言語の極限同定みたいな話

Hattori Kazuhiro

参考文献

- wikipedia/Language_identification_in_the_limit
- wikipedia/Grammar_induction
- paper/gold67limit.pdf
- paper/Angluin80
- paper/Shinohara83.pdf
- paper/Arimura94.pdf

諸定義

- アルファベットとはシンボルの有限の集合
  - 文脈によってはシンボルは 2 つ以上に限ることがある
  - $\Sigma = \{0, 1, \cdots\}$
- テキストとはアルファベットの上の文字列
  - $\Sigma^* = \{\epsilon, 0, 1, \cdot, 00, 01, 10, 11, \cdots, 010, \cdots\}$
    - 空文字列を ϵ と書く
- 言語とはテキストの集合
  - $L \subseteq \Sigma^*$

E. Mark Gold:

"Language Identification in the Limit", 1967

動機

- AI に言葉を話させたい (作文をさせたい)
- 言葉を話すためにはその言語を学習しなければならない
  1. AI (計算機) は言語を学習できるか
  2. どの時点で学習したと言えるか

言葉を話せる AI

- 人間の場合は、英語のルールを正しく書き下せなくても、“英語を話せる”と言う
- AI の場合でも、同様の“英語を話せる”とかいえるようなモデルが作れるはず (模倣)
  - すなわち言語のルールを書き下すことなく学習ができるはず
  - 英語に限らず自然言語一般のルールを学習させたい
    - 人間は任意の自然言語を学習可能
  - 自然言語をよりも単純な言語はより用意に学習できる
    - 自然言語を lower bound とする
    - より複雑な言語は別にどうでもいい
  - 少なくとも自然言語を含むような単純な言語の枠組みを設定する

言葉を話せる AI

- 人間の場合は、英語のルールを正しく書き下せなくても、“英語を話せる”と言う
- AI の場合でも、同様の“英語を話せる”とかいえるようなモデルが作れるはず (模倣)
  - すなわち言語のルールを書き下すことなく学習ができるはず
  - 英語に限らず自然言語一般のルールを学習させたい
    - 人間は任意の自然言語を学習可能
  - 自然言語をよりも単純な言語はより用意に学習できる
    - 自然言語を lower bound とする
    - より複雑な言語は別にどうでもいい
  - 少なくとも自然言語を含むような単純な言語の枠組みを設定する

言語の枠組み: 言語クラス

何の枠組みもなしに言語を学習するってのは難しすぎる

言語クラス $C = \{L_1, L_2, \cdots\}$ [1]

(実際は添字付きとは限らない)

- C はありえる言語全体 [2]
  - 例えば「ありえる自然言語全体」を設定する
  - 例えば「文脈自由文法全体」を設定する
    - その中には実在する自然言語 (英語など) が含まれる
- 言語の学習とは言語クラス C からそれっぽい言語を選ぶことだとする

[1] 個別の言語について、それが学習可能か？ を議論せず、言語クラスに対して議論する

[2] 学習したい言語が正規言語だと分かってるとき、正規言語全体というクラスを指定すればいい

情報提示 (information presentation)

何はともあれ、情報を提示されないと学習できない
Gold **は** Text **、**Informant **の** 2 **つがあると考えた**
現代語訳すると

- Text $\Rightarrow$ **正提示**
- Informant $\Rightarrow$ **完全提示**

紛らわしいので現代語を使います

情報提示/正提示

- **正提示**: 言語に出現するテキストの (無限) 列

$$w_1, w_2, \cdots$$

  - 言語に含まれるテキストはこの列のどこかで必ず出現することが保証される:
    $(\forall w \in L, \exists i \in \mathbb{N}, w_i = w)$ [3]

- これは次に相当する
  - 子供が、大人から常に正しいテキストを一つずつ聞く

[3] 謎の操作 $content$ を使って次のようにも書く:

$$content(w_1, w_2, \cdots) = L$$

情報提示/正提示

- 正提示: 言語に出現するテキストの (無限) 列

$$w_1, w_2, \cdots$$

  - 言語に含まれるテキストはこの列のどこかで必ず出現することが保証される:
    $(\forall w \in L, \exists i \in \mathbb{N}, w_i = w)$ [3]

- これは次に相当する
  - 子供が、大人から常に正しいテキストを一つずつ聞く

[3] 謎の操作 $content$ を使って次のようにも書く:

$$content(w_1, w_2, \cdots) = L$$

情報提示/完全提示

- **完全提示**: 何でもいいから任意のテキストの列で、そのテキストが言語に含まれるかのラベル (Bool) が付いている

$$(x_1, I_1), (x_2, I_2), \cdots$$

- $I_i \wedge (x_i \in L)$ or $\neg I_i \wedge (x_i \notin L)$
- ただし $content(x_1, x_2, \cdots) = \Sigma^*$
- 次に相当
  - 子供が大人から常に正しいテキストを聞く $(I_i = true)$
  - 子供はたまに作文をする (話す)
    - 文法エラーをすると、リアクションから誤りだと知る $(I_i = false)$

情報提示/完全提示

- 完全提示: 何でもいいから任意のテキストの列で、そのテキストが言語に含まれるかのラベル (Bool) が付いている

$$(x_1, I_1), (x_2, I_2), \cdots$$

- $I_i \wedge (x_i \in L)$ or $\neg I_i \wedge (x_i \notin L)$
- ただし $content(x_1, x_2, \cdots) = \Sigma^*$

- 次に相当
  - 子供が大人から常に正しいテキストを聞く $(I_i = true)$
  - 子供はたまに作文をする (話す)
    - 文法エラーをすると、リアクションから誤りだと知る $(I_i = false)$

言語学習モデル (guessing machine)

1. どちらかの提示により、学習者は $i_1, i_2, \cdots$ を逐次的に受け取る
2. 学習者は謎の手続き G によって言語を推論する

時刻 t における推論:

$$g_t = G(i_1, i_2, \cdots, i_t)$$

g_t の型は特定の一つの言語を説明するもの、或いは、言語そのもの $(L(g_t) \in C$ 或いは $g_t \in C)$

極限における同定

原理的に有限個の情報からは学習できない、という言語クラスはいくらでもあるので極限を考える

$$g_t = G(i_1, i_2, \cdots, i_t)$$

が極限 $t \to \infty$ で同定してかつ正しい言語を指し示すことを「極限同定」 という

- N.B. 学習者は「いつ自分は正しく言語を学習できたか」を知る必要はないし、知る術はない

言語クラスの学習可能性

Gold が提示した枠組みにおける「学習可能性」とは次のこと

- 「ある言語クラスが学習可能」を示すためには、
  - ある guessing machine を構成して、
  - その推論結果が常に極限で同定できていることを言えば良い

完全提示から正規言語は学習可能

- 言語からオートマトン (g_t) を復元する
- 言語から正規表現を得られる
  - 但し正例と負例を無限に作る必要がある
    - 例えば、任意のテキストについて、言語に含まれるかどうか判断する神託
    - あるいは手元にあるオートマトンと同値なオートマトンを作りたい?

libalf: Automata Learning Framework

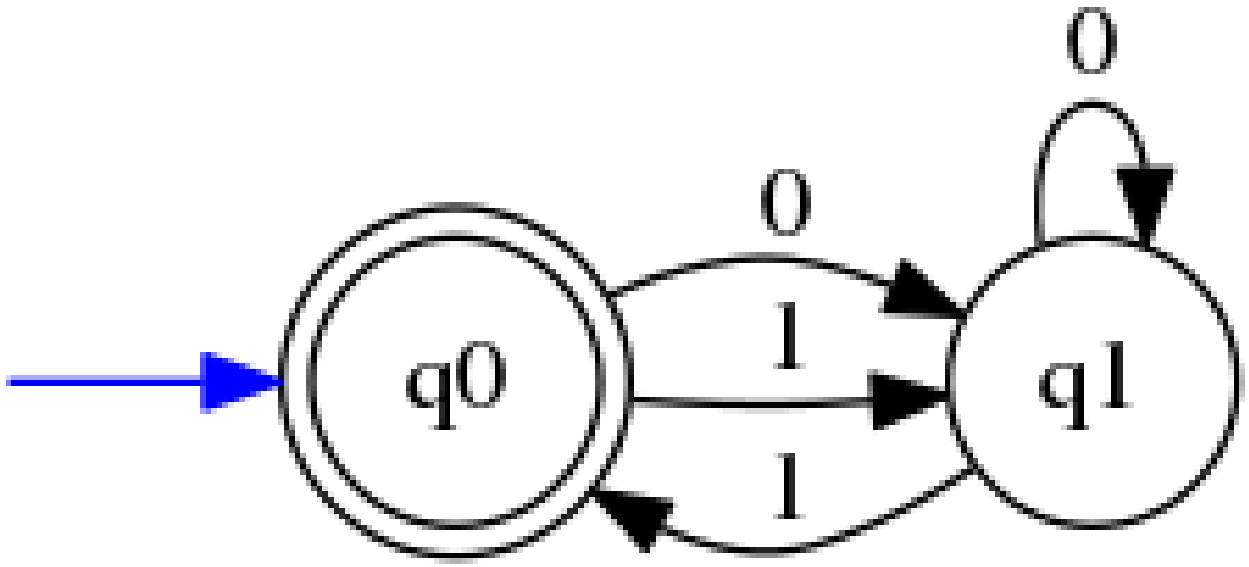
オートマトンを復元してくれるライブラリ:
libalf (libalf.informatik.rwth-aachen.de) **とかゆう実装がある**

Algorithm	offline	online	target model
Angluin's L*		X	DFA
L* (adding counter-examples to columns)		X	DFA
Kearns / Vazirani		X	DFA
Rivest / Schapire		X	DFA
NL*		X	NFA
Regular positive negative inference (RPNI)	X		DFA
DeLeTe2	X		NFA
Biermann & Feldman's algorithm	X		NFA
Biermann & Feldman's algorithm (using SAT-solving)	X		DFA

libalf/demo/online

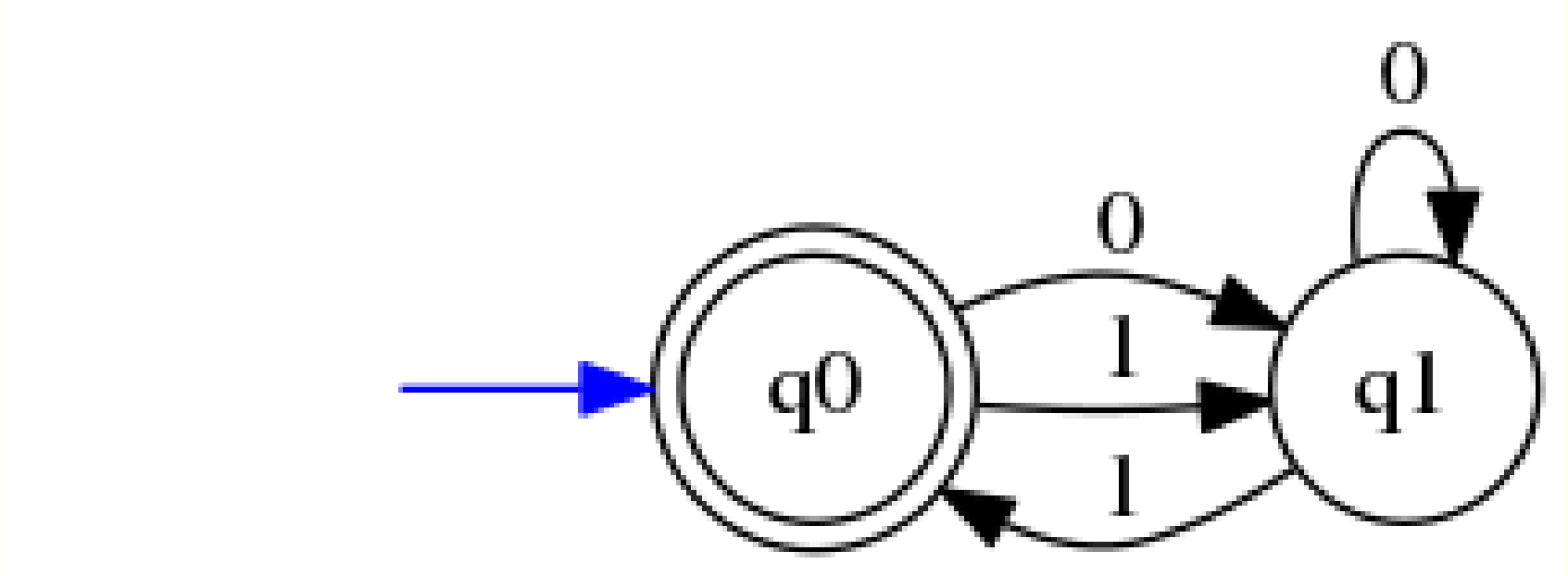
```
Please insert the alphabet size (between 1 and 10): 2
Please classify the word '' (0/1): 1
Please classify the word '0' (0/1): 0
Please classify the word '1' (0/1): 0
Please classify the word '00' (0/1): 0
Please classify the word '01' (0/1): 1

Conjecture:

digraph finite_automaton {
        graph[fontsize=8]
        rankdir=LR;
        size=8;

        node [shape=doublecircle, style="", color=black]; q0;
        node [shape=circle, style="", color=black]; q1;
        node [shape=plaintext, label="", style=""]; iq0;
        iq0 -> q0 [color=blue];
        q0 -> q1 [label="0"];
        q0 -> q1 [label="1"];
        q1 -> q1 [label="0"];
        q1 -> q0 [label="1"];
};
Please specify whether the conjecture is equivalent (y/n): ^[[D
```

out



僕「0^n1^n を受理させたい!!」
上は 11 を受理している! おかしい!!

out

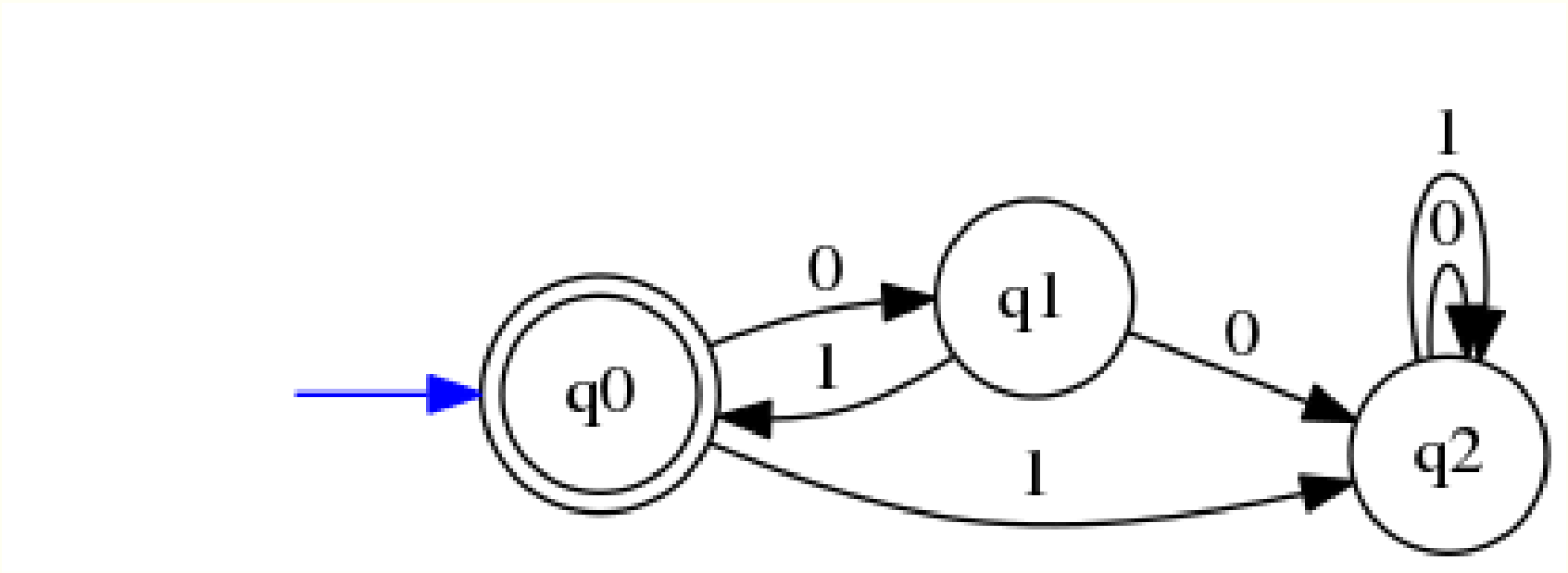


僕「0^n1^n を受理させたい!!」
上は 11 を受理している! おかしい!!

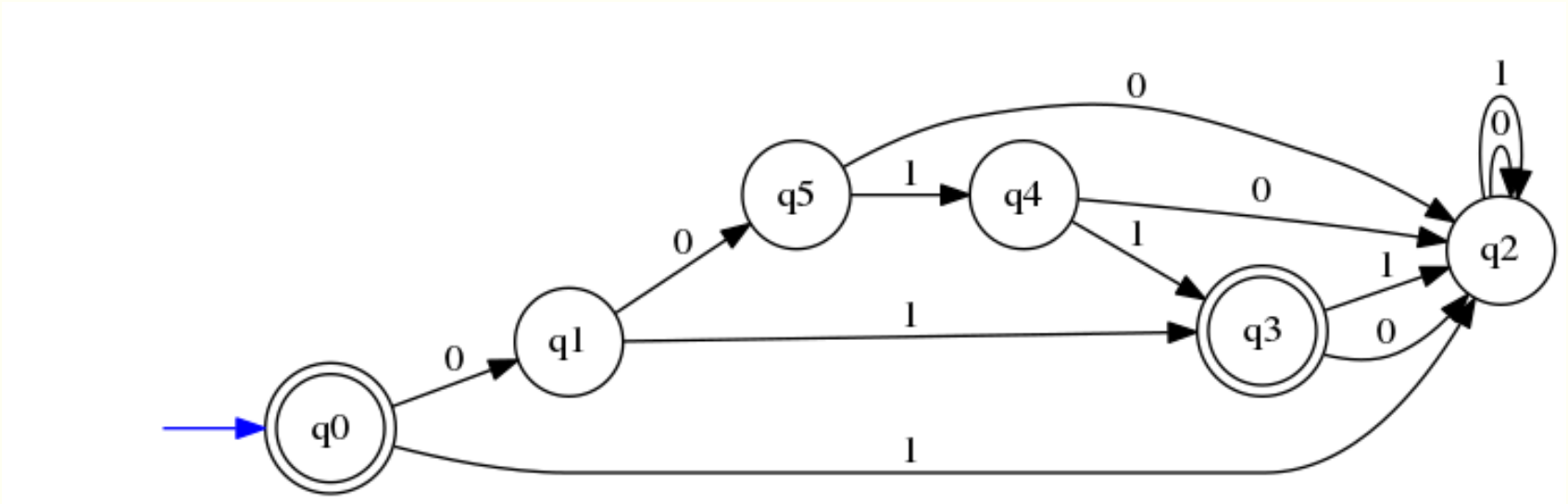
cont

```
Please specify whether the conjecture is equivalent (y/n): n
Please enter a counter example: 11
Please classify the word '11' (0/1): 0
Please classify the word '110' (0/1): 0
Please classify the word '111' (0/1): 0
Please classify the word '10' (0/1): 0
Please classify the word '001' (0/1): 0
Please classify the word '011' (0/1): 0
Please classify the word '1101' (0/1): 0
Please classify the word '1111' (0/1): 0
Please classify the word '101' (0/1): 0
```

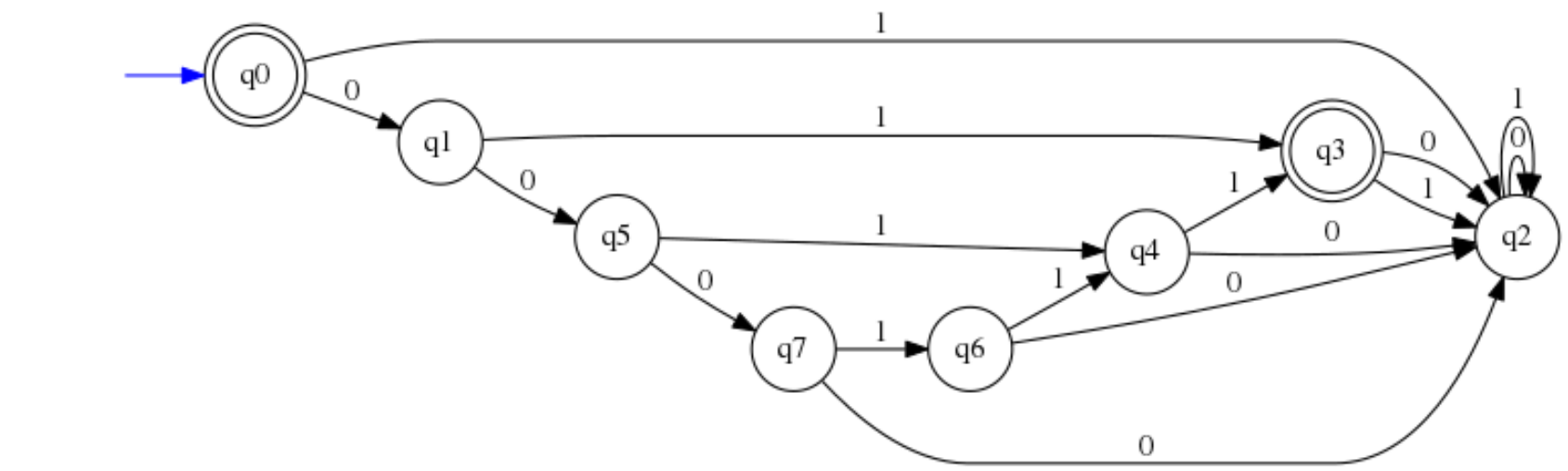
out



out

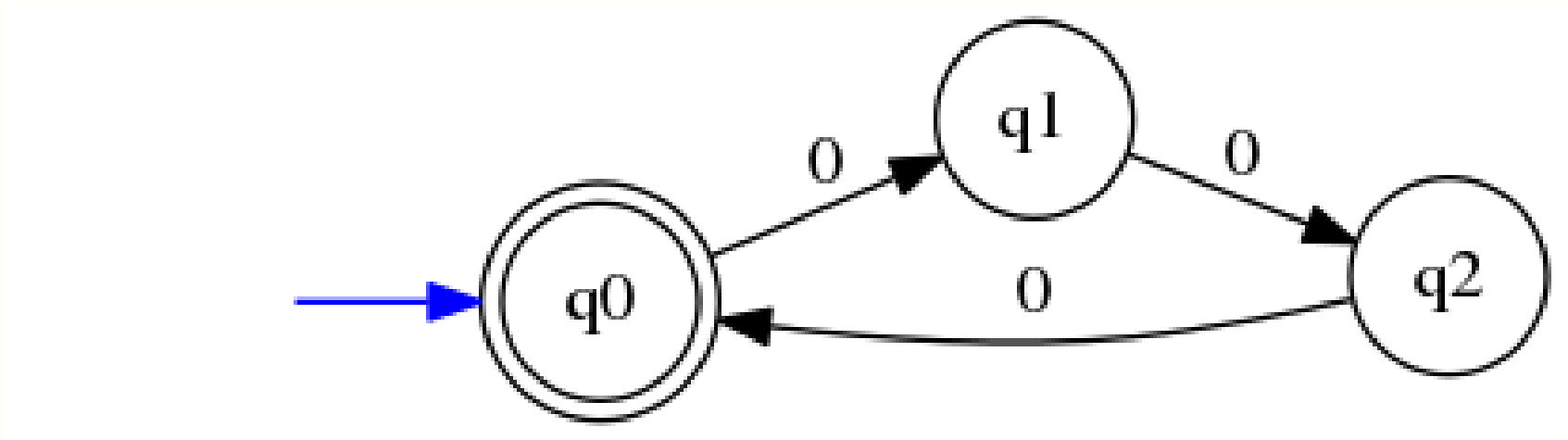


out



0^n1^n は正規言語じゃない?!?!?

/(000)*/



$\epsilon, 0, 00, 000, 0000, 00000$ を神託した時点で上を出力

正提示から正規言語を極限同定することは不可能

- **正規言語全体は大きすぎる (超有限) ため**
- **どの時刻 $t \in \mathbb{N}$ においても一つの言語に絞ることができない**

Gold **曰く、“正提示は完全提示に較べて弱い”**

子供の言語学習 (acquisition of grammar by children)

ちょいちょい「子供が学習するときは〜」という話が出るが、実際はどうなのか

psycholinguist **曰く** [McNeill, 1966]:
「子供が文法誤りをしたとき、それを指摘することは滅多にない」

- **完全提示ってのはちょっと仮定が強すぎるのでは？**
  - **でも我々は自然言語を学習する**

自然言語は正提示から学習可能説

- 多くの自明な言語クラスなら正提示から学習可能であることを Gold は示した
- 英語は文脈自由文法だと言われているが、実際には、全ての文脈自由文法が自然言語になるわけではない
  - もっと制限がある (もっと小さなクラスである)
  - 例えば、学習可能性の結論として: あり得る自然言語のクラスは、無限言語を少なくともひとつは含み、全ての有限言語を含むことはない (超有限ではない)

子供は、我々がわからない方法で負例を学習する説

- 例えば、発言をして、思ったような反応が得られなかったとき
- だとすると完全提示からの学習をしてもいい
- 原始再帰的言語は完全提示から学習可能であることを Gold は示している
  - 文脈依存文法も原始再帰的言語の一つ
  - 英語は完全提示から学習可能

Dana Angluin

"On the Complexity of Minimum Inference of Regular Sets", 1978
"Finding Patterns Common to a Set of Strings", 1979

Angluin さんのやったこと

正提示から学習可能な非自明な言語クラスの発掘

1. **パターン言語**

パターン: $\mathcal{PAT}$

- 有限アルファベット $\Sigma = \{0, 1, \cdots\}$
- 無限変数 $X = \{x_1, x_2, \cdots\}$

$$\mathcal{PAT} = (\Sigma \cup X)^+$$

- 代入、$\preceq$ (less-general-than)
  - 代入 $[x_i/p]$ は、パターン中に出現する全ての 変数 x_i を 空でない パターン p に置き換える操作

$$\begin{aligned} x_1x_1 &\succeq \underline{x_2}\ \underline{x_2} && [x_1/x_2] \\ &\succeq \underline{ax_1}\ \underline{ax_1} && [x_2/ax_1] \\ &\succeq a\underline{bc}a\underline{bc} && [x_1/bc] \end{aligned}$$

$$ax_1ax_1 \not\succeq aa$$

パターン言語

$$L(p) = \{w \in \Sigma^* | w \preceq p\}$$

例えば、

$$L(x_1 x_1) = \{ww | w \in \Sigma^+\}$$

パターン言語は正提示から学習可能

パターン言語のクラス

$$C = \{L(p) | p \in \mathcal{PAT}\}$$

学習過程 (推論機械 G)

- $L = L(p) \in C$ を学習したい
- 正提示 $s_1, s_2, \cdots (s_i \in L)$ を受け取る

推論機械はパターンを出力する:

$$p_t = G(s_1 \cdots s_t)$$

大雑把にいうと

- $\forall t, content(s_1, \cdots, s_t) \subseteq L(p_t)$ (無矛盾)
- $L(p_1) \subseteq L(p_2) \subseteq \cdots$ (保守的)
- 上 2 つを満たす為には次のようにすれば十分 (極小言語戦略)
  - $p_t = \arg\min_p L(p)\ s.t.\ content \subseteq L(p)$

ってやると、

$$L = L(\lim_t p_t)$$

になる.

学習過程 (推論機械 G)

- $L = L(p) \in C$ を学習したい
- 正提示 $s_1, s_2, \cdots (s_i \in L)$ を受け取る

推論機械はパターンを出力する:

$$p_t = G(s_1 \cdots s_t)$$

大雑把にいうと

- $\forall t, content(s_1, \cdots, s_t) \subseteq L(p_t)$ (無矛盾)
- $L(p_1) \subseteq L(p_2) \subseteq \cdots$ (保守的)
- 上 2 つを満たす為には次のようにすれば十分 (極小言語戦略)
  - $p_t = \arg\min_p L(p)\ s.t.\ content \subseteq L(p)$

ってやると、

$$L = L(\lim_t p_t)$$

になる.

皆が興味がなさそうだったら飛ばすページ

Prop. 有限の厚みを持つ言語クラスは極小言語によって極限同定可能である

言語クラスが有限の厚みを持つとは $\Longleftrightarrow$ 任意のテキストの有限集合 S について $\{L | S \subseteq L\}$ が有限であること

1. 無矛盾かつ保守的な推論による推論の列: $g_1, g_2, \ldots$ は収束する
   - 1.1 有限の厚みより、正提示の 1 つ目を含む言語は有限個しかない
     - 1.1.1 推論も有限個しかない
   - 1.2 保守性より $L(g_1) \subseteq L(g_2) \subseteq \cdots$ (g_i に半順序がつく)
   - 1.3 推論列はどこかで停まるか、極大を定める

パターン言語は正提示から学習可能であることの証明

1. 先に挙げた極小言語戦略による推論機械を構成する
2. パターン言語クラスは有限の厚みを持つことを示す
   - $\forall w, \{p | w \preceq p\}$ が有限であることを言えばよい
   - hint: $q \preceq p \Rightarrow |q| \geq |p|$
3. 先の Prop. から極限同定可能

Shinohara: “Polynomial Time Inference of Extended Regular Pattern Languages”, 1991

パターン言語を消去可能パターン言語に拡張

- **空の代入を許す**:

$$ax_1ax_1 \succeq aa$$

消去可能パターン言語であって正則なものは、正提示から学習可能であることを示した

- N.B. **パターン言語が正則** $\iff$ **一つのパターンに出現する同じ変数** x_i **は高々一つ (出現する変数が全て異なる)**

使い道

```
$
Author:   Angluin, D.
Title:    Inductive Inference of Formal Languages from Positive
          Data
Journal:  Inform. Contr. 45
Year:     1980
$
Author:   Maier, D.
Title:    The Complexity of Some Problems on Subsequences and
          Supersequences
Journal:  JACM 25
Year:     1978
$
...;

 Author:   <w>Title:    <x>Journal:  <y>Year:     <z>,
```

Arimura, Shinohara: “Finding Minimal Generalizations of Unions of Pattern Languages and . . . ”, 1994

パターンの和言語も正提示から学習可能であることを示した

- **大体一つのパターンの言語なんて単純すぎる**
- $L(p_1, \cdots, p_k) = L(p_1) \cup \cdots \cup L(p_k)$
  - **ただしパターンの数に上限 k を設定する**

Takeuchi, Sato: “誤情報を含む正則パターン言語の多項式時間推論”, 1998

タイトルの通り

- 学習したい言語からちょっとくらいズレた提示があっても良いようにする
- 言語の「近傍系」を学習していく
  - パターンの上の一種のハミング距離を取る

(一定の) 誤情報を含む正提示から学習可能

Ng, Shinohara: "Inferring Unions of the Pattern Languages by the Most Fitting Covers", 2005

- 今見てきた極小言語戦略は提示 $s_1, \cdots, s_t$ を含む中で極小な言語を推論として出力する
  - 極小とは言語の包含関係に関する
  - 言語は無限集合なので大きさを比較するのは難しい
- 要するに小さければ良い
  - 実際の言語ではテキストの長さには上限が (たぶん) ある
    - $\Rightarrow$ 任意の自然言語は有限集合である
    - 大きさに関する極小を取ることができる

性質を調べたものであって、この場合の推論方法を示したわけでも学習可能性を言ったものでもない

おわりに

- 誰も自然言語のことを忘れてパターン言語だけをやっている
- 完全提示からの学習なんて誰もやってない (ことはないけど)
- というか確率を導入すべきだ
- もはや誰も手をつけない分野
  - ちょっと古ければ論文は大量にある
  - 2000 年以降もたま〜にぽつぽつ出てる