Syntax and parsing – an overview

Lilja Øvrelid
INF5830
Fall 2013

Overview

- INF5830 so far
  - general methodology
  - statistical methods
  - words, frequencies
- The rest of the course
  - theoretical background and practical experience with two NLP tasks
  - “deeper processing”: syntactic and semantic analysis
    - data-driven dependency parsing
    - semantic role labeling (SRL)
  - experimental methodology
  - supervised machine learning; classification
  - evaluation
  - academic writing
Part I: Data-driven dependency parsing

- Syntax and parsing (today)
- Dependency grammar (10/10)
- Dependency parsing (17/10)
- Experimental methodology (22/10)
- Project A (written report due Oct. 30th):
  - training and evaluation of parsers for several languages
  - MaltParser: freely available software for data-driven dependency parsing

Part II: Semantic Role Labeling (SRL)

- Semantic roles, theoretical (24/10)
- Semantic role labeling, practical/computational (31/10)
- Project B (written report due Nov. 20th)
  - CoNLL 2008: syntactic and semantic parsing of English
  - solve part of this task: semantic argument classification
  - feature engineering (using syntactic analysis)
  - supervised machine learning
  - evaluation
Lectures and groups

❖ Curriculum: largely research literature
  › “Classics” from linguistics, e.g. Zwicky, Fillmore
  › Computational linguistics research literature, e.g. Nivre, Gildea & Jurafsky
❖ Lectures: introduction to topics, synthesis of curriculum
❖ Group teaching:
  › focused reading groups (please prepare!!)
  › practical sessions related to obligatory assignments

Today’s lecture

❖ (Repetition of) basic principles of syntax:
  › form vs function
  › constituents and phrases
  › context-free grammars
❖ Brief overview of syntactic parsing
  › traditional, grammar-driven parsing
  › statistical/data-driven parsing
Syntax

- Syntax: rules for constructing grammatical sentences and determining their meaning
- “Who does what to whom?”
- Wealth of theories: some differences, a lot in common
  - Government and Binding (GB)
  - Minimalist Program (MP)
  - Head-driven phrase structure grammar (HPSG)
  - Lexical Functional Grammar (LFG)
  - Categorial Grammar
  - Dependency Grammar
  - ...

Why bother?

- Theoretical syntacticians concerned with **grammaticality**
  - *The President nominated a new Supreme Court justice*
  - *President the new Supreme justice Court nominated*
- Relevant for some NLP applications:
  - text generation
  - grammar checking
- But mostly want systems that are robust and can handle realistic (noisy) language
Why bother?

- Parsing provides “scaffolding” for semantic analysis
- Direct, down-stream usage of syntactic information
  - opinion mining
  - information extraction
  - syntax-informed statistical machine translation
  - sentence compression
  - etc.

Syntax and parsing – an overview

Generative grammar

- Noam Chomsky: *Syntactic Structures* (1957)
- inspired by natural sciences: language as a set of sentences
  - set: a collection of objects, e.g.: \( \{a, b, c\} \), \( a \in \{a, b, c\} \), \( d \notin \{a, b, c\} \)
- a grammar should generate the set of all grammatical sentences in a language:
  - *Alle nordmenn liker ikke fotball*
  - *Liker ikke alle nordmenn fotball*
  - *Nordmenn alle ikke liker fotball*
  - *Liker nordmenn fotball ikke alle*
Grammar

► Should contain:
  ▶ the form of syntactic units
  ▶ the role (function) this unit plays in relation to other unit
► One form may have several functions
  ▶ She bought a nice house.
  ▶ The house is nice.
► Form and function for larger units than words
  ▶ A nice house is an important asset.
  ▶ We visited a nice house.

Constituents

► The words in a sentence are organized into groupings
► function as a whole
► relate to other words as a unit
  ▶ The dog ate my homework
  ▶ The dog ate my homework
► linguistic tests of constituency
Constituents

▶ Linguistic tests:
  ▶ “stand alone”:
    ▶ What did the dog eat?
    ▶ My homework
    ▶ *ate my
  ▶ “replaced by pronoun”
    ▶ Where is your homework?
    ▶ The dog ate it
    ▶ The dog ate my homework and the cat did too
  ▶ “moves as a unit”:
    ▶ It was my homework that the dog ate.
    ▶ My homework was eaten by the dog.

Heads

Important syntactic concept (we will get back to this!)
  ▶ Within most constituents, one element is distinguished: the head
  ▶ Determines internal structure
  ▶ Determines external distribution (possible functions)
Form and function

- **Syntactic form** - constituents are described using parts of speech and phrases
  - phrases - larger constituents above word level
  - phrases named after the **head** - central, obligatory member
  - e.g. NP, VP, PP

- **Syntactic function** - constituents are described by their role in the sentence as a whole
  - Subject
  - (Direct and Indirect) Object
  - Adverbial

Phrases: syntactic categories

- Constituent: head + (argument(s)/adjunct(s))
- Substitutable
  - *The dog ate the birthday cake*
  - *The dog ate the delicious birthday cake*
  - *The dog ate the delicious birthday cake that was meant for Bea*
Arguments vs. adjuncts

- Subconstituents which are not heads: arguments or adjuncts
  - arguments: selected by the head and complete the meaning
  - adjuncts: not selected by the head and refine the meaning
- Different PoS may take argument(s):
  - John *invited* Mary to the event
  - John’s *invitation* of Mary to the event caused quite a stir
  - Mary found the book *under* the couch
- Adjuncts are not obligatory and may often iterate
  - John ran on Sunday / with Mary / in the park

Noun phrase (NP)

- Head is a noun
- Typically functions as subject or object
- Examples:
  - determiner + noun: *the dog, en hund*
  - proper name: *Barack Obama, Japan*
  - pronoun: *he, they, han, henne*
- Agreement – e.g. number, gender, definiteness
- Head determines agreement
Prepositional phrase (PP)

- Head is a preposition
- Followed by an NP (prepositional argument)
- Examples:
  - prep + NP in the garden, over the rooftops

Verb phrase (VP)

- Head is a verb (finite/non-finite)
- All elements of the sentence except the subject
- Examples:
  - verb sleeps, danced
  - verb + NP: ate the cake
  - verb + NP + NP: gave him the cake
  - verb + NP + PP put all the papers in the drawer
Phrase structure grammars

- Capture constituent status and ordering
- Formal model: context-free grammar
  1. \( S \rightarrow NP \ VP \)
  2. \( NP \rightarrow D \ N \)
  3. \( VP \rightarrow V \ NP \)
- Syntactic structure as phrase structure trees

Context-free grammars (CFGs)

- Formally, a CFG is a 4-tuple \( < N, \Sigma, R, S > \), where
  - \( N \) is a set of non-terminal symbols (syntactic categories)
  - \( \Sigma \) is a set of terminal symbols (words)
  - \( R \) is a set of rules \( A \rightarrow \alpha \), where
    - \( A \) is a non-terminal
    - \( \alpha \) is a string of symbols taken from the set \( (\Sigma \cup N)^* \)
  - \( S \) is a designated start symbol
Subcategorization

- An important aspect captured by grammars
- Arguments of verbs may be classified at several levels:
  - grammatical relations (functions): subject, object, indirect object, etc.
  - semantic roles: agent, patient, recipient, etc.
- Subcategorization frame: classification of verbs according to the types of arguments they take (form and function)
  - Intransitive verb. NP[subject]. John ran.
  - Transitive verb. NP[subject], NP[object]. John saw Mary.
  - Ditransitive verb. NP[subject], NP[direct object], NP[indirect object]. John gave Mary a book.

Syntactic functions

- Describe the arguments of verbs
- English (SVO-language):
  - the subject precedes the main verb
  - the object follows it
- Typological variation
  - fixed word order (grammatical functions are identified by word order)
  - free word order languages (Latin, Russian)
    - case
    - agreement
  - what about Norwegian?
**Syntax**

**Syntactic functions**

- Subjects are special (cross-linguistically)
- Typical subject properties (here English):
  - Agreement on present tense verbs
  - Nominative case
  - To the left of the verb
  - Raising/control

**Functional analysis**

- **Predicate**: finite verb form (+++ infinitive, participles)

  My old pal bought a car in Bergen yesterday
  My old pal |bought| a car in Bergen yesterday
  pred
Functional analysis

- **Predicate**: finite verb form (++ infinitive, participles)

My old pal has bought a car in Bergen
My old pal |has bought| a car in Bergen
    pred

My pal  likes to drive
My pal  |likes to drive|
    pred
**Functional analysis**

- **Subject:** who or what pred?

  My old pal | bought | a car in Bergen yesterday

  | My old pal | bought | a car in Bergen yesterday
  subj        pred

- **Direct object:** who or what did subj pred?

  | My old pal | bought | a car in Bergen
  subj        pred

  | My old pal | bought | a car | in Bergen
  subj        pred    d.obj
Functional analysis

- **Indirect object**: who or what did subj pred d.obj?

<table>
<thead>
<tr>
<th>Mary</th>
<th>gave</th>
<th>them</th>
<th>a gift</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>pred</td>
<td>d.obj</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mary</th>
<th>gave</th>
<th>them</th>
<th>a gift</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>pred</td>
<td>i.obj</td>
<td>d.obj</td>
</tr>
</tbody>
</table>

- **Adverbial (locative)**: where did subj pred (d.obj)?

<table>
<thead>
<tr>
<th>My old pal</th>
<th>bought</th>
<th>a car</th>
<th>in Bergen</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>pred</td>
<td>d.obj</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>My old pal</th>
<th>bought</th>
<th>a car</th>
<th>in Bergen</th>
</tr>
</thead>
<tbody>
<tr>
<td>subj</td>
<td>pred</td>
<td>d.obj</td>
<td>loc</td>
</tr>
</tbody>
</table>
Functional analysis

- **Adverbial (temporal):** when, for how long or how often did subj pred (d.obj) (ral)?

```plaintext
| My old pal | bought | a car | in Bergen | yesterday |
| subj        | pred   | d.obj | loc       |           |
| My old pal | bought | a car | in Bergen | yesterday |
| subj        | pred   | d.obj | loc       | tmp       |
```

Defining syntactic functions

- **Thematic (semantic) criteria**
  - tendency for subjects to be agents and objects to be patients
  - but not perfect correspondence
    - Mary was loved by all
    - The wind broke the window

- **Structural criteria**
  - Subj: sister of VP, daughter of S
  - Obj: sister of V, daughter of VP

- **Morphological criteria:**
  - English: subject agrees with finite verb
  - many languages identify syntactic functions via case (nominative, accusative, dative, etc.)
Syntactic parsing

- automatically determining the syntactic structure for a given sentence
- search through all possible trees for a sentence
- bottom-up vs top-down approaches

Top-down

- builds structure from root of tree (S) to leaves
- operates with a list of constituents to be built and rewrites them by matching their category to a LHS of the grammar rules
- several ways of rewriting: search problem
- depth-first vs breadth-first search
  - [ S ]
  - [ NP VP ]
  - [ DT NN VP ] [ NP PP VP ]
  - [ a NN VP ] [ NP PP VP ] [ DT N VP ]
  - etc.
**Bottom-up**

- starts with the words and tries to build the trees from them and up
- if a sequence in the goal list matches the RHS of a rule we may substitute
- if RHS if several rules match: search problem
- standard presentation is as shift-reduce parsing
  - [ the ] [ the woman reports ]
  - [ DT ] [ woman reports ]
  - [ DT woman ] [ reports ]
  - [ DT NN ] [ reports ]
  - [ NP ] [ reports ]
  - [ NP reports ] [ ]
  - etc.

**Ambiguities**

- more than one possible structure for a sentence
- natural languages are hugely ambiguous
- a very common problem

<table>
<thead>
<tr>
<th>PoS-ambiguities</th>
<th>Attachment ambiguities</th>
</tr>
</thead>
<tbody>
<tr>
<td>VB</td>
<td>in effort</td>
</tr>
<tr>
<td>VBZ</td>
<td>to control</td>
</tr>
<tr>
<td>VBP</td>
<td>inflation</td>
</tr>
<tr>
<td>VBZ</td>
<td></td>
</tr>
<tr>
<td>NNP</td>
<td></td>
</tr>
<tr>
<td>NNS</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td></td>
</tr>
<tr>
<td>NNS</td>
<td></td>
</tr>
<tr>
<td>CD</td>
<td></td>
</tr>
<tr>
<td>NN</td>
<td></td>
</tr>
</tbody>
</table>

Fed raises interest rates 0.5 %
Back in the days (90s)

- Parsers assigned linguistically detailed syntactic structures (based on linguistic theories)
- Grammar-driven parsing: possible trees defined by the grammar
- Problems with coverage
  - only around 70% of all sentences were assigned an analysis
- Most sentences were assigned very many analyses by a grammar
  - no way of choosing between them

Enter data-driven (statistical) parsing

- Today data-driven/statistical parsing is available for a range of languages and syntactic frameworks
- Data-driven approaches: possible trees defined by the treebank (may also involve a grammar)
- Produce one analysis (hopefully the most likely one) for any sentence
- And get most of them correct
- Still an active field of research, improvements are still possible!
Statistics in parsing

- classical NLP parsing:
  - symbolic grammar and lexicon
  - proof systems to prove parses from words
- ambiguity problem is very large
  - minimal grammar on previous sentence: 36 parses
  - large broad-coverage grammar: millions of parses
- use probabilities to pick the most likely parse

Treebanks

- need data to estimate probabilities
- collection of sentences manually annotated with the correct parse ⇒ a treebank
- Penn Treebank: treebanks from Brown, Switchboard, ATIS og Wall Street Journal corpora
- Treebanks for other languages
  - Prague Dependency Treebank (czech)
  - Negra/Tuba-DZ (German)
  - Penn (Chinese)
  - Norwegian: under development
Syntactic parsing

Treebanks
Eksempel fra Penn Treebank (WSJ)

( (S
   (PP-LOC (IN In)
    (NP
     (NP (NNP Thursday) (POS 's))
     (NN edition))
   (, ,)
   (NP-SBJ (PRP it))
   (VP (VBD was)
    (VP (ADVP-MNR (RB incorrectly))
     (VBN indicated)
     (S
      (NP-SBJ (DT the) (NN union))
      (VP (VBD had)
       (VP (VBN paid)
        (NP (DT a) (NN fee))
        (PP-DTV (TO to)
         (NP
          (NML (JJ former) (NNP House) (NNP Speaker))
          (NNP Jim) (NNP Wright))))))))

Syntax and parsing – an overview 43(45)

Probabilistic Context-free grammars (PCFGs)

- Formally, a PCFG is a 5-tuple $\langle N, \Sigma, R, S, P \rangle$, where
  - $N$ is a set of non-terminal symbols (syntactic categories)
  - $\Sigma$ is a set of terminal symbols (words)
  - $R$ is a set of rules $A \rightarrow \alpha$, where
    - $A$ is a non-terminal
    - $\alpha$ is a string of symbols taken from the set $(\Sigma \cup N)^*$
  - $S$ is a designated start symbol
  - $P(R)$ gives the probability of each rule
Next week: Dependency Grammar

- An alternative to phrase structure representations
- Syntactic functions are central
- Claimed to be closer to semantic analysis

Small birds sing loud songs

```
Small   birds   sing   loud   songs
  nmod    sbj     nmod
  ▼       ▼       ▼
  nmod    nmod
  ▼       ▼
  obj
```