Acquisition of Language Symbols

Pankaj Prateek (pratikkr@)
Mentor: Dr. Amitabha Mukerjee (amit@)

Dept. of CSE, IIT Kanpur

Previous Work

The baby designer model learns patterns in an apprenticeship situation. When presented with a set of functional constraints and variable set governing them, it explores the design space, using domain-general learning algorithms to discover patterns in the better performing designs. These patterns get transformed to chunks in case they occur frequently. In this process, knowledge of language and labels for such concepts and patterns is not required.

On exposure to language, these implicit associations get transformed to rules in the symbolic space, thus incorporating labels.

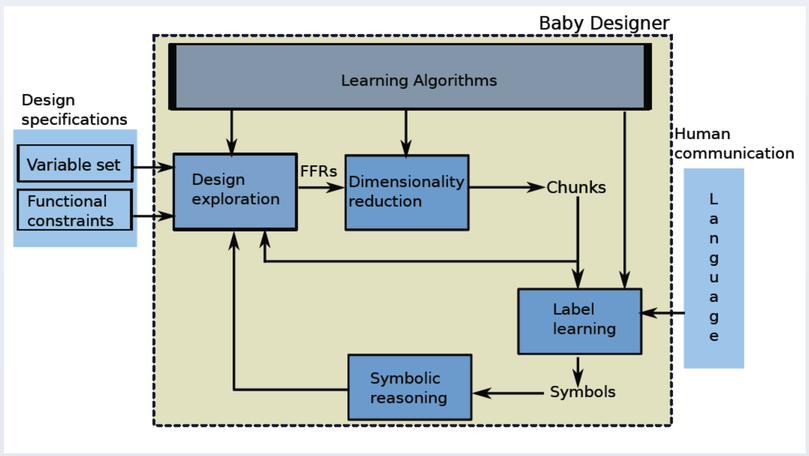
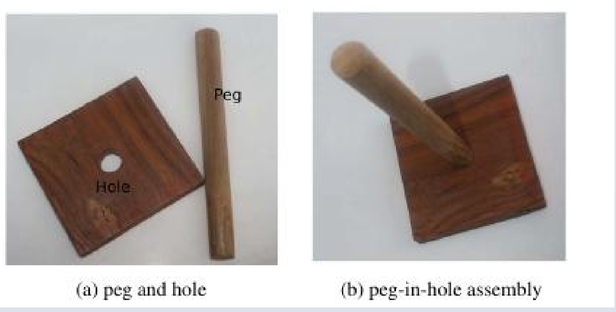


Fig 1: Architecture of the Baby Designer

Due to difference in experience and language, the symbols acquired by different agents differ. In their paper[1], Dr. Mukerjee and Madan Dabbereu have considered how agents map the chunks (low-dimensional characterizations) to language based on human commentary produced in the same context. This was explored by learning labels in English and Telugu in the simple domain of tight and loose fits using the peg-in-hole assembly. This project aims to extend their work to Hindi.



(a) peg and hole (b) peg-in-hole assembly

Fig 2: Peg-in-hole Assembly

Apparatus

A wooden block with 5 Holes (A (tapering), B, C, D, E) and 5 aluminium/steel pegs (1,2,3,4,5) of diameters (9mm, 12.25mm, 15.75mm, 19mm, 25mm).

Loose-fit situations – A:1, C:5, D:3

Tight-fit situations – A:2, B:4, E:3

Methodology

Experiment 1:State

This experiment focused at collecting spoken Hindi data in situations when the peg is already inserted in the hole. The pre-assembled setup was kept on a table and the participants were asked not to lift the assembly from the table.

Each participant was provided with the following instruction:

“यह एक छेद है और यह एक पैग है. यह पैग पहले से इस छेद में डाली हुई है. आपको इन दोनों के बीच हो रहे संपर्क का हिन्दी में वर्णन करना है. यह ध्यान रखिये कि यह ब्लॉक इस टेबल से उठे नही और जितना हो सके हिन्दी के अलावा किसी और भाषा का प्रयोग ना करिये.”

Experiment 2: Action

This experiment focused at collecting spoken Hindi data in situations where the participant is asked to actively insert the peg in the hole. There was no restriction as in the previous experiment.

Each participant was provided with the following instruction:

“यह एक छेद है और यह एक पैग है. आपको यह पैग इस छेद में डालनी है और इन दोनों के बीच हो रहे संपर्क का हिन्दी में वर्णन करना है. जितना हो सके हिन्दी के अलावा किसी और भाषा का प्रयोग ना करिये.”

Associating Linguistic Lables

The association of a word w_i with a concept C_j can be measured using conditional probability. But the direction of association could be either p(C/w) or p(w/C).

Since only two concepts are involved:

$$\frac{p(C_T/w)}{p(C_L/w)} = \frac{p\left(\frac{w}{C_T}\right).p(C_T)}{p\left(\frac{w}{C_L}\right).p(C_L)}$$

The number of instances of C_L and C_T are almost same in the set, so the direction of association doesn’t matter.

For strongest association with C_T, compute $\max_i\{\frac{p\left(\frac{W_i}{C_T}\right).p(C_T)}{p\left(\frac{W_i}{C_L}\right).p(C_L)}\}$

Results

Words like “tight” and “loose” are readily associated with [tight] and [loose] fits, respectively, based on uninformed word association from Hindi Language. Thus the words have top associations with the concepts.

Unlike Telugu[1], Hindi is richly inflected, it incorporates foreign words very easily into the language. This can be seen from the fact that the Hindi native correspondences for “tight”, viz. , “तंग” and “कसा” were used only occasionally by the participants even after they were specifically told to adhere to Hindi in their narrations. This was the same with Hindi speaking population which had very less contact with English language.

State Profiled: [loose] corpus

Term	f_T	$p(\frac{w}{C_T})$	f_L	$p(\frac{w}{C_L})$	$f_{T,L}$	$p(w)$	$\frac{p(\frac{w}{C_T})}{p(\frac{w}{C_L})}$
Without Stemming							
loose	1	0.00190	4	0.00877	5	0.00509	4.61403
घूमे	1	0.00190	4	0.00877	5	0.00509	4.61403
ताकत	2	0.00380	7	0.01535	9	0.00916	4.03728
comparison	1	0.00190	3	0.00657	4	0.00407	3.46052
peg	1	0.00190	3	0.00657	4	0.00407	3.46052
size	1	0.00190	3	0.00657	4	0.00407	3.46052
इसमें	2	0.00380	6	0.01315	8	0.00814	3.46052
कम	1	0.00190	3	0.00657	4	0.00407	3.46052
बड़ा	2	0.00380	6	0.01315	8	0.00814	3.46052
दीला	3	0.00570	7	0.01535	10	0.01018	2.69152
With Stemming							
loose	1	0.00190	4	0.00877	5	0.00509	4.61403
comparison	1	0.00190	3	0.00657	4	0.00407	3.46052
peg	1	0.00190	3	0.00657	4	0.00407	3.46052
कम	1	0.00190	3	0.00657	4	0.00407	3.46052
बड़ा	3	0.00570	9	0.01973	12	0.01221	3.46052
दीला	3	0.00570	7	0.01535	10	0.01018	2.69152
पा	3	0.00570	7	0.01535	10	0.01018	2.69152
कारण	1	0.00190	2	0.00438	3	0.00305	2.30701
गैप	1	0.00190	2	0.00438	3	0.00305	2.30701
छेद	2	0.00380	4	0.00877	6	0.00610	2.30701

Table 1: Hindi State Profiled [loose] corpus: Top 10 words by conditional ratio

State Profiled: [tight] corpus

Term	f_T	$p(\frac{w}{C_T})$	f_L	$p(\frac{w}{C_L})$	$f_{T,L}$	$p(w)$	$\frac{p(\frac{w}{C_T})}{p(\frac{w}{C_L})}$
Without Stemming							
tight	9	0.01711	1	0.00219	10	0.01018	7.80228
रही	7	0.01330	1	0.00219	8	0.00814	6.06844
लग	5	0.00950	1	0.00219	6	0.00610	4.33460
महसुस	4	0.00760	1	0.00219	5	0.00509	3.46768
movement	3	0.00570	1	0.00219	4	0.00407	2.60076
इसलिये	3	0.00570	1	0.00219	4	0.00407	2.60076
पर	6	0.01140	2	0.00438	8	0.00814	2.60076
जो	5	0.00950	2	0.00438	7	0.00712	2.16730
यहाँ	5	0.00950	2	0.00438	7	0.00712	2.16730
थोड़ा	16	0.03041	7	0.01535	23	0.023421	1.98153
With Stemming							
tight	9	0.01711	1	0.00219	10	0.01018	7.80228
महसुस	4	0.00760	1	0.00219	5	0.00509	3.46768
movement	3	0.00570	1	0.00219	4	0.00407	2.60076
इसलिये	3	0.00570	1	0.00219	4	0.00407	2.60076
पर	6	0.01140	2	0.00438	8	0.00814	2.60076
जो	5	0.00950	2	0.00438	7	0.00712	2.16730
यहाँ	5	0.00950	2	0.00438	7	0.00712	2.16730
लग	5	0.00950	2	0.00438	7	0.00712	2.16730
थोड़ा	18	0.03422	8	0.01754	26	0.02647	1.95057
pole	6	0.01140	3	0.00657	9	0.00916	1.73384

Table 2: Hindi Action Profiled [tight] corpus: Top 10 words by conditional ratio

Action Profiled: [loose] corpus

Term	f_T	$p(\frac{w}{C_T})$	f_L	$p(\frac{w}{C_L})$	$f_{T,L}$	$p(w)$	$\frac{p(\frac{w}{C_T})}{p(\frac{w}{C_L})}$
Without Stemming							
loose	0	0	10	0.01730	10	0.00754	Large Value
इसका	1	0.00133	3	0.00519	4	0.00301	3.88235
था	1	0.00133	3	0.00519	4	0.00301	3.88235
पे	1	0.00133	3	0.00519	4	0.00301	3.88235
बड़ा	3	0.00401	7	0.01211	10	0.00754	3.01960
जा	4	0.00534	8	0.01384	12	0.00904	2.58823
fit	1	0.00133	2	0.00346	3	0.00226	2.58823
इसलिये	1	0.00133	2	0.00346	3	0.00226	2.58823
इस	1	0.00133	2	0.00346	3	0.00226	2.58823
मतलब	1	0.00133	2	0.00346	3	0.00226	2.58823
With Stemming							
loose	0	0	10	0.01730	10	0.00718	Very Large
पड़ा	1	0.00133	6	0.01038	7	0.00502	7.76470
को	1	0.00133	5	0.00865	6	0.00431	6.47058
दोनो	1	0.00133	5	0.00865	6	0.00431	6.47058
पुरा	1	0.00133	4	0.00692	5	0.00359	5.17647
उस	2	0.00267	6	0.01038	8	0.00574	3.88235
बिल्कुल	1	0.00133	3	0.00519	4	0.00287	3.88235
घर्षन	2	0.00267	5	0.00865	7	0.00502	3.23529
नही	7	0.00935	17	0.02941	24	0.01724	3.14285
fit	1	0.00133	2	0.00346	3	0.00215	2.58823

Table 3: Hindi Action Profiled [loose] corpus: Top 10 words by conditional ratio

Action Profiled: [tight] corpus

Term	f_T	$p(\frac{w}{C_T})$	f_L	$p(\frac{w}{C_L})$	$f_{T,L}$	$p(w)$	$\frac{p(\frac{w}{C_T})}{p(\frac{w}{C_L})}$
Without Stemming							
tight	10	0.01336	1	0.00173	11	0.00829	7.72727
इसको	7	0.00935	1	0.00173	8	0.00603	5.40909
को	5	0.00668	1	0.00173	6	0.00452	3.86363
रही	9	0.01203	2	0.00346	11	0.00829	3.47727
दोनो	4	0.00534	1	0.00173	5	0.00377	3.09090
same	3	0.00401	1	0.00173	4	0.00301	2.31818
गयी	3	0.00401	1	0.00173	4	0.00301	2.31818
थोड़ी	3	0.00401	1	0.00173	4	0.00301	2.31818
बिल्कुल	3	0.00401	1	0.00173	4	0.00301	2.31818
घर्षन	5	0.00668	2	0.00346	7	0.00527	1.93181
With Stemming							
tight	10	0.01336	1	0.00173	11	0.00790	7.72727
same	3	0.00401	1	0.00173	4	0.00287	2.31818
पे	3	0.00401	1	0.00173	4	0.00287	2.31818
size	7	0.00935	3	0.00519	10	0.00718	1.80303
बड़ा	11	0.01470	5	0.00865	16	0.01149	1.70000
ज्यादा	6	0.00802	3	0.00519	9	0.00646	1.54545
equal	2	0.00267	1	0.00173	3	0.00215	1.54545
friction	2	0.00267	1	0.00173	3	0.00215	1.54545
peg	4	0.00534	2	0.00346	6	0.00431	1.54545
इसलिये	2	0.00267	1	0.00173	3	0.00215	1.54545

Table 4: Hindi Action Profiled [tight] corpus: Top 10 words by conditional ratio

References

1. [Madan Dabbeeru, Amitabha Mukherjee]. "Using Symbol Emergence to Discover Multi-Lingual Translations in Design". Proceedings of the ASME 2010 International Design Engineering Technical Conferences \& Computers and Information in Engineering Conference Proceedings of IDETC/DTM 2010, 2010.

2. [Dabbeeru, Madan Mohan and Mukerjee, Amitabha]. "Learning concepts and language for a baby designer". Design Computing and Cognition'10, 2011.

3. [S V P Gopi Srinath, Nikhil Joshi, Prabhat Mudgal, Amitabha Mukerjee]. "Learning grounded semantics of Hindi nouns from video surveillance and user commentary". Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, 2010.

4. [Madan Dabbeeru, Amitabha Mukerjee]. "Computational models of tacit knowledge."

5. [De Saussure, Ferdinand]. "Nature of the linguistic sign". Course In General Linguistics, 1916.