Linguistic structure prediction: the case of light verbs in Hindi

Ashwini Vaidya

IIT Delhi

24/2

Predicting linguistic structure

- Language is often seen as a bag of words
- But this fails to take into account the *structural* representation of language
- Basic idea: Some linguistic structure must be inferred or predicted for correct interpretation [Smith, 2011]

Motivation

Disambiguation

nirnay lenaa	'decision take; decide'	Light verb construction
kaagaz lenaa	'paper take'	Ordinary noun and verb

- This distinction has been shown to aid NLP applications like parsing [Begum et al., 2011] and MT [Pal et al., 2011]

Light verb constructions

- A multi-word predicate consisting of a predicating noun and a light verb e.g. *give* a *sigh*, *take* a *walk*
- In Hindi light verb constructions (LVC) can consist of both **noun + light verb** and verb + light verb
- Unlike English, Hindi is verb-final: preverbal noun + light verb e.g. *nirnay lenaa* 'decision take; decide'

Insights from previous work

- MWE identification: association measures, linguistic knowledge and parallel corpora have been used.

Insights from previous work

- MWE identification: association measures, linguistic knowledge and parallel corpora have been used.
- LVCs (as compared to other multi-word expressions) benefit from linguistic features [Vincze et al., 2011]

Insights from previous work

- MWE identification: association measures, linguistic knowledge and parallel corpora have been used.
- LVCs (as compared to other multi-word expressions) benefit from linguistic features [Vincze et al., 2011]
- Tu and Roth (2011) found that linguistic and statistical (collocational) features perform at par for English

Insights from previous work

- MWE identification: association measures, linguistic knowledge and parallel corpora have been used.
- LVCs (as compared to other multi-word expressions) benefit from linguistic features [Vincze et al., 2011]
- Tu and Roth (2011) found that linguistic and statistical (collocational) features perform at par for English
- Recent work on English LVCs emphasizes importance of semantic features [Chen et al., 2015]

Insights from previous work

- Hindi: Begum et al. (2011) have used linguistic features for Hindi
- Hindi: comparing word embeddings and wordnet features [Singh et al., 2015]
- Wordnet features perform better, compounds easier to identify than LVCs

Linguistic challenges for Hindi

- Linguistic notion of an LVC differs across languages
- Most predicating nouns in Hindi LVC are not nominalizations
- Diagnostic feature for English: replacing principle *take* a *walk* $\rightarrow$ *walk*
- Only a handful of predicating nouns are nominalizations in Hindi

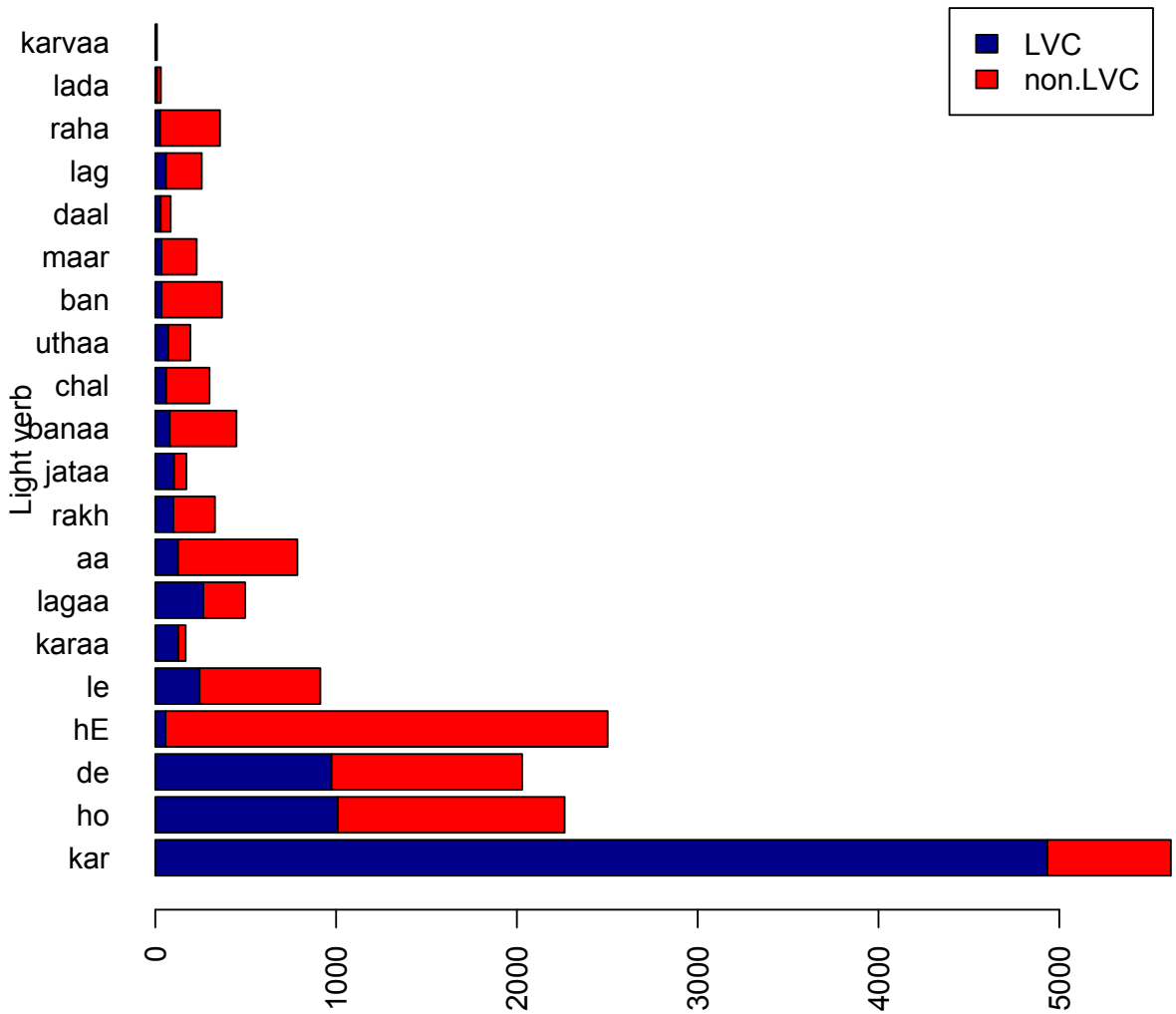
Linguistic challenges for Hindi

- Light verbs in Hindi act as 'verbalizers' and are used to introduce new predicates [Butt, 2010]
- Borrowed words are common: e.g. *email kar* 'email do; email'
- Necessary to understand the behaviour of LVCs in a large corpus

LVC Data

- Use the Hindi Treebank [Palmer et al., 2009] to study LVC behaviour in Hindi
- Hindi Treebank annotates LVCs with a specific label `pof` (part-of)
- Treebank has 47,163 predicates, of which 37% are annotated as LVC

20 most commonly occurring light verbs



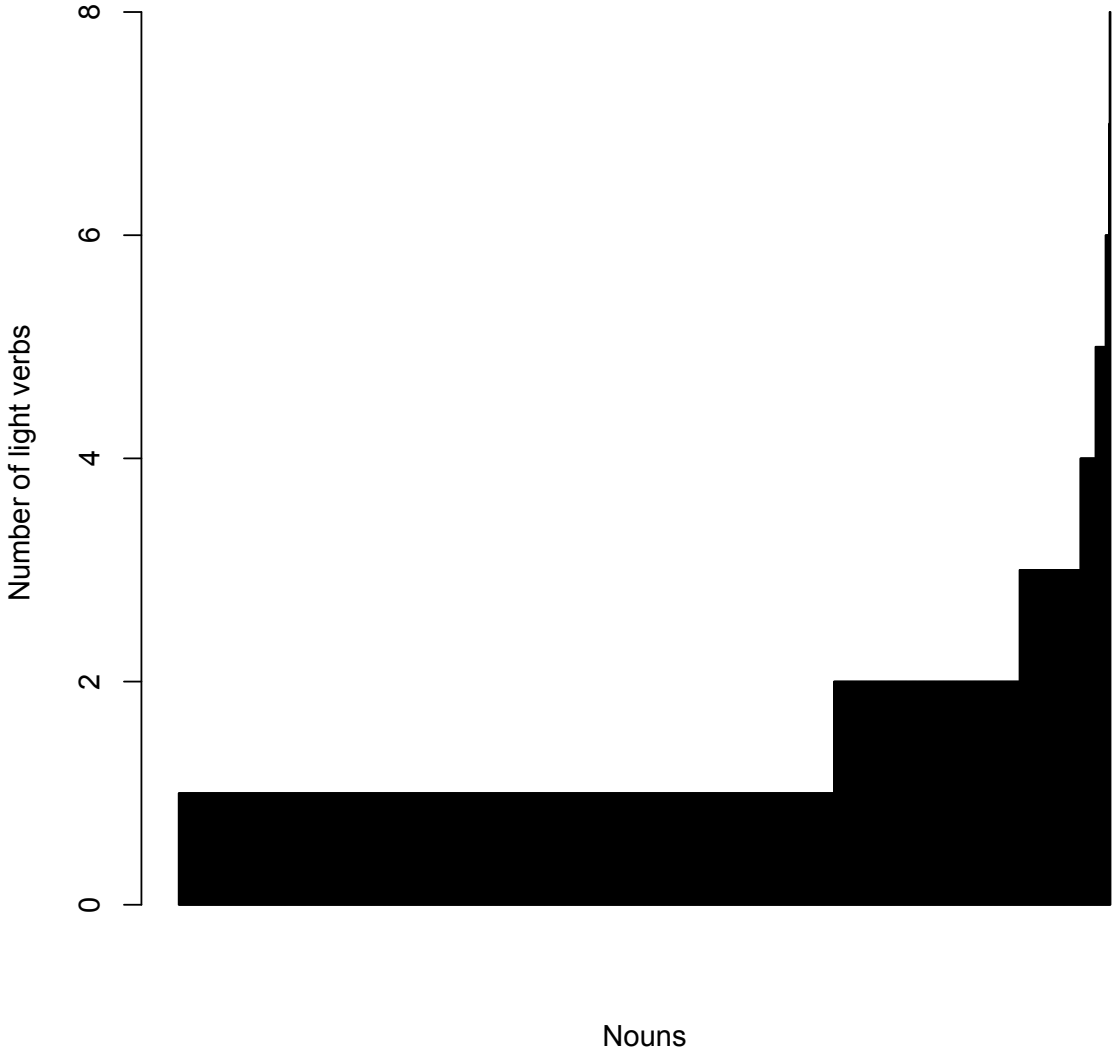
Light verb

Varying distributions

- Most diagnostic tests (and linguistic features) geared towards *kar* [Bhattacharyya et al., 2007, Mohanan, 1997]
- *kar* 'do' is used more often in an LVC as compared to non-LVC
- *aa* 'come' shows the opposite pattern, where it occurs more often in a non-LVC

- Features should generalize to more than one light verb in the data

Number of light verbs with a noun



Predicating nominal

- Nearly 3000 unique nominals in the Hindi Treebank that occur in LVCs
- Show alternations with different light verbs e.g. *ishaara kar* 'make a sign' vs. *ishaara de* 'give a sign'
- Some nouns e.g *bharosa* 'trust' will alternate with 8 light verbs, others with only one e.g *maut* 'death'
- Collocational measures based on bigrams *alone* may ignore low frequency alternations

Proposed linguistic feature- I

- Associate each light verb with a set of ontological properties (Hindi WordNet)
- Based on combinatorial behaviour of LVC [Sulger and Vaidya, 2014]

Light verb	**Ontological property of Noun (WordNet)**	**Example**
de 'give'	Communication_Action_Abstract_Inanimate	*sandesh denaa* 'message give'
kar 'do'	Physical_Action_Abstract_Inanimate	*tareef karnaa* 'praise do'

Proposed linguistic feature- II

- Predicating nouns in an LVC contribute an argument
- Post-position on the non-subject argument: we include *par* 'on', *se* 'with' *ko* 'to' and *kii* 'of'
- Useful in identifying a predicating noun irrespective of the light verb

(1) samir=ne **ghadii=kii** chorii k-ii
Samir.M.Sg-Erg watch.F.Sg-gen theft.F do-Perf.F
'Samir stole the watch'

Feature set

Type	No	Feature
Lexical	1	Verb lemma
	2	Noun lemma
Morpho-syntactic	3	Postposition on predicating noun
	4	**Arguments of eventive noun**
Collocational	5	Log-likelihood value
	6	Pointwise Mutual Information value
Semantic	7	Ontological category of noun
	8	**Light verb and ontological property**

Table: Features used for LVC detection

Experimental setup

Dataset

- Hindi Treebank data with the *pof* label to identify LVC and non-LVC
- Top 20 most frequently occurring LVCs used (90% of all annotations)
- Test set drawn from news ('Test') and literary criticism ('ICON')

	Train	**Development**	**Test** (Treebank)	**Test** (ICON)
News	14282	3500	1708	2757
LVC	6739	1665	790	1056
non-LVC	7543	1835	918	1701

Table: Training, Development and Test instances, with the number of light and non-light verbs

Model and feature selection

- Logistic regression and SVM with RBF kernel (10 fold CV to find best values for C and γ)
- Performed feature selection, resulting in 1638 features
- Verb lemma as a baseline: given a light verb lemma, predict whether it is +/- LVC

ICON: overall performance

- Number of unseen nouns in this test set: 612/2591

ICON	**Logistic Regression**			**SVM with RBF**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
LVC	87.77	76.13	81.54	88.42	76.7	82.15
Non-LVC	86.31	93.41	89.72	86.63	93.76	90.06
Accuracy	86.79			**87.23**		
Begum et. al. (2011)	85.28					
Verb lemma Baseline	75.87			75.66		

Table: Precision, recall and F1 scores for the ICON test set

ICON: Individual light verbs

Individual LVs	LVs in test data	Precision	Recall	F1	Baseline	Accuracy
kar 'do'	650	96.54	95.07	95.80	81.23	93.23
ho 'be'	454	72.26	83.49	77.47	54.62	77.97
de 'give'	216	85.36	70.00	76.92	53.7	80.5
le 'take'	110	91.66	52.38	66.66	61.81	80
LF-TR	263	85.71	42.85	57.14	73.0	86.31
LF-INT	1064	83.33	16.12	27.02	88.34	89.84

Table: Precision, Recall and F1 for individual light verbs in the ICON test set, using Logistic Regression. The baseline accuracy uses the verb lemma as the feature.

Treebank test set (News)

NEWS	**Logistic Regression**			**SVM with RBF**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
LVC	86.36	91.39	88.80	85.8	90.25	87.97
Non-LVC	92.20	87.58	89.83	91.22	87.14	89.13
Accuracy	89.34			88.58		
Verb lemma Baseline	80.97			80.91		

Table: Precision, recall and F1 scores for the News test set

- Number of unseen nouns in this test set: 180/1634

Individual LVs	LVs in test data	Precision	Recall	F1	Baseline	Accuracy
kar 'do'	205	99.40	99.20	99.30	92.12	98.71
ho 'be'	191	77.77	69.30	73.29	47.12	73.29
de 'give'	546	61.32	79.26	69.14	60.0	71.70
le 'take'	88	50.00	69.23	58.06	85.22	85.22
LF_TR	209	65.82	82.53	73.23	72.24	81.81
LF_INT	469	84.21	57.14	68.08	94.02	96.80

Table: Precision, Recall and F1 for individual light verbs in the News set using Logistic Regression. The verb lemma feature is the baseline.

Alternating nouns

Alternating nouns: nouns which occur with >=2 light verbs
E.g. *sahyog denaa*, *sahyog lenaa* or *sahyog karnaa*

- 15% LVCs in the news test set are alternating nouns (135/790)
- Each of these occur with a different surface structure and variable frequency
- 88% of these alternating nouns are correctly predicted by the classifier

Most informative features

- 25 most informative features
- About half of these features were semantic

+LVC	**-LVC**
kar 'do'	Postposition on noun
shuru 'beginning'	*hai* 'be'
Log-likelihood	'Group, Noun' e.g. *dal* 'group'
kar + 'Physical, Action,Abstract,Noun'	'Period, Time' e.g. *October*
PMI	*ban* 'make'
Postposition on argument	'Person, Mammal, Animate, Noun' e.g. *abhineta* 'actor'

Confidence of the model

- More confident about predicting *non* LVCs
- Positive class is more distributed, model less certain about these
- 'Confusing' examples idiomatic *bojh daal* 'weight put; to inconvenience (someone) '
- Intransitive LVCs e.g. *bhojan kar* 'meal do; eat'
- Utilize the LVC classifier to improve annotation quality

Summary

- Linguistic features, especially semantic features useful for generalizing across light verbs
- Propose to use this work to update LVC annotations
  - More precise definition of 'true' LVC: 'argument taking predicating noun' + light verb
  - Separate cases of idiomatic LVCs, treat intransitive and incorporation as non-LVCs

Begum, R., Jindal, K., Jain, A., Husain, S., and Sharma, D. M. (2011).
Identification of Conjunct Verbs in Hindi and their effect on Parsing Accuracy.
In *In Proceedings of the 12th CICLing, Tokyo, Japan*.

Bhattacharyya, P., Chakrabarti, D., and Sarma, V. (2007).
Complex Predicates in Indian languages and Wordnets.
Language Resources and Evaluation, 40(3-4):331–355.

Butt, M. (2010).
The Light Verb Jungle: Still Hacking Away.
In Amberber, M., Harvey, M., and Baker, B., editors, *Complex Predicates in Cross-Linguistic Perspective*, pages 48–78. Cambridge University Press.

Chen, W., Bonial, C., and Palmer, M. (2015).
English Light Verb Construction Identification Using Lexical Knowledge.

In *Proceedings of the AAAI-15, Austin, TX, USA*.


Mohanan, T. (1997).
Multidimensionality of representation- NV complex predicates in Hindi.
In Alsina, A., Bresnan, J., and Sells, P., editors, *Complex Predicates*. CSLI Publications, Stanford.


Pal, S., Chakraborty, T., and Bandopadhyay, S. (2011).
Handling Multi-word expressions in Phrase-based Statisical Machine Translation.
In *Proceedings of the Workshop on Multiword Expressions (MWE 2011), 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*.


Palmer, M., Bhatt, R., Narasimhan, B., Rambow, O., Sharma, D. M., and Xia, F. (2009).
Hindi Syntax: Annotating Dependency, Lexical Predicate-Argument Structure, and Phrase Structure.

In *Proceedings of ICON-2009: 7th International Conference on Natural Language Processing*, Hyderabad.


Singh, D., Bhingardive, S., Patel, K., and Bhattacharyya, P. (2015).
Detection of Multiword Expressions for Hindi Language using Word Embeddings and WordNet-based Features.
In *Proceedings of ICON-2015*.


Smith, N. A. (2011).
Linguistic Structure Prediction.
Morgan and Claypool.


Sulger, S. and Vaidya, A. (2014).
Towards Identifying Hindi/Urdu Noun Templates in Support of a Large-Scale lfg Grammar.
In *Proceedings of the Fifth Workshop on South and Southeast Asian Natural Language Processing at COLING 2014*.


Vincze, V., T, I. N., and Berend, G. (2011).

Detecting noun compounds and light verb constructions: a contrastive study.
In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)*.