Spectral Methods in our SPiCe'16 Submission

Farhana Ferdousi Liza and Marek Grzes
School of Computing, University of Kent, UK

University of Kent

Our Team

Hidden Markov Models

Natural Language Processing

Spectral Learning

Our Score

The highest score among methods that did not use Neural Networks

Position	Team Name	Score
1	shib	10.5036139488
2	ushitora	10.4169257879
3	ToBeWhatYouWhatToBe	9.9987525940
4	Markov_s_Principle	9.4928018749
5	vha	9.4351277351
6	ZZZZZZZZ	9.2227905095
7	uwtacoma	9.1279414594
8	Ping	9.0090818107
9	Rafael-UoL	7.5950857401
10	MarlonTree	6.5513240695
11	codeBlue	4.7569596767
12	JGR	3.2075991929
13	Hunter	2.7801739872
14	dolboeb	2.4963775873
15	ValarMorghulis	2.4496953487
16	TubularBell	2.1567180753
17	TeamEigen	1.7499862611

Initial Attempts

Spectral learning for HMMs (Hsu et al. 2012)

Observable Operator Model for HMMs

$$Pr(x_1, \ldots, x_t) = 1_m^\top A_{x_t} \ldots A_{x_1} \pi.$$

Empirical moment calculation:

$$\begin{aligned} P_1 &\in \mathbb{R}^n & [P_1]_i &= Pr(x_1 = i) \\ P_{2,1} &\in \mathbb{R}^{n \times n} & [P_{2,1}]_{ij} &= Pr(x_2 = i, x_1 = j) \\ P_{3,x,1} &\in \mathbb{R}^{n \times n} & [P_{3,x,1}]_{ij} &= Pr(x_3 = i, x_2 = x, x_1 = j). \end{aligned}$$

$P_{2,1} = U\Sigma V^*$

U defines an m-dimensional subspace that preserves the state dynamics.

Transformed operators for HMMs

$$\hat{b}_1 = \hat{U}^\top \hat{P}_1;\ \hat{b}_\infty = (\hat{P}_{2,1}^\top \hat{U})^+ \hat{P}_1;\ \hat{B}_x = \hat{U}^\top \hat{P}_{3,x,1} (\hat{U}^\top \hat{P}_{2,1})^+ \ \ \forall_x \in [n]$$

$$\hat{Pr}(x_1, \ldots, x_t) = b_\infty^\top B_{x_t} \ldots B_{x_1} b_1$$

$$\lim_{N \to \infty} \sum_{x_1, \ldots, x_t} |Pr(x_1, \ldots, x_t) - \hat{Pr}(x_1, \ldots, x_t)| = 0.$$

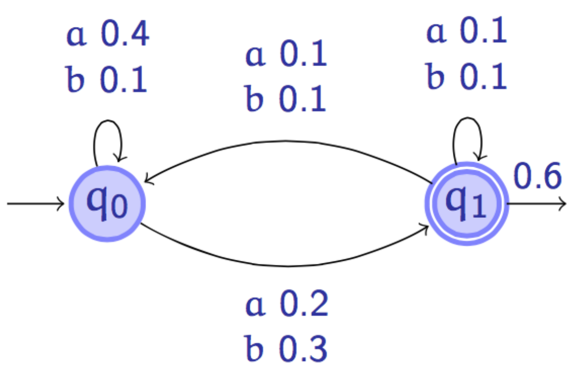
The Main Parameters of the Method

- The number of hidden states

Main Methods

Weighted Finite Automata and Sequence Prediction

Example with 2 states and alphabet $\Sigma = \{a, b\}$



Operator Representation

$$\boldsymbol{\alpha}_0 = \begin{bmatrix} 1.0 \\ 0.0 \end{bmatrix}$$

$$\boldsymbol{\alpha}_\infty = \begin{bmatrix} 0.0 \\ 0.6 \end{bmatrix}$$

$$\mathbf{A}_a = \begin{bmatrix} 0.4 & 0.2 \\ 0.1 & 0.1 \end{bmatrix}$$

$$\mathbf{A}_b = \begin{bmatrix} 0.1 & 0.3 \\ 0.1 & 0.1 \end{bmatrix}$$

$$f(ab) = 0.4 \times 0.3 \times 0.6 + 0.2 \times 0.1 \times 0.6 = 0.084$$

Balle et. al. (EMNLP 2014)

Hankel Matrix

$$S = \left\{ \begin{array}{c} aa,\ b,\ bab,\ a, \\ bbab,\ abb,\ babba,\ abbb, \\ ab,\ a,\ aabba,\ baa, \\ abbab,\ baba,\ bb,\ a \end{array} \right\} \longrightarrow \hat{\mathbf{H}} = \begin{array}{c} \\ \lambda \\ a \\ b \\ ba \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \left[\begin{array}{cc} 1.31 & 1.56 \\ .19 & .62 \\ .56 & .50 \\ .06 & .31 \end{array} \right] \end{array}$$

Balle et al. (EMNLP 2014)

The Basis

$\mathcal{S} = \{a, b\}$, $\mathcal{P} = \{a, b\}$

$\Sigma^\star$	λ	a	b	aa	ab	...
λ	1	0.3	0.7	0.05	0.25	...
a	0.3	0.05	0.25	0.02	0.03	...
b	0.7	0.6	0.1	0.03	0.2	...
aa	0.05	0.02	0.03	0.017	0.003	...
ab	0.25	0.23	0.02	0.11	0.12	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$\mathbf{h}_{\lambda,\mathcal{S}}$, $\mathbf{h}_{\mathcal{P},\lambda}$, $\mathbf{H}$, $\mathbf{H}_a$

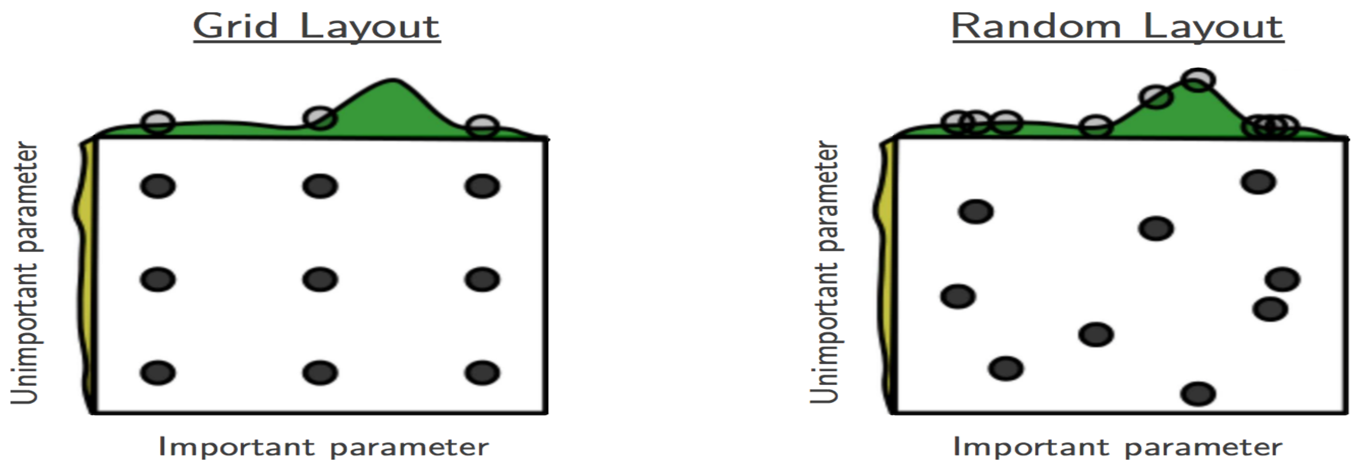
Balle et al. (EMNLP 2014)

The Main Parameters of the Method

- The number of hidden states
- The basis
  - The basis can be chosen from a sub-block of the Hankel matrix where the rows and columns correspond to the substrings and the cells correspond to the frequencies of the substrings in the data.
  - Therefore, the maximum length of the substrings can be considered as a parameter

Parameter Tuning

- A combination of (manual) coordinate ascent and random search
- Why random search?



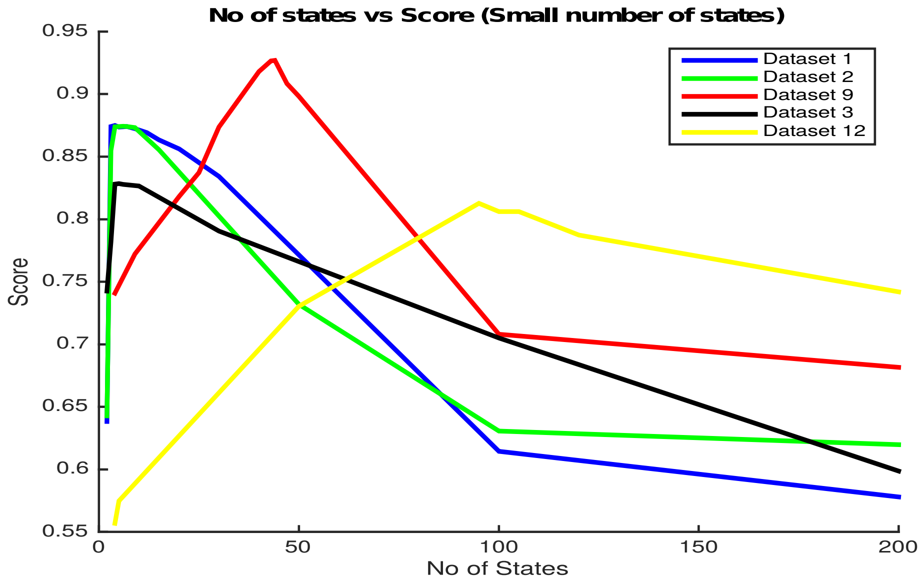
(BERGSTRA AND BENGIO (2012))

Other Methods

- 3-gram model with smoothing worked better than spectral learning on 3 problems

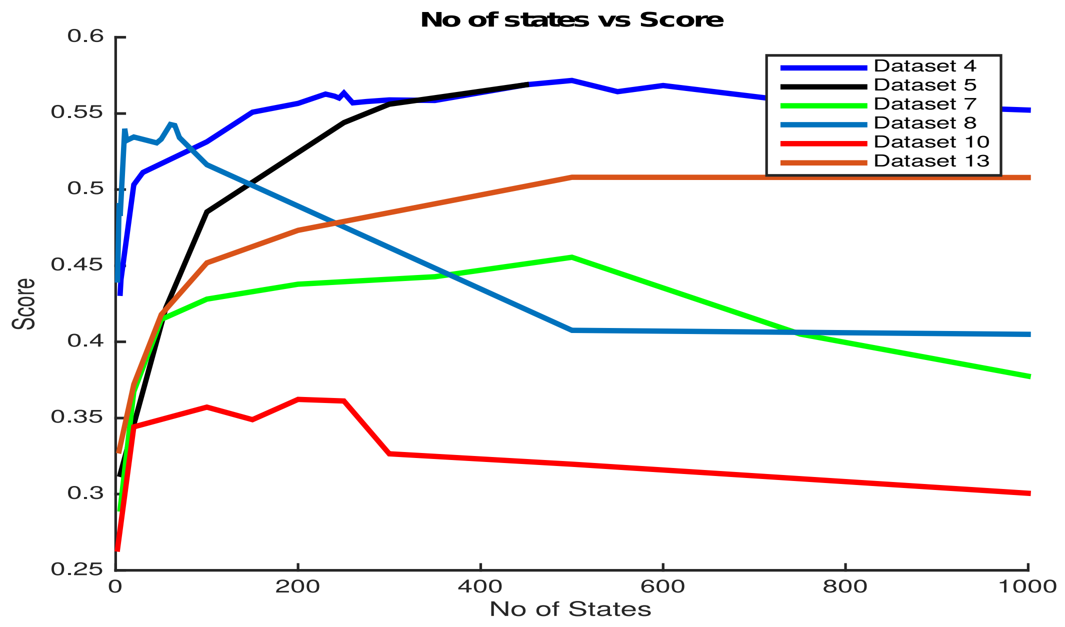
Experimental results (1)

- The Spectral Method did well on problems 1, 2, 3, 9,12
- Presumably, those problems have small numbers of hidden states



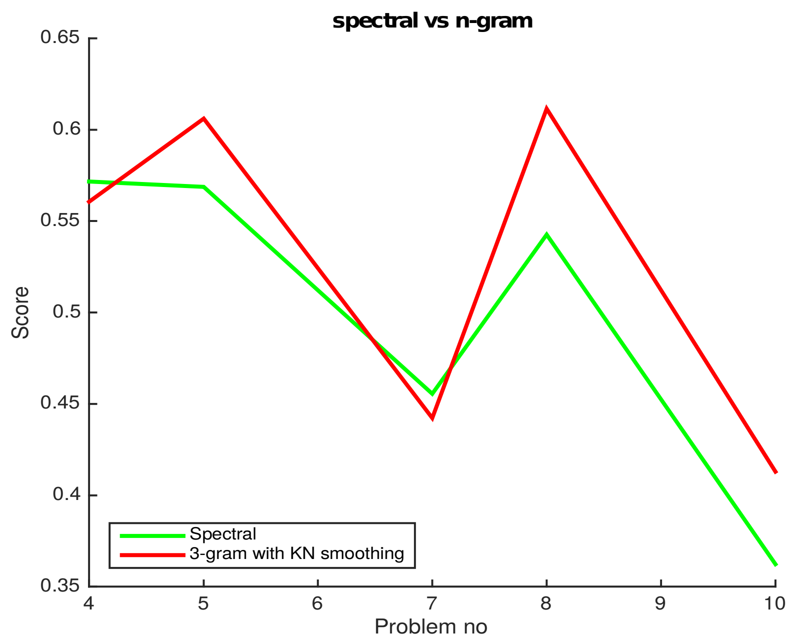
Experimental result (2)

- Score prediction is invariant to changes in the number of states on problems 4, 5, 7, 8,10,11,13



Experimental result (3)

- On problems 5, 8 and 10, 3-gram with smoothing gave slightly batter results than the corresponding spectral approach



The Final Parameter Values for WFA

Problem No	Rank	lrows	lcolumns	Score
1	4	5	5	**0.8789916635**
2	6	5	5	**0.8731489778**
3	5	10	3	**0.8248148561**
4	500	5	5	**0.5272911191**
5	3-gram with Kneser–Ney smoothing			**0.6142422557**
6	300	6	7	**0.8096061349**
7	500	4	4	**0.4474728703**
8	3-gram with Kneser–Ney smoothing			**0.6235375404**
9	57	8	7	**0.9324635267**
10	3-gram with Kneser–Ney smoothing			**0.3965168893**
11	100	5	5	**0.4147772193**
12	95	4	4	**0.8113699555**
13	500	5	5	**0.4990697801**
14	2	10	10	**0.4649848044**
15	3	6	6	**0.2899561226**

THE UK'S EUROPEAN UNIVERSITY

www.kent.ac.uk

University of Kent