Computational Linguistics: Summing up

Raffaella Bernardi

KRDB, Free University of Bozen-Bolzano

P.zza Domenicani, Room: 2.28, e-mail: bernardi@inf.unibz.it

Contents

1. What have you learned?

1. Which are the different natural language levels?
2. What's the goal of Computational Linguists?
3. What's the main challenge CLs need to tackle?
4. Which are the formal tools suitable for the different NL levels?
5. Where does NL fit within the Chomsky's hierarchy?
6. Which must be the expressivity of the used FG?
7. What's the consequence on the practical level?
8. Why top down and bottom up parsing approaches are not satisfactory?

1.1. Syntax-Semantics

1. What's the connection between Syntax and Semantics of NL?

2. How do we build the meaning of a sentence?

3. Which formalism can we use for representing NL meaning?

4. Which FG capture the link between Syntax and Semantics?

5. Which other FGs have we seen?

6. Which are their main advantages?

7. Which are the complexity of these FGs?

2. History of CL

Now, we will zoom out and look at the whole field of Computational Linguistics: to understand current research and directions, it is also important to know the past (what has been tried, what succeeded, what failed and why.)

How old is CL? "Computational linguistics, or natural language processing (NLP), is nearly as old as serious computing. Work began more than forty years ago, and one can see it going through successive phases, roughly ten year periods from the late fifties onwards." (Sparck Jones 1994)

2.1. Phases

1. The first phase, beginning in the late fifties, was linguistically oriented, focusing on machine translation, with people learning, painfully, how to do things computationally.

2. The second phase, from the late sixties to the late seventies, recognised the role of real world knowledge, was strongly motivated by AI, and drove NLP from this.

3. The third phase, dominating the eighties, acknowledged the specific modulating or controlling function for language relative to the world, and tried to capture this, in its necessarily systematic aspect, in grammatico-logical models for NLP.

4. The fourth phase, that we are in now, while taking the grammatico-logical skeleton for granted, recognises the significance of actual language usage, both idiosyncrastic and habitual, as a constraint on performance, and is therefore heavily into data mining from corpora.

2.2. What have Computational linguists achieved?

But what can we actually do now, given NLP's necessary concerns both with generic capabilities like syntactic parsing and with particular tasks like translation, i.e. with both subsystem and whole system functions?

Read:

Karen Sparck Jones. "Natural language processing: she needs something old and something new (maybe something borrowed and something blue, too)"

2.3. A different summary from someone I don’t recall the name of

Two foundational paradigms

- Automaton
- Probabilistic Models

2.4. Early Roots: 1940’s and 1950’s

2.4.1. Automaton

- Turing’s (1936) model of algorithmic computation
- Kleene’s (1951, 1956) finite automata and regular expressions
- Shannon (1948) applied probabilistic models of discrete Markov processes to automata for language
- Chomsky (1956): The first who considered finite-state machines as a way to characterize a grammar
- Led to the field of **Formal Language Theory** which used algebra and set theory to define formal languages as sequences of symbols.

2.4.2. Probabilistic Model

- algorithms for speech and language processing
- Shannon: the “noisy channel” model
- Shannon: borrowing of “entropy” from thermodynamics to measure the information content of a language

2.5. Two Camps: 1957-1970

2.5.1. Symbolic paradigm

- Chomsky:
  - Formal language theory, generative syntax, parsing
  - Linguists and computer scientists
  - Earliest complete parsing systems: Zelig Harris, UPenn
- Artificial intelligence
  - Created in the summer of 1956
  - Two-month workshop at Dartmouth
  - Focus of the field initially was the work on reasoning and logic (Newell and Simon)
  - Early natural language systems were built: 1. Worked in a single domain 2. Used pattern matching and keyword search

2.5.2. Stochastic paradigm

- Took hold in statistics and EE
- Late 50's: applied Bayesian methods to OCR
- Mosteller and Wallace (1964): applied Bayesian methods to the problem of authorship attribution for The Federalist papers.

2.6. Additional Developments: 1960's

1. First serious testable psychological models of human language processing. Based on transformational grammar
2. First on-line corpora
    - The Brown corpus of American English
    - 1 million word collection
    - Samples from 500 written texts
    - Different genres (news, novels, non-fiction, academic,.)
    - Assembled at Brown University (1963-64, Kucera and Francis)
    - William Wang's (1967) DOC (Dictionary on Computer) – On-line Chinese dialect dictionary

2.7. 1970-1983

Explosion of research

1. Stochastic paradigm: Developed speech recognition algorithms
   - HMM's
   - Developed independently by Jelinek et al. at IBM and Baker at CMU
2. Logic-based paradigm
   - Prolog, definite-clause grammars (Pereira and Warren, 1980)
   - Functional grammar (Kay, 1979) and LFG 1970-1983
3. Natural language understanding
   - SHRDLU (Winograd, 1972)
   - The Yale School: Focused on human conceptual knowledge and memory organization
   - Logic-based LUNAR question-answering system (Woods, 1973)
4. Discourse modeling paradigm

2.8. Revival of Empiricism and FSM's: 1983-1993

1. Finite-state models
    - Phonology and morphology (Kaplan and Kay, 1981)
    - Syntax (Church, 1980)
2. Return of empiricism
    - Rise of probabilistic models in speech and language processing
    - Largely influenced by work in speech recognition at IBM
3. Considerable work on natural language generation

2.9. Reunion of a Sort: 1994-1999

1. Probabilistic and data-driven models had become quite standard
2. Increases in speed and memory of computers allowed commercial exploitation of speech and language processing: Spelling and grammar checking
3. Rise of the Web emphasized the need for language based information retrieval and information extraction

3. State-of-the-Art works … I am aware of

Morphology

- Xerox: http://www.xrce.xerox.com/competencies/content-analysis/demos/english
- Morph-it: http://sslmitdev-online.sslmit.unibo.it/linguistics/morph-it php: Morphological resource for Italian.

Syntax http://aclweb.org/aclwiki/index.php?title=Parsing_(State_of_the_art)

Semantics CCG+Boxer: http://svn.ask.it.usyd.edu.au/trac/candc/wiki (FOL)

More: http://aclweb.org/aclwiki/index.php?title=State_of_the_art

4. Applications and projects (TN e BZ)

- QUALL-ME: http://qallme.itc.it/ (Magnini)
- CACAO: http://www.cacaoproject.eu/home (me)
- MIVaS: student lexicon evaluation (me)
- Anaphora (Poesio)
- Semantic Clustering via Latent Semantic Analysis or Strudel Model (Baroni and Poesio)
- Lexical Resources extracted from Corpora (Baroni)
- Speech (Riccardi)
- Lexicography (Abel)
- . . . (FBK, CIMeC, EURAC)

5. Main Conferences, mailing list etc.

- ACL, EACL, NAACL
- Coling
- LREC
- HLT
- ESSLLI Summer School
- ICoS
- IWCS
- . . .

- Corpora mailing list
- ACL
- . . .

More info at:

http://aclweb.org/aclwiki/index.php?title=Main_Page

6. CL Projects (6th of June, Room: see RIS)

Faisal Chowhdury	Underspecification	09:00-09:20
Tsvetan Dunchev	Semantics in Prolog	09:20-09:40
Marco Trevisan	ACE	09:40-10:00
cappuccino break		
Ana	TAG	10:10-10:30
Manfred Gerstgrasse	Incremental parsing	10:30-10:50
nguyen trung kien	CG and ND	10:50-11:10
coffee break		
Dinh Le Thanh	Question type tagger for BoB	11:20-11:40
Anja Roubickova	Machine Translation	11:40-12:00
Stanislav Skotnick	CF efficient parsing	12:00-12:20
lunch break		
Auste	Chatter bot	14:00-14:20
Abhinav	Finite State Automata	14:20-14:40
Philipp Volgger	Brill's algorithm and Tiger Corpus	14:40-15:00
tea break		
Alessandro Ercolani	LSA	15:10-15:30
Martins Zalcmanis	Lexical Semantics	15:30-15:50