Syntactic Formalisms for Parsing Natural Languages

Aleš Horák, Miloš Jakubíček, Vojtěch Kovář
(based on slides by Juyeon Kang)

ia161@nlp.fi.muni.cz

Autumn 2013

CZ.1.07/2.2.00/28.0041
Centrum interaktivních a multimediálních studijních opor pro inovaci výuky a efektivní učení

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Course objective

Introducing

- theoretical backgrounds on parsing
- parsing methods focused on syntax
- practical implementation methods
- possible applications and evaluations

Course syllabus

PART I : Theoretical backgrounds

- Historical overview
- State of the art parsing methods and trends
- Advanced syntactic formalisms

PART II : Practical applications

- Applications & Use Cases
- Practical Implementations
- Parsing Evaluation

Course format

- Weekly lectures (2 hours)
- Final written exam
- Two homework assignments
- Grading
  - Final exam: 60 points
  - Each homework: 20 points
  - For each homework 10 % top scoring individuals receive 5 bonus points
  - Points required for colloquium: 60 points

Introductive and Historical Overview on Natural Languages Parsing

IA161
Syntactic Formalisms for
Parsing Natural Languages

Main points

- Introduction to Natural Language Processing
- Issues in Syntax
- What is a parsing?
- Overview of Parsing methods and trends

Why natural language processing ?

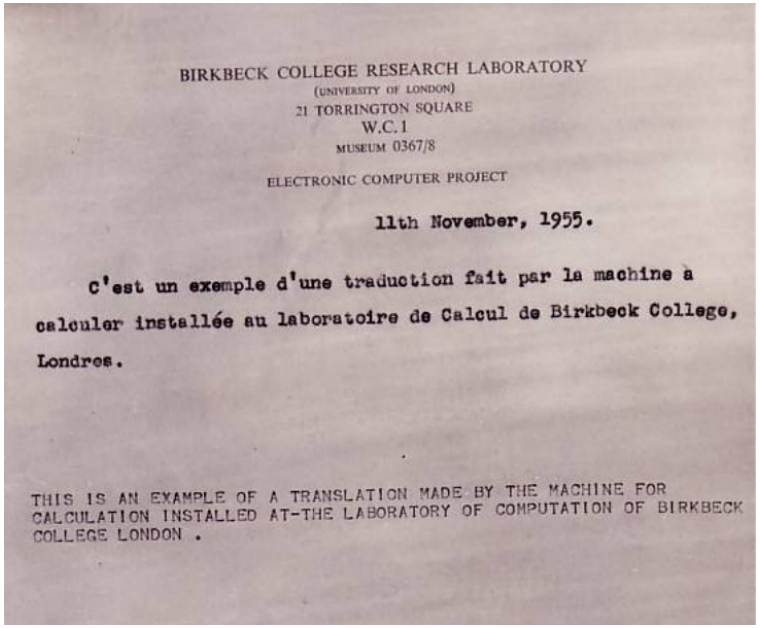
- Huge amounts of data from Internet and Intranet
- Applications for processing large amounts of texts need NLP expertise
  - Classify text into categories
  - Index and search large texts
  - Automatic translation
  - Speech recognition
  - Information extraction
  - Automatic summarization
  - Question answering
  - Knowledge acquisition
  - Text generation/dialogues

History of Natural Language Processing

- 1948 – 1st NLP application?

- dictionary look-up system by Andrew Booth, for machine translation purposes
- developed at Birkbeck College, London University

History of Natural Language Processing



BIRKBECK COLLEGE RESEARCH LABORATORY
(UNIVERSITY OF LONDON)
21 TORRINGTON SQUARE
W.C.1
MUSEUM 0367/8

ELECTRONIC COMPUTER PROJECT

11th November, 1955.

C'est un exemple d'une traduction fait par la machine à calculer installée au laboratoire de Calcul de Birkbeck College, Londres.

THIS IS AN EXAMPLE OF A TRANSLATION MADE BY THE MACHINE FOR CALCULATION INSTALLED AT THE LABORATORY OF COMPUTATION OF BIRKBECK COLLEGE LONDON .

History of Natural Language Processing

- 1949 – Warren Weaver

- Natural Sciences Division Director in the Rockefeller Foundation
- Mathematician, Science Advocate
- WWII code breaker
- He viewed Russian as English in code – the ”Translation” memorandum

Also knowing nothing official about, but having guessed and inferred considerable about powerful new mechanized methods in cryptography – methods which I believe succeed even when one does not know what language has been coded – one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say “This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

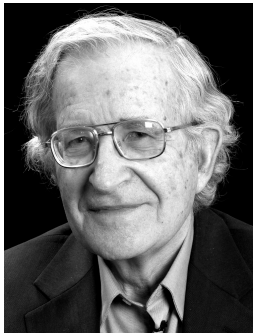
History of Natural Language Processing

- 1966 – Over-promised under-delivered
  - Machine Translation worked only word by word
  - NLP brought the first hostility of research funding agencies
    - NLP gave AI a bad name before AI had a name.
  - All funding of NLP came to a grinding halt due to the infamous ALPAC report.
    - Public spent 20 million with very limited outcomes.
- 1966–1976 – “A lost decade”
- Revival in 1980’s
  - Martin Kay: The Proper Place of Men and Machines in Language Translation

NLP looked to Linguistics

Linguistics is language described, not prescribed.
Linguistics had few applicable theories for Machine Translation

- 1957 – Noam Chomsky's Syntactic Structures revolutionized Linguistics as it applies to Machine Translation.



  - Rule based system of syntactic structures.
  - Believed there are features common to all languages that enable people to speak creatively and freely.
  - Hypothesized all children go through the same stages of language development regardless of the language they are learning – a concept of an innate Universal Grammar (never proven)
  - One of the most prominent persons of NLP in 20th century, though very controversial.

NLP looked to Linguistics

- 1958 – Bar-Hillel report
  - Concluded Fully-Automatic High-Quality Translation (FAHQT) could not be accomplished without human knowledge.
- 1968 – Case Grammar (Fillmore)
  - "The case for case" paper
  - Later evolved into Frame Semantics
- 1970 – Augmented Transition Networks (Woods)
  - Procedural Semantics – Theory of the "meaning" of sentence.
  - Augmented Transition Network (ATN) parser

Successful NLP Systems

- 1970 – Shrdlu (Terry Winograd)

Human: *Pick up a big red block.*
Computer: OK.
Human: *Grasp the pyramid.*
Computer: I don't understand which pyramid you mean.
Human (changing their mind): *Find a block which is taller than the one you are holding and put it into the box.*
Computer: By "it", I assume you mean the block which is taller than the one I am holding.
Computer: OK.
Human: *What does the box contain?*
Computer: The blue pyramid and the blue block.
Human: *What is the pyramid supported by?*
Computer: The box.
Human: *How many blocks are not in the box?*
Computer: Four of them.
Human: *Is at least one of them narrower than the one which I told you to pick up?*
Computer: Yes, the red cube.

Successful NLP Systems II

- 1973 – Lunar question answering system (Woods)

WHAT IS THE AVERAGE CONCENTRATION OF ALUMINUM IN HIGH ALKALI ROCKS?
WHAT SAMPLES CONTAIN P200?
GIVE ME THE MODAL ANALYSES OF P200 IN THOSE SAMPLES
GIVE ME EU DETERMINATIONS IN SAMPLES WHICH CONTAIN ILM

Successful NLP Systems III

- 1976 – TAUM-METEO (University of Montreal)
  - prototype MT system for translating weather forecasts between English and French
- 1985 – METEO (John Chandioux)
  - successor of TAUM-METEO
  - in operational use at Environnement Canada forecasts until 30th of September 2001
- 1970 – SYSTRAN
  - provided translations for US Air Force's Foreign Technology Division
  - adopted by XEROX (1978)
  - still developed, present in wide range of systems
- Google language tools
- Microsoft spell check

Major Issues in NLP

Ambiguity in Language:

- Syntactic (structural)
- Semantic (word sense)
- Referential

Ambiguity Makes NLP difficult

- Structural/Syntactic ambiguity
  - I saw the Grand Canyon flying to New York.
  - I saw the sheep grazing in the field.
- Word Sense ambiguity
  - The man went to the bank to get some cash.
  - The man went to the bank and jumped in the river.
- Referential ambiguity
  - Steve hated Paul. He hit him.
  - He = Steve ? or he = Paul ?

Linguistics levels of analysis

- Speech
- Written language
  - Phonetics
  - Phonology
  - Morphology
  - Syntax
  - Semantics
  - Beyond: pragmatic, cognitive, logic...

Each level has an input and output representation, output from one level is the input to the next, sometimes levels might be skipped (merged) or split.

Issues in syntax

- Propagation of errors from lower levels – mainly morphology, need to correctly identify the part of speech (POS)
  "The man did his homework"
  - Who did what?
    man=noun; *did*=verb; *his*=genitive; *homework*=noun
- Identify collocations
  - *Mother in law, hot dog, ...*

More issues in Syntax

- Anaphora resolution
  "The son of my professor entered my class. He scared me."
- Preposition attachment
  "I saw the man in the park with a telescope."

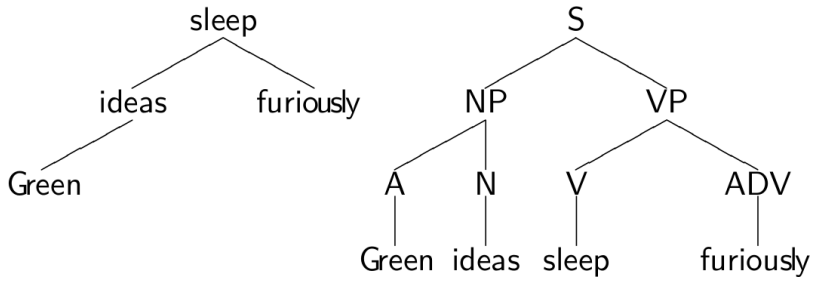
Syntax input and output

- **Input:** sequence of pairs (lemma, (morphological) tag)
- **Output:** sentence structure (tree) with annotated nodes (all lemmas, (morpho-syntactic tags, functions) of various forms
- Deals with:
  - The relation between lemmas & morphological categories and the sentence structure use syntactic categories such as subject, verb, object,...

Syntactic representation

- Tree structure
- Two main ideas for the tree
  - **Phrase structure** (derivation tree)
    - Using bracketed grouping
    - Brackets annotated by phrase type
    - Heads (often) explicitly marked
  - **Dependency structure**
    - Basic relation: head (governor) – dependent
    - Links annotated by syntactic functions
    - Phrase structure implicitly present

Dependency Tree vs. PS Tree



Shallow parsing

"the man chased the bear"

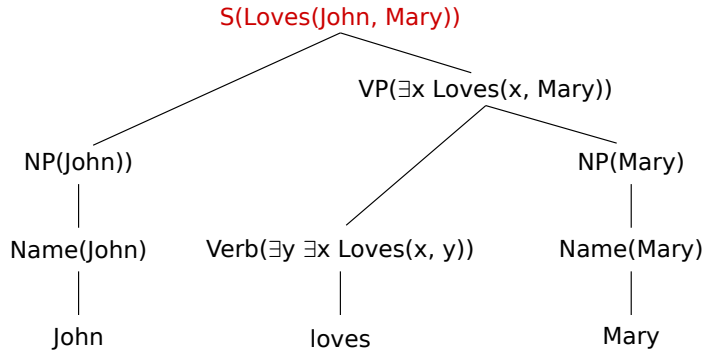
"the man" "chased the bear"

Subject -- Predicate

- Identify basic structures
  - NP-[the man] VP-[chased the bear]

Full parsing

"John loves Mary"



Help figuring out automatically questions *like who did what and when?*

What is a natural language parsing ?

- One of the most commonly researched tasks in Natural Language Processing (NLP)
  - Parsing, in traditional sense, is what happens when a student takes the words of a sentences one by one, assigns each to a part of speech, specifies its grammatical categories, and lists the grammatical relations between words (identifying subject and various types of object for a verb, specifying the word with which some other word agrees, and so on).

Characteristics of parsing

Much of the history of parsing until a few decades ago can be understood as the direct consequence of the history of theories of grammar:

- Parsing is done by human beings, rather than by physical machines or abstract machine
- What is parsed is a bit of natural language, rather than a bit of some language-like symbolic system
- Parsing is heuristic rather than algorithmic

New notions of parsing

In the second half of 20^{th} century the parsing has come to be extended to a large collection of operations in relation with theoretical linguistics, formal language theory, computer science, artificial intelligence and psycholinguistics:

- Parsing is the syntactic analysis of languages.
- The objective of Natural Language Parsing is
  - to determine parts of sentences (such as verbs, noun phrases, or relative clauses), and the relationships between them (such as subject or object).
- Unlike parsing of formally defined artificial languages (such as Java or predicate logic), parsing of natural languages presents problems due to ambiguity, and the productive and creative use of language.

Parsing

- The grammar for Natural Language is ambiguous and typical sentences have multiple possible analyses (syntactically and semantically).
- Some parsing tools (i.e. grammatical, morphologic, syntactic, statistic, probabilistic, heuristic, ...) help to find the most plausible parse tree of a given sentence.

Practical function of a parsing

- Parsing can tell us when a sentence is in a language defined by a grammar
- Parsing makes the extraction of the information possible by identifying relations between words, or phrases in sentences.

Practical function of a parsing

- Parsers are being used in a number of disciplines:
  - In computer science
    - Compiler construction, database interfaces, self-describing databases, artificial intelligence...
  - In linguistics
    - Text analysis, corpora analysis, machine translation...
  - In document preparation and conversion
  - In typesetting chemical formulae
  - In chromosome recognition

Practical function of a parsing

- However,
  - Many different possible syntactic formalisms:
    - Regular expressions, Context-free grammars, Context-sensitive grammars, ...
  - Many different ways of representing the results of parsing:
    - Parse tree, Chart, Graph, ...

Historical overview of parsing methods

- Basically two ways to parse a sentence
  - **Top-down** vs. **Bottom-up**

    We can characterize the search strategy of parsing algorithms in terms of the **direction** in which a structure is built:
    from the words upwards (bottom-up) or
    from the root node downwards (top-down)

Historical overview of parsing methods

- **Directionality** in these two ways

Directional vs. Non-directional

  - Non-directional top-down methods by S. Unger (1968)
  - Non-directional bottom-up methods by CYK
  - Directional top-down methods:
    - The predict/match automaton, Depth-first search (backtrack), Breadth-first search (Greibach), Recursive descent, Definite Clause grammars
  - Directional bottom-up methods:
    - The shift/reduce automaton, Depth-first search (backtrack), Breadth-first search, restricted by Earley(1970)

Historical overview of parsing methods

- Methods originating at parsing of formal languages
  - Linear directional top-down methods:
    - LL(K)
  - Linear directional bottom-up methods:
    - Precedence, bounded-context, LR (k), LALR(1), SLR(1)
- Methods specifically devised for parsing of natural languages
  - Generalized LR (Masaru Tomita)
  - Chart parsing (Martin Kay)

Summary

- Natural language parsing as one of the NLP domain
- Extended notion of parsing in relation with different fields
- Ambiguity of language
- What is it to “parse”?
- Overview of basic parsing methods

References I

- H. Bunt, J. Carroll & G. Satta (eds.): *New Developments in Parsing Technology*, Kluwer, Dordrecht/Boston/London 2004
- H. Bunt, P. Merlo, & J. Nivre (eds.): *Trends in Parsing Technology: Dependency Parsing, Domain Adaptation, and Deep Parsing*, Springer Dordrecht, Heidelberg/London/New York 2010
- H. Bunt, M. Tamita (eds.): *Recent advances in parsing technology*, Kluwer, Boston, 1996
- G. Dick: *Parsing techniques: a practical guide*, Springer, 2008
- Roger G. Johnson: *Andrew D. Booth – Britain’s Other “Fourth Man”. In: History of Computing. Learning from the Past*, Springer Berlin Heidelberg, 2010.
- J. Hutchins: *From First Conception to First Demonstration: the Nascent Years of Machine Translation, 1947–1954. A Chronology. In: Machine Translation, Volume 12, Issue 3*, Kluwer, 1997.

References II

- J. Hutchins: *Milestones no.6: Bar-Hillel and the nonfeasibility of FAHQT. In: International Journal of Language and Documentation no.1*, 1999.
- M. Kay: *The proper place of men and machines in language translation. In: Machine Translation, Volume 12, Issue 1–2*, Kluwer 1997 (reprint of 1980).
- More on history of MT: `http://www.hutchinsweb.me.uk/history.htm`