Inferring Descriptive Generalisations of Formal Languages

Dominik D. Freydenberger\textsuperscript{1}  Daniel Reidenbach\textsuperscript{2}

\textsuperscript{1}Goethe University, Frankfurt

\textsuperscript{2}Loughborough University, Loughborough

COLT 2010
Our goal:
Learning patterns common to a set of strings.

- **pattern**: word consisting of **terminals** \((\in \Sigma)\) and **variables** \((\in X)\)
- \(\text{Pat}_{\Sigma} := (\Sigma \cup X)^+\): set of all patterns over \(\Sigma\)
- **substitution**: terminal-preserving morphism \(\sigma : \text{Pat}_{\Sigma} \to \Sigma^*\)
  
  \((\forall a \in \Sigma : \sigma(a) = a)\)
- **language of a pattern** \(\alpha \in \text{Pat}_{\Sigma}\): set of all images of \(\alpha\) under substitutions (write: \(L(\alpha)\))

**Example**

\[
L_{\text{NE},\Sigma}(x \ a \ y \ x) = \{v \ a \ w \ v \mid v, w \in \Sigma^+\},
L_{\text{E},\Sigma}(x \ a \ y \ x) = \{v \ a \ w \ v \mid v, w \in \Sigma^*\}.
\]
The classical model

Identification in the limit of indexed families from positive data (Gold ’67)

- **Indexed family (of recursive languages):** \( \mathcal{L} = (L_i)_{i \in \mathbb{N}} \), where \( w \in L_i \) is uniformly decidable
- **Text** of a language \( L \): a total function \( t : \mathbb{N} \rightarrow \Sigma^* \) with \( \{ t(i) \mid i \in \mathbb{N} \} = L \)
- Set of all texts of \( L \): \( \text{text}(L) \)
- \( \mathcal{L} \in \text{LIM-TEXT} \) if there exists a computable function \( S \) such that, for every \( i \) and for every \( t \in \text{text}(L_i) \), \( S(t^n) \) converges to a \( j \) with \( L_j = L_i \)

- **NE-patterns** (yes, Angluin ’80)
- **E-patterns** (not if \( |\Sigma| \in \{2, 3, 4\} \), Reidenbach ’06, ’08)
- Terminal-free E-patterns (only if \( |\Sigma| \neq 2 \), Reidenbach ’06)
Inferring descriptive generalisations

Descriptive patterns

Definition

- Let $\mathcal{P}_\Sigma$ be a class of pattern languages over $\Sigma$.
- A pattern $\delta$ is $\mathcal{P}_\Sigma$-descriptive of a language $L$ if:
  1. $L(\delta) \in \mathcal{P}_\Sigma$,
  2. $L(\delta) \supseteq L$,
  3. there is no $L(\gamma) \in \mathcal{P}_\Sigma$ with $L(\delta) \supset L(\gamma) \supseteq L$.
- We write: $\delta \in D_{\mathcal{P}_\Sigma}(L)$

In other words: $L(\delta)$ is (one of) the closest generalisation(s) of $L$ in $\mathcal{P}_\Sigma$, and $\delta$ is (one of) the best description(s) of $L$.

Our approach:

Learning of such generalisations.
Definition

- Let $P_\Sigma$ be a class of pattern languages over $\Sigma$.
- Let $L$ be a class of nonempty languages over $\Sigma$.
- $L$ can be $P_\Sigma$-descriptively generalised ($L \in DG_{P_\Sigma}$) if there is a computable function $S$ such that, for every $L \in L$ and for every $t \in \text{text}(L)$, $S(t^n)$ converges to a $\delta \in D_{P_\Sigma}(L)$.

Main conceptual differences to LIM-TEXT:

- Infer generalisations instead of exact descriptions of the languages.
- Choose hypothesis space separate from language class.

Interesting phenomenon:

- One language can have several descriptive patterns,
- One pattern can be descriptive of several languages.
Characterisation theorem (for indexed families)

**Theorem**

Let $\Sigma$ be an alphabet, let $\mathcal{L} = (L_i)_{i \in \mathbb{N}}$ be an indexed family over $\Sigma$, and let $\mathcal{P}_\Sigma$ be a class of pattern languages. $\mathcal{L} = (L_i)_{i \in \mathbb{N}} \in \text{DG}_{\mathcal{P}_\Sigma}$ if and only if there are effective procedures $d$ and $f$ satisfying the following conditions:

(i) For every $i \in \mathbb{N}$, there exists a $\delta_d(i) \in D_{\mathcal{P}_\Sigma}(L_i)$ such that $d$ enumerates a sequence of patterns $d_{i,0}, d_{i,1}, d_{i,2}, \ldots$ satisfying, for all but finitely many $j \in \mathbb{N}$, $d_{i,j} = \delta_d(i)$.

(ii) For every $i \in \mathbb{N}$, $f$ enumerates a finite set $F_i \subseteq L_i$ such that, for every $j \in \mathbb{N}$ with $F_i \subseteq L_j$, if $\delta_d(i) \not\in D_{\mathcal{P}_\Sigma}(L_j)$, then there is a $w \in L_j$ with $w \not\in L_i$.

- $d$ is an enumeration of an appropriate subset of the hypothesis space
- $f$ is similar to Angluin’s telltales
Remarks

- Characterisation shows significant connection to Angluin’s characterisation of indexed families in LIM-TEXT.
- Main differences:
  - our model requires an enumeration of a subset of the hypothesis space,
  - we do not need to distinguish all $L_i, L_j$ with $L_i \neq L_j$,
  - the strategy in our proof might discard a correct hypothesis.
- Our strategy does not test membership or inclusion of pattern languages, but only membership for the indexed family.
Further topics

**Further directions in our paper:**

1. More general: Inductive inference with hypotheses validity relation (model $\text{HYP}$).
2. Less general: Consider a smaller class of patterns and a fixed strategy.
Inferring ePAT$_{tf, \Sigma}$-descriptive patterns

- ePAT$_{tf, \Sigma}$: The class of all E-pattern languages that are generated from terminalfree patterns.
- Inclusion for ePAT$_{tf, \Sigma}$ is well understood and decidable.
- Strategy Canon: For every finite set $S$, return the pattern $\delta \in D_{ePAT_{tf, \Sigma}} (S)$ that is minimal w.r.t. the length-lexicographical order.
- Telling set of $L$: A finite set $T \subseteq L$ with $D_{ePAT_{tf, \Sigma}} (T) \cap D_{ePAT_{tf, \Sigma}} (L) \neq \emptyset$.

**Theorem**

Let $\Sigma$ be an alphabet with $|\Sigma| \geq 2$. For every language $L \subseteq \Sigma^*$, and every text $t \in \text{text}(L)$, Canon converges correctly on $t$ if and only if $L$ has a telling set.
Telling set languages

\(\mathcal{TSL}_\Sigma\): the class of all languages over \(\Sigma\) that have a telling set

\(\mathcal{TSL}_\Sigma \in \text{DG}_{e\text{PAT}_{tf,\Sigma}}\), using Canon as strategy

Some properties of \(\mathcal{TSL}_\Sigma\):

- contains every DTF0L language \(\Rightarrow\) superfinite
- is not countable
- does not contain all of REG
- contains all \(e\text{PAT}_{tf,\Sigma}\)-languages (if \(|\Sigma| \neq 2\))
- does not contain all \(e\text{PAT}_{tf,\Sigma}\)-languages (if \(|\Sigma| = 2\))