A Semantics-First Approach for Word-learning using Visuo-Linguistic corpus

Nikhil Sudhakar Joshi
(Y9111015)

Supervisor:
Prof. Amitabha Mukerjee

Department of Computer Science
Indian Institute of Technology, Kanpur

14th June,2011

Outline

What does it mean to learn a language?

- Rationalists'
  - Innatism
  - Learning a generative syntax
- Empiricists' view
  - Language comes from usage
  - Learning mapping between language and the world.

Language and thought

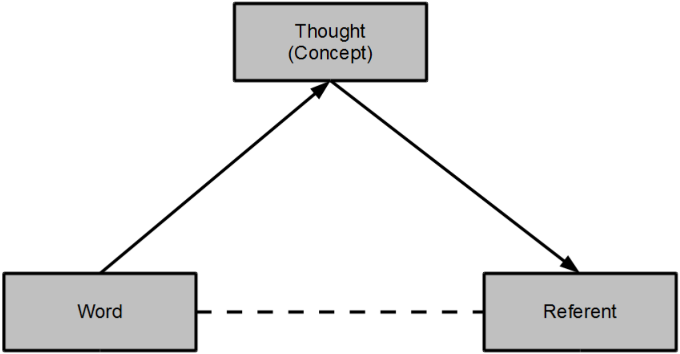


Figure: **Semiotic Triangle**: Odgen and Richards, 1923

Cognitive Grammar

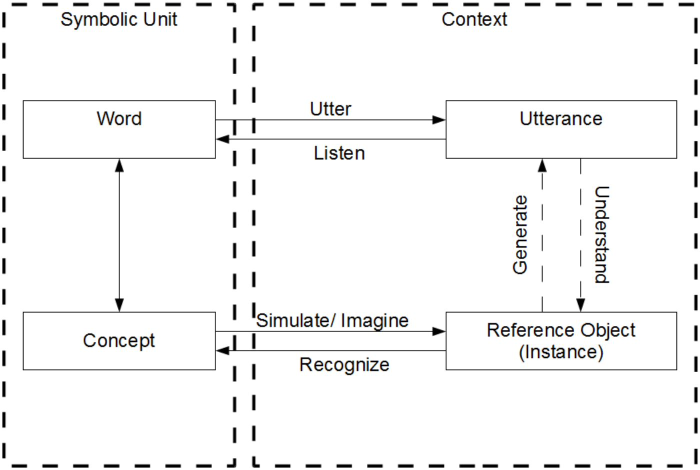


Figure: **Cognitive Grammar**: Processes (Langacker,1987)

What is word-learning?

- What does it mean to learn a word?
- Bootstrapping problem in word learning?
  - Conceptual development affects word-learning
  - Linguistic usage affects conceptual structure
  - How and where to start?
- Symbol Grounding problem (Harnad, 1990)

Approaches to initial word learning

- Nativists (Chomsky)
  - Word-to-semantics mapping is taken to be inborn
- Piagetian (Piagets)
  - Words and their semantics are learnt simultaneously
- Semantics-First (Mandler, Quinn)
  - Based on Preverbal conceptual development
  - Semantics is learnt first before the words

Related Work

- Learning words from pictures (Barnard,2003)
- Learning words from sensorimotor data (Oates,2000)
- Learning visually grounded semantics (Roy-Pentland,2002; Yu-Ballard,2004; Guha-Mukerjee,2008)
- Shortcomings :
  - Scenes with simple objects and descriptions in constrained language
  - Provision for Feedback
  - Predefined or hand-coded semantics
  - Lack of referential ambiguity

Our approach

- Semantics-first approach
- Visuo-Linguistic corpus
  - A complex and real traffic scenario with multiple objects
  - Free and unconstrained narratives with full sentences
- Attention Model
  - Handles referential uncertainty
- Learn linguistic units for concepts learnt before

Symbol Learning Framework

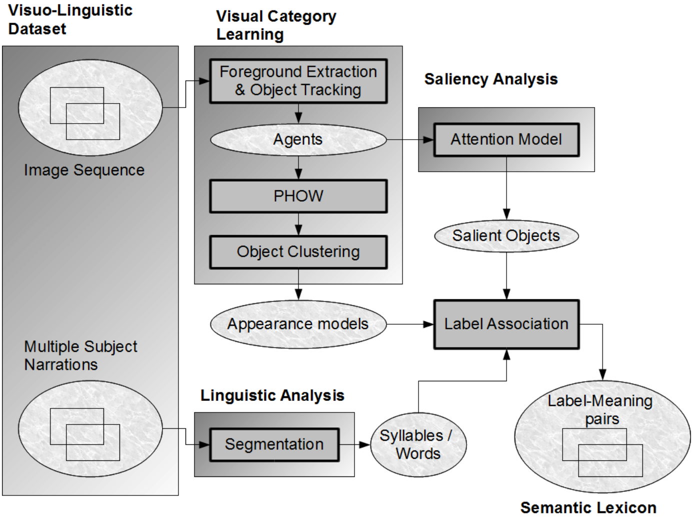


Figure: **Symbol Learning Framework**

Key Assumptions

- Availability of Visuo-Linguistic corpus
- Video shot from static camera (Stable background)
- Shared attention between speaker and listener (theory of mind)
- Linguistic focus follows perceptual focus.

Collecting Narrations

- Collected Narrations in three different ways from 44 subjects.
  - Free Unconstrained Narratives (11 subjects)
  - Feedback based Narratives (20 subjects)
  - Child Directed Narratives (13 subjects)
- Transcribed the narrations into text
- Time-stamped the narrations at sentence boundaries and long pauses.

Discovering visual categories

- Discovering object categories (Gopi, 2010)
  - Foreground Extraction
  - Object Tracking (Guha, 2008)
  - Object Clustering
- Discovering motion categories
  - Modeling trajectory
  - Trajectory clustering

Foreground Extraction

- Learning stable patterns of the background
- Background subtraction

Figure: **Segmented object blobs after background subtraction.**

Object Tracking

- Mean-shift based object tracking
- Considers overlapping sequences of blobs.

Figure: **Agents as sequences of isolated foreground blobs.**

PHOW and Object Clustering

- A code-book of 300 SIFT words
- Blobs are projected as PHOW over 300 SIFT words
- Object models clustered into 30 clusters using *k-means* algorithm

Figure: **Representative views from all agents in cluster C0**

Object clustering:Results

Class: # agents	Clusters	Purity
H:52	C1,C2,C4,C10, C11,C12,C14,C21	51/63 (81%)
M:36	C3,C8,C9,C22, C23,C24,C26	35/48 (73%)
B:32	C5,C6,C7,C15, C20,C28	22/25 (88%)
T:21	C0,C16,C17, C18,C25	15/27 (56%)
L:12	C12,C29	11/13 (83%)
C:16	C19	9/10 (90%)
N:8	C27	2/4 (50%)

Table: *Purity of Clusters from k- means ($k = 30$).*

Learning Trajectories

- Trajectory of agent *a*

$$T_a = \{(t_i, f_i) | i = 1, 2, ..., n\}$$

- Features: Position and velocity $f_i = (x_i, y_i, vx_i, vy_i)$
- 10 points per trajectory.
- 192 agents clustered into 7 clusters using *k-means*

Figure: **Representative Trajectories**

Learning Trajectories:Results

Ground-Truth Cluster	LR	RL	T	C	N	Total	% Purity
C1 (RL)	0	20	0	0	1	21	95
C2 (LR)	15	0	1	0	1	17	88
C3 (LR)	20	0	2	0	1	23	87
C4 (RL)	0	26	8	1	3	38	68
C5 (LR)	21	2	4	8	4	39	54
C6 (LR)	13	8	4	2	7	34	38
C7 (T)	0	3	14	3	0	20	70
Total	69	59	33	14	17	192	

Table: **Ground-Truth distribution of Trajectory clusters**:

Attention Model

- Tries to predict the perceptual focus based on visual saliency
- Two Types of Attention Models
  - Top-Down (Task dependent)
  - Bottom-Up (Task independent)
- Bottom-Up Attention Model based on
  - Blob size
  - Velocity
  - Confidence measure (Singh-Maji, 2006)

Associating language labels

- Align co-occurring sentences with most salient objects
- Segmenting sentences into smaller linguistic units
- Associate Linguistic units with the appropriate concept
- Maximally associated unit is a label for the concept.

Association Measure

- Utterance probability of a linguistic unit l for the speaker s at time t

$$P(l|s,t) = \begin{cases} 1 & \text{if } l \text{ is uttered by s at time t} \\ 0 & \text{otherwise} \end{cases}$$

- Attention Probability

$$P(c|t) = \begin{cases} 1 & \text{if } c \text{ is visually salient at time t} \\ 0 & \text{otherwise} \end{cases}$$

- Joint Probability

$$J(l,c) = \frac{1}{T} * \sum_{t=1}^{T} P(c|t) * P(l|t)$$

where

$$P(l|t) = \frac{1}{|S|} \sum_{s \in S} P(l|s,t)$$

Issues to be addressed

- What should be the linguistic unit of association?
- What should be the association measure?
- Which of the linguistic units should be associated?
- Usefulness and necessity of attention model
- When can we say a word is learnt?

Summary of experimental settings

L	M	T	G	A	V	D
W	DJ	$T+$	$G+$	$A+$	obj	ADULT-1
S	CP	$T-$	$G-$	$A-$	traj	ADULT-2
P_w	MI					CDS
P_s						ALL

Table: **Parameters of experimentation**:

Word-level Association: Results

(W, CP, T+, G+, A+, obj, ALL)						
	$\mathbf{k = 1}$		$\mathbf{k = 2}$		$\mathbf{k = 3}$	
C	***l***	**CP**	***l***	**CP**	***l***	**CP**
T	Tempo	4.46	ek Trak	2.52	dAe.N se bAe.N	1.08
	kAr	4.33	ek Tempo	2.16	Ai Ai TI	0.87
	pe	4.25	ek kAr	1.84	do OTo aur	0.79
B	sAikal	1.95	ek sAikal	1.14	sAikal jA rahI	0.32
	moTarsAikal	0.79	aur sAikal	0.32	gais silinDar le	0.32
	pe	0.63	lefTsAiD	0.32	silinDar le ke	0.32
M	pe	8.60	ek Tempo	4.39	sAmAn le ke	1.45
	bAik	7.12	ek bAik	3.27	aur ek bAik	1.44
	Tempo	6.56	ek OTo	3.19	ek sAikal pe	1.03
L	Trak	17.29	ek Trak	10.67	se ek Trak	1.74
	pe	3.24	tIn sAikalwAle	2.01	ek Trak nikalA	1.47
	sAikal	2.84	Trak gayA	1.76	niilii ra.ng kI	1.25
H	saD.ak	7.50	krOs kar	3.93	krOs kar rahA	3.04
	krOs	6.68	ek Tempo	2.68	roD krOs kar	1.46
	roD	6.54	roD krOs	2.52	lAl sharT me.n	1.16
C	kAr	7.76	ek kAr	4.89	bAe.N se dAe.N	1.41
	gADI	3.99	ek gADI	2.31	kAr jA rahI	1.12
	nikalii	2.81	krOs kar	1.44	krOs kar rahA	1.11

Table: **Word-level Associations**

Syllabic-level Association

- Words in an utterance merged across word boundaries.
- Identify syllables in continuous utterance
  - Syllable as vowel terminated string.
- Associate Syllabic k-grams with visual categories.

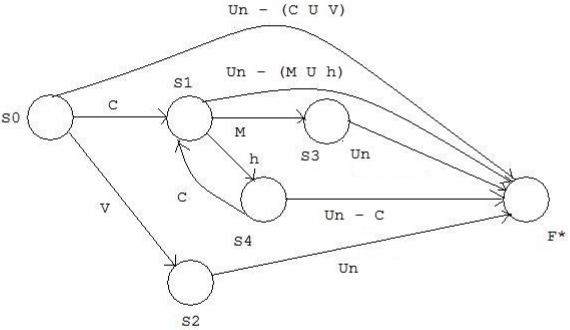


Figure: **FSM**: To identify syllabic units

Syllabic-level Association: Results

	$k = 2$		$k = 3$		$k = 4$	
Concept	***l***	**CP**	***l***	**CP**	***l***	**CP**
TEMPO	ik	12.23	taraf	6.81	sAikal	5.79
	jAr	9.23	rek	6.45	aurek	5.22
	kal	8.76	sAik	6.32	jArahAhai	4.52
BICYCLE	ik	3.4	sAik	3.06	sAikal	2.9
	sAi	3.06	ikal	2.9	eksAi	1.3
	kal	2.91	eksA	1.3	ksAik	1.3
MOTORCYCLE	ik	19.09	sAik	10.54	sAikal	9.24
	jAr	13.43	ikal	9.24	rsAik	7.21
	sAi	12.41	bAik	8.88	jArahAhai	6.02
TRUCK	Trak	19.23	ekTra	11.83	ekTrak	11.83
	kTra	11.83	kTrak	11.83	sAikal	6.41
	jAr	10.2	rahehai.n	8.61	jArahAhai	4.67
HUMAN	hlhai	14.37	saD.ak	7.66	sAikal	6.35
	jAr	10.86	aAdamI	7.62	jArahAhai	6.29
	kal	10.78	taraf	7.26	ekaAd	5.2
CAR	kkA	5.32	ekkA	5.32	ekkAr	5.15
	jAr	4.51	kkAr	5.15	ekaur	2.63
	hlhai	4.33	rahlhai	4.33	jArahlhai	2.46

Table: **Poly-syllabic Associations**:

Phrase-level Association

- Fragment
  - k-gram l_k is a fragment with respect to an n-gram l_n $(n > k)$ for a category c if l_k is contained in l_n and $\frac{M(l_n, c)}{M(l_k, c)} > \tau$
- Independent Unit
  - A smaller k-gram l_k is independent of a higher n-gram l_n w.r.t. a concept c if l_k is not a *fragment* of l_n.
- Unit Independence Conjecture
  - Only those smaller k-grams l_k which are independent of all higher n-grams l_n w.r.t a concept c can be labels for c.

Phrase-level Association: Results

(P_w, CP / MI , T+, G+, A+, obj, ALL)				
Concept (c)	**CP**		**MI**	
	l	$M(l, c)$	l	$M(l, c)$
TEMPO	Tempo	4.46	kAr	7.41
	kAr	4.33	bAik	7.34
	pe	4.25	Tempo	6.54
BICYCLE	sAikal	1.95	sAikal	1.34
	ek sAikal	1.14	ek sAikal	0.96
	moTarsAikal	0.79	gais silinDar	0.53
MOTORCYCLE	pe	8.60	pe	12.88
	bAik	7.12	bAik	11.64
	Tempo	6.56	skUTar	8.99
TRUCK	Trak	17.29	Trak	15.01
	ek Trak	10.67	ek Trak	9.91
	pe	3.24	tIn sAikalwAle	2.37
HUMAN	saD.ak	7.50	saD.ak	27.90
	krOs	6.68	krOs	20.76
	roD	6.54	roD	18.19
CAR	kAr	7.76	kAr	9.30
	ek kAr	4.89	ek kAr	6.61
	gADI	3.99	gADI	4.38

Table: **Phrase-level Associations**: Top3 word k-grams ($k = 1$ to 4

Phrase-level Association: Results

(P_s, CP / MI, T+, G+, A+, obj, ALL)				
Concept (c)	CP		MI	
	l	$M(l, c)$	l	$M(l, c)$
TEMPO	ik	12.23	ik	29.68
	jAr	9.23	jAr	21.35
	kal	8.76	kal	19.70
BICYCLE	sAikal	2.90	sAikal	2.81
	jAr	1.62	eksAi	1.60
	eksAi	1.30	ksAik	1.60
MOTORCYCLE	ik	19.09	ik	39.35
	D	15.08	D	28.61
	jAr	13.43	Tar	26.42
TRUCK	Trak	19.23	Trak	22.55
	ekTrak	11.83	ekTrak	14.70
	jAr	10.20	jAr	9.42
HUMAN	hAhai	14.37	hAhai	62.35
	D	13.85	D	53.54
	jAr	10.86	wAlA	46.14
CAR	ekkAr	5.15	ekkAr	9.38
	jAr	4.51	gADI	6.12
	rahlhai	4.33	rahlhai	5.05

Table: **Syllabic phrase level Association**

Different Association Measures

- Dominance Weighted Joint Probability (DJ)
    - Considers distribution of joint probability of label and a concept over concept space
    - favours peaky distributions over flat ones
- Conditional Probability (CP) of label given concept.
- Mutual Information (MI)

Association Measures: Results

(W, M*, T-, G+, A+, obj, ALL)						
Concept (c)	**DJ**		**CP**		**MI**	
	l	$M(l, c)$	l	$M(l, c)$	l	$M(l, c)$
TEMPO	hai	1.11	hai	21.62	hai	21.50
	aur	0.90	aur	15.74	aur	17.04
	jA	0.52	se	10.35	se	10.32
BICYCLE	gais	0.03	ek	2.58	sAikal	0.86
	sAikal	0.02	hai	1.96	gais	0.45
	uspar	0.02	sAikal	1.95	silinDar	0.32
MOTORCYCLE	ek	1.40	ek	36.83	ek	31.58
	hai	1.22	hai	30.37	hai	28.32
	skUTar	0.76	aur	15.80	jA	14.90
TRUCK	Trak	0.41	ek	30.99	ek	12.21
	ek	0.25	hai	21.93	Trak	12.11
	Ta.Nkar	0.21	Trak	17.29	hai	8.52
HUMAN	hai	5.84	ek	34.11	hai	62.17
	ek	5.39	hai	31.32	ek	58.05
	rahA	3.61	aur	15.64	rahA	33.77
CAR	kAr	0.40	ek	18.20	kAr	7.43
	camcamAtI	0.23	hai	12.58	ek	6.29
	mahAshay	0.22	kAr	7.76	hai	4.12

Table: **Word-level Associations for different probability measures**

Association Measures: Results

(W, M*, T+, G+, A+, obj, ALL)						
Concept (c)	DJ		CP		MI	
	l	$M(l, c)$	l	$M(l, c)$	l	$M(l, c)$
TEMPO	bAik	0.42	Tempo	4.46	mArutI	2.24
	kAr	0.40	kAr	4.33	bAik	1.92
	piilii	0.36	pe	4.25	kAr	1.72
BICYCLE	sAikal	0.02	sAikal	1.95	silinDar	0.16
	uspar	0.02	moTarsAikal	0.79	I.njin	0.14
	I.njin	0.02	pe	0.63	Dilaks	0.14
MOTORCYCLE	skUTar	0.76	pe	8.60	bAik	4.65
	bAik	0.70	bAik	7.12	pe	4.43
	a.ndar	0.54	Tempo	6.56	skUTar	4.30
TRUCK	Trak	0.41	Trak	17.29	Trak	6.22
	Ta.Nkar	0.21	pe	3.24	peTrol	1.47
	peTrol	0.16	sAikal	2.84	Ta.Nkar	1.13
HUMAN	saD.ak	2.74	saD.ak	7.50	saD.ak	12.51
	biThAke	2.12	krOs	6.68	krOs	7.06
	rikshAwAlA	1.84	roD	6.54	biThAke	5.81
CAR	kAr	0.40	kAr	7.76	kAr	3.61
	camcamAtI	0.23	gADI	3.99	gADI	1.46
	mahAshay	0.22	nikalii	2.81	nikalii	1.33

Consistent Dominance and confidence in learning

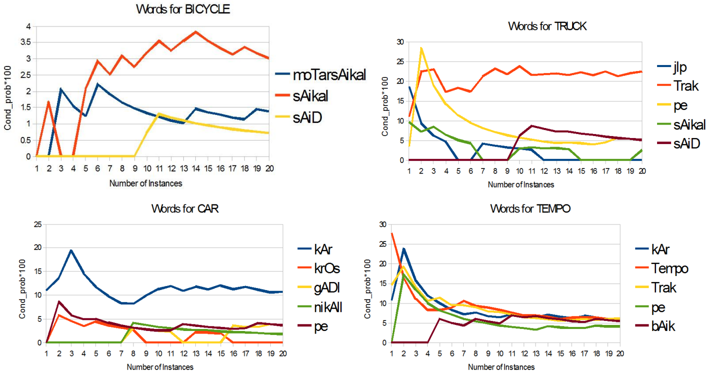


Figure: **Increasing usage**: Effect on word-level associations.

Random order and stability of learning

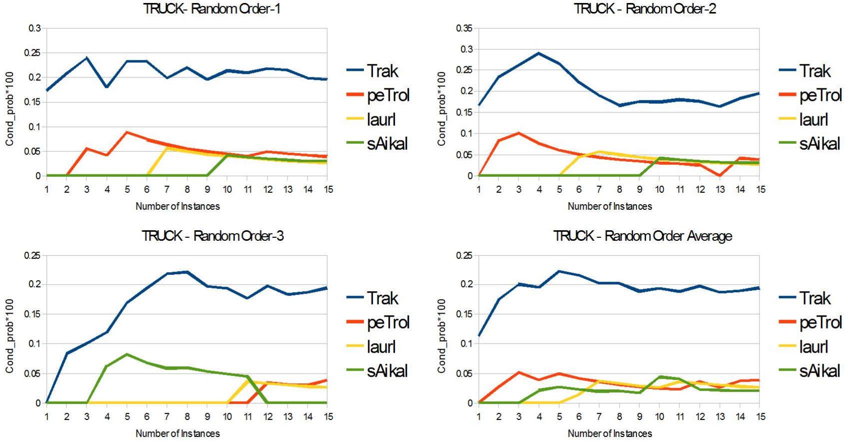


Figure: **Random usage**: Effect on word-level associations for TRUCK.

Labels for trajectories: Results

(W, CP, T*, G-, A+, traj, ALL)				
Concept (c)	Top 1000 retained (T-)		Top 1000 removed (T+)	
	$k = 3$	CP	$k = 3$	CP
C1	jA rahA hai	2.84	aur ek bAik	0.79
	jA rahe hai.n	1.52	krOs kar rahA	0.78
	jA rahI hai	1.39	geT kI taraf	0.61
C2	jA rahA hai	3.96	krOs kar rahA	2.24
	kar rahA hai	2.76	lAl sharT me.n	1.17
	jA rahI hai	2.31	roD krOs kar	0.92
C3	jA rahA hai	6.49	bAe.N se dAe.N	2.12
	jA rahI hai	2.39	pUch rahA hai	1.47
	bAe.N se dAe.N	2.12	pAr hotI hai	1.24
C4	jA rahA hai	3.35	roD krOs kar	2.18
	jA rahI hai	3.15	krOs kar rahA	2.16
	kar rahA hai	3.03	saD.ak krOs kar	0.84
C5	jA rahA hai	3.92	roD krOs kar	2.43
	jA rahI hai	3.62	krOs kar rahA	2.38
	kar rahA hai	2.95	geT kI taraf	1.54
C6	jA rahe hai.n	3.19	ek aur Tempo	0.75
	jA rahI hai	2.44	blaik kalar kI	0.74
	jA rahA hai	1.70	pe ek aAdamI	0.66
C7	jA rahI hai	5.28	geT kI taraf	1.36
	jA rahA hai	4.47	TI ke geT	1.20

Incremental Analysis

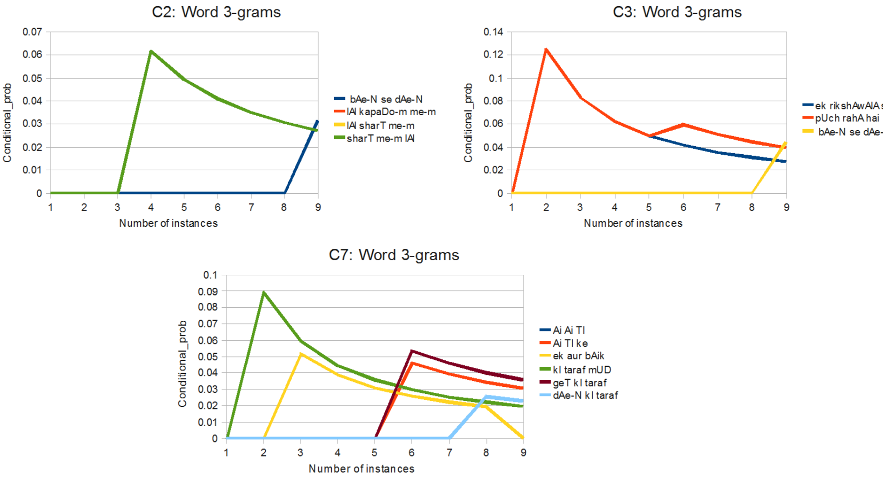
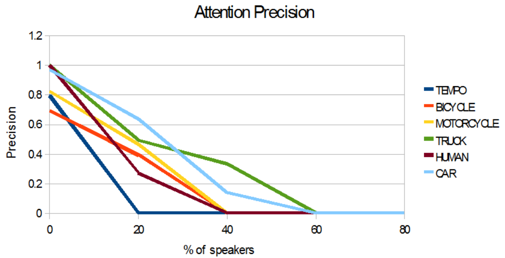
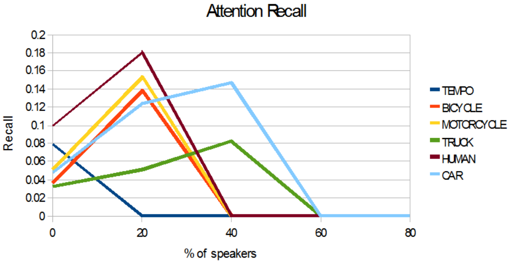


Figure: **Incremental Analysis of Trajectory labels**

Cluster	*l*	CP	*l*	MI
C0 (T)	kAr	4.98	bAik	2.4
	bAik	4.96	moTarsAikal	2.13
	Tempo	4.5	kAr	2.01
C8 (M)	bAik	14.22	skUTar	4.57
	skUTar	12.66	bAik	3.7
	pe	12.53	pe	2.27
C15 (B)	sAikalwAle	8.82	sAikal	3.64
	sAikal	7.12	sAikalwAle	3.63
	dAe.N	6.85	dAe.N	1.67
C19 (C)	kAr	8.27	kAr	3.78
	gADI	4.05	gADI	1.4
	nikalii	2.85	nikalii	1.27
C22 (M)	roD	6.82	roD	0.41
	pe	2.68	khAll	0.3
	skUTar	1.92	laDkl	0.29
C25 (T)	Tempo	18.33	Tempo	5.62
	pe	11.75	mUD	3.05
	sAikal	6.87	pe	2.75
C28 (B)	Tempo	12.36	Tempo	3.02
	sAikal	8.48	sAikalwAle	3.01
	sAikalwAle	6.27	mUD	1.83
C29 (L)	Trak	26.4	Trak	4.83
	pe	8.02	sAmAn	1.51
	sAmAn	5.95	Ore.nj	1.25

Evaluating usefulness





Evaluating necessity

(W, CP, T+, G+, A*, obj, ALL)				
Concept (c)	With Attention (A+)		Without Attention (A-)	
	l	CP	l	CP
TEMPO	Tempo	4.46	Tempo	9.71
	kAr	4.33	pe	5.79
	pe	4.25	OTo	5.51
BICYCLE	sAikal	1.95	sAikal	1.63
	moTarsAikal	0.79	moTarsAikal	0.69
	pe	0.63	pe	0.59
MOTORCYCLE	pe	8.60	pe	7.17
	bAik	7.12	bAik	6.25
	Tempo	6.56	roD	5.55
TRUCK	Trak	17.29	Trak	14.39
	pe	3.24	sAikal	4.40
	sAikal	2.84	pe	3.87
HUMAN	saD.ak	7.50	krOs	6.76
	krOs	6.68	roD	6.68
	roD	6.54	saD.ak	6.32
CAR	kAr	7.76	vain	7.89
	gADI	3.99	kAr	7.73
	nikalii	2.81	gADI	5.68

Table: **Word-level Association with and without attention model**

Conclusion

- Learnt words as labels for several object categories based on
  - Minimally supervised object discovery from complex 3D-Video
  - Bottom-up attention model
  - Multiple narrations describing the scene
- Attempted to learn directional concepts and their labels
- Knowledge of word-boundaries is not a pre-requisite to word-learning.
- Consistent Dominance to assess the confidence in word learning.
- Tried to evaluate the usefulness and necessity of attention model.

Future-Work

- Language-driven attention model
- Learning synonyms.
- Symoblic theft.
- Incremental word-learning.
- Proving domain and language independence of the approach.