Ogden’s Lemma, Multiple Context-Free Grammars, and the Control Language Hierarchy

Makoto Kanazawa
National Institute of Informatics and SOKENDAI
Japan

Multiple Context-Free Grammars

- Independently by Vijay-Shanker, Weir, and Joshi (1987)
- Many equivalent models
- Often thought to be an adequate formalization of mildly context-sensitive grammars (Joshi 1985)

Arising from concerns in computational linguistics. CFGs are almost good enough for NL grammars, but not quite; a mild extension of CFGs is needed. Several criteria were put forward as to what constitutes a “mild” extension.

Which properties of CFGs are shared by/ generalize to MCFGs?
Multiple Context-Free Grammars

\[ A(\alpha_1, \ldots, \alpha_q) \leftarrow B_1(x_1, \ldots, x_{1,q_1}), \ldots, B_n(x_{n,1}, \ldots, x_{n,q_n}) \]

\( n \geq 0, q, q_i \geq 1, \alpha_i \in (\Sigma \cup \{x_{i,j} \mid i \in [1,n], j \in [1,q_i]\})^* \)

- each \( x_{i,j} \) occurs exactly once in \((\alpha_1, \ldots, \alpha_q)\)

\[ q = \dim(A) \quad \text{(dimension of A)} \]
\[ \dim(S) = 1 \]
\[ L(G) = \{ w \in \Sigma^* \mid G \vdash S(w) \} \]

It’s best to think of an MCFG as a kind of logic program. Each rule is a definite clause. Nonterminals are predicates on strings.

\[
\begin{align*}
S(x_1 # x_2) & \leftarrow D(x_1, x_2) \\
D(\varepsilon, \varepsilon) & \leftarrow \\
D(x_1, y_1, y_2, x_2) & \leftarrow E(x_1, x_2), D(y_1, y_2) \\
E(a_1 x_1 a, a x_2 a) & \leftarrow D(x_1, x_2)
\end{align*}
\]

\[ m \text{-MCFG} = \text{MCFG with nonterminal dimension not exceeding } m \]
\[ 1 \text{-MCFG} = \text{CFG} \]

Derivation tree for \( w = \text{proof of } S(w) \)

The languages of MCFGs form an infinite hierarchy.
Derivation trees of MCFGs are similar to those of CFGs. When the same nonterminal occurs twice on the same path of a derivation tree,...

You can decompose the derivation tree into three parts, and the middle part can be iterated any number of times, including zero times. In the overall derivation tree, the variables $x_1, x_2, y_1, y_2$ are instantiated by ...

The number of iterated substrings (factors) larger than two.

For MCFGs, need to consider a generalized form of the condition of the pumping lemma. Not straightforward; open question for a long time.

Iterative Properties

L is **k-iterative** iff $\exists p \forall z \in L(|z| \geq p \Rightarrow$

$$\exists u_1 \ldots u_{k+1} V_1 \ldots V_k (Z = u_1 V_1 \ldots u_k V_k u_{k+1} \land V_1 \ldots V_k \neq \varepsilon \land \forall n \geq 0(u_1 V_1^n \ldots u_k V_k^n u_{k+1} \in L))$$

$L \in \text{CFL} \Rightarrow L$ is 2-iterative

$L \in m-\text{MCFL} \Rightarrow L$ is $2m$-iterative?
Difficulty with Pumping

The middle part of the derivation tree may look like this.

Or like this.

The pumping lemma fails for 3-MCFGs.
Pumping Lemma for Subclasses

Well-Nestedness

Difficulty with Pumping

Pumping possible for special cases. Well-nested MCFGs.

Has a natural equivalent characterization: $\text{yCFT}_{sp}$

Pumping not easy to prove even form well-nested MCFGs: this situation can still arise.
A very simple example. The only choice you can make is the number of times you use the second rule. Actually 2-iterative, but no straightforward connection between the iterated substrings and parts of derivation trees.

My proof of the pumping lemma for m-MCFL\textsubscript{wn} and 2-MCFL is not straightforward.

Ogden’s Lemma for CFL

\[ L \in \text{CFL} \Rightarrow \]

\[ \exists p \forall z \in L \text{(at least } p \text{ positions of } z \text{ are marked} \Rightarrow \]

\[ \exists u_1v_1v_2u_2v_2u_3 \text{ such that } z = u_1v_1u_2v_2u_3 \land (u_1, v_1, u_2, v_2, u_3 \text{ each contain a marked position} \land v_1v_2v_3 \text{ contains no more than } p \text{ marked positions} \land \forall n \geq 0 (u_1v_1^n u_2v_2^n u_3 \in L)) \]

Ogden 1968
L has the **weak Ogden property** iff
\[ \exists p \forall z \in L \text{ (at least } p \text{ positions of } z \text{ are marked } \Rightarrow \]
\[ \exists k \geq 1 \exists u_1 \ldots u_{k+1} v_1 \ldots v_k (\]
\[ z = u_1 v_1 \ldots u_k v_k u_{k+1} \wedge \]
\[ \exists i (v_i \text{ contains a marked position }) \wedge \]
\[ \forall n \geq 0 (u_1 v_1^n \ldots u_k v_k^n u_{k+1} \in L) \]

There are various ways of generalizing Ogden’s lemma suitable for MCFGs. At least this much should be implied.

The Failure of Ogden’s Lemma

This is the first new result in this talk.

A language for which the weak Ogden property fails.
non-branching 3-MCFG

A(ε) ← A(bx_1)
A(x_1) ← A(x_1)
B(x_1, bx_3) ← B(x_1, x_3)
C(x_1, x_2, bx_3) ← C(x_1, x_2, x_3)
C(x_1, bx_2, bx_3) ← C(x_1, x_2, x_3)
D(x_1, x_3) ← D(x_1, x_3)
S(x_1, x_3) ← S(x_1, x_3)

A(ε) ← A(bx_1)
A(x_1) ← A(x_1)
B(x_1, bx_3) ← B(x_1, x_3)
C(x_1, x_2, bx_3) ← C(x_1, x_2, x_3)
C(x_1, bx_2, bx_3) ← C(x_1, x_2, x_3)
D(x_1, x_3) ← D(x_1, x_3)
S(x_1, x_3) ← S(x_1, x_3)

B(x_1, ε) ← A(x_1)
B(x_1, bx_2) ← B(x_1, x_2)
C(x_1, x_2, ε) ← C(x_1, x_2, x_3)
C(x_1, bx_2, x_3) ← C(x_1, x_2, x_3)
D(x_1, x_3) ← D(x_1, x_3)
S(x_1, x_3) ← S(x_1, x_3)

B(a x_1, b x_2) ← B(x_1, x_2)
C(x_1, x_2) ← C(x_1, x_2)
D(x_1, x_3) ← D(x_1, x_3)
S(x_1, x_3) ← S(x_1, x_3)

C(x_1, a x_2, bx_3) ← C(x_1, x_2, x_3)
C(x_1, bx_2, bx_3) ← C(x_1, x_2, x_3)
D(x_1, x_3) ← D(x_1, x_3)
S(x_1, x_3) ← S(x_1, x_3)

C(x_1, ε, bx_3) ← C(x_1, x_2, x_3)
C(x_1, bx_2, bx_3) ← C(x_1, x_2, x_3)
D(x_1, x_3) ← D(x_1, x_3)
S(x_1, x_3) ← S(x_1, x_3)

D(x_1, bx_3) ← D(x_1, x_3)
D(x_1, bx_3) ← D(x_1, x_3)
S(x_1, bx_3) ← S(x_1, x_3)
S(x_1, bx_3) ← S(x_1, x_3)

2-MCFG

\{ a_1 b_1 \ldots a_i b_i \ldots a_n b_{n-1} \mid n \geq 3, i_0, \ldots, i_n \geq 0 \}

Mark the positions of $.

\{ a_1 b_1 \ldots a_i b_i \ldots a_n b_{n-1} \mid n \geq 3, i_0, \ldots, i_n \geq 0 \}

is 2-iterative

a_1 \ldots a_i b \ldots a_n b_{n-1}

Subclasses of MCFL that are known to have an Ogden property.

Weir\’s (1992) Control Language Hierarchy

\[ C_k \subseteq 2^{k-1}\text{-MCFL} \]
(Kanazawa and Salvati 2007)

\[ C_2 = 2\text{-MCFL} \]

Generalization of Ogden\’s lemma

\[ C_1 = \text{CFL} \]
(Palis and Shende 1995)
Control Grammars

Languages in each level of the control language hierarchy are given by "control grammars".

Control Language Hierarchy

\[ C_1 = CFL \]
\[ C_{k+1} = \{ L(G, K) | K \in C_k \} \]

Ogden’s Lemma for \( C_k \)

It is quite straightforward to prove an Ogden’s lemma for \( C_k \).

\[ \exists p \forall z \in L \text{ (at least } p \text{ positions of } z \text{ are marked } \Rightarrow \]
\[ \exists u_1 \ldots u_{2k+1} v_1 \ldots v_p ( \]
\[ z = u_1 v_1 \ldots u_p v_p u_{2k+1} \land \]
\[ \exists i(u_i, v_i, u_{i+1} \text{ each contain a marked position}) \land \]
\[ v_{i-1}^p \ldots u_{2k} v_{2k-1}^i, \text{ contains no more than } p \]
\[ \text{marked positions } \land \]
\[ \forall n \geq 0(u_1 v_1^p u_2 v_2^p \ldots u_n v_n^p u_{2k+1} \in L) \]

Palis and Shende 1995

- \( k = 1 \) gives Ogden’s (1968) original lemma.
Proper MCFGs

Approach this existing result from the MCFG formalism.
Sufficient condition for an Ogden property.

Slight variation of an earlier example.

Ogden’s Lemma for m-MCFL-prop

Constrain all of $v_1, \ldots, v_{2m}$

$\exists p \forall z \in L$ (at least $p$ positions of $z$ are marked) \Rightarrow

$\exists u_1 \ldots u_{2m+1} v_1 \ldots v_{2m} (z = u_1 v_1 \ldots u_{2m} v_{2m} u_{2m+1} \land$

$\exists (u, v, u_1$ each contain a marked position) \land

$\forall (u_1 v_1 u_2 v_2, \ldots, u_{2m} v_{2m} u_{2m+1} \in L)$

$m = 1$ gives Ogden’s (1968) original lemma.
The earlier result is subsumed by the present result.

Generates strings with nonterminals.

The construction doubles the dimension, preserves properness.
The different requirement shows the properness of the inclusion.

Summary

- Pumping doesn't imply Ogden: There is no Ogden-like theorem for $3-$MCFL$_{wn}$ n 2-MCFL
- There is a natural Ogden's lemma for m-MCFL$_{prop}$
- Covers Weir's control language hierarchy