English texts
- How do we get translation model?

Statistical MT

- Use of large volumes of example translations – so called parallel corpora
  - E.g. Canadian Hansard (parallel French and English texts)
- Align texts – automatically
  - at sentence level – relatively easy – use sentence lengths
  - at word level – hard
    - significant hand-annotation impossible
    - languages vary a lot in word ordering and lexicalisation
- Solution: use expectation maximization (EM) algorithm to iteratively compute word alignment probabilities

My hovercraft is full of eels*

Mon aéroglisseur est plein d'anguilles

Would you like to go for a ride in a hovercraft?

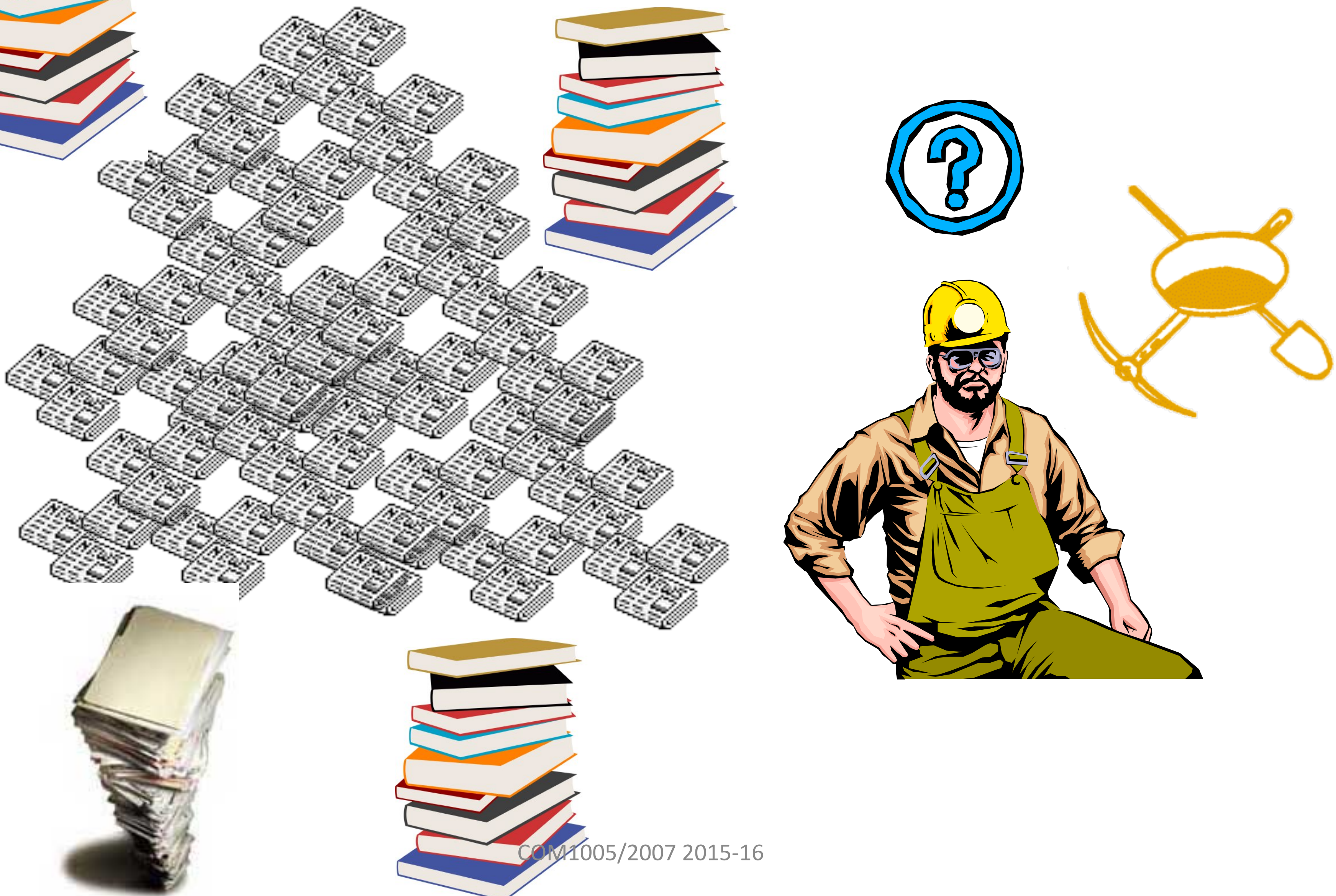
Voulez-vous aller faire un tour dans un aéroglisseur?

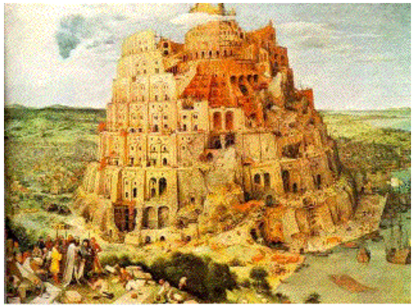
*For the source of this example, and the perils of translation see: http://www.youtube.com/watch?v=G6D1YI-41ao

Text Processing Applications

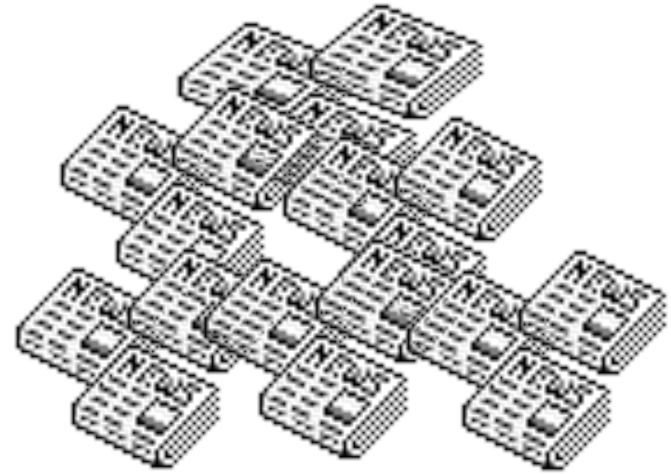
- Spelling correction
- Word prediction in text messaging
- Machine translation
- Document retrieval (mono/cross-lingual)
- Information Extraction/Text Mining
- Question Answering
- Plagiarism/Reuse Detection
- Summarisation
- Categorisation
- Text generation
- Dialogue Systems
- …

Text Mining: Scenario





Text Mining Scenario Components: Texts



- Genres
  - Newspapers
  - Company reports
  - Web pages/blogs
  - Books
  - Scientific papers
  - Legal documents
- E-Formats
  - Word Documents
  - PDF/Postscript
  - HTML/SGML/XML
- Languages
  - English … French … Greek … Russian … Chinese … Hindi … Sanskrit … Linear B
  - Character encodings: ASCII, ISO 8859, Unicode, …

Text Mining Scenario Components: Users

- User domain of interest
  - Business – competitor intelligence, corporate intranet/memory
  - Scientists/academics – access to literature
  - Military/police intelligence – open source intelligence, intranet
  - Journalists – news archives
- User level of domain expertise
  - Novice/expert
- User linguistic competence
  - Adult/child
  - Native/non-native language speaker
  - Uni/multi-lingual

Text Mining Scenario Components: Information Access Needs

- Ad hoc searching
  - Specific questions: "What year did the Berlin Wall come down?"
  - General background/context: "Tell me about Zakopane"
- Stable intelligence gathering
  - Scenario-related: "Build a database recording new projects in the energy sector: the players, location, energy type, start date, capitilisation"
  - Entity-related: "Build a database of key scientists in the pharma industry: name, employer, position, start and end dates"
- Current awareness
  - Alerting: "Let me know when any papers are published on the crystallographic structure of any lipase"
  - Document selection: "Assemble articles on drug approvals"

Text Mining Scenario Components: Information Access Needs

- Summarisation
  - Single/multi-document: "Summarise the Bulger trial"
- Knowledge discovery

  - Trends/correlations in time series data, e.g. commodity price changes
  - Transitive linkages, e.g between businesses, people, enzymes

Text Mining Scenario Components: Tools

- Information retrieval
- Document clustering/routing/categorisation
- Information extraction
- Summarisation
- Agents
- Web crawlers
- KDD/data mining

What is Information Extraction?

- The Information Extraction (IE) task: from each text in a set of unstructured natural language texts identify information about predefined classes of entities and relationships and record this information in a structured form by either
  - Annotating the source text, e.g. using XML tags; or
  - Filling in a data structure separate from the text, e.g a template or database record

- E.g from financial newswire stories identify those dealing with management succession events and from these extract details of organisations and persons, the post being assumed or vacated, the reason for vacancy, etc.

- IE may also be described as
  - The activity of populating a structured information repository (database) from an unstructured, or free text, information source.
  - The activity of creating a semantically annotated text collection ("The Semantic Web")

Information Extraction: Uses

- The resulting structured information source is then used for some other purpose:
  - Searching/analysis using conventional database queries
  - data-mining
  - generating a summary (perhaps in another language)
  - constructing indices into/within/between the source texts.

Information Extraction: Example

Who's News: @ Burns Fry Ltd.

04/13/94

WALL STREET JOURNAL (J), PAGE B10 <CO>

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Information Extraction: Example

Who's News: @ Burns Fry Ltd.

04/13/94

WALL STREET JOURNAL (J), PAGE B10 <CO>

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Information Extraction: Example

Who's News: @ Burns Fry Ltd.

04/13/94

WALL STREET JOURNAL (J), PAGE B10 <CO>

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Information Extraction: Example

Who's News: @ Burns Fry Ltd.

04/13/94

WALL STREET JOURNAL (J), PAGE B10 <CO>

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Information Extraction: Example

Who's News: @ Burns Fry Ltd.

04/13/94

WALL STREET JOURNAL (J), PAGE B10 <CO>

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Information Extraction: Example

Who's News: @ Burns Fry Ltd.

04/13/94

WALL STREET JOURNAL (J), PAGE B10 <CO>

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

Information Extraction: Example

BURNS FRY Ltd. (Toronto) -- Donald Wright, 46 years old, was named executive vice president and director of fixed income at this brokerage firm. Mr. Wright resigned as president of Merrill Lynch Canada Inc., a unit of Merrill Lynch & Co., to succeed Mark Kassirer, 48, who left Burns Fry last month. A Merrill Lynch spokeswoman said it hasn't named a successor to Mr. Wright, who is expected to begin his new position by the end of the month.

<TEMPLATE-9404130062> :=
DOC_NR: "9404130062“
CONTENT: <SUCCESSION_EVENT-1>

<SUCCESSION_EVENT-1> :=
SUCCESSION_ORG: <ORGANIZATION-1>
POST: "executive vice president"
IN_AND_OUT: <IN_AND_OUT-1> <IN_AND_OUT-2>
VACANCY_REASON: OTH_UNK

<IN_AND_OUT-1> :=
IO_PERSON: <PERSON-1>
NEW_STATUS: OUT
ON_THE_JOB: NO

<IN_AND_OUT-2> :=
IO_PERSON: <PERSON-2>
NEW_STATUS: IN
ON_THE_JOB: NO
OTHER_ORG: <ORGANIZATION-2>
REL_OTHER_ORG: OUTSIDE_ORG

<ORGANIZATION-1> :=
ORG_NAME: "Burns Fry Ltd.“
ORG_ALIAS: "Burns Fry“
ORG_DESCRIPTOR: "this brokerage firm“
ORG_TYPE: COMPANY
ORG_LOCALE: Toronto CITY
ORG_COUNTRY: Canada

<ORGANIZATION-2> :=
ORG_NAME: "Merrill Lynch Canada Inc."
ORG_ALIAS: "Merrill Lynch"
ORG_DESCRIPTOR: "a unit of Merrill Lynch & Co."
ORG_TYPE: COMPANY

<PERSON-1> :=
PER_NAME: "Mark Kassirer"

<PERSON-2> :=
PER_NAME: "Donald Wright"
PER_ALIAS: "Wright"
PER_TITLE: "Mr."

Information Extraction: Techniques

- Hobbs (2000):

  An IE system is a cascade of transducers or modules that at each step add structure and often lose information, hopefully irrelevant, by applying rules that are acquired manually and/or automatically.

- A typical cascade or pipeline:
  - Text zoning
  - Tokenisation
  - Sentence splitting
  - Part of speech tagging
  - Pattern matching/lexical lookup for entity recognition
  - Pattern matching for relation extraction

Information Extraction: Techniques

Person Position Organization

Mr. Wright, executive vice president of Merrill Lynch Canada Inc.,

is-employed-by

Information Extraction: Techniques

Person Position Organization

Mr. Wright, executive vice president of Merrill Lynch Canada Inc.,

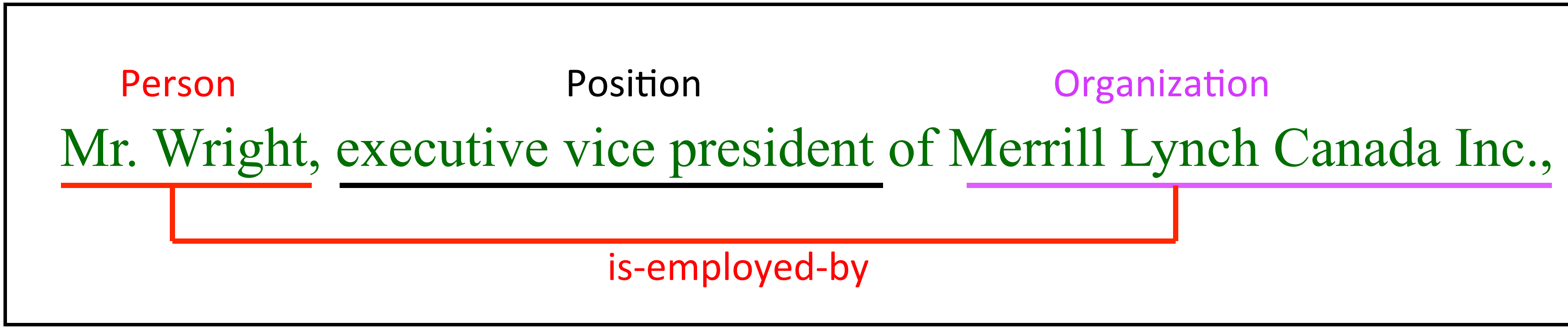
is-employed-by

- **Knowledge engineering approaches** – systems using hand-authored rules
  - "deep" – linguistically inspired "language understanding" systems
  - "shallow" – systems engineered to the IE task, typically using pattern action rules:

Pattern: "Mr. $Uppercase-initial-word"
Action: add-entity(person("Mr. $Uppercase-initial-word))

Pattern: "$Person , $Position of $Organization"
Action: add-relation(is-employed-by($Person,$Organization))

Information Extraction: Techniques

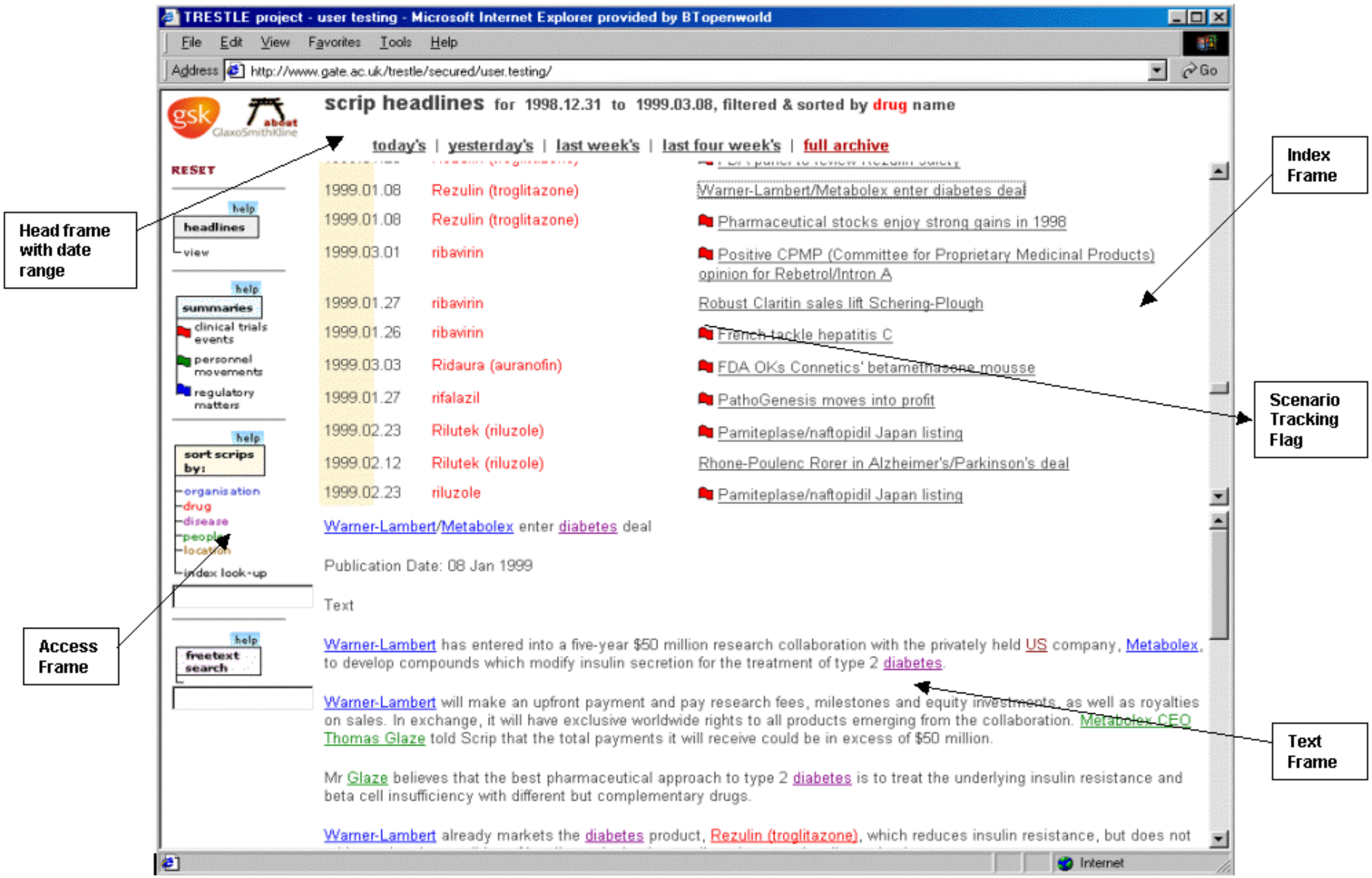


- **Supervised learning approaches**
  - Systems supplied with texts with manually annotated entities + relations
  - For each entity/relation create a training instance
    - k words either side of an entity mention
    - k words to the left of entity 1 and to the right of entity 2 plus the words in between
  - Training instances represented in terms of features
    - words, parts of speech, orthographic characteristics, syntactic information
  - Systems may learn
    - patterns that match extraction targets
    - Classifiers that classify every token as beginning, inside or outside a tag type
  - Learning techniques include: covering algorithms, HMMs, SVMs

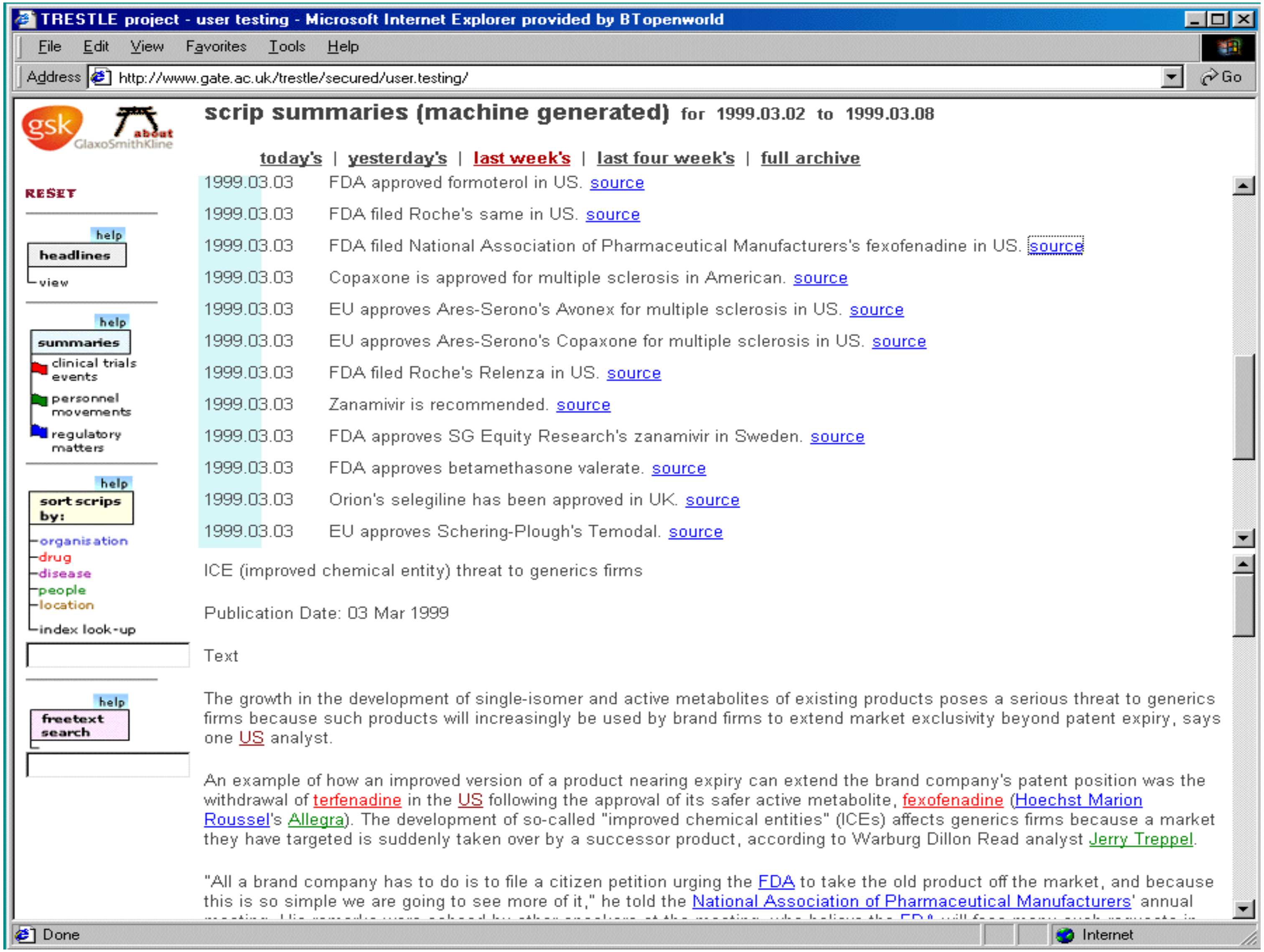
Information Extraction: Techniques

- **Bootstrapping/minimal supervision approaches**
  - Rely on redundancy in large corpora: same information repeated many times in many ways
  - Start with a small number of "seed" tuples known to stand in a given relation (e.g. "is-located-in") and iteratively
    - Label corpus with their occurrences (" … Microsoft … Redmond … ")
    - Extract the contexts of known tuple occurrences (" <left>Microsoft<middle>Redmond<right> ")
    - Group similar contexts to derive matching rules
    - Apply rules to corpus
    - Rank rules by number of tuples matched and "goodness" of matched tuples
    - Select top n rules
    - Re-label corpus with selected rules and extract new tuples to add to seed tuples

NEAT: Named Entity Access to Text



SCAT: Scenario Access To Text

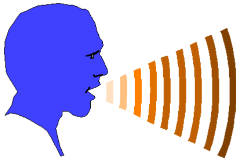


Text Processing Applications

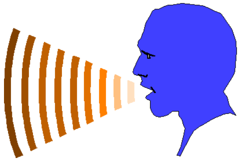
- Spelling correction
- Word prediction in text messaging
- Machine translation
- Document retrieval (mono/cross-lingual)
- Information Extraction/Text Mining
- Question Answering
- Plagiarism/Reuse Detection
- Summarisation
- Categorisation
- Text generation
- Dialogue Systems
- …

Dialogue Systems

- Conversational computers a staple of science fiction for generations – HAL, R2D2, CP3O, Deep Thought, Marvin, Data …
- Computers capable of talking with humans in open ended dialogue still a far way off
- But conversational agents that participate in limited task-based dialogues are approaching viability
- Motivated by scenarios where voice is required
  - Telephone information services/bookings/call centres
  - Hands-free environments -- equipment operation
  - Support for the disabled
- Example: Amities -- Automated Multilingual Interaction with Information and Services



Dialogue Systems: Task-oriented Dialogues



S1: Hello. You've reached the [Communicator]. Tell me your full name

U2: hi I'd like to fly to seattle tuesday morning

S3: Traveling to Seattle on Tuesday, August eleventh in the morning. Your full name?

U4: [John Doe]

S5: Hello, [John], I'm glad to hear from you again. You can now ask me for help.

S6: OK. Let's see I have a United flight departing Pittsburgh at ten forty five a.m. arriving into Seattle at one fifty seven p.m. Is that OK?

U7: Yes

S8: Will you return to Pittsburgh from Seattle?

U9: no i will go to Los Angeles

S10: Traveling to Los Angeles. on what date

U11: Sorry can we go back to talk about the previous leg again

S12: For the first leg, you are on a United flight 1446 departing Pittsburgh at ten forty five A.M. tomorrow, arriving in Seattle at one fifty seven P.M. Do you want to take that?

U13: actually i want to leave on wednesday

S14: a flight on wednesday.

S15: I have a flight departing Pittsburgh at ten forty five a.m., arrives Seattle at one fifty seven p.m. Is that OK?

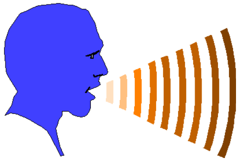
U16: Yes

S17: OK and on what day do you want to depart Seattle?

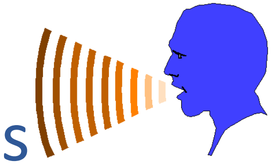
U18: next Monday

S19: Traveling on Monday, August sixteenth.

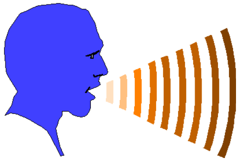
CMU Communicator System, Xu & Rudnicky, 2000



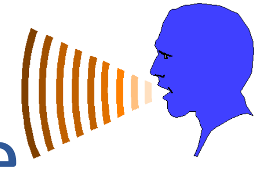
Dialogue Systems: Task-oriented Dialogues



- Some popular domains for task-oriented dialogue systems:
  – Travel enquiries/bookings (airlines, trains, buses,boats)
  – Telephone call routing/banking
  – Tutoring systems
  – Remote access to email
  – Weather/movie/restaurant information
- Task-oriented dialogues
  – Constrain vocabulary and grammar
  – Define a natural "shape" for the dialogue, based on a a small set of mutually understood goals
  – Do not require a large amount of arbitrary world knowledge
  – Allow the system largely to control the dialogue or regain control after any confusion
  – Have near term commercial promise

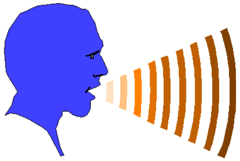


Additional Challenges of Dialogue

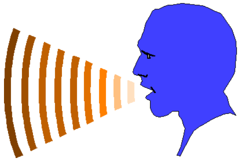


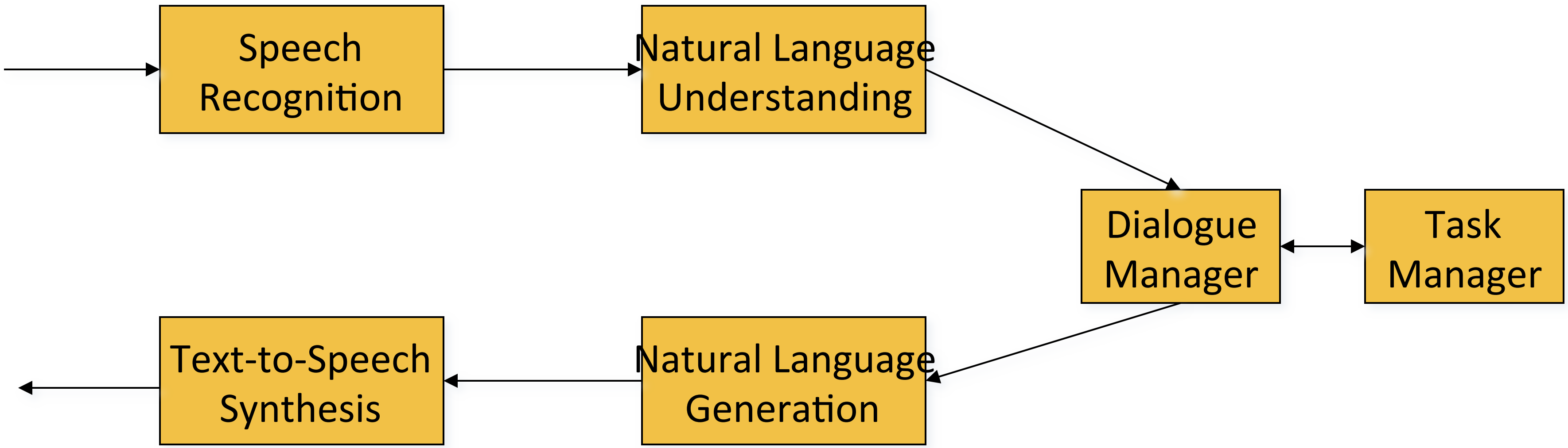
- Speech recognition
- Real-time
- Form of joint activity between two or more participants
  - Turn taking
  - Common ground -- set of things mutually believed -- must be established and maintained
  - Inference plays a very significant role -- conversation guided by implicit maxims/heuristics
    - Be exactly as informative as required
    - Try to make your contribution one that is true
    - Be relevant
    - Be perspicuous

Much of what goes on in dialogue to do with establishing maintaining joint activity and NOT information going from A to B

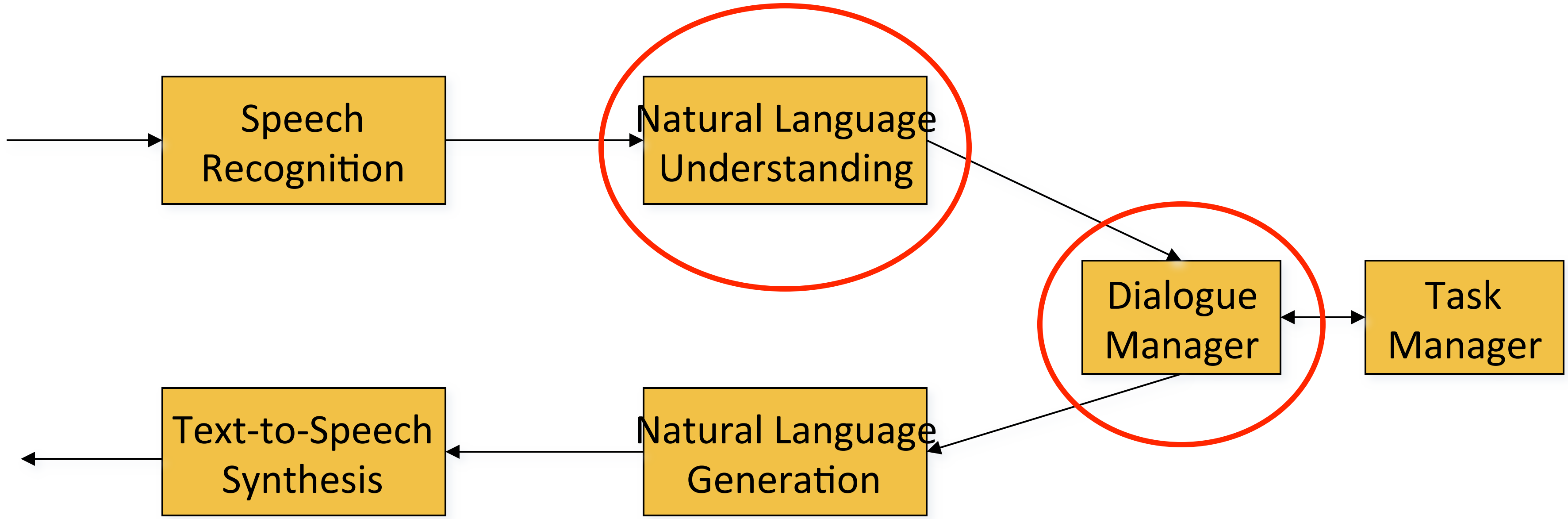


Typical Dialogue System Architecture





Typical Dialogue System Architecture

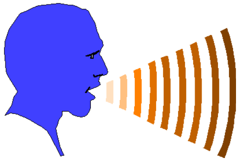


Dialogue Systems: Techniques (1)

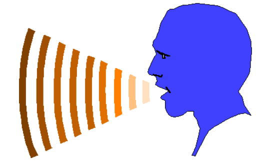
- Simplest task-oriented dialogue systems revolve around filling a form or frame

Slot	Question
Origin City	"From what city are you leaving?"
Destination City	"Where are you going?"
Departure Time	"When would you like to leave?"
Arrival Time	"When do you want to arrive?"

- Grammars/language models used to recognize input tailored for frame
- States of dialogue manager correspond to slots in frame
  - Actions are slot/state specific
  - Users responses are interpreted as appropriate for the current slot/state



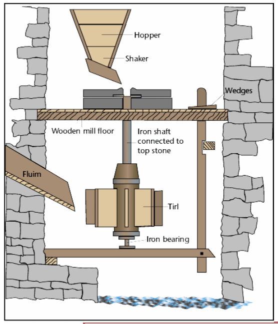
Dialogue Systems: Techniques (2)



- To move beyond form-filling applications a more sophisticated model is needed
- One approach proposes
  - Supplement the NLU with a dialogue act tagger – determines whether user utterance is a question, a confirmation, request for clarification, assertion, etc.
  - Split discourse manager into
    - A representation of the current information state – discourse context, beliefs, goals, user model, task context, etc.
    - A behavioural agent which contains rules to update the information state and a control structure
- Such an approach highly non-deterministic
  - Many ways to realize dialogue system's goals
  - Recent work on use of re-inforcement learning + Markov decision processes to optimize choice of next action by dialogue system and on learning parameters of the decision process from data

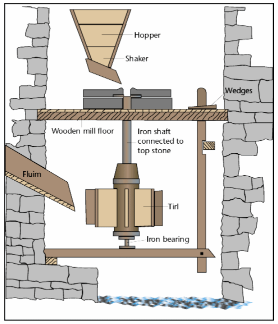
Outline

- The Significance of Language
- Turing and the Imitation Game
- Why Language Processing is so hard for a Computer
- Two Frameworks for the Computational Analysis of Language
  - Symbolic
  - Statistical
- Applying Partial Knowledge: Text Processing Applications
- Summing Up and The Road Ahead



Summing Up

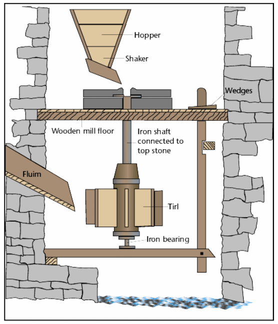
- We started by asking two questions:
  1. Why has the Turing Test not been passed yet?
  2. What have researchers been doing for the past 50 years to try to get computers to understand natural language?
- What can we now say about these questions?



Summing Up

1. Why has the Turing Test not been passed yet?

- Human language is very complex and our understanding of it is still limited
- Interpreting language requires a huge amount of extra-linguistic knowledge. Not clear
  - What the extent/nature of this knowledge is
  - How to get it into a computer/cause a computer to acquire it



Summing Up

2. What have researchers been doing for the past 50 years to try to get computers to understand natural language?

- Working in predominantly two frameworks
  - one deriving from descriptive linguistics
  - one deriving from information theory and statistical sequence analysis techniques

  to build computational models that emulate some human language processing capabilities
- This work has led to
  - Wide range of insights into properties of human languages
  - Host of practical applications

The Way Forward I

Refining/extending existing techniques – "more of the same"

- Plenty of scope for refining and extending current techniques
  - In MT: seek to exploit comparable corpora – vastly greater volumes
  - In IE: lots of on-going work to address specific problems such as
    - Extraction of spatial and temporal information
    - Weakly supervised learning – learning from a minimal set of annotated examples
    - Generating training data automatically using existing structured sources, e.g. Wikipedia Infoboxes
  - In Dialogue: further exploring data-driven techniques for learning dialogue models from examples (e.g. partially observable Markov decision processes)

The Way Forward II

Are our modelling/explanatory frameworks adequate?

- Descriptive linguistics based on theoretical notions evolved within linguistics in the process of systematic description of natural languages
  - Little sign linguists are moving towards a consensus
  - Like systematic description in other sciences, seek deeper explanatory framework
- Statistical sequence analysis eschews theoretical notions from linguistics to concentrate on regularities in the data
  - Language understanding a function not just of the language (signal) being processed, but of the understander too
- Suggest current approaches need to be informed by insights from
  - Cognitive and developmental psychology
  - Neuroscience and study of the language system in the brain

The Way Forward III

Getting world knowledge into the machine

- Understanding language requires understanding a lot about the world
- We get this knowledge through perception and experience and through cultural transmission
- Efforts to hand code world knowledge for use by computers have failed (pace Cyc)
- Current efforts are addressing
  - Induction of ontologies from texts
  - Population of “fact repositories” from on-line sources
- Can a way be found to bootstrap the knowledge a machine needs to interpret text from those texts themselves?

Questions?