Corpora and Computational linguistics

Current trends in Computational Linguistics III

Ashwini Vaidya

Workshop on Experimental and Empirical Methods in Linguistics

What is a corpus?

- A collection of language data

What is a corpus?

- A collection of language data
- Ideally: balanced, representative of various domains, large size

What is a corpus?

- A collection of language data
- Ideally: balanced, representative of various domains, large size
- Includes text from books, newspapers, but it may also be annotated:, dictionaries, parse trees etc.

Linguistic Structure

- Annotated corpora contain labels for **linguistic structure**
  - Words
  - Syntax
  - Semantics
  - Coreference ...
- We are interested in the *relationships* of words to each other

Linguistic Structure

- Computational linguistics (CL) research has focused on the transformation of text into a representation of structure
- Words $\rightarrow$ part of speech tags
  - The boy went to school $\rightarrow$ The/Det boy/NN went/VB to/PRP school/NN
  - Syntax $\rightarrow$ parse trees
- Eventually, these structures would correspond more closely with the meaning of a text

Annotated corpora and Lexical Resources

- Brown, Switchboard, EMILLE and many others
- Penn Treebank, Prague Dependency Treebank, Stanford Dependencies
- WordNet, FrameNet, VerbNet

Corpora and Annotated Corpora

- The corpus and annotated corpora is essential in Computational Linguistics and NLP
- We build and test our tools on a large amount of linguistic data that represents the problem we're trying to solve
- In these fields, corpus work is a given, but in linguistic theory — not always

Corpora and Annotated Corpora

- Corpus work and annotation is descriptive
- Role of descriptive grammars, typological studies
- Corpora complement and support this type of descriptive work

Corpora and Annotated Corpora

- Corpus linguistics spurs deeper research into a problem
  - Light verb constructions (Chen et al, 2015)
  - Caused motion constructions (Hwang 2014, 2015)
  - Stative constructions in Arabic (Mansouri,2016)
  - Dependency parsing and semantic role labelling (Choi, 2012)

English Light Verb Construction Identification Using Lexical Knowledge. 2015 WT Chen, C Bonial, M Palmer AAAI, 2368-2374
Identification and representation of caused motion constructions. J.D Hwang (thesis) CU Boulder, 2014
Stative and Stativizing Constructions in Arabic News Reports: A Corpus-Based Study, 2016, Aous Mansouri (thesis)
Optimization of Natural Language Processing Components for Robustness and Scalability. Choi, J. D. Ph.D. Thesis, University of Colorado Boulder, 2012.
Identification of Caused Motion Construction. 2015. JD Hwang, M Palmer * SEM@ NAACL-HLT, 51-60

A place for corpus linguistics

- de Marneffe and Potts (2017)
- Intuitions of linguists, native speakers are more privileged in corpus linguistics and psycholinguistics: starting point for any study
- Humans are biased (over-exposure, cognitive load, etc) whether in the armchair or in the lab: need for greater care
- Theoretical study $\rightarrow$ corpora $\rightarrow$ psycholinguistics

Christopher Potts and Marie Catherine de Marneffe. 2017. Developing linguistic theories using annotated corpora, in The Handbook of Linguistic Annotation eds. James Pustejovsky and Nancy Ide

de Marneffe and Potts (2017)

- But corpus study is a behaviourist stance about language!
- A corpus is never the ultimate object of study: role of inference and generalization
- CHILDES: gain insights into the developmental errors made by children learning language
- Systematic errors (and their analysis) do influence theories in L1 and L2 acquisition

MacWhinney, Brian. 2000.The CHILDES project: Tools for analyzing talk Mahwah, NJ: Lawrence Erlbaum Associates 3rd edn

de Marneffe and Potts(2017)

- Statistical quality of the generalizations: gradient not categorical
- Again, these would need to be interpreted: are categorical restrictions underlying them? What does this imply for our theories?
- These are actually the most exciting 'contact' areas, fertile for further research!

Limitations

- A single corpus cannot be representative of all language
- In every corpus,we need to think of the design of the corpora as well as quantity (currently, emphasis is on the latter)
- Model evaluations should also take into account properties/limitations of the corpus

Romantics vs. Revolutionaries?

Steedman (2011)

In every field in which progress beckons, romantics and revolutionaries find themselves in an uneasy alliance. The role of the romantics is to define the often unattainable goal. That of the revolutionaries is to advance towards it. Each needs the other, and constantly fears they are forsaken. Sometimes they are right.

- More cross-fertilization in linguistic theories/computational grammars than anyone cares to admit (LFG, TAG)
- Fillmore's case for case (1968) to Frame Semantics (1982) for FrameNet

Romantics and Revolutionaries: What theoretical and computational linguists need to know about each other*but were afraid to ask. 2011. Mark Steedman. Linguistic Issues in language technology, Vol6, Issue 11
Frederick Newmeyer. Linguistic Theory in America: The First Quarter Century of Transformational Generative Grammar. New York: Academic Press. Second edition 1986

“I, for one welcome our new computer overlords”

IBM Watson vs Ken Jennings and Brad Rutter

Question Answering systems

Jeopardy: Formulate the question

William Wilkinson's "An account of the principalities of Wallachia and Moldovia" inspired this author's most famous novel.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1221, Prague, June.

Question Answering systems

Jeopardy: Formulate the question

William Wilkinson's "An account of the principalities of Wallachia and Moldovia" inspired this author's most famous novel.

Who is Bram Stoker?

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1221, Prague, June.

Question Answering systems

Jeopardy: Formulate the question

William Wilkinson's "An account of the principalities of Wallachia and Moldovia" inspired this author's most famous novel.

Who is Bram Stoker?

- Question Answering improves with knowledge of 'Who did what to whom' (Shen and Lapata, 2007)
- This knowledge forms a small part of a pipeline that includes many other components (e.g. Watson's DeepQA)

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1221, Prague, June.

Generalize over different surface realizations

- John broke the window [`Agent/S, Theme/O`]
- The window broke [`Theme/O`]
- The window was broken by John [`Agent/O, Theme/S`]
- The rock broke the window [`Instrument/S, Theme/O`]

A corpus of semantic roles

- Every verbal predicate in a sentence is annotated with semantic roles
- Atif$_{ARG0}$ kitaab$_{ARG1}$ paRhegaa$_{parh.01}$

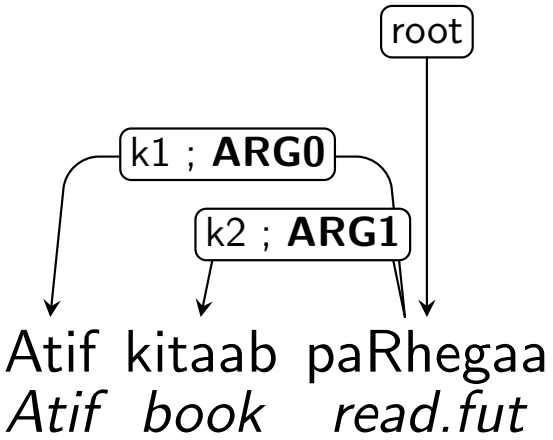

Figure: Dependency tree with PropBank annotations `ARG0` and `ARG1`

Shallow semantic representation

John rings the bell
$\text{ring}_{ring.01}(\text{John}_{ARG0}, \text{bell}_{ARG1})$

- There was an event of ringing, two participants: John and bell
- ARG0: agent-like role, ARG1: theme-like role
- (Actual labels don't matter `RINGER`, `THING RUNG`)

Task: semantic role labelling

1. Identify the predicate in the sentence (verbal)

Task: semantic role labelling

1. Identify the predicate in the sentence (verbal)

John **rang** the bell

Task: semantic role labelling

1. Identify the predicate in the sentence (verbal)

John **rang** the bell

2. Disambiguate the sense of the predicate

Task: semantic role labelling

1. Identify the predicate in the sentence (verbal)

John **rang** the bell

2. Disambiguate the sense of the predicate

✓ John **rang**$_{ring.01}$ the bell
Tall aspen trees **ring** $_{ring.02}$ the lake

Task: semantic role labelling

1. Identify the predicate in the sentence (verbal)

John **rang** the bell

2. Disambiguate the sense of the predicate

✓ John **rang**$_{ring.01}$ the bell
Tall aspen trees **ring** $_{ring.02}$ the lake

3. Identify the arguments

Task: semantic role labelling

1. Identify the predicate in the sentence (verbal)

John **rang** the bell

2. Disambiguate the sense of the predicate

✓ John **rang**$_{ring.01}$ the bell
Tall aspen trees **ring** $_{ring.02}$ the lake

3. Identify the arguments

John **rang**$_{ring.01}$ the bell

Semantic Role Labelling: Recently

- Many other efforts to derive semantic roles via Wordnet, Wikipedia and Wikitionary (Hartmann et al, 2016; Exner et al, 2016;)
- Unsupervised approaches to learning semantic roles, Swier and Stevenson, (2004) or Collobert (2011) which is not task-specific
- AMR (Abstract Meaning Representation) is a new graph-based representation of sentences tries to collapse separate levels of annotation (Banarescu et al, 2013)

Highlight: Linguistic challenges

1. Hindi SRL 'PropBank' (Bhat et al, 2017)
2. Predicate argument-structure annotation
3. Extending to closely related languages

Riyaz Ahmad Bhat, Rajesh Bhatt, Annahita Farudi, Prescott Klassen, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Ashwini Vaidya, Sri Ramagurumurthy Vishnu, and Fei Xia, 2017. The Hindi/Urdu Treebank Project, Handbook of Linguistic Annotation (edited by Nancy Ide and James Pustejovsky), Springer Press.

Step1 :Frame files

Table: A frame file

parh.01	'Read'
ARG0	Reader
ARG1	The thing read

- Frames for 700 simple verbs in Hindi
- This work was done manually for Hindi (also advances on inferring these frames automatically (Raza, 2010)

Inferring Subcat Frames of Verbs in Urdu. 2010. Ghulam Raza, COLING 2010

Step 2: Numbered arguments and modifiers

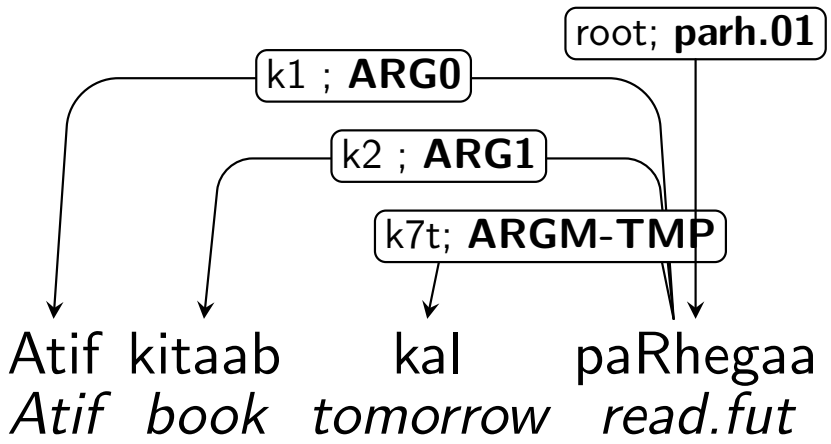

Figure: Dependency tree with PropBank annotations

Elided arguments

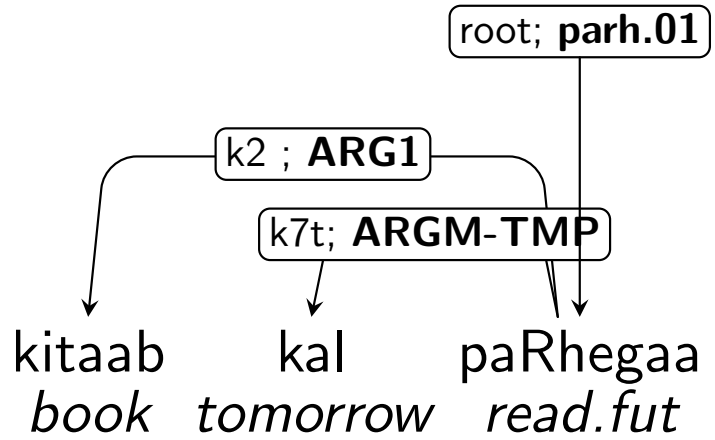

Figure: Dependency tree with elided argument

Elided arguments

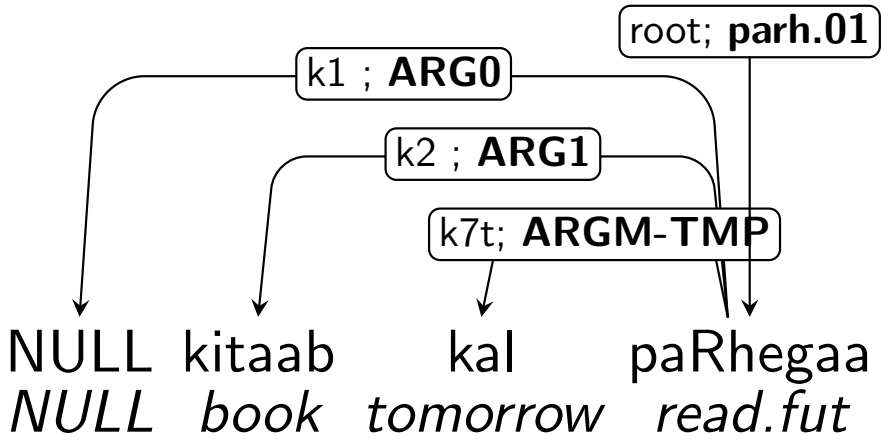

Figure: Dependency tree with elided argument

ECs for missing arguments

मोहन_ने$_i$	*PRO*$_i$	किताब	पढ़नी	चाही
Mohan_ERG	(he)	book	read_INF	want_PERF

Figure: An example of PRO: "*Mohan wanted PRO to read the book*" in Hindi.

इसके	बाद	(*pro*)	गाडी	भी	खरीदी
This_GEN	after	(he)	car	also	buy_PST

Figure: An example of pro: "*After this, pro also bought a car*" in Hindi.

मोहन_ने$_i$	किताब	पढ़ी	और	*GAP*$_i$	सो_गया
Mohan_ERG	book	read_PERF	and	(he)	sleep_go_PERF

Figure: An example of GAP: "*Mohan read the book and GAP slept*" in Hindi.

REL	भागता_ हुआ	लड़का	सेब	खा_रहा_है
(who)	run_INF_be	boy	apple	eat_CONT_be

Figure: An example of REL: "*The REL running boy is eating an apple*" in Hindi.

Elided arguments

- EC insertion allows easier recovery of the predicate-argument structure
- Use dependency and valence properties of the verb to predict ECs (Vaidya et al 2012)
- EC insertion as a sequential tagging problem to predict Chinese ECs (Yang and Xue, 2010)

Empty Argument Insertion in the Hindi PropBank. 2012. Ashwini Vaidya, Jinho D. Choi, Martha Palmer, Bhuvana Narasimhan, LREC 2012
Chasing the ghost: recovering empty categories in the Chinese Treebank. 2010. Yaquin Yang, Nianwen Xue. COLING 2010

Elided arguments (Hindi)

Corpus	PRO	REL	GAP	pro
37K	490	176	33	96
47K	553	238	24	76

Table: Numbers of empty arguments in our corpus.

Elided arguments (Hindi)

Corpus	PRO	REL	GAP	pro
37K	490	176	33	96
47K	553	238	24	76

Table: Numbers of empty arguments in our corpus.

	Type	Precision	Recall	F1-score
37K/2710	PRO	87.79	93.69	90.64
	REL	100.00	94.32	97.08
47K	PRO	85.02	91.91	88.33
	REL	100.00	89.92	94.69

Table: Insertion accuracies of PRO (in %).

Elided arguments

- Many statistical parsers ignore EC, but they do make a difference in some tasks e.g. inferring valency information for subcategorization frames
- Recent dependency formalisms such as Universal Dependencies don't posit EC [1]

[1] http://universaldependencies.org/

Multi-word predicates

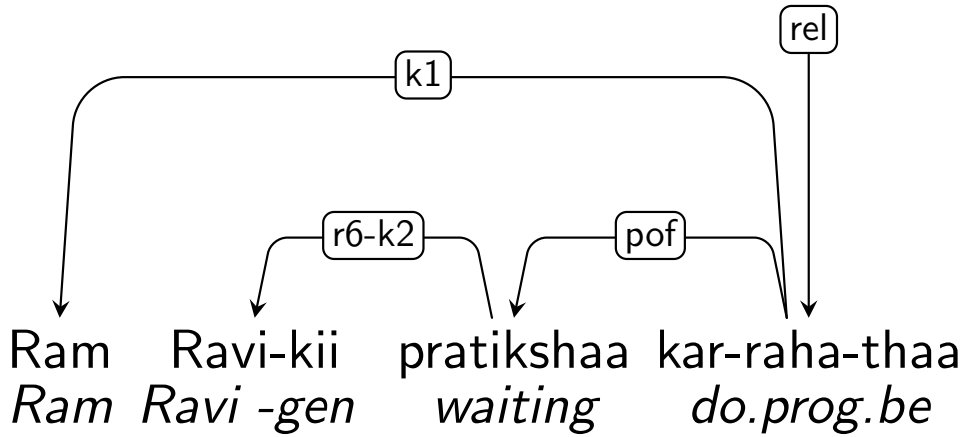

Figure: Dependency tree for 'Ram was waiting for Ravi'. LVC: *pratikshaa* 'waiting' *kar* 'do'.

Light verb constructions

- Verbal multiword expression with two elements: light verb and predicating noun

PropBank Annotation of Multilingual Light Verb Constructions. 2010. Jena D. Hwang, Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Yuping Zhou, Nianwen Xue, and Martha Palmer. LAW 2010

Light verb constructions

- Verbal multiword expression with two elements: light verb and predicating noun

Ram-ne *cycle-kii* chorii kii
Ram-erg cycle-gen theft do.prf
'Ram stole a cycle (did theft of a cycle)'

PropBank Annotation of Multilingual Light Verb Constructions. 2010. Jena D. Hwang, Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Yuping Zhou, Nianwen Xue, and Martha Palmer. LAW 2010

Light verb constructions

- Verbal multiword expression with two elements: light verb and predicating noun

  Ram-ne *cycle-kii* chorii kii
  Ram-erg cycle-gen theft do.prf
  'Ram stole a cycle (did theft of a cycle)'

- Consider the noun in the LVC as the main predicate (Hwang et al, 2010)

PropBank Annotation of Multilingual Light Verb Constructions. 2010. Jena D. Hwang, Archna Bhatia, Clare Bonial, Aous Mansouri, Ashwini Vaidya, Yuping Zhou, Nianwen Xue, and Martha Palmer. LAW 2010

Many other verbal multiwords in Hindi!

- Idiomatic *god lenaa* 'lap take; adopt'
- Stacked LVCs *shaadi kar denaa* 'marriage do give; get married off'
- V+V LVCs *likh denaa* 'write give; write (completely)'

Identifying LVCs

- Association measures, linguistic knowledge and parallel corpora have been used.

English Light Verb Construction Identification Using Lexical Knowledge Wei-Te Chen, Claire Bonial, Martha Palmer, 2015, Proceedings of AAAI
Vincze, Veronika, Istvan Nagy T, and Gabor Berend. 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).
Ashwini Vaidya, Sumeet Agarwal, Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. Proceedings of COLING 2016 pdf

Identifying LVCs

- Association measures, linguistic knowledge and parallel corpora have been used.
- LVCs (as compared to other multi-word expressions) benefit from linguistic features (Vincze and Nagy 2011)

English Light Verb Construction Identification Using Lexical Knowledge Wei-Te Chen, Claire Bonial, Martha Palmer, 2015, Proceedings of AAAI
Vincze, Veronika, Istvan Nagy T, and Gabor Berend. 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).
Ashwini Vaidya, Sumeet Agarwal, Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. Proceedings of COLING 2016 pdf

Identifying LVCs

- Association measures, linguistic knowledge and parallel corpora have been used.
- LVCs (as compared to other multi-word expressions) benefit from linguistic features (Vincze and Nagy 2011)
- Tu and Roth (2011) found that linguistic and statistical (collocational) features perform at par for English

English Light Verb Construction Identification Using Lexical Knowledge Wei-Te Chen, Claire Bonial, Martha Palmer, 2015, Proceedings of AAAI
Vincze, Veronika, Istvan Nagy T, and Gabor Berend. 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).
Ashwini Vaidya, Sumeet Agarwal, Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. Proceedings of COLING 2016 pdf

Identifying LVCs

- Association measures, linguistic knowledge and parallel corpora have been used.
- LVCs (as compared to other multi-word expressions) benefit from linguistic features (Vincze and Nagy 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)

English Light Verb Construction Identification Using Lexical Knowledge Wei-Te Chen, Claire Bonial, Martha Palmer, 2015, Proceedings of AAAI
Vincze, Veronika, Istvan Nagy T, and Gabor Berend. 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).
Ashwini Vaidya, Sumeet Agarwal, Martha Palmer. 2016. Linguistic features for Hindi light verb construction identification. Proceedings of COLING 2016 pdf

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'

Task

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

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

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

Granularity of annotation labels

- The syntactic label for LVCs is too broad and all-inclusive in the Hindi Treebank
- Labels need to be more narrow, multiwords behave differently (Fazly, 2007)
- Larger problem: label granularity (affects many lexical resources)
- Annotator reliability: distinctions in the label $\downarrow$ (as compared to ambiguity of the target items) (Brown et al, 2010)

Fazly, Afsaneh and Suzanne Stevenson. 2007. Automatic Acquisition of Knowledge about Multi- word Predicates. In Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language,Information and Computation.
Susan Windisch Brown, Travis Rood, and Martha Palmer. 2010. Number or Nuance: Which Factors Restrict Reliable Word Sense Annotation? Proceedings of LREC 2010

Highlight: Linguistic challenges

1. Predicate argument-structure annotation
2. Extending to closely related languages

Extending to closely related languages

- Porting manually created frames (subcat frames) from Hindi to Urdu
- Urdu takes a large amount of its vocabulary from Arabic and Persian
- Given a verb, identify its source language using n-gram models (Bhat et al, 2014)

Riyaz Ahmad Bhat, Naman Jain, Dipti Misra Sharma, Ashwini Vaidya, Martha Palmer, James Babani, Tafseer Ahmed. 2014. Adapting Predicate Frames for Urdu PropBanking. In Proceedings of the EMNLP Workshop on Language Technology for Closely Related Languages and Language Variants. Pg 4755

Frames from Hindi to Urdu

	Simple Verbs		Nominal Predicates	
Language	**Total**	**Unique**	**Total**	**Unique**
Arabic	12	1	6,780	765
Hindi	7,332	441	1,203	258
Persian	69	3	2,276	352
Total	7,413	445	10,259	1,375

- Table shows predicates from the Urdu Treebank (Simple verbs are almost identical); even borrowings except *farmaanaa* 'say'
- Several nominal predicates (LVC) are not shared, many are borrowed from Arabic and Persian
- Many sense divergences from Arabic; work under progress for Persian

Frames from Hindi to Marathi?

- Cognates : words with a common ancestor e.g. *Hund* 'German; dog' and *Hound* 'English; dog'
- Leverage the notion of cognacy to find similar verbs in closely related language E.g. *jalnaa* (burn; Hindi) and *jaLne* (burn; Marathi) (Singh and Surana, 2007)
- Subcategorization frames from Hindi could then be shared with Marathi
- Cognacy detection using LSTM networks

Anil Kumar Singh and Harshit Surana. 2007. Study of cognates among South Asian languages for the purpose of building lexical resources. Journal of Language Technology.
Shantanu Kumar, Ashwini Vaidya and Sumeet Agarwal.2017. Discovering Cognates Using LSTM Networks (submitted)