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., $\mathcal{L} \subset PTIME$.
   (d) The languages in $\mathcal{L}$ have the *constant growth property*.
2. A formalism F is *mildly context-sensitive* iff the set $\{L \mid L = L(G)$ for some grammar G 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 $\{w^k \mid w \in T^*\} \in \mathcal{L}$ 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: $\{a^{2^n} \mid n \geq 0\}$ 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 *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 $c \in C$.*

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 $|w| \geq c$: $w = xv_1yv_2z$ with

- $|v_1v_2| \geq 1$,
- $|v_1yv_2| \leq c$, and
- for all $i \geq 0$: $xv_1^iyv_2^iz \in L$.

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.
  $\{c^n d^n \mid n \geq 0\} \cup \{a^{2^n} \mid n \geq 0\}$ is of constant growth.
- 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.
  $\{c^n d^n \mid n \geq 0\} \cup \{a^{2^n} \mid n \geq 0\}$ is not semilinear.
  $\{a^n b^n \mid n \geq 0\} \cup \{(aa)^n b^n \mid n \geq 0\}$ is semilinear.

Semilinearity (2)

First, we introduce Parikh mappings. These are functions that count for each letter of an (ordered) alphabet the occurrences of this letter in a word w.

Example: $w = aababaab$, a the first letter and b the second of the alphabet. Parikh image of w: $\langle |w|_a, |w|_b \rangle = \langle 5, 3 \rangle$.

Definition 2 (Parikh mapping) *Let $X = \{a_1, \ldots, a_n\}$ be an alphabet with a fixed order of the elements. The Parikh mapping $p : X^* \to \mathbb{N}^n$ is defined as follows:*

- *For all $w \in X^* : p(w) := \langle |w|_{a_1}, \ldots, |w|_{a_n} \rangle$ where $|w|_{a_i}$ is the number of occurrences of a_i in w.*
- *For all languages $L \subseteq X^* : p(L) := \{p(w) \mid w \in L\}$ is the Parikh image of L.*

Semilinearity (3)

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.

Ex.: $\{ww \mid w \in \{a,b\}^*\}$ and $\{ww^R \mid w \in \{a,b\}^*\}$ 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 (4)

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 imgage is the union of finitely many linear sets.

Ex.: $\{a^n b^n \mid n \geq 0\} \cup \{b^n c^n \mid n \geq 0\}$, a the first, b the second and c the third terminal.

Parikh image:
$\{\langle 0,0,0 \rangle + n\langle 1,1,0 \rangle \mid n \geq 0\} \cup \{\langle 0,0,0 \rangle + n\langle 0,1,1 \rangle \mid n \geq 0\}$

Semilinearity (5)

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

The set $\{x_0 + n_1 x_1 + \cdots + n_m x_m \mid n_i \in \mathbb{N}$ 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 $M_1, \dots, M_l$. Then the constant growth property holds for L with

$$c_0 := max\{\Sigma_{i=1}^n y_i \mid \text{there are } x_1, \dots, x_m \text{ such that } \{\langle y_1, \dots, y_n\rangle + n_1 x_1 + \cdots + n_m x_m \mid n_i \in \mathbb{N}\} \text{ is one of the sets } M_1, \dots, M_l\} \text{ and}$$

$$C := \{\Sigma_{i=1}^n y_i \mid \text{there are } x_1, \dots, x_m \text{ such that } \{x_1 + n_1\langle y_1, \dots, y_n\rangle + \cdots + n_m x_m \mid n_i \in \mathbb{N}\} \text{ is one of the sets } M_1, \dots, M_l\}.$$

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 $\mathcal{L}$ of all TALs

- contains all CFLs,
- is a subset of PTIME (parsing is $\mathcal{O}(n^6)$),
- 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

- $|v_1 v_2 w_1 w_2 w_3 w_4| \leq c$,
- $|w_1 w_2 w_3 w_4| \geq 1$,
- $w = x v_1 y v_2 z$, and
- $x w_1^n v_1 w_2^n y w_3^n v_2 w_4^n z \in L(G)$ for all $n \geq 0$.

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

MCS and TAG (3)

Every TAL L is semilinear:

- Take the CFG that describes the set of derivation trees;
- Add to the righthand side of every production all terminals that label nodes in the elementary trees of the righthand side.

The result is a CFG that is letter equivalent to the original TAG.

Example: TAG for copy language, elementary trees α, β_a and β_b.

Letter equivalent CFG:

$$\begin{array}{llll} S \to \alpha & \alpha \to \varepsilon & \alpha \to aa\beta_a & \alpha \to bb\beta_b \\ & \beta_a \to \varepsilon & \beta_a \to aa\beta_a & \beta_a \to bb\beta_b \\ & \beta_b \to \varepsilon & \beta_b \to aa\beta_a & \beta_b \to bb\beta_b \end{array}$$

References

[Joshi, 1985] Joshi, A. K. (1985). Tree adjoining grammars: How much contextsensitivity is required to provide reasonable structural descriptions? In Dowty, D., Karttunen, L., and Zwicky, A., editors, *Natural Language Parsing*, pages 206–250. Cambridge University Press.

[Kracht, 2003] Kracht, M. (2003). *The Mathematics of Language*. Number 63 in Studies in Generative Grammar. Mouton de Gruyter, Berlin.

[Michaelis and Kracht, 1996] Michaelis, J. and Kracht, M. (1996). Semilinearity as a Syntactic Invariant. In *Logical Aspects of Computational Linguistics*, Nancy.

[Parikh, 1966] Parikh, R. (1966). On context-free languages. *Journal of the ACM*, 13:570–581.

[Radzinski, 1991] Radzinski, D. (1991). Chinese number-names, tree adjoining languages, and mild context-sensitivity. *Computational Linguistics*, 17:277–299.

[Weir, 1988] Weir, D. J. (1988). *Characterizing Mildly Context-Sensitive Grammar Formalisms*. PhD thesis, University of Pennsylvania.