CMSC35000-1 Introduction to Artificial Intelligence **Winter 2005**

Lecture 10: 3/2/05

Lecturer: Partha Niyogi *Scribe: Hoang Trinh*

10.1 Introduction

Let Σ be the set of words, $\Sigma = \{$the ball run and ...$\}$

Let N be the set of nonterminals, $N = \{$S, V, N, Adj, Pr, ...$\}$

Consider the English language, $L_{eng} \in \Sigma^*$, having the following rule:

$$\alpha \to \beta$$
$$\alpha, \beta \in (\Sigma \cup N)^*$$

Review the levels in Chomsky's hierarchy: with $x \in \Sigma, \beta \in N$

- Type 1 (Regular language):
$$A \to Bx$$
$$A \to x$$
- Type 2 (Context-free language):
$$A \to \alpha$$
- Type 3 (Context-sensitve language):
$$\beta A \gamma \to \alpha$$

The following sentences belong to the English language above:

The rat died (NV)

(The rat (the cat (the dog chased) ate) died) ($N^\alpha V^\beta$)

Claim: $L_{eng} \cap \{N^\alpha V^\beta\} = \{N^k V^k\}$

Question:

- Is English language context-free or not?
- Is English language regular or not?

10.2 The logical problem of language acquisition

Let us consider a language L_{eng} with a grammar $g_{eng} \in G$

Sentences: $s_1, s_2, ...$

Grammar: $g_1, g_2, ...$

10.2.1 The Central Dogma

1. All languages can be learnt
2. Learning is from positive examples
3. Learning does not depend on the precise order of examples

T is called a text corpus of a language L if:

T = $s_1, s_2, .., s_n$ such that:

- each $s \in L$ occurs at least once in T
- no $s \notin L$ occurs in T

Learning algorithm:

Let A be an algorithm that learns grammar G from a set of data sequences D.

$$A : D \rightarrow G$$

D = $\bigcup_{k \geq 1} D_k$ with

$$D_k = \{(s_1, s_2, .., s_k) \text{ such that } s_i \in \Sigma^*\}$$

D_k is the set of all data streams of length k

G is the set of grammar to be learnt by A. $A(\alpha) \in G$ with $\alpha \in D$

$\alpha \in D \Rightarrow \alpha \in D_j$ for some j

Let t_k be the first k elements of the sequence $T = s_1, s_2, .., s_n$

i.e $t(k) = s_1, s_2, .., s_k$ therefore:

$$t_k \in D \quad \forall k$$

A learns g on text T if $A(t_k) \rightarrow g$

$A(t_k) \rightarrow g$ if $\exists N$ s.t $\forall n > N$

$$L_A(t_n) = L_g$$

Note that:

A learns g if $\forall t$ from L_g, A learns g on text t

A learns G if $\forall g$ from G, A learns g

Theorem: If g is learnable by A then there exists a locking sequence σ for g

$$\sigma = s_1, s_2, ..., s_k s_i \in L_g$$

σ is called a locking sequence for g if:

$L_{A(\sigma)} = L_g$ and $\forall$ extension $\alpha = (s_1^{'}, s_2^{'}, ..., s_m^{'})$ with $s_m^{'} \in L_g$, we have:

$$L_{A(\sigma\circ\alpha)} = L_g$$

Prove:

Suppose not, i.e g is learnable yet no locking sequence.

Take any text t for g

$$t = s_1, s_2, ...$$

We will form a new text t':

Start at $q_1 = s_1$

Look at $A(q_1)$. If $L_{A(q_1)} \neq L_g$ then

$$q_2 = q_1 \circ s_2$$

else if $L_{A(q_1)} = L_g$

$$q_2 = q_1 \circ \alpha \circ s_2$$

(because there is no locking sequence $\Rightarrow \exists\alpha$ s.t $L_{A(q_1\circ\alpha} \neq L_g$)

Now consider the text $t' = q_1, q_2, ...$

Obviously t' is a text corpus of L_g because:

- all elements of L_g occur at least once in t'
- No element $\notin L_g$ in t'

We see that the text t' changes its mind infinitely often about $g \Rightarrow g$ is not learnable $\Rightarrow$ contradiction.

(theorem proved)

10.2.2 Gold Theorem

Gold Theorem: *(Gold 1967)*

If the family L (Superfinite family) consists of all the finite languages and at least 1 infinite language, then it is not learnable.

Proof:

Suppose not, i.e L_∞ is learnable, then by the Theorem, a locking sequence exists:

$$\exists\sigma_{L_\infty} = s_1, s_2, ...s_k s_i \in L_\infty$$

Consider $L = \bigcup_i \{s_i\}$

Consider a text t for L that begins with σ_{L_∞}

$$t = \sigma_{L_\infty} s_1^{'} s_2^{'} s_3^{'} ...$$

with $s_i^{'} \in L \subset L_\infty$

$L_{A(t_k)} = L_\infty$ with $\forall k \geq |\sigma_{L_\infty}|$

Therfore L is not learnable $\Rightarrow$ Contradiction

10.2.3 Questions

1. $L = \{L_1, L_2\}$ such that $L_1 \subset L_2$
   Is L learnable?
2. $L = \{\text{all finite languages}\}$
   Is L learnable?

Chomsky says the class G of all natural languages must be a subset of the set of all context-free languages (if natural languages are really context-free).