Introduction to Computational Linguistics

J. Savoy
Université de Neuchâtel

C.D. Manning & H. Schütze: *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge (MA), 1999.

N. Indurkhya & F.J. Damereau (Ed): *Handbook of Natural Language Processing*. CRC, 2nd Ed., 2010.

P.M. Nugues: *An Introduction to Language Processing with Perl and Prolog*. Springer, Berlin, 2006

Outline

- **Description :**
  Problèmes, questions et applications du traitement de la langue naturelle. Comptage statistique (loi de Zipf), modèle de langue et applications à l'analyse de corpus. Classification automatique (méthode Naïve Bayes). Principes de la recherche d'information. Moteur de recherche sur Internet et applications
- **Date :**
  vendredi 27 mai et 3 juin, de 13h15 à 16h30
- **Enseignant :**
  Jacques.Savoy@UniNE.ch

Outline

- **Computational Linguistics**
- Turing Test
- The real problem
- Technologies & Examples

Linguistics

- What is Linguistics?
  - The origins of language
  - Animals and human language
  - The development of writing
  - The sound patterns of language
  - Morphology
  - Phrases and sentences: grammar
  - Syntax
  - Semantics
  - Pragmatics
  - Discourse analysis
  - Language and the brain
  - Languages history and change

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

Computational Linguistics

- 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
  - Extract information / patterns from them
- Methods
  - Logic & grammar-based approaches
  - Statistics & machine learning (ML) methods

5

Computational Linguistics

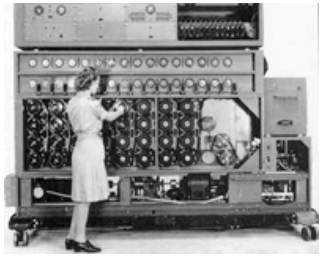
- Process / topics
  - Text segmentation
  - Part-of-speech (POS) tagging
  - Parsing
  - Word-Sense Disambiguation (WSD)
  - Natural Language (NL) Generation
  - Speech Recognition
  - Text-to-Speech Synthesis
  - Text Summarization
  - Evaluation

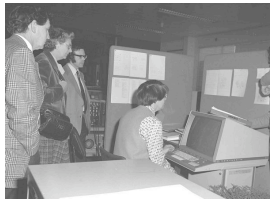
N. Indurkhya & F.J. Damereau (Ed): Handbook of Natural Language Processing. CRC, 2nd Ed., 2010

6

Computational Linguistics

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





7

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 the answer to a question (facts, yes/no, lists, definition)? (question/answering)
  - Can you retrieve documents on a given topic?
  - Can we categorize incoming messages into predefined categories? (spam filter)
  - Can you represent (index) this document collection?
  - Can we correct the spelling of a document? (OCR)

8

Computational Linguistics

- CL methods
  - Program a computer
  - Efficiency (speed)
  - Effectiveness (quality)
  - Reliable and robust processing (errors)
  - Represent text / document
  - Interest in both recognition and generation
  - Semantics (the *Holy Grail* search)
- Web: huge amount, open, lot of opportunities
- « Der Teufel liegt im Detail »

9

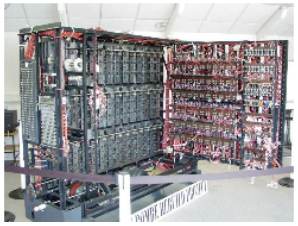
Outline

- Computational Linguistics
- **Turing Test**
- The real problem
- Technologies & Examples

10

Alan M. Turing (1912 – 1954)

- English mathematician and logician
- "*On Computable Numbers with an Application to the Entscheidungs-problem*" (1936). The Turing machine: the first universal programmable computer
- Code breaker (at Bletchley Park, with his "Bombe" – Colossus)



11

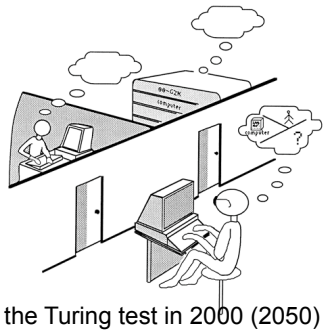
Turing Test

- Alan Turing: Computing Machinery and Intelligence" (1950)
- Can machines think?
  - "Can machines do what we (as thinking entities) can do?"
  - Natural language conversation
  - How can we prove that a machine can think?
- The Turing Test

12

Turing Test

- Imitation game
  - Natural language understanding
  - Knowledge representation
  - Automatic reasoning (inference)
  - Machine learning
- Current technology cannot pass this test
  When?



A. Turing: A computer will pass the Turing test in 2000 (2050) with 10^9 bits of memory (around 120 MB) in five-minute test

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 ?
…
Principle: <pattern> generates <answer>
see www.alicebot.org/oig/

Outline

- Computational Linguistics
- Turing Test
- **The real problem**
- Technologies & Examples

"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, MySQL, 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

17

"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

18

What's the world's most used database?

- 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)
- Internet (Feb. 2002)
  visible Web: 167 TB
  Hidden Web: 91,500 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/

19

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?

20

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 (NL) Understanding so complex?

Why is Natural Language Understanding difficult?

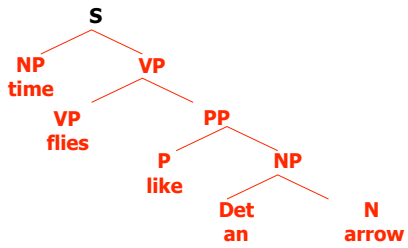
1. Infinite diversity of sentences
   a. the vocabulary is not completely known (Out-Of-Vocabulary problem OOV)
   b. the set of constructions is itself not completely predetermined
   c. the set of senses attributed to each word is also not completely predetermined
      Java: an island, coffee, a dance, a domestic fowl, a computer programming language
      BSE: Bovine Spongiform Encephalopathy, Bombay Stock Exchange (or Boston, Beirut, Bahrain), Breast Self-Examination, Bachelor of Science in Engineering, Basic Service Element, etc.
2. Tolerance of errors (robustness)

Why is Natural Language Understanding difficult?

3. Implicit elements
   a. Anaphoric references
      « Anne promised that she would be on time. »
   b. Polysemy
      « 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. Contractions, 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*

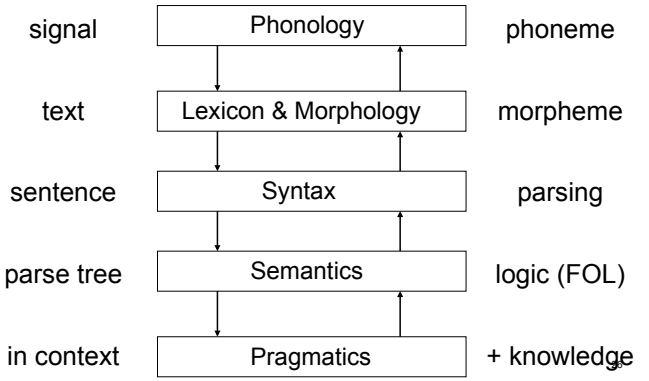


Where are the ambiguities?

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

25

Domains of NLP
(Recognition & Synthesis)



26

Outline

- Computational Linguistics
- Turing Test
- The real problem
- **Technologies & Examples**
  - 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

27

Terms and technologies

- Word processing / Desktop publishing
  - Spelling detection / correction (OCR technology)
  - Dictionary access
  - Internationalization / Translations aids
  - Controlled vocabularies.
- Some success
  - Word processing
  - Google (information retrieval)
  - Machine Translation with Google (?)

28

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
  - Question Answering (QA) – NL querying
  - Thesaurus/key phrase/terminology generation

29

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

30

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!

31

Outline

- Computational Linguistics
- Turing Test
- The real problem
- Technologies & **Examples**
  - 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

32

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

33

Email handling

"The web's okay to use; it's my email that is out of control"

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"

35

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*?

36

Email response: “eCRM”

electronic Customer Relationship Marketing

“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...”

37

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)

38

Small devices

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

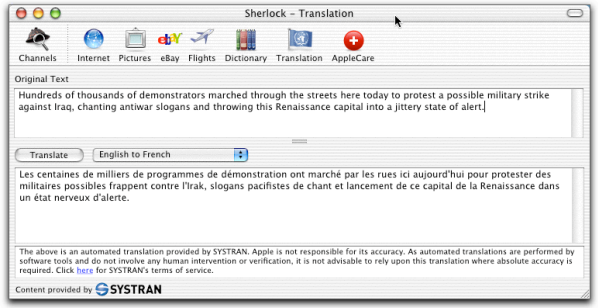
Small devices

- 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”

40

Machine translation

- High quality MT is still a distant goal



41

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.

42

Information Extraction

- 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

43

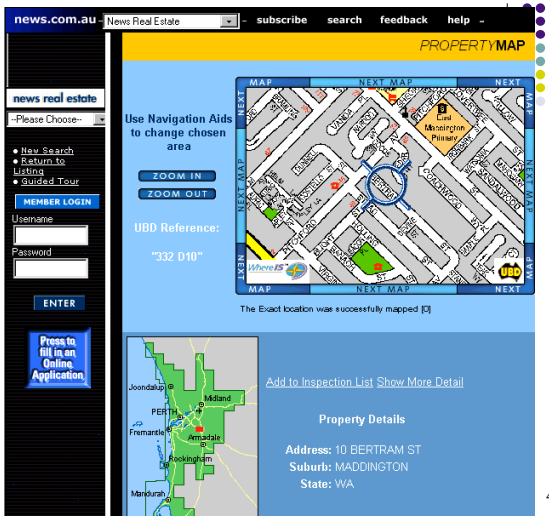
Classified Advertisements (Real Estate)

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
```

44



45

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)

46

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, http:\\trec.nist.gov)

47

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?"

48

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?

49

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.*
- Mixed approach (linguistic and statistics)

50

References

- P.M. Nugues: *An Introduction to Language Processing with Perl and Prolog*. Springer, Berlin, 2006.
- C.D. Manning & H. Schütze : *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge (MA), 1999.
- N. Indurkhya & F.J. Damereau (Ed): *Handbook of Natural Language Processing*. CRC, 2nd Ed., 2010.
- G. Gaznar & C. Mellish: *Natural Language Processing in PROLOG: An Introduction to Computational Linguistics*. Addison-Wesley, 1989.
- Journal
  Computational Linguistics (www.aclweb.org)
  Journal of the American Society for Information Science & Technology
  Journal of Quantitative Linguistics
- Conferences
  ACL (The Association for Computational Linguistics)

51

Sources

- Main providers
  - LDC Linguistic Data Consortium (UPenn)
  - ELRA European Linguistic Resources Association (Paris)
- Public
  - BNC British National Corpus
  - Project Gutenberg
  - **Text Encoding Initiative** (TEI, www.tei-c.org)
    to make explicit what is implicit.

52