Complexity and Character of Human Languages

Informatics 2A: Lecture 25

Shay Cohen

19 November 2015

1. Human Language Complexity
   - Chomsky Hierarchy
   - Strong and Weak Adequacy

2. Mildly Context-Sensitive Grammars

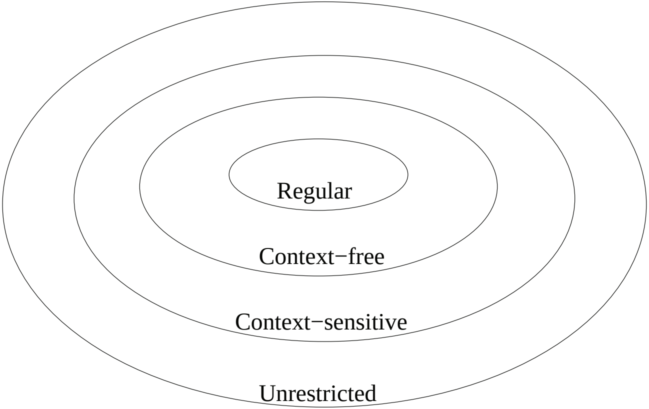
Reading: J&M. Chapter 16.3–16.4.

Review

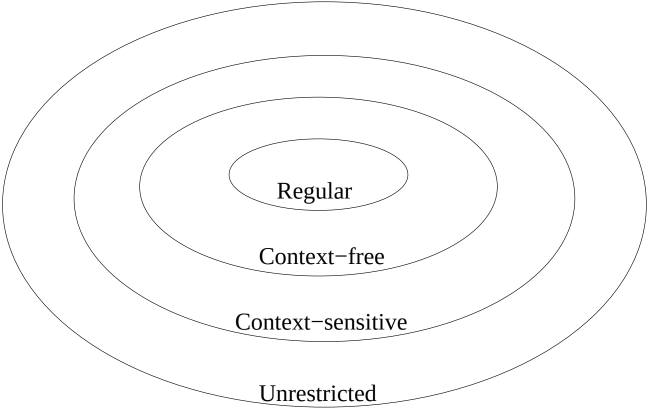
Chomsky Hierarchy: classifies languages on scale of complexity:

- Regular languages: those whose phrases can be 'recognized' by a finite state machine.
- Context-free languages: the set of languages accepted by pushdown automata. Many aspects of PLs and NLs can be described at this level;
- Context-sensitive languages: equivalent with a linear bounded nondeterministic Turing machine, also called a linear bounded automaton. Need this to capture e.g. *typing rules* in PLs.
- Unrestricted languages: *all* languages that can in principle be defined via mechanical rules.

Review



Review



Where do human languages fit within this complexity hierarchy?

Recursion

The potential infiniteness of the language faculty has been recognized by Galileo, Descartes, von Humboldt.

Discrete Infinity

- Sentences are built up by discrete units
- There are 6-word sentences, and 7-word sentences, but no 6.5 word sentences
- There is no longest sentence!
- There is no non-arbitrary upper bound to sentence length!

Mary thinks that John thinks that George thinks that Mary thinks that this course is boring!
I ate lunch and slept and watched tv and went to the bathroom and had a coffee and got dressed . . .

Strong and Weak Adequacy

Questions about the formal complexity of language are about the computational power of syntax, as represented by a grammar that's adequate for it.

A strongly adequate grammar

- generates all and only the strings of the language;
- assigns them the "right" structures — ones that support a correct representation of meaning. (See previous lecture.)

A weakly adequate grammar

generates all and only the strings of a language but doesn't necessarily give a correct (insightful) account of their structures.

Is Natural Language Regular?

It is generally agreed that NLs are not (in principle) regular.

Centre-embedding

[The cat$_1$ likes tuna fish$_1$].
[The cat$_1$ [the dog$_2$ chased$_2$] likes tuna fish$_1$].
[The cat$_1$ [the dog$_2$ [the rat$_3$ bit$_3$] chased$_2$] likes tuna fish$_1$].

Is Natural Language Regular?

It is generally agreed that NLs are not (in principle) regular.

Centre-embedding

[The cat$_1$ likes tuna fish$_1$].
[The cat$_1$ [the dog$_2$ chased$_2$] likes tuna fish$_1$].
[The cat$_1$ [the dog$_2$ [the rat$_3$ bit$_3$] chased$_2$] likes tuna fish$_1$].

Idea of proof

(the+noun)^n (transitive verb)$^{n-1}$ likes tuna fish.
A $=$ { the cat, the dog, the rat, the elephant, the kangaroo $\ldots$ }
B $=$ { chased, bit, admired, ate, befriended $\ldots$ }
Intersect `/A* B* likes tuna fish/` with English
$L = x^n y^{n-1}$ *likes tuna fish*$, x \in A, y \in B$
Use pumping lemma to show L is not regular

Another example

Courtesy of an anonymous Inf2a student in the 2012 exam . . .

John, Andrew and Mark were wearing T-shirts
that were red, blue and yellow respectively.

Using this idea, can encode the language $\{a^n b^n \mid n \geq 2\}$.

Is Natural Language Context Free?

It seems NLs aren't always context free! E.g. in Swiss German, some verbs (e.g. *let*, *paint*) take an object in accusative form, while others (e.g. *help*) take it in dative form.

Crossing dependencies

...das mer	d'chind	em Hans	es huus	lönd	hälfe	aastriiche
...that we	the children	Hans	the house	let	help	paint
	NP-ACC	NP-DAT	NP-ACC	V-ACC	V-DAT	V-ACC

...that we let the children help Hans paint the house

Abstracting out the key feature here, we see that the same sequence over $\{a, d\}$ (in this case *ada*) must 'appear twice'.

But it turns out that $\{ss \mid s \in \{a, d\}^*\}$ isn't context-free (see a later lecture). Hence neither is Swiss German!

Weaker examples

These ‘crossing dependencies’ are non-context-free in a very strong sense: no CFG is even weakly adequate for modelling them.

Other phenomena can *in theory* be modelled using CFGs, though it seems unnatural to do so. E.g. a versus an in English.

a banana	an apple
a large apple	an exceptionally large banana

Over-simplifying a bit: a before consonants, an before vowels.

In theory, we could use a context-free grammar:

$$\begin{array}{rclcl@{\qquad}rclcl}
\mathrm{NP} & \rightarrow & \mathbf{a}\ \mathrm{NP1}^c & & & \mathrm{NP} & \rightarrow & \mathbf{an}\ \mathrm{NP1}^v & & \\
\mathrm{NP1}^c & \rightarrow & \mathrm{N}^c & \mid & \mathrm{AP}^c\ \mathrm{NP1} & \mathrm{NP1}^v & \rightarrow & \mathrm{N}^v & \mid & \mathrm{AP}^v\ \mathrm{NP1} \\
\mathrm{AP}^c & \rightarrow & \mathrm{A}^c & \mid & \mathrm{Adv}^c\ \mathrm{AP} & \mathrm{AP}^v & \rightarrow & \mathrm{A}^v & \mid & \mathrm{Adv}^v\ \mathrm{AP}
\end{array}$$

But more natural to use context-sensitive rules, e.g.

$$\begin{array}{lcl}
\mathrm{DET}\ [\text{c-word}] & \rightarrow & \mathbf{a}\ [\text{c-word}] \\
\mathrm{DET}\ [\text{v-word}] & \rightarrow & \mathbf{an}\ [\text{v-word}]
\end{array}$$

Mild context sensitivity

A set $\mathcal{L}$ of languages is mildly context-sensitive if:

- $\mathcal{L}$ contains all context-free languages.
- $\mathcal{L}$ can describe cross-serial dependencies. There is an $n \geq 2$ such that $\{w^k | w \in T^*\} \in \mathcal{L}$ for all $k \leq n$.
- The languages in $\mathcal{L}$ are polynomially parsable.
- The languages in $\mathcal{L}$ have the constant growth property.

Let X be an alphabet and $L \subseteq X^*$. L has constant growth property iff there is a constant $c_0 > 0$ and a finite set of constants $C \subset \mathbb{N} \setminus \{0\}$ such that for all $w \in L$ with $|w| > c_0$, there is a $w' \in L$ with $|w| = |w'| + c$ for some $c \in C$

Example: the language $\{a^{2^n} | n \in \mathbb{N}\}$ does not have the constant growth property.

Combinatory Categorial Grammars

CCGs are more powerful than CFGs, but less powerful than arbitrary CSGs.

They satisfy the criteria for mildly context-sensitive languages, i.e. the set of languages defined by CCGs is mildly context-sensitive.

The set of categories (nonterminals) in CCG is compositional, defined by a set of atomic units such as S, NP and PP.

There are combination rules that tell us how to generate new categories from older ones in a derivation.

Linear Indexed Grammars

Linear indexed grammars (LIGs) are more powerful than CFGs, but much less powerful than an arbitrary CSGs. Think of them as mildly context sensitive grammars. These seem to suffice for NL phenomena.

Definition

An indexed grammar has three disjoint sets of symbols: terminals, non-terminals and indices.

An index is a stack of symbols that can be passed from the LHS of a rule to its RHS, allowing counting and recording what rules were applied in what order.

Summary

- The ‘narrow’ language faculty involves a computational system that generates syntactic representations that can be mapped onto meanings.
- This raises the question of the complexity of this system (its position in the Chomsky hierarchy).
- A weakly adequate grammar generates the correct strings, while a strongly adequate one also generates the correct structures.
- NLs appear to surpass the power of context-free languages, but only just.
- The mild form of context-sensitivity captured by LIGs seems weakly adequate for NL structures.

Next Lecture: Models of human parsing.