Mathematics for linguists

Gerhard Jäger
gerhard.jaeger@uni-tuebingen.de

Uni Tübingen, WS 2009/2010

November 26, 2009

The pumping lemma

- Let L be an **infinite** regular language over a finite alphabete Σ.
- There is a NFA M that accepts L.
- There is a number n such that M has n states.
- Almost all words in L consist of more than n letters.
  - Let $\vec{x} \in L$, with $l(\vec{x}) > n$.
  - There is a run of M that recognizes $\vec{x}$.
  - Since M has n states and $l(\vec{x}) > n$, at least one state of M is visited more than once. Let q be the state that is visited more than once.
  - $\vec{x}$ can be represented as $\vec{y} \cdot \vec{z} \cdot \vec{w}$, such that
    - between the initial state and q the string $\vec{y}$ is accepted,
    - between the first and the second visit of q the string $\vec{z}$ is accepted, and
    - between the second visit of q and the final state, the string $\vec{w}$ is accepted.

The pumping lemma

- Therefore:
  - the loop from q to q, during which $\vec{x}$ is accepted, can be repeated arbitrarily many times.
- Hence: $\vec{y} \cdot \vec{z}^{\,i} \cdot \vec{w} \in L$, for arbitrary $i \geq 0$.

The pumping lemma

These considerations hold for arbitrary infinite regular languages.

Theorem

Let L be an infinite regular language over the alphabet Σ. Then there is a number n, such that all words $\vec{x} \in L$ with $l(\vec{x}) > n$ can be decomposed into $\vec{x} = \vec{y} \cdot \vec{z} \cdot \vec{w}$, such that the following facts hold:

1. $l(\vec{z}) \geq 1$,
2. $l(\vec{y}) + l(\vec{z}) \leq n$, *and*
3. *for all* $i \in \mathbb{N}$: $\vec{y} \cdot \vec{z}^i \cdot \vec{w} \in L$.

Applications of the pumping lemma

The pumping lemma is useful if one wants to prove that a given language is **not** regular.

- Example: $L = \{a^m b^m | m > 0\}$ is not regular.
- Proof:
  - Suppose L is regular.
  - Then there is an n with the properties that are mentioned in the pumping lemma (the number of of states of the automaton that accepts L).
  - $a^n b^n \in L$.
  - $a^n b^n = \vec{x} \cdot \vec{y} \cdot \vec{z}$, with $l(\vec{x} \cdot \vec{y}) \leq n$, $l(\vec{y}) \geq 1$, and $\vec{x} \cdot \vec{z} \in L$.
  - $\vec{y} = a^j$, for some $j \geq 1$.
  - Hence $\vec{x} \cdot \vec{z} = a^{n-j} b^n \in L$, which is a contradiction to the definition of L.
  - Hence L is not regular.

Applications of the pumping lemma

- Example: $L = \{a^n b^m | m \geq n > 0\}$ is not regular.
- Proof:
  - Suppose L is regular.
  - Then there is an $n > 0$ with the properties that are mentioned in the pumping lemma.
  - $a^n b^n \in L$.
  - $a^n b^n = \vec{x} \cdot \vec{y} \cdot \vec{z}$, with $l(\vec{x} \cdot \vec{y}) \leq n$, $l(\vec{y}) \geq 1$, and $\vec{x} \cdot \vec{y}\vec{z} \in L$.
  - $\vec{y} = a^j$, for some $j \geq 1$.
  - Hence $\vec{x} \cdot \vec{y}^{(n+1)\cdot m} \cdot \vec{z} \in L$, and this is a contradiction to the definition of L.
  - Hence L is not regular.

Applications of the pumping lemma

- In a similar way it is possible to show that for a Σ with at least two elements, the following languages are not regular:
  - $\{\vec{w} \cdot \vec{w} | \vec{w} \in \Sigma^*\}$ (the "copy language")
  - $\{\vec{w} \cdot \vec{w}^R | \vec{w} \in \Sigma^*\}$ (the "mirror language" or "palindrome language")
- Somewhat more complex:

$$L = \{\vec{x} \in \{a, b\}^* | \text{number of } a \text{ in } \vec{x} = \text{number of } b \text{ in } \vec{x}\}$$

Applications of the pumping lemma

To prove that L is not regular, the following insight is important:

Theorem

If L_1 and L_2 are regular, then $L_1 \cap L_2$ is regular.

First we show that the complement of a regular language is also regular. This is almost obvious: If a DFA M accepts L, then you only have to turn the non-final states into final states and vice versa to get a DFA that accepts the complement $\overline{L} = \Sigma^* - L$.
During the last lecture it was shown that the union of two regular languages is also regular.
Thus, if L_1 and L_2 are regular, then $\overline{L_1}$ and $\overline{L_2}$ are also regular, and therefore alsoe $\overline{L_1} \cap \overline{L_2}$, and therefore alsoe $\overline{\overline{L_1} \cap \overline{L_2}}$.
According to de Morgan's law, this equals $L_1 \cap L_2$.

Applications of the pumping lemma

- Proof that $L = \{\vec{x} \in \{a, b\}^* | \text{number of } a \text{ in } \vec{x} = \text{number of } b \text{ in } \vec{x}\}$ is not regular:
  - a^*b^* is regular, because this language can be described by a regular expression.
  - Suppose L is regular. Then $L \cap a^*b^* = \{a^n b^n | n \geq 0\}$ must also be regular.
  - It was shown above that this language is not regular. Hence L is not regular either.

Is English regular?

With the help of the pumping lemma it is possible to show that natural languages are not regular. One possible argument for English runs as follows:

- It is possible to construct arbitrarily long sentences in English with the expressions "either ... or ...":

Either it rains or it snows.

Either John believes that either it rains or it snows, or the sun is shining.

Either it seems that either John believes that either it rains or it snows, or the sun is shining, or today is Thursday.

...

Is English regular?

- For every *either* in an English sentence, there is a corresponding or. The number of occurrences of or is thus at least as large as the number of occurrences of either.
- Regular languages are closed under the deletion of single elements from Σ: If I delete all occurrences of a given symbol — let's say a — in all words of a regular language L, the resulting language is again regular. (Proof: In a regular expression that describes L, replace all occurrences of a by ϵ.)

Is English regular?

- Suppose English is regular. More specifically, this means that the set E of all grammatical sentences of English is a regular language over the alphabet Σ (= the set of all morphemes of English).
- Then the language E', that is the result of deleting all morphemes except *either* and *or* in all English sentences, is also regular.
- $E' = \{\vec{x} \in \{\textit{either}, \textit{or}\}^* | \text{number of eithers in } \vec{x} \leq \text{number of ors in } \vec{x}\}$

Is English regular?

- *either** *or** is a regular language.
- Hence *either** *or** $\cap\, E' = \{\textit{either}^n\, \textit{or}^m | m \geq n\}$ is regular.
- Since we proved above that this language is **not** regular, we have derived a contradiction. So we proved that the original assumption — that E is regular — must be false.

Recursive constructions like the English *either ... or ...* can probably be found in all natural languages.[1] Hence Type-3 grammars are insufficient to describe natural languages.

[1] There are claims that the South American language Pirahã does not have such constructions, but this is heavily contested; see http://de.wikipedia.org/wiki/Piraha.