Introduction to Computational Linguistics (CL)

J. Savoy
Université de Neuchâtel

Some slides from C.D. Manning

Outline

- **The CL domain**
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Question / Answering

Linguistics

- What is Linguistics?

  - The origins of language
  - Animals and human language
  - The development of writing
  - The sounds of language
  - The sound patterns of language
  - Words and word-formation processes
  - Morphology
  - Phrases and sentences: grammar
  - Syntax
  - Semantics

G. Yule: The Study of Language. Cambridge University Press, 2008

Linguistics

- Pragmatics
- Discourse analysis
- Language and the brain
- First language acquisition
- Second language acquisition / learning
- Gestures and sign languages
- Languages history and change
- Language and regional variation
- Language and social variation
- Language and culture

G. Yule: The Study of Language. Cambridge University Press, 2008

Computational Linguistics

- In Computational Linguistics (CL)?
- Related domains
  - Mathematics: probability theory, statistics, information theory
  - Computer science: representation & processing
  - Linguistics
- Why today?
  - Huge amount of texts available on-line
  - Need tools to process them

Computational Linguistics

- Processes
  - Text segmentation
  - Part-of-speech (POS) tagging
  - Parsing
  - Word-Sense Disambiguation (WSD)
  - Anaphora Resolution
  - Natural Language Generation
  - Speech Recognition
  - Text-to-Speech Synthesis
  - Evaluation

R. Mitkov (Ed): The Oxford Handbook of Computational Linguistics of Language. Oxford University Press, 2003

Computational Linguistics

- Methods & Resources
  - Finite-State (FS) Technology
  - Statistical Methods
  - Logic Programming
  - Machine Learning (ML)
  - Corpus Linguistics
  - Ontologies
  - Sublanguage and Controlled Languages

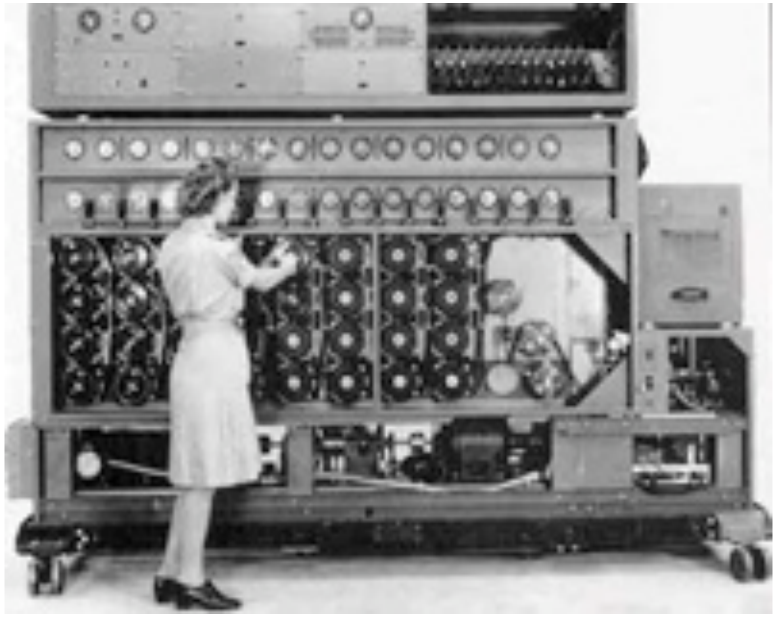
R. Mitkov (Ed): The Oxford Handbook of Computational Linguistics of Language. Oxford University Press, 2003

Outline

- The CL domain
- **A. Turing**
- Numbers and data formats
- Database
- The real problems
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Question / Answering

Computational Linguistics

- Meaning of “to compute a text”? Beyond a simple text-processing!



Computational Linguistics

- CL research questions
  - Can we infer the meaning by computing a document?
  - Can we translate automatically a document?
  - Can we summarize (automatically) a document?
  - Can we find automatically the answer to a question (facts, yes/no, lists, definition)?
  - Can you retrieve documents on a given topic?
  - Can you represent (index) this document collection?
  - Can we correct the spelling of a document?

Computational Linguistics

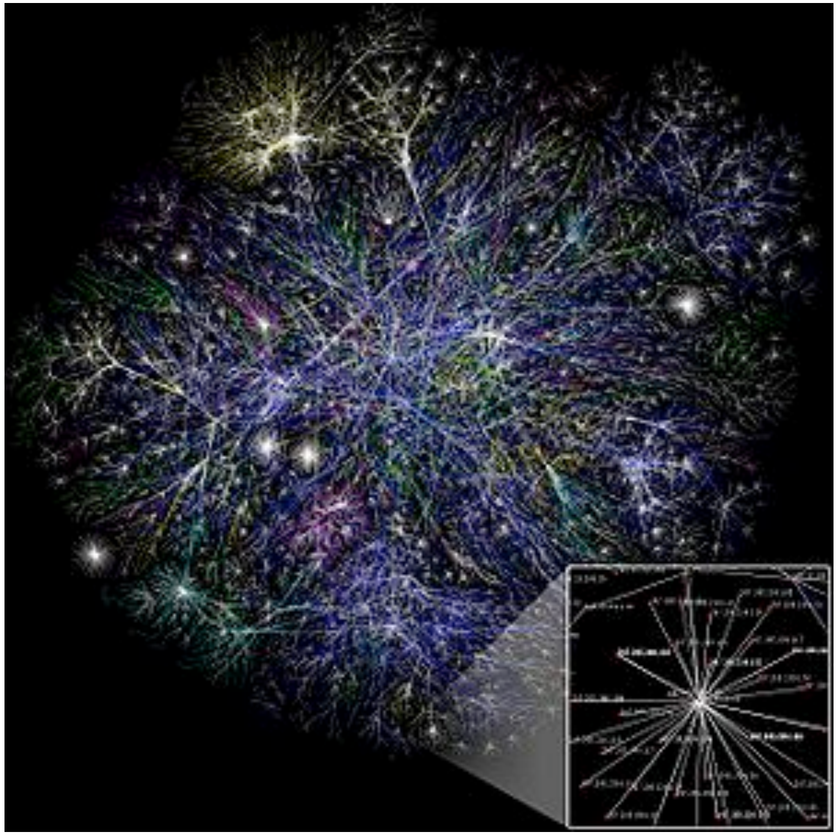
- CL methods
  - Program a computer
  - Efficiency (speed)
  - Effectiveness (quality)
  - Reliable and robust processing (errors)
  - Represent text / document
- « Der Teufel liegt im Detail »

Computational Linguistics

- The main differences?
  - Deal with large amount of data
  - Discover *mean* effects (usually not rare event)
  - Mainly data-driven
  - Interest in both recognition and generation
  - Semantics (the Holy *Grail* search)
- Focus
  - Written
  - Available on electronic support
  - Lexical, morphology

Computational Linguistics

- The current context
  - Web
  - Open-source
  - Lot of opportunities



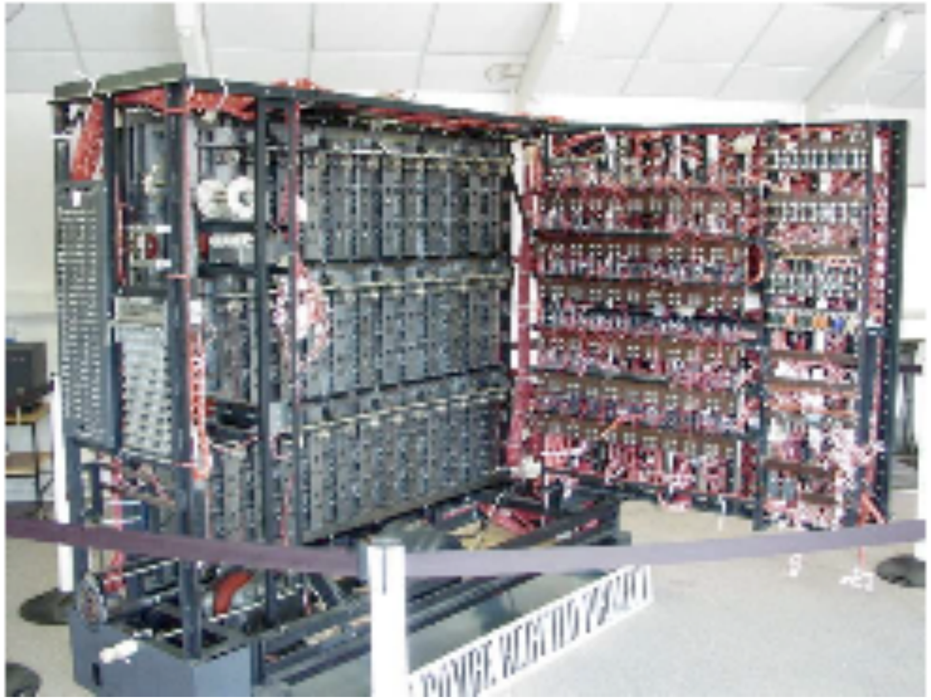
Alan M. Turing (1912 – 1954)

- Alan Turing: English mathematician and logician
- "*On Computable Numbers with an Application to the Entscheidungsproblem*" (1936) The Turing machine: the first universal programmable computer

From Computer Desktop Encyclopedia
Reproduced with permission.
© 2001 The Computer Museum History Center

Alan M. Turing (1912 – 1954)

- Code breaker (at Bletchley Park, with his“Bombe” – Colossus - to decipher Enigma cryptosystem)



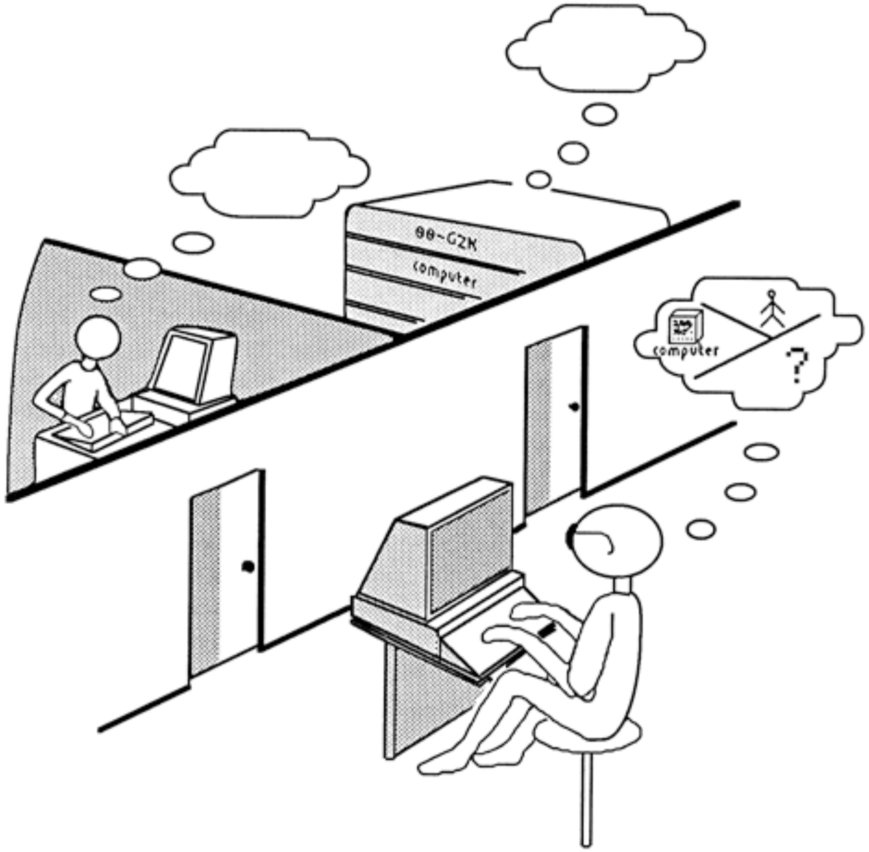
Turing Test

- Alan Turing: Computing Machinery and Intelligence” (1950)

- Can machines think?
  - Need to define “machine” & “intelligence”
  - “Can machines do what we (as thinking entities) can do?”
  - Natural language conversation
- Current technology cannot pass this test
  When?

Turing Test

- The principle of the test
  - Natural language understanding
  - Knowledge representation
  - Automatic reasoning (inference)
  - Machine learning



Turing Test (Eliza)

Example with ELIZA (J. Weizenbaum, MIT, 1966)

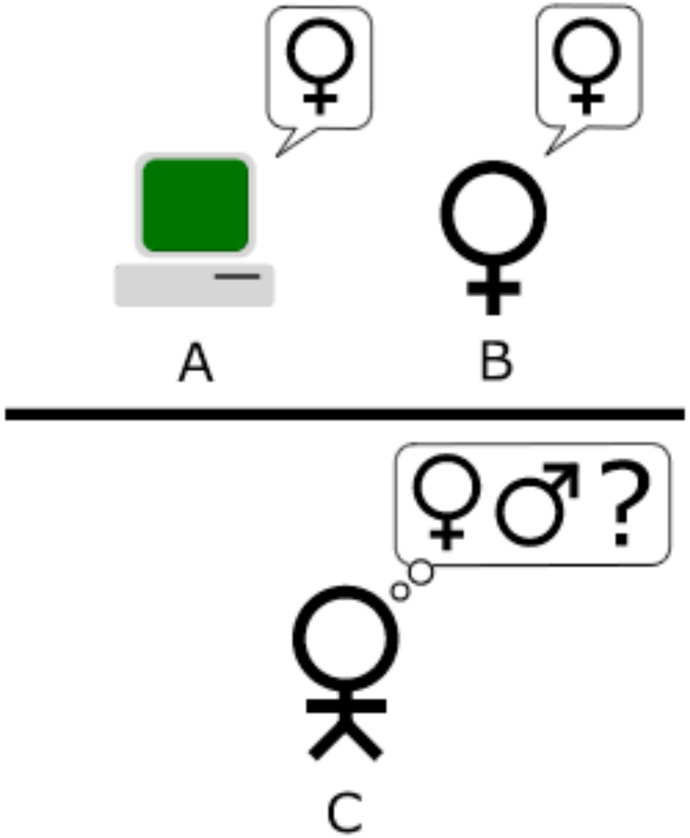
Girl : Men are all alike.
Eliza : In what way ?
Girl : They’re always bugging us about something or other.
Eliza : Can you think of a specific example?
Girl : Well, my boyfriend made me come here.
Eliza : Your boyfriend made you come here?
Girl : He says I’m depressed much of time.
Eliza : I am sorry to hear you are depressed.
Girl : It’s true. I am unhappy.
Eliza : Do you think coming here will help you not to be unhappy?
Girl : I need some help, that much seems certain.
Eliza : What would it mean to you if you got some help?

Turing Test (Eliza)

Girl : Perhaps I could learn to get along with my mother.
Eliza : Tell me more about your family.
…
Girl : My father hated me.
Eliza : Tell me more about your family.
…
Girl : My boyfriend made me come here.
Eliza : Your boyfriend made your come here ?
…
Girl : You don’t seem quite certain.
Eliza : Why do you remember … just now ?
…

Turing Test

- Variant
  - Imitation game
- Full Test
  - Vision
  - Robotic
- Restricted test
  - Politician
  - Schizophrenia (Parry, 1972)



Turing Test

- J. Searle (*Minds, Brains and Programs*, 1980)
  Passing the test will prove nothing.
  It is not sufficient to say that computers are intelligent
  - Syntactic nature of computers (Chinese room) (such as `Eliza`, manipulating symbols without understand them)
  - Semantic nature of human beings
- R. French
  This not necessary (the flight of a bird sees with a radar)

Turing Test

- Loebner Price (1991)
  - ALICE
  - Ai Jabberwacky
- A. Turing: A computer will pass the Turing test in 2000 with 10^9 bits of memory (around 120 MB) in five-minute test
- We think to pass it around 2029

Outline

- The CL domain
- A. Turing
- **Numbers and data formats**
- Database
- The real problems
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - **Question / Answering**

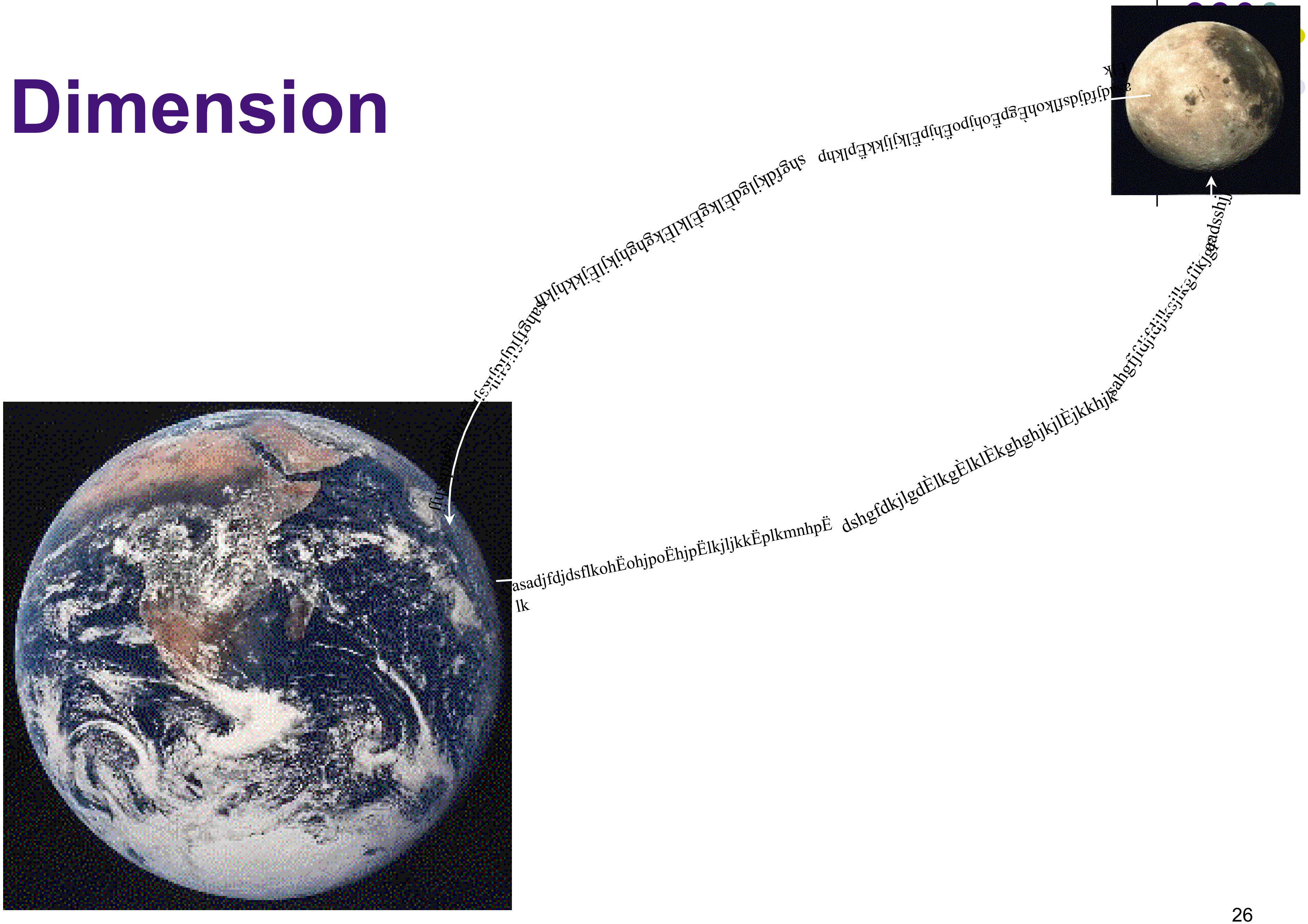
Dimension

Unit	1	1	1
Kilo	K	1000	10^{3}
Mega	M	1000000	10^{6}
Giga	G	1000000000	10^{9}
Tera	T	1000000000000	10^{12}
Peta	P	1000000000000000	10^{15}
milli	m	0.0001	10^{-3}
micro	µ	0.0000001	10^{-6}
nano	n	0.0000000001	10^{-9}
pico	p	0.0000000000001	10^{-12}

Dimension

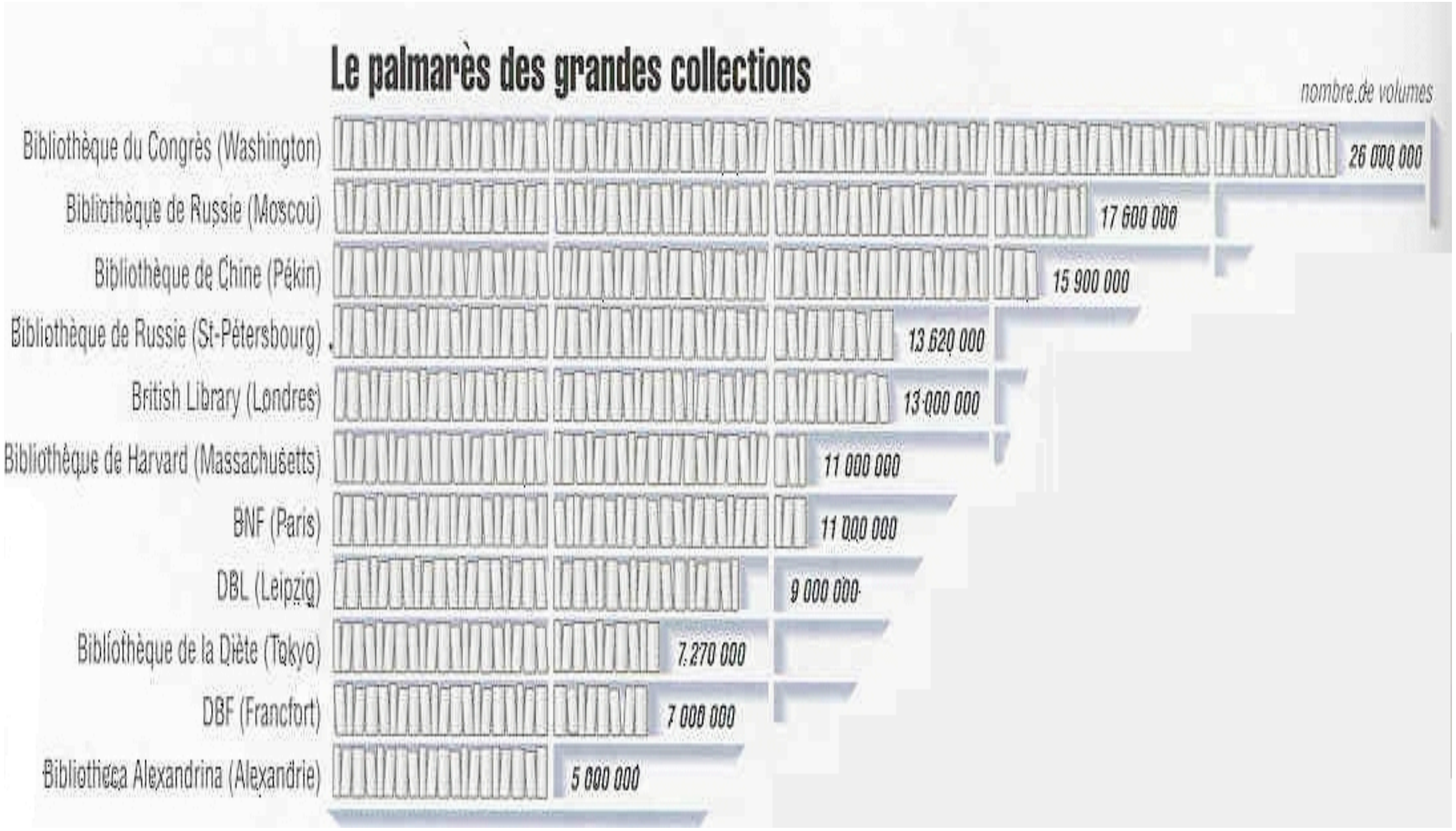
- Difference between 1 bit (Binary Digit) and 1 byte (= 8 bits = 1 character).
- Convention: “a” (97) = 0011 0001
- 1 page = 400 words, 5 char/word = 2,000 char/page
- 1 romance = 500 pages = 1,000,000 char = 1 MB
- One terabyte (1 TB = 1,000 x 1,000 MB) of text means?
- Count 1 mm per character, we can go from:
  - Geneva to Paris?
  - Geneva to Montreal?
  - Geneva to Canberra? (and return?)
  - 1 m = 1 K char, 1 Km = 1M char

Dimension


asadjfdjdsflkohËohjpoËhjpËlkjljkkËplkmnhpË dshgfdkjlgdÈlkgÈlklÈkghghjkjlÈjkkhjksahgfjfdjfdjlksjlksjlkgfikjgadsshj
lk

Dimension

- Large libraries across the world



Dimension

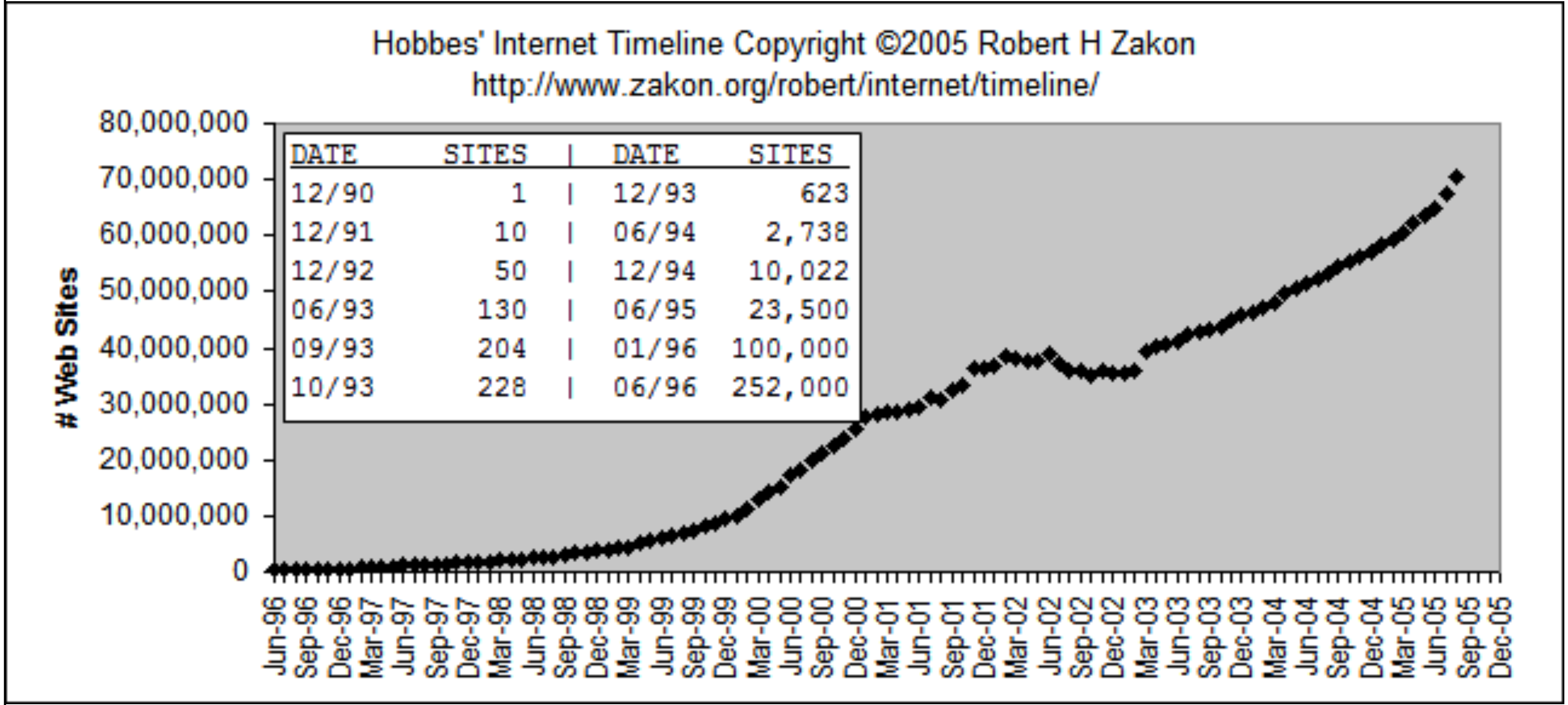
- 1 romance = 500 pages = 1,000,000 char = 1 MB = 10^6 char
- British Library = 13 MB = 13 M x 10^6 characters
- In characters? 13 x 10^6 x 10^6 = 13 x 10^{12} = 13 TB
- Other media
  - 1 symphony = 14 MB to 30 MB (MP3)
  - 1 film = 4 GB (4.1 GB on a DVD)
- And the Web (estimation in 2002)
  - visible: 167 TB
  - Hidden Web: 91,850 TB
  - E-mails: 440,606 TB

Lyman P., Varian H. R. *How much information? 2003*, available at the web site www.sims.berkeley.edu /how-much-info/

In short

- 1 romance = 1 MB = 10^6 char
- 1 symphony = 14 MB
- 1 film = 4 GB
- Web visible: 167 TB
- Hidden Web: 91,850 TB
- will continue to exist ...

In short



Hobbes' Internet Timeline Copyright ©2005 Robert H Zakon
http://www.zakon.org/robert/internet/timeline/

DATE	SITES	DATE	SITES
12/90	1	12/93	623
12/91	10	06/94	2,738
12/92	50	12/94	10,022
06/93	130	06/95	23,500
09/93	204	01/96	100,000
10/93	228	06/96	252,000

Sources

- What are the available resources in CL
  - Spoken documents
  - Written documents
  - Dictionaries (bilingual) & Terminological resources
- Main providers
  - LDC Linguistic Data Consortium (UPenn)
  - ELRA European Linguistic Resources Association (Paris)
- Public
  - BNC British National Corpus
  - Project Gutenberg
  - Various Digital Libraries (DL) projects

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- **Database**
- The real problems
- Technology
    - OCR
    - NL/DB interface, Web / IR search
    - Product Info, e-mail
    - Text categorization, clustering
    - Small devices, chat rooms
    - Information Extraction (IE)
    - Machine Translation (MT)
    - Question / Answering

What’s the world’s most used database?

- Oracle (SQL)?
- Excel?
- Perhaps, Microsoft Word?
- Largest database (Feb. 2007)

1. World Data Centre for Climate (220 PB)
2. National Energy Research Scientific Computing Center (2.8 PB)
3. AT&T (312 TB)
4. Google
5. Sprint
6. ChoicePoint (LexisNexis)
7. YouTube (45 TB)
8. Amazon (42 TB)
9. Central Intelligence Agency
10. Library of Congress (20 TB)

“Databases” in 1990

- Database systems (mostly relational) are the pervasive form of information technology providing efficient access to structured, tabular data primarily for governments and corporations: Oracle, Sybase, Informix, etc.

ID	Name	Address	Salary
1253	Tintin	Moulinsart 10	5,780
2345	Tournesol	Liberty 3	10,090
345	Dupont	Central 6a	5,600
674	Dupond	Central 6b	5,600

“Databases” in 1990

- Database systems (mostly relational)

- (Text) Information Retrieval systems is a small market dominated by a few large systems providing information to specialized markets (legal, news, medical, Intellectual Property IP (patents), corporate info): Westlaw, Medline, Lexis/Nexis
- Commercial NLP market basically nonexistent
  - mainly research area

“Databases” in 2010

- A lot of new things seem important:
  - Internet, Web search, Portals, Peer-to-Peer, Agents, XML/Metadata, Data mining
- Is everything the same (new buzzwords), different, or just a mess?
- There is more of everything, it’s more distributed, and it’s *less structured.*
- Large textbases and information retrieval are a crucial component of modern information systems, and have a big impact on everyday people (web search, portals, email)

Linguistic data is ubiquitous

- Most of the information in most companies, organizations, etc. is material in human languages (reports, customer email, web pages, discussion papers, text, sound, video) – not stuff in traditional databases
  - Estimates: 70%, 90%? (all depends how you measure). Most of it.
- Most of that information is now available in digital form:
  - Estimate for companies in 1998: about 60% More like 90% now?

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- **The real problems**
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Question / Answering

The problem

- When people see text, they understand its meaning (by and large)
- When computers see text, they get only character sequences (and perhaps HTML tags)
- We'd like computer agents to see *meanings* or be able to intelligently process text
- Why is Natural Language Understanding (NLU) so complex?

Why is Natural Language Understanding difficult?

1. Infinite diversity of sentences
   a. the vocabulary is not completely known (Out-Of-Vocabulary problem OOV)
   b. the set of constructions is itself not completely predetermined
   c. the set of senses attributed to each word is also not completely predetermined
2. Tolerance of errors (robustness)

Why is Natural Language Understanding difficult?

Not always so easy

- "Three witches watch three Swatch watches. Which witch watch which Swatch watch?"
  ("Trois sorcières regardent trois montres Swatch. Quelle sorcière regarde quelle montre Swatch ?")
- "From two to two to two two" (1H58 à 2H02)
- "Three Swedish switched witches watch three Swiss Swatch watch switches. Which Swedish switched witch watches which Swiss Swatch watch switch?"

  ("Trois sorcières suédoises et transsexuelles regardent les boutons de trois montres Swatch suisses. Quelle sorcière suédoise transsexuelle regarde quel bouton de quelle montre Swatch suisse ?")

Why is Natural Language Understanding difficult?

3. Implicit elements

a. contraction

« Anne promised that she would be on time. »

b. anaphoric reference

« Mr Major arrived in France today. The prime minister will meet the President tomorrow. The Conservative leader will then travel to Moscow where he will meet Mr Gorbachev. Mrs Major will join her husband in Russia, where this son of a circus artist is relatively unknown figure. »

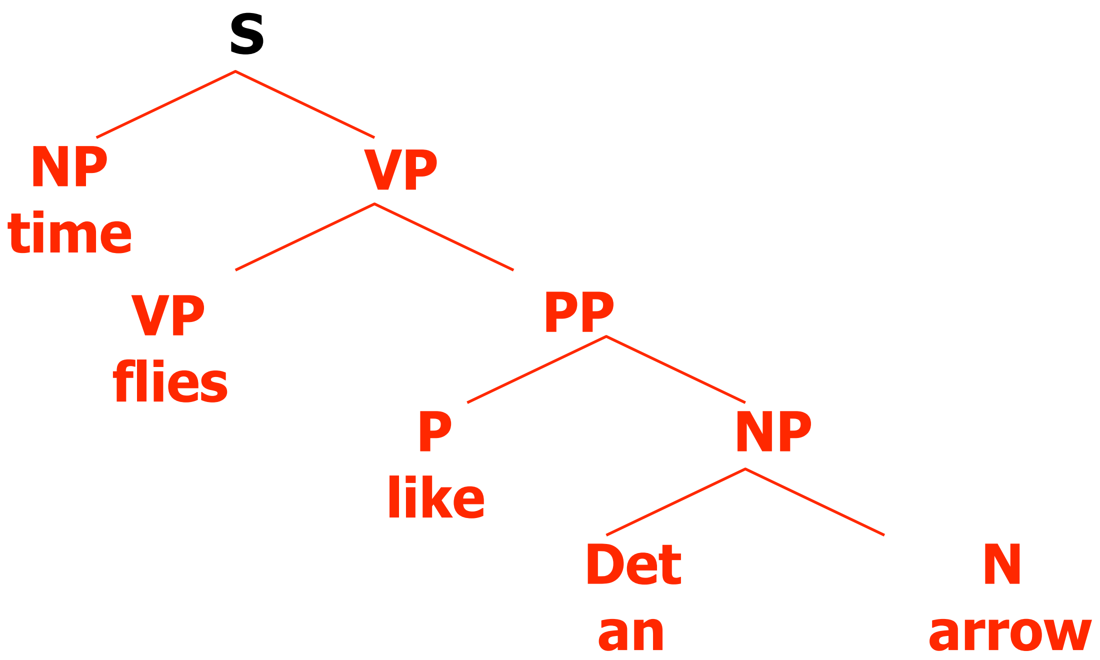
c. ellipses

« John is having dinner with Mary tomorrow night, and Paul with Susan . »

Why is Natural Language Understanding difficult?

4. The hidden structure of language is highly ambiguous

Structures for: *Times flies like an arrow*

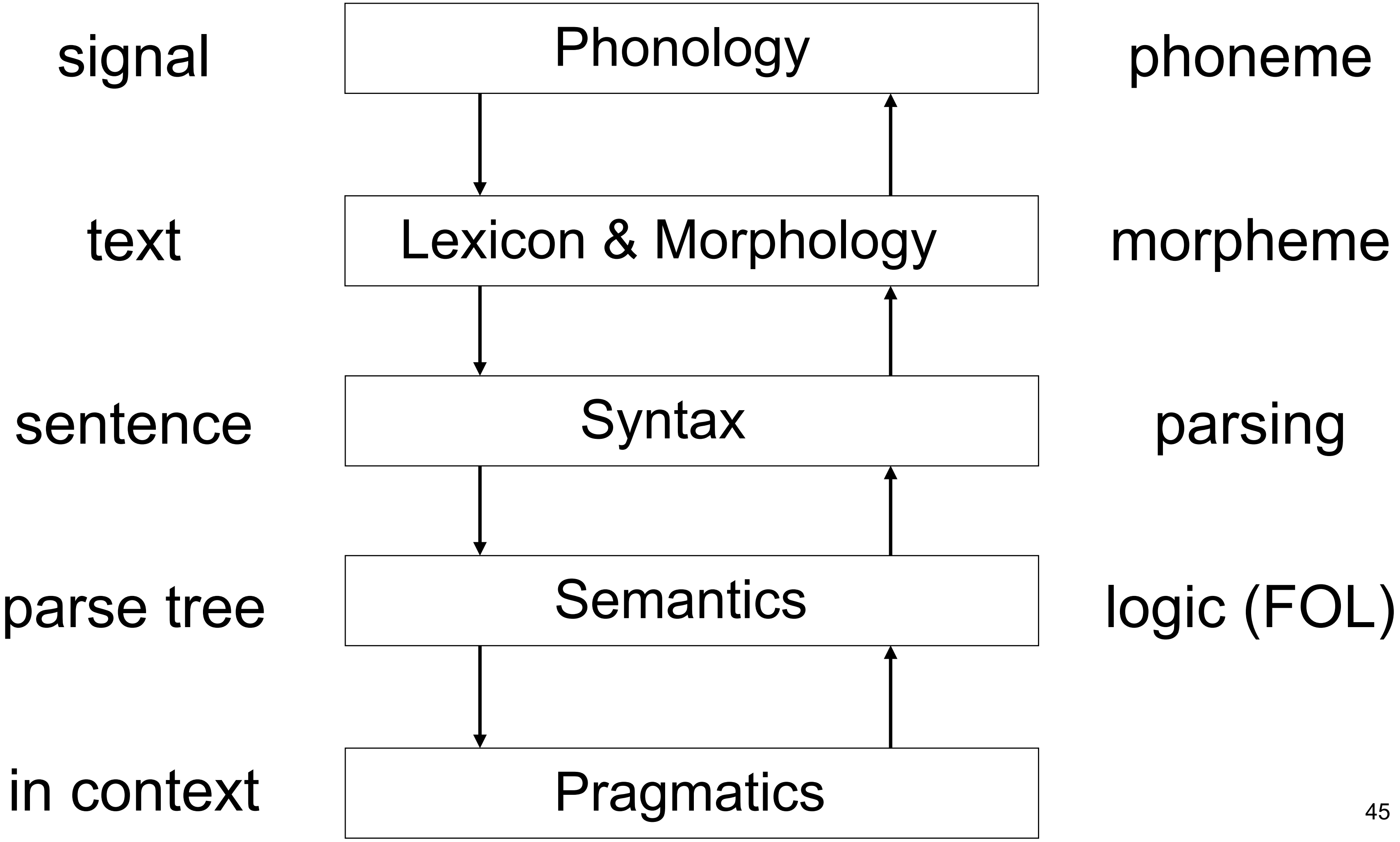


Where are the ambiguities?

- Syntactic attachments could be complex and lead to ambiguities
  ”The old woman was a witness of sexual relationship between two cars”
- Part of speech ambiguities
  ("saw" as a tool or a verb form)
- Word sense ambiguities & homographs
  "bat" (baseball vs. mammals)
  "PRC" vs. "China"
- Other examples

  "The ink is in the pen"
  "The pig is in the pen"

Domains of NLP
(Recognition & Synthesis)



signal	Phonology	phoneme
text	Lexicon & Morphology	morpheme
sentence	Syntax	parsing
parse tree	Semantics	logic (FOL)
in context	Pragmatics	

Linguistics

- The human language is based on a double articulation principle
  - A limited number of phonemes may generate an infinite number of words
  - A limited number of morphemes may generate a infinite number of (sentences) meanings
- The human language is based on an arbitrary relationship between forms (strings, significant) and meaning (signifié) (Saussure)
- But here: robustness, redondancy

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- **Technology**
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Question / Answering

Practical applied NLP goals

Use language technology to add value to *data* by:

- interpretation
- transformation
- value filtering
- augmentation (providing metadata)

Two motivations:

- The amount of information in textual form
- Information integration needs NLP methods for coping with ambiguity and context

Terms and technologies

- Word processing / Desktop publishing
  - Spelling detection / correction
  - Dictionary access
  - Internationalization / Translations aids
  - Controlled vocabularies.
- Some success
  - Word processing
  - Google
  - TREC / MUC (NIST)
  - MT with Systran

Terms and technologies

- Text processing
  - Stuff like TextPad (Emacs), Perl, grep. Semantics and structure blind, but does what you tell it in a nice enough way. Still useful.
- Information Retrieval (IR)
  - Implies that the computer will try to find documents which are relevant to a user while understanding nothing (big collections)
- Intelligent Information Access (IIA)
  - Use of clever techniques to help users satisfy an information need (search or UI innovations)

Terms and technologies

- Locating small stuff. Useful nuggets of information that a user wants:
  - Information Extraction (IE): Database filling
    - The relevant bits of text will be found, and the computer will understand enough to satisfy the user’s communicative goals
  - Wrapper Generation (WG) [or Wrapper Induction]
    - Producing filters
  - Question Answering (QA) – NL querying
  - Thesaurus/key phrase/terminology generation

Terms and technologies

- Big Stuff. Information Management
  Overviews of data (condense the data):
  - Summarization
    - Of one document or a collection of related documents (cross-document summarization)
  - Categorization (documents)
    - Including text filtering and routing
  - Clustering (collections)
- Text segmentation: subparts of big texts
- Topic detection and tracking (business intelligence)
  - Combines IE, categorization, segmentation

Terms and technologies

- Digital libraries (DL)
  with text, sound, images, pictures, video
  with different natural languages (Europe)
- Text (Data) Mining (DM)
  - Extracting nuggets from text. Opportunistic.
  - Unexpected connections that one can discover between bits of human recorded knowledge.
- Natural Language Understanding (NLU)
  - Implies an attempt to completely understand the text…
- Machine translation (MT), Speech recognition, etc.
  - Now available wherever software is sold!

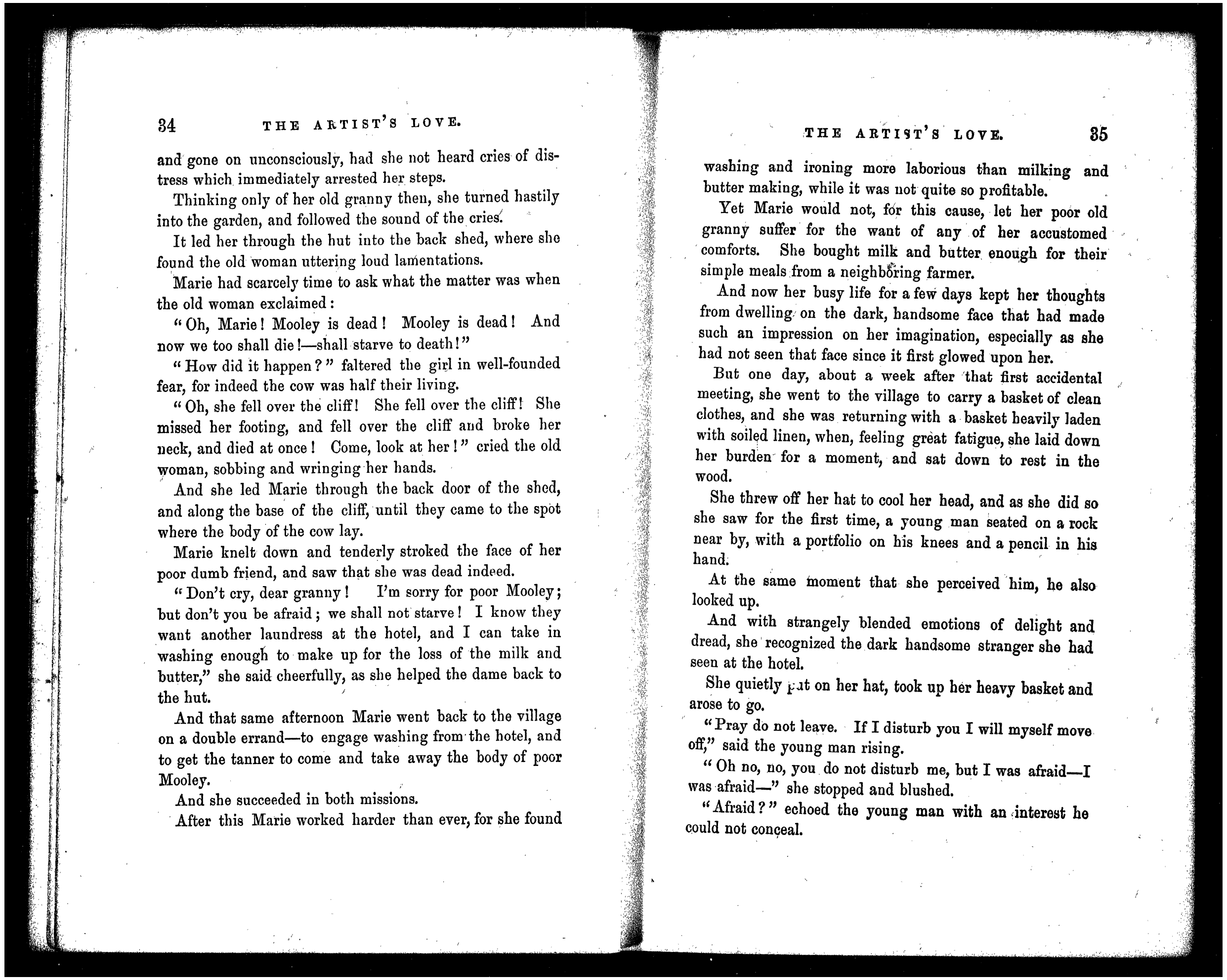
Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
  - **OCR**
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Question / Answering

Added value

- The encoding problem
  - character representation (ASCII, ISO-Latin1, EUR-JP, UniCode, etc.)
    Different standards for the same language
  - input & output devices
  - RTF (Word), PDF, SGML, HTML, XML
- Digitalizing (OCR)
  Is it always the (perfect) solution?

OCR, Typical page



and gone on unconsciously, had she not heard cries of distress which immediately arrested her steps.

Thinking only of her old granny then, she turned hastily into the garden, and followed the sound of the cries.

It led her through the hut into the back shed, where she found the old woman uttering loud lamentations.

Marie had scarcely time to ask what the matter was when the old woman exclaimed:

"Oh, Marie! Mooley is dead! Mooley is dead! And now we too shall die!—shall starve to death!"

"How did it happen?" faltered the girl in well-founded fear, for indeed the cow was half their living.

"Oh, she fell over the cliff! She fell over the cliff! She missed her footing, and fell over the cliff and broke her neck, and died at once! Come, look at her!" cried the old woman, sobbing and wringing her hands.

And she led Marie through the back door of the shed, and along the base of the cliff, until they came to the spot where the body of the cow lay.

Marie knelt down and tenderly stroked the face of her poor dumb friend, and saw that she was dead indeed.

"Don't cry, dear granny! I'm sorry for poor Mooley; but don't you be afraid; we shall not starve! I know they want another laundress at the hotel, and I can take in washing enough to make up for the loss of the milk and butter," she said cheerfully, as she helped the dame back to the hut.

And that same afternoon Marie went back to the village on a double errand—to engage washing from the hotel, and to get the tanner to come and take away the body of poor Mooley.

And she succeeded in both missions.

After this Marie worked harder than ever, for she found

washing and ironing more laborious than milking and butter making, while it was not quite so profitable.

Yet Marie would not, for this cause, let her poor old granny suffer for the want of any of her accustomed comforts. She bought milk and butter enough for their simple meals from a neighboring farmer.

And now her busy life for a few days kept her thoughts from dwelling on the dark, handsome face that had made such an impression on her imagination, especially as she had not seen that face since it first glowed upon her.

But one day, about a week after that first accidental meeting, she went to the village to carry a basket of clean clothes, and she was returning with a basket heavily laden with soiled linen, when, feeling great fatigue, she laid down her burden for a moment, and sat down to rest in the wood.

She threw off her hat to cool her head, and as she did so she saw for the first time, a young man seated on a rock near by, with a portfolio on his knees and a pencil in his hand.

At the same moment that she perceived him, he also looked up.

And with strangely blended emotions of delight and dread, she recognized the dark handsome stranger she had seen at the hotel.

She quietly put on her hat, took up her heavy basket and arose to go.

"Pray do not leave. If I disturb you I will myself move off," said the young man rising.

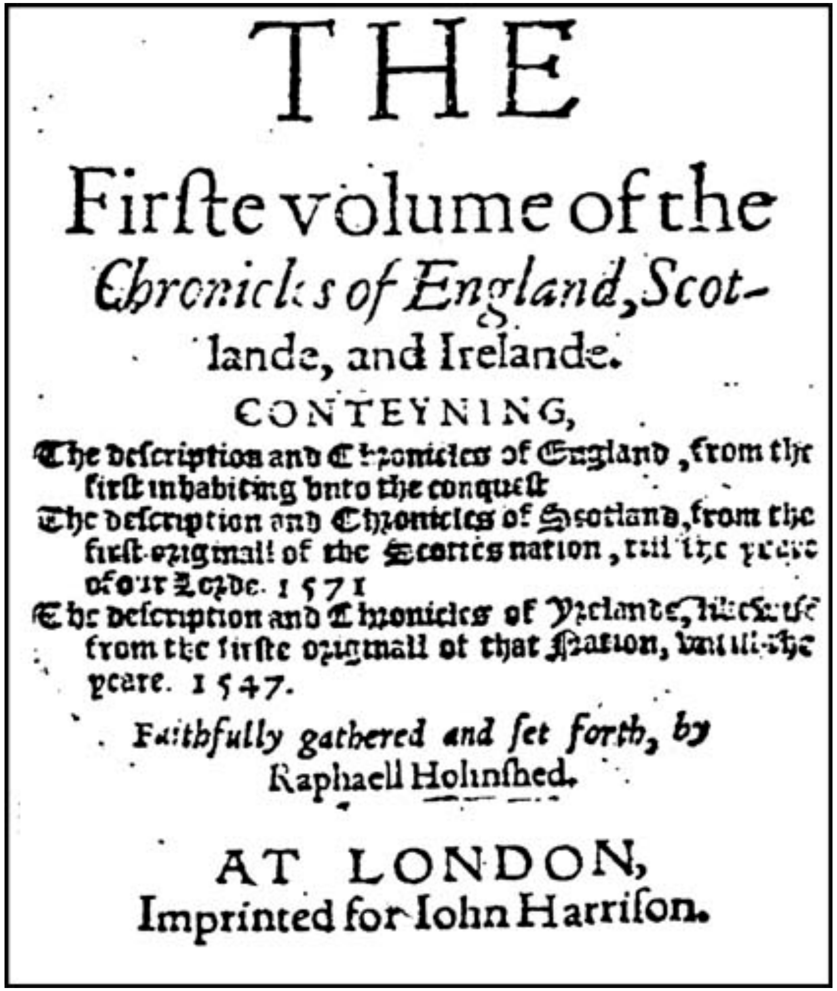
"Oh no, no, you do not disturb me, but I was afraid—I was afraid—" she stopped and blushed.

"Afraid?" echoed the young man with an interest he could not conceal.

OCR result

```
34      THE  ARTIST'S  LOVE.
and gone on unconsciously, had she not heard cries- of dis-
tress which immediately arrested her steps.
Thinking only of her old granny then, she turned hastily
into the garden, and followed the sound of the cries.
It led her through the hut into the back shed, where she
found the old woman uttering loud lamentations.
Marie had scarcely time to ask what the matter was when
the old woman exclaimed:
"Oh, Marie! Mooley is dead! Mooley is dead! And
now we too shall die!-shall starve to death!"
"How did it happen?" faltered the girl in well-founded
fear, for indeed the cow was half their living.
"Oh, she fell over the cliff! She fell over the cliff! She
missed her footing, and fell over the cliff and broke her
neck, and died at once! Come, look at her!" cried the old
woman, sobbing and wringing her hands.
And she led Marie through the back door of the shed,
and along the base of the cliff, until they came to the spot
where the body of the cow lay.
Marie knelt down and tenderly stroked the face of her
poor dumb friend, and saw that she was dead indeed.
```

OCR, Typical page



THE

Firſte volume of the
Chronicles of England, Scot-
lande, and Irelande.

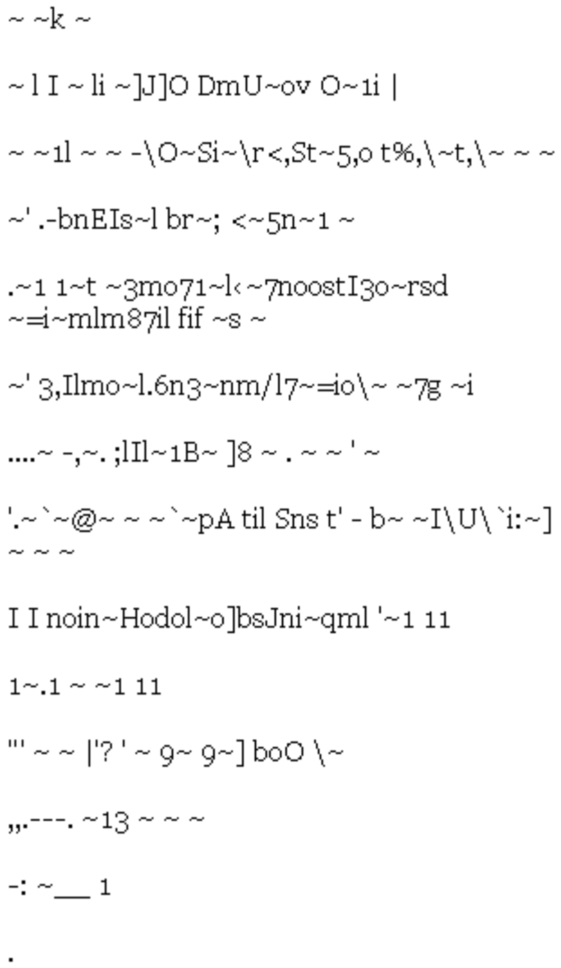
CONTEYNING,

The deſcription and Chronicles of England, from the
firſt inhabiting vnto the conqueſt
The deſcription and Chronicles of Scotland, from the
firſt originall of the Scottes nation, till the yeare
of our Lorde. 1571
The deſcription and Chronicles of Yrelande, likewiſe
from the firſte originall of that Nation, vntill the
yeare. 1547.

Faithfully gathered and ſet forth, by
Raphaell Holinſhed.

AT LONDON,
Imprinted for Iohn Harriſon.

OCR result



```
~ ~k ~

~ l I ~ li ~]J]O DmU~ov O~1i |

~ ~1l ~ ~ -\O~Si~\r<,St~5,o t%,\~t,\~ ~ ~

~' .-bnEIs~l br~; <~5n~1 ~

.~1 1~t ~3mo71~l‹ ~7noostI3o~rsd
~=i~mlm87il fif ~s ~

~' 3,Ilmo~l.6n3~nm/l7~=io\~ ~7g ~i

....~ -,~. ;lIl~1B~ ]8 ~ . ~ ~ ' ~

'.~`~@~ ~ ~`~pA til Sns t' - b~ ~I\U\`i:~]
~ ~ ~

I I noin~Hodol~o]bsJni~qml '~1 11

1~.1 ~ ~1 11

"' ~ ~ |'? ' ~ 9~ 9~] boO \~

,,.---. ~13 ~ ~ ~

-: ~___ 1

.
```

Added value

- The encoding problem
  - character representation (ASCII, ISO-Latin1, EUR-JP, Unicode, etc.)
  - input & output devices
  - RTF (Word), PDF, SGML, HTML, XML (www.tei-c.org/P4X/SG.html)
- Digitalizing (OCR)
- **Text Encoding Initiative** (TEI, www.tei-c.org) to make explicit what is implicit.
- Preservation (media disappear, data remain)

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
  - OCR
  - **NL/DB interface, Web / IR search**
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Question / Answering

Natural Language Interfaces to Databases

- This was going to be the big application of NLP in the 1980s
  - How many service calls did we receive from US last month?
  - I am listing the total service calls to US for November 2008.
  - The total for November 2008 was 1756.
- It has been recently integrated into MS SQL Server (English Query)
- Problems: need largely hand-built custom semantic support

NLP for IR / Web search?

- It’s a no-brainer that NLP should be useful and used for web search (and IR in general):
  - Search for ‘Jaguar’
    - the computer should know or ask whether you’re interested in big cats [scarce on the web], cars, or, perhaps a molecule geometry and solvation energy package, or a package for fast network I/O in Java
  - Navigation vs. transaction vs. information
  - Google and stemming*:*
    - Search for "*probabilistic model*", and "*probabilistic models*" (or "Jeu de Nim" and "Nîmes")

NLP for IR / Web search?

- Word sense disambiguation (WSD) technology generally works well in text categorization
- Synonyms can be found or listed
  - in limited context?
- Lots of people have been suggested solutions (startups)
  - iPhrase "Traditional keyword search technology is hopelessly outdated"
  - The future is here?

NLP for IR / Web search?

- But in practice it’s an idea that hasn’t gotten much traction
  - Correctly finding linguistic base forms is usually straightforward, but produces little advantage over crude stemming which just slightly over equivalence classes words
  - Word sense disambiguation only helps on average in IR if over 90% accurate (Sanderson SIGIR'94), and that’s about where we are
  - Syntactic phrases should help, but people have been able to get most of the mileage with “statistical phrases” – which have been aggressively integrated into systems recently

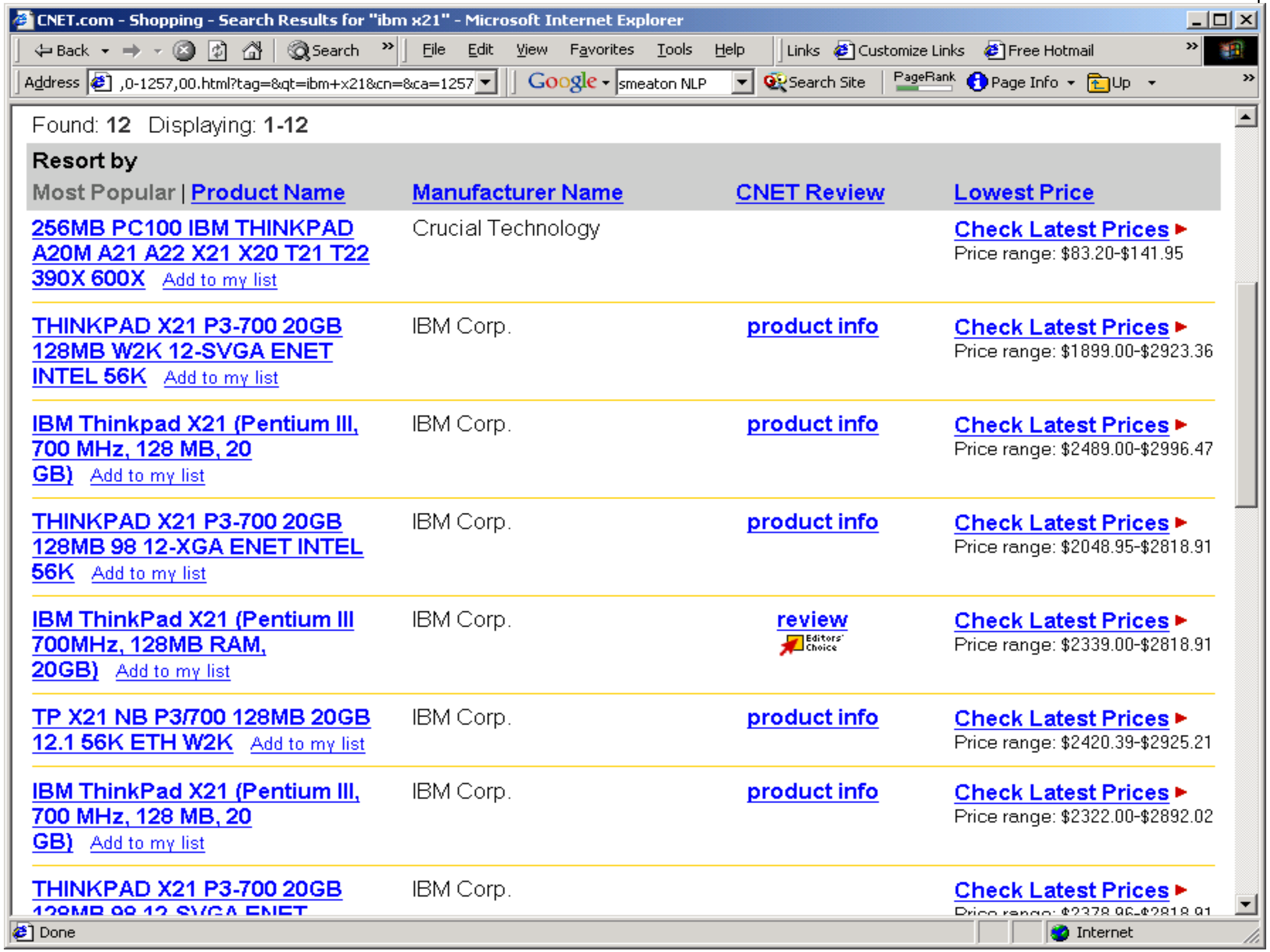
NLP for IR / Web search?

- People can easily scan among results (on their 21” monitor) … if you’re above the fold
- Much more progress has been made in link analysis, and use of anchor text, etc.
- Anchor text gives human-provided synonyms
- Link or click stream analysis gives a form of pragmatics: what do people find correct or important (in a default context)
- Focus on short, popular queries, news, etc.
- Using human intelligence always beats artificial intelligence (Turing test)

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - **Product Info, e-mail**
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - **Question / Answering**

Product information



Found: **12** Displaying: **1-12**

Resort by

Most Popular |

Product Name	Manufacturer Name	CNET Review	Lowest Price
256MB PC100 IBM THINKPAD A20M A21 A22 X21 X20 T21 T22 390X 600X Add to my list	Crucial Technology		Check Latest Prices ▸ Price range: $83.20-$141.95
THINKPAD X21 P3-700 20GB 128MB W2K 12-SVGA ENET INTEL 56K Add to my list	IBM Corp.	product info	Check Latest Prices ▸ Price range: $1899.00-$2923.36
IBM Thinkpad X21 (Pentium III, 700 MHz, 128 MB, 20 GB) Add to my list	IBM Corp.	product info	Check Latest Prices ▸ Price range: $2489.00-$2996.47
THINKPAD X21 P3-700 20GB 128MB 98 12-XGA ENET INTEL 56K Add to my list	IBM Corp.	product info	Check Latest Prices ▸ Price range: $2048.95-$2818.91
IBM ThinkPad X21 (Pentium III 700MHz, 128MB RAM, 20GB) Add to my list	IBM Corp.	review Editors' Choice	Check Latest Prices ▸ Price range: $2339.00-$2818.91
TP X21 NB P3/700 128MB 20GB 12.1 56K ETH W2K Add to my list	IBM Corp.	product info	Check Latest Prices ▸ Price range: $2420.39-$2925.21
IBM ThinkPad X21 (Pentium III, 700 MHz, 128 MB, 20 GB) Add to my list	IBM Corp.	product info	Check Latest Prices ▸ Price range: $2322.00-$2892.02
THINKPAD X21 P3-700 20GB 128MB 98 12-SVGA ENET	IBM Corp.		Check Latest Prices ▸ Price range: $2378.96-$2818.91

Product information/ Comparison shopping, etc.

- Need to learn to extract info from online vendors
- Can exploit uniformity of layout, and (partial) knowledge of domain by querying with known products
- Example: E-commerce agent.
  Most commerce is currently done *manually*. But there is no reason to suppose that certain forms of commerce could not be safely delegated to agents
  - “finding the cheapest copy Office 2007 from online stores”
  - “flight from Zurich to New York with veggie meal, window seat”
  - Gives convenient aggregation of online content
- Bug: not popular with vendors

Email handling

- Big point of pain for many people
- There just aren't enough hours in the day!
  - even if you're not a customer service rep
- What kind of tools are there to provide an electronic secretary?
  - Negotiating routine correspondence
  - Scheduling meetings
  - Filtering junk
  - Summarizing content
- "The web's okay to use; it's my email that is out of control"

Email handling

“The web’s okay to use; it’s my email that is out of control”

Copyright 2003 by Randy Glasbergen.
www.glasbergen.com

“I found a solution to your spam problem. I’ve set up your e-mail to automatically delete any message with a vowel in it.”

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - **Text categorization, clustering**
  - Small devices, chat rooms
  - Information Extraction (IE)
  - Machine Translation (MT)
  - Question / Answering

Text Categorization is a task with many potential uses

- Take a document and assign it a label representing its content (MeSH heading, ACM keyword, Yahoo! category). Categories are *pre-defined*.
- Classic example: decide if a newspaper article is about politics, business, or sports?
- There are many other uses for the same technology:
  - Is this page a laser printer product page?
  - Does this company accept overseas orders?
  - What kind of job does this job posting describe?
  - What kind of position does this list of responsibilities describe?
  - What position does this this list of skills best fit?
  - Is this the “computer” or “harbor” sense of *port*?

Text Categorization

Three phases

1. Feature extraction & selection
   - Lexical analysis
   - Select the most important features
2. Document representation
   - Set of terms?
   - Combining with binary features
3. Induction
   Usually, simple machine learning (ML) algorithms are used to compare the representation and to take the final decision (required a training set)

Text Categorization

- Accuracy is more dependent on:
  - Naturalness of classes.
  - Quality of features extracted and amount of training data available.
- Accuracy typically ranges from 65% to 97% depending on the situation
  - Non-semantic bearing terms included in the feature set may degrade the effectiveness.
- Multi-lingual text categorization:
  - Simple multiple independent monolingual systems?

Email response: “eCRM”

electronic Customer
Relationship
Marketing

© 2009 by Randy Glasbergen.
www.glasbergen.com

“If you’d like to press 1, press 3.
If you’d like to press 3, press 8.
If you’d like to press 8, press 5...”

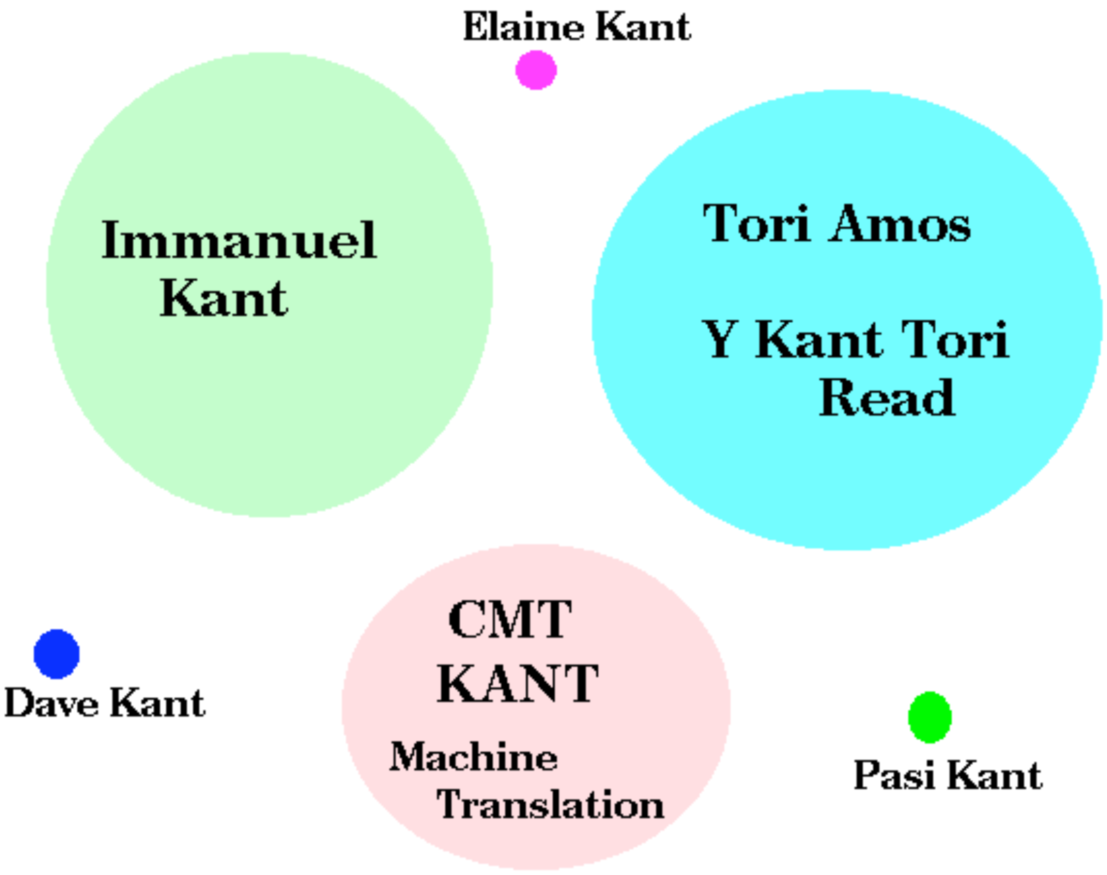
Email response: “eCRM”

- electronic Customer Relationship Marketing
- Automated systems which attempt to categorize incoming email, and to automatically respond to users with standard, or frequently seen questions
- Most but not all are more sophisticated than just keyword matching
- Generally use text classification techniques
- Can save real money by doing 50% of the task close to 100% right (e.g., Bell Canada)

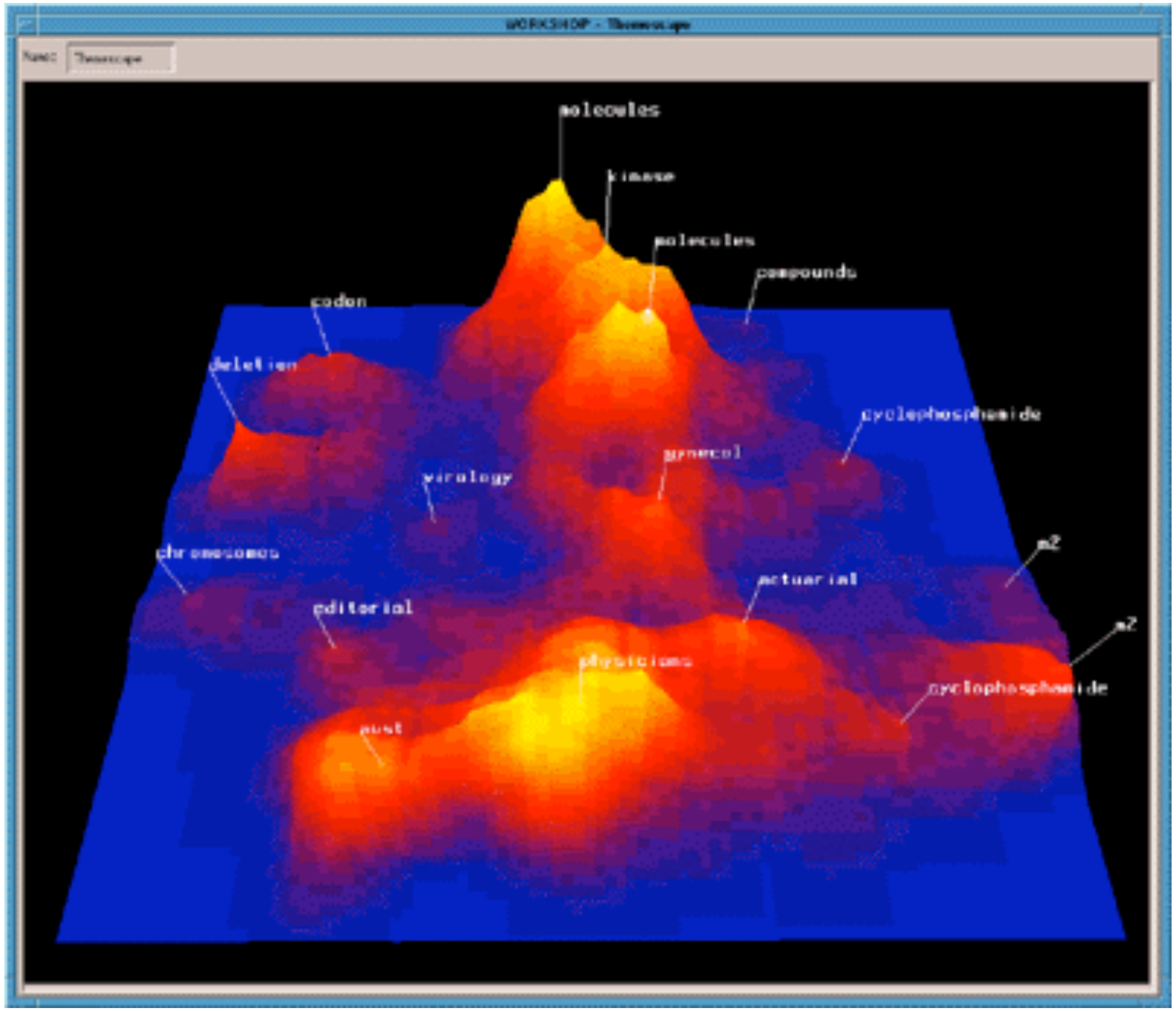
Text Clustering in Browsing, Search and Organization

- Scatter/Gather Clustering
  - Cutting, Pedersen, Karger, Tukey ’92, ’93
- Cluster sets of documents into general “themes”, like a table of contents
- Display the contents of the clusters by showing topical terms and typical titles
- User chooses subsets of the clusters and re-clusters the documents within them
- Resulting new groups have different “themes”

Clustering (of query *Kant*)



Clustering a Multi-Dimensional Document Space (image from Wise et al. 95)



Clustering

- June 11, 2001: The latest KDnuggets Poll asked: What types of analysis did you do in the past 12 months.
  - The results, multiple choices allowed, indicate that a wide variety of tasks is performed by data miners. Clustering was by far the most frequent (22%), followed by Direct Marketing (14%), and Cross-Sell Models (12%)
- Clustering of results can work well in certain domains (e.g., biomedical literature)
- But it doesn’t seem compelling for the average user, it appears (Altavista, Northern Light)

Citeseer/ResearchIndex

- An online repository of papers, with citations, etc. Specialized search with semantics in it
- Great product; research people love it
- However it’s fairly low tech. NLP could improve on it:
  - Better parsing of bibliographic entries
  - Better linking from author names to web pages
  - Better resolution of cases of name identity
    - E.g., by also using the paper content

Chat rooms/groups/discussion forums/usenet

- Many of these are public on the web
- The signal to noise ratio is very low
- But there's still lots of good information there
- Some of it has commercial value
  - What problems have users had with your product?
  - Why did people end up buying product X rather than your product Y
- Some of it is time sensitive
  - Rumors on chat rooms can affect stock price
    - Regardless of whether they are factual or not

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
    - OCR
    - NL/DB interface, Web / IR search
    - Product Info, e-mail
    - Text categorization, clustering
    - **Small devices, chat rooms**
    - Information Extraction (IE)
    - Machine Translation (MT)
    - Question / Answering

Small devices

- With a big monitor, humans can scan for the right information
- On a small screen, there's *hugely* more value from a system that can show you what you want:
  - phone number
  - business hours
  - email summary
    - "Call me at 11 to finalize this"

Small devices

© 2000 Randy Glasbergen.
www.glasbergen.com

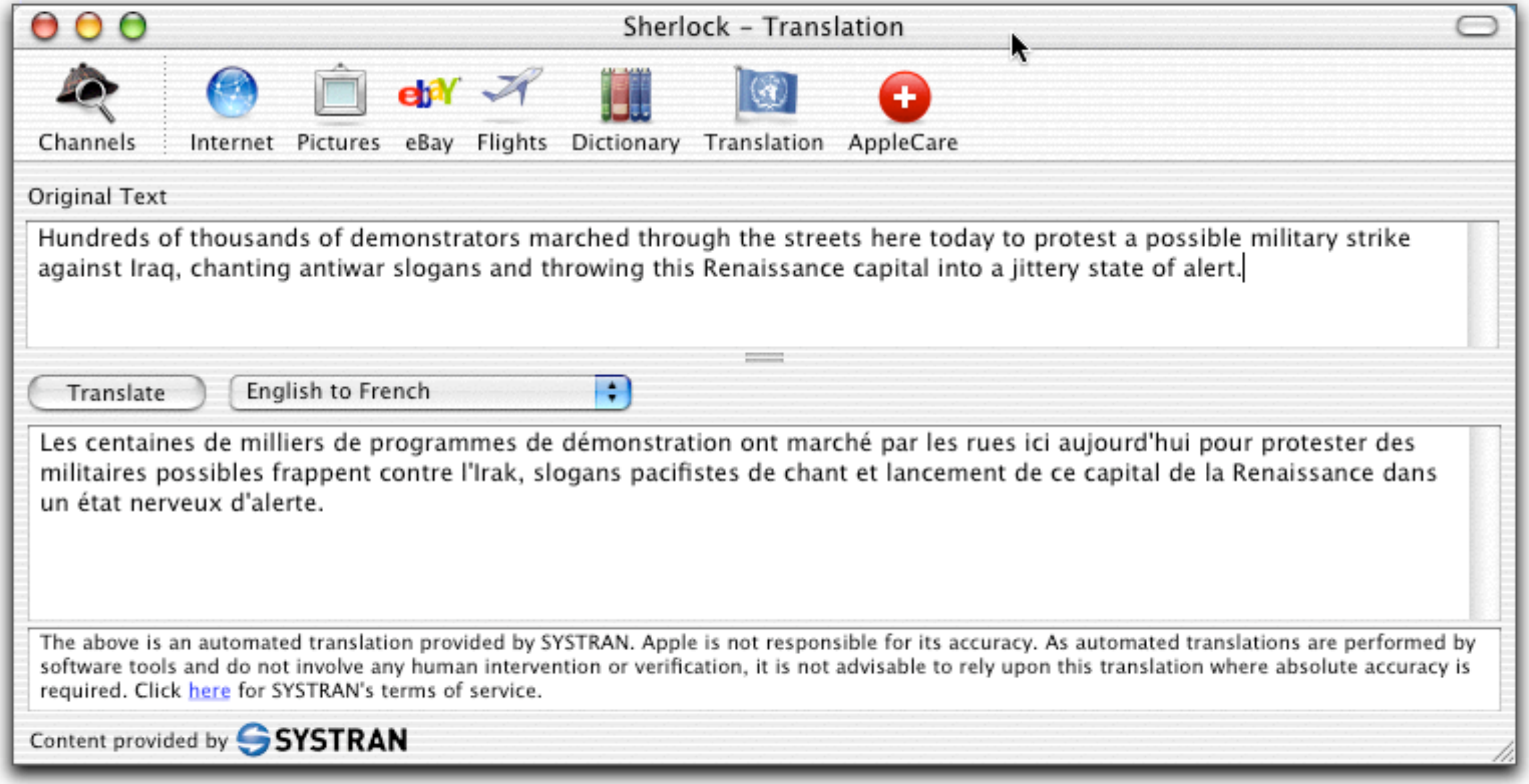
"E-mail, voice mail, web pages, stock quotes, news, banking...that's a lot of responsibility for such a little guy!"

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - Information Extraction (IE)
  - **Machine Translation (MT)**
  - Question / Answering

Machine translation

- High quality MT is still a distant goal



Original Text

Hundreds of thousands of demonstrators marched through the streets here today to protest a possible military strike against Iraq, chanting antiwar slogans and throwing this Renaissance capital into a jittery state of alert.

English to French

Les centaines de milliers de programmes de démonstration ont marché par les rues ici aujourd'hui pour protester des militaires possibles frappent contre l'Irak, slogans pacifistes de chant et lancement de ce capital de la Renaissance dans un état nerveux d'alerte.

The above is an automated translation provided by SYSTRAN. Apple is not responsible for its accuracy. As automated translations are performed by software tools and do not involve any human intervention or verification, it is not advisable to rely upon this translation where absolute accuracy is required. Click here for SYSTRAN's terms of service.

Content provided by SYSTRAN

Machine translation

- High quality MT is still a distant goal
- But MT is effective for scanning content
- And for machine-assisted human translation
- Dictionary use accounts for about half of a traditional translator's time.
  (word in context)
- Printed lexical resources are not up-to-date
- Electronic lexical resources ease access to terminological data.

Machine translation

- Various resources could be used
  - Official Journal of the European Union
  - United Nations
  - Wikipedia
  - Web

  to have a parallel corpus (statistical translation model)
- More difficult to translate proper nouns
- “Translation memory” systems: remember previously translated documents, allowing automatic recycling of translations

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
  - OCR
  - NL/DB interface, Web / IR search
  - Product Info, e-mail
  - Text categorization, clustering
  - Small devices, chat rooms
  - **Information Extraction (IE)**
  - Machine Translation (MT)
  - Question / Answering

Task: Information Extraction

Suppositions:

- A lot of information that *could* be represented in a structured semantically clear format isn't
- It may be costly, not desired, or not in one's control (screen scraping) to change this.

- Goal: being able to answer semantic queries using "unstructured" natural language sources

Information Extraction

- Information extraction systems
  - Find and understand relevant parts of texts.
  - Produce a structured representation of the relevant information: *relations* (in the DB sense)
  - Combine knowledge about language and the application domain
  - Automatically extract the desired information
- When is IE appropriate?
  - Clear, factual information (who did what to whom and when?)
  - *Only a small portion of the text is relevant.*
  - ***Some errors can be tolerated***

Examples of Existing IE Systems

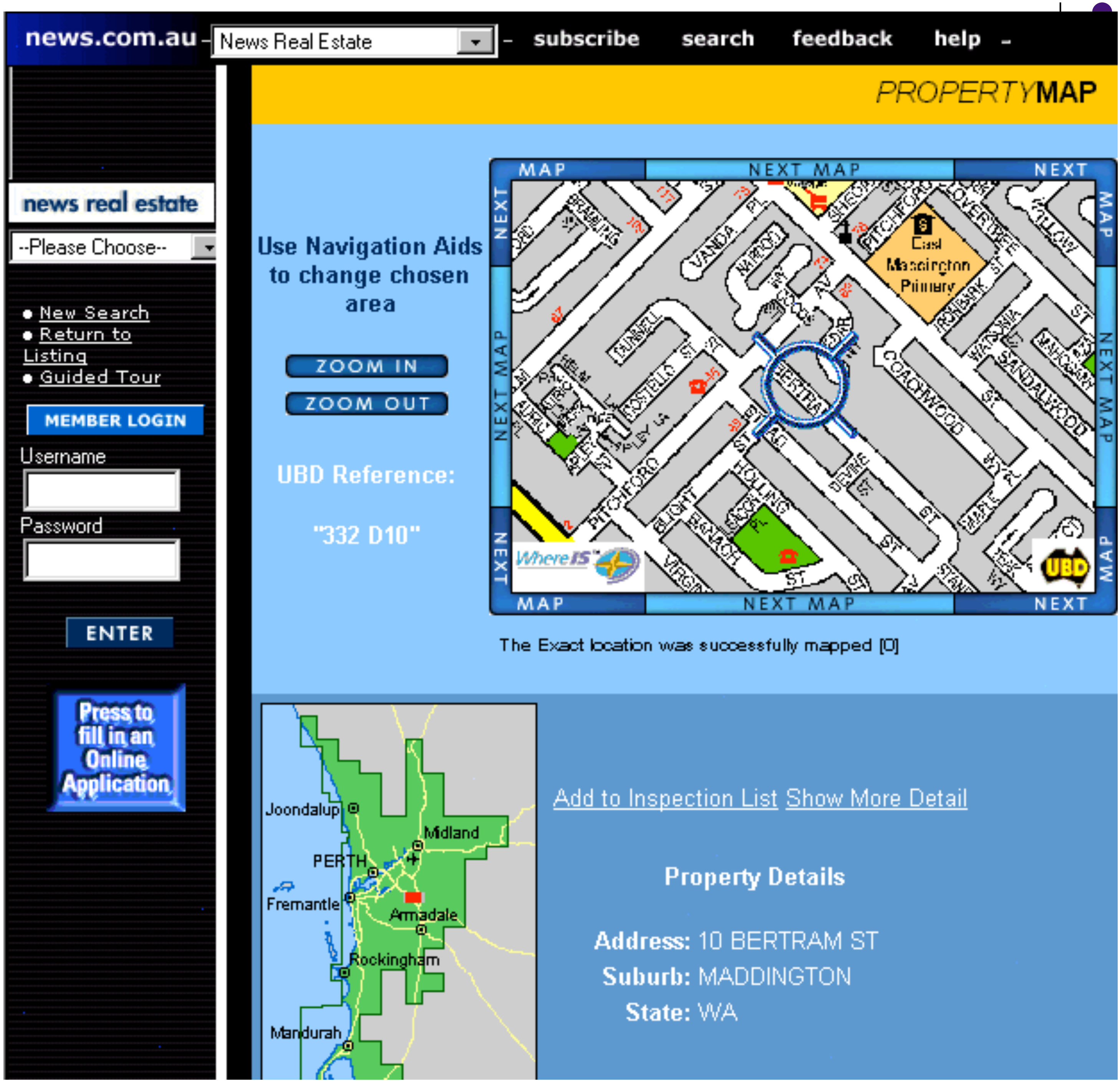
- Systems to summarize medical patient records by extracting diagnoses, symptoms, physical findings, test results, and therapeutic treatments.
- Gathering earnings, profits, board members, etc. from company reports
- Verification of construction industry specifications documents (are the quantities correct/reasonable?)
- Real estate advertisements
- Building job databases from textual job vacancy postings
- Extracting protein interaction with gene from biomed texts

Classified Advertisements (Real Estate)

Background:

- Advertisements are plain text
- Lowest common denominator: only thing that 70+ newspapers with 20+ publishing systems can all handle

```
<ADNUM> 2067206v1 </ADNUM>
<DATE> March 02, 1998 </DATE>
<ADTITLE> MADDINGTON $89,000
  </ADTITLE>
<ADTEXT>
OPEN 1.00 - 1.45<BR>
U 11 / 10 BERTRAM ST<BR>
NEW TO MARKET Beautiful<BR>
3 brm freestanding<BR>
villa, close to shops & bus
  <BR>
Owner moved to Melbourne<BR>
ideally suit 1st home
  buyer,<BR>
investor & 55 and over. <BR>
Brian Hazelden 0418 958 996
```



Use Navigation Aids to change chosen area

UBD Reference: "332 D10"

The Exact location was successfully mapped [0]

Property Details

Address: 10 BERTRAM ST
Suburb: MADDINGTON
State: WA

Why doesn’t text search (IR) work?

What you search for in real estate advertisements:

- Suburbs. You might think easy, but:
  - Real estate agents: Coldwell Banker, Mosman
  - Phrases: Only 45 minutes from Parramatta
  - Multiple property ads have different suburbs
- Money: want a range not a textual match
  - Multiple amounts: was $155K, now $145K
  - Variations: offers in the high 700s [*but not* rents for $270]
- Bedrooms: similar issues (br, bdr, beds, B/R)

Outline

- The CL domain
- A. Turing
- Numbers and data formats
- Database
- The real problems
- Technology
    - OCR
    - NL/DB interface, Web / IR search
    - Product Info, e-mail
    - Text categorization, clustering
    - Small devices, chat rooms
    - Information Extraction (IE)
    - Machine Translation (MT)
    - **Question / Answering**

Question / Answering

- With massive collections of on-line documents, manual translation of knowledge is impractical: we want answers from textbases
- Understand the question and answer within 50 byte snippets of text drawn from a text collection, and required to contain at least one concept of the semantic category of the expected answer type.
- Various evaluation campaigns in the last years (TREC)

Question / Answering

- Factual
  “Who was President of the United States in 1878?”
  “Who is the *Norwegian* king?”
  “How many scandals was Tapie implicated in while **boss** at Marseille?”

- Definition
  “What is a quasar?”
  “What is Bollywood?”

- List
  “List the names of casinos owned by Native Americans?”

Question / Answering

- Using Google
  - No good answers in the top 10
  - Using only the important terms
    e.g., Q= president United States 1978
  - Using the exact match
    e.g., Q="president of the United States" 1978
- Using Yahoo!
  - Same results
  - But "Who was president of the United States in 1978?", we found the answer in positions 8 and 10.
- Get reciprocal points for highest correct answer.

Question / Answering

- Question variability
- Name a film in which Jude Law acted.
  Jude Law was in what movie?
  Jude Law acted in which film?
  What is a film starring Jude Law?
  What film was Jude Law in?
- What was the name of the first Russian astronaut to do a spacewalk?
  Name the first Russian astronaut to do a spacewalk.
  Who was the first Russian to do a spacewalk?
  Who was the first Russian astronaut to walk in space?
- Other examples
  What is Colin Powell best known for? vs.
  Who is Colin Powell?

Question / Answering

- Two main strategies: Many patterns vs. deeper representation
- Select the passages according to the query words and their lexical variations + possibly the presence of an entity of the expected type
  - Variants found using the Web or WordNet
  - Measure the proximity between the entities
- Extracting the answer
  - Using patterns

    "When was Mozart born?" →

    word with uppercase, then "(" [1-9]([0-9])*3"-" [1-9]([0-9])")"

Question / Answering

- Translate retrieved passages and question into a formal representation (e.g., Semantic network, First order logic)
- Example:
  “Which company created the Internet Browser Mosaic?”

    organization AT(x2) & company NN(X2) & create VB(e1,x2,x6) & Internet NN(x3) & browser NN(x4) & Mosaic NN(x5) & nn NNC(x6, x3, x4, x5)

  - X for subject or object

  - E for possible
- Semantic information in WordNet could be useful

Conclusion

- Complete human-level natural language understanding is still a distant goal
- But there are now practical and usable partial NLU systems applicable to many problems
- An important design decision is in finding an appropriate match between (parts of) the application domain and the available methods
- *But, used with care, statistical NLP methods have opened up new possibilities for high performance text understanding systems.*

Introduction to Computational Linguistics (CL)

J. Savoy

Université de Neuchâtel

C. D. Manning & H. Schütze : *Foundations of statistical natural language processing*. The MIT Press, Cambridge (MA)

P. M. Nugues: *An introduction to language processing with Perl and Prolog*. Springer, Berlin