Mildly Context-Sensitive Grammar

Formalisms:

Mild Context-Sensitivity

Laura Kallmeyer
Heinrich-Heine-Universität Düsseldorf
Sommersemester 2011

Overview

1. Mild Context-Sensitivity
2. Cross-Serial Dependencies
3. Constant Growth
4. Semilinearity
5. MCS and TAG

Mild Context-Sensitivity (1)

- We know that CFGs are not powerful enough to describe all natural language phenomena.
- Question: How much context-sensitivity is necessary to deal with natural languages?
- In an attempt to characterize the amount of context-sensitivity required, [Joshi, 1985] introduced the notion of mild context-sensitivity (MCS).
- MCS is a term that refers to classes of languages, not to formalisms.

Mild Context-Sensitivity (2)

1. A set \( \mathcal{L} \) of languages is mildly context-sensitive iff
   (a) \( \mathcal{L} \) contains all context-free languages.
   (b) \( \mathcal{L} \) can describe a limited amount of cross-serial dependencies.
   (c) The languages in \( \mathcal{L} \) are polynomially parsable, i.e., \( \mathcal{L} \subseteq \text{PTIME} \).
   (d) The languages in \( \mathcal{L} \) have the constant growth property.
2. A formalism \( F \) is mildly context-sensitive iff the set \( \{ L | L = L(G) \text{ for some grammar } G \text{ of the formalism } F \} \) is mildly context-sensitive.
Cross-Serial Dependencies
The second property (limited amount of cross-serial dependencies) is a little unclear. It can be taken to mean the following:
There is an $n \geq 2$ such that $\{u^k \mid w \in T^*\} \in \mathcal{L}$ for all $k \leq n$.

Constant Growth (1)
The constant growth property roughly means that, if we order the words of a language according to their length, then the length grows in a linear way.

Example: $\{a^n \mid n \geq 0\}$ is not of constant growth.

The following definition is from [Weir, 1988].

Definition 1 (Constant Growth Property) Let $X$ be an alphabet and $L \subseteq X^*$. $L$ has the constant growth property if there is a constant $c_0 > 0$ and a finite set of constants $C \subseteq \mathbb{N} \setminus \{0\}$ such that for all $w \in L$ with $|w| > c_0$, there is a $w' \in L$ with $|w| = |w'| + c$ for some $c \in C$.

Constant Growth (2)
How can we show the constant growth property for a given language?

• Via a pumping lemma. The maximal size of the pumped material is the maximal length difference we encounter in the language.

• Via letter-equivalence with a context-free language. This shows the semilinearity of the language, a property that is stronger than constant growth.

Constant Growth (3)
Example: Pumping Lemma for a CFL $L$: There is a $c > 0$ such that for all $w \in L$ with $|w| \geq c$: $w = x_1y_1z$ with

- $|v_1v_2| \geq 1$,
- $|v_1yv_2| \leq c$, and
- for all $i \geq 0$: $x_1^i y_1^i z \in L$.

Consequently, $L$ is of constant growth with $c_0 = c$ and $C = \{1, 2, \ldots, c\}$. 
Semilinearity (1)

Semilinearity is a language property that is stronger than constant-growth.

- Constant growth is only an existential property: For a language to be of constant growth, it is enough to have an infinite sequence \( w_1, w_2, \ldots \) in the language with \( |w_{i+1}| - |w_i| \leq c \). Besides this, there can be other words in the language that arise from some exponential process.

- Semilinearity is a universal property: every word in the language is part of a sequence where the counts of the different terminals in these words are linear combinations of specific initial counts.

\[ \{c^n d^n \mid n \geq 0\} \cup \{a^n | a \mid n \geq 0\} \text{ is of constant growth.} \]

Semilinearity (2)

Two words are letter equivalent if they contain equal number of occurrences of each terminal symbol, and two languages are letter equivalent if every string in one language is letter equivalent to a string in the other language and vice-versa.

Example:

\[ \{ww \mid w \in \{a, b\}^*\} \text{ and } \{ww^R \mid w \in \{a, b\}^*\} \text{ are letter equivalent.} \]

Definition 3 (Letter equivalent) Let \( X \) be an alphabet.

1. Two words \( w_1, w_2 \in X^* \) are letter equivalent if there is a Parikh mapping \( p \) such that \( p(w_1) = p(w_2) \).

2. Two languages \( L_1, L_2 \subseteq X^* \) are letter equivalent if there is a Parikh mapping \( p \) such that \( p(L_1) = p(L_2) \).

Semilinearity (3)

We define for \( \langle a_1, \ldots, a_n \rangle, \langle b_1, \ldots, b_n \rangle \in \mathbb{N}^n \) and \( m \in \mathbb{N} \) that

- \( \langle a_1, \ldots, a_n \rangle + \langle b_1, \ldots, b_n \rangle := \langle a_1 + b_1, \ldots, a_n + b_n \rangle \), and

- \( m \langle a_1, \ldots, a_n \rangle := \langle ma_1, \ldots, ma_n \rangle \).

A language is semilinear if its Parikh image is the union of finitely many linear sets.

Example:

\[ \{a^n b^n \mid n \geq 0\} \cup \{b^n c^n \mid n \geq 0\}, a \text{ the first, } b \text{ the second and } c \text{ the third terminal.} \]

Parikh image:

- \( \{0,0,n(1,1,0) \mid n \geq 0\} \cup \{0,0,n(0,1,1) \mid n \geq 0\} \)
Semilinearity (5)

Definition 4 (Semilinear) 1. Let $x_0, \ldots, x_m$ with $m \geq 0$ be in $\mathbb{N}^n$ for some $n \geq 0$.

The set \[ \{ x_0 + n_1 x_1 + \cdots + n_m x_m \mid n_i \in \mathbb{N} \text{ for } 1 \leq i \leq m \} \]

is a linear subset of $\mathbb{N}^n$.

2. The union of finitely many linear subsets of $\mathbb{N}^n$ is a semilinear subset of $\mathbb{N}^n$.

3. A language $L \subseteq X^*$ is semilinear iff there is a Parikh mapping $p$ such that $p(L)$ is a semilinear subset of $\mathbb{N}^n$ for some $n \geq 0$.

Semilinearity (6)

Proposition 1 The constant growth property holds for semilinear languages.

Assume $L \subseteq X^*$ is semilinear and $p(L)$ is a semilinear Parikh image of $L$ where $p(L)$ is the union of the linear sets $M_1, \ldots, M_l$. Then the constant growth property holds for $L$ with

\[ c_0 := \max \{ \sum_{i=1}^n y_i \mid \text{there are } x_1, \ldots, x_m \text{ such that } \{ y_1, \ldots, y_n \} + n_1 x_1 + \cdots + n_m x_m \mid n_i \in \mathbb{N} \} \]

is one of the sets $M_1, \ldots, M_l$ and

\[ C := \{ \sum_{i=1}^n y_i \mid \text{there are } x_1, \ldots, x_m \text{ such that } \{ x_1 + n_1 (y_1, \ldots, y_n) + \cdots + n_m x_m \mid n_i \in \mathbb{N} \} \]

is one of the sets $M_1, \ldots, M_l$.

Semilinearity (7)

Parikh has shown that a language is semilinear if and only if it is letter equivalent to a regular language. The proof is given in [Kracht, 2003, p. 151]. As a consequence, we obtain that context-free languages are semilinear.

Proposition 2 (Parikh Theorem)

Each context-free language is semilinear [Parikh, 1966].

Furthermore, each language that is letter equivalent to a semilinear language is semilinear as well since the Parikh images of the two languages are equal. Therefore, in order to show the semilinearity (and constant growth) of a language, it is sufficient to show letter equivalence to a context-free language.

Semilinearity (8)

Joshi’s hypothesis that natural languages are mildly context-sensitive has been questioned only by two natural language phenomena that have been claimed to be non-semilinear:

- Case stacking in Old Georgian [Michaelis and Kracht, 1996].

  The analyses of Old Georgian, however, are based on very few data since there are no speakers of Old Georgian today.

- Chinese number names [Radzinski, 1991]. It is however not totally clear to what extent this constitutes a syntactic phenomenon.

Therefore, even with these counterexamples, there is still good reason to assume that natural languages are mildly context-sensitive. Furthermore, non-semilinearity does not entail non-constant-growth.
MCS and TAG (1)
The set $L$ of all TALs
- contains all CFLs,
- is a subset of PTIME (parsing is $O(n^6)$),
- and contains the copy language, i.e., can generate a limited amount of cross-serial dependencies.

MCS and TAG (2)
Every TAL is of constant growth:
Pumping Lemma for a TAL $L$: There is a $c > 0$ such that for all $w \in L$ with $|w| \geq c$ there are $x, y, z, v_1, v_2, w_1, w_2, w_3, w_4 \in T^*$ such that
- $|v_1 v_2 w_1 w_3 w_4| \leq c$,
- $|w_1 w_2 w_3 w_4| \geq 1$,
- $w = x v_1 y v_2 z$, and
- $x v_1^n v_2^n y v_1^n v_2^n w_3 w_4 z \in L(G)$ for all $n \geq 0$.

Consequently, $L$ is of constant growth with $c_0 = 2c$ and $C = \{1, 2, \ldots, c\}$.

MCS and TAG (3)
Every TAL $L$ is semilinear:
- Take the CFG that describes the set of derivation trees;
- Add to the right-hand side of every production all terminals that label nodes in the elementary trees of the right-hand side.
The result is a CFG that is letter equivalent to the original TAG.

Example: TAL for copy language, elementary trees $\alpha, \beta_a$ and $\beta_b$.
Letter equivalent CFG:

$$
S \rightarrow \alpha \\
\alpha \rightarrow \varepsilon \quad \alpha \rightarrow aa\beta_a \quad \alpha \rightarrow bb\beta_b \\
\beta_a \rightarrow \varepsilon \quad \beta_a \rightarrow aa\beta_a \quad \beta_a \rightarrow bb\beta_b \\
\beta_b \rightarrow \varepsilon \quad \beta_b \rightarrow aa\beta_a \quad \beta_b \rightarrow bb\beta_b
$$

References


