NLP and ML: Synergy or Divergence?

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay and IIT Patna
www.cse.iitb.ac.in/~pb

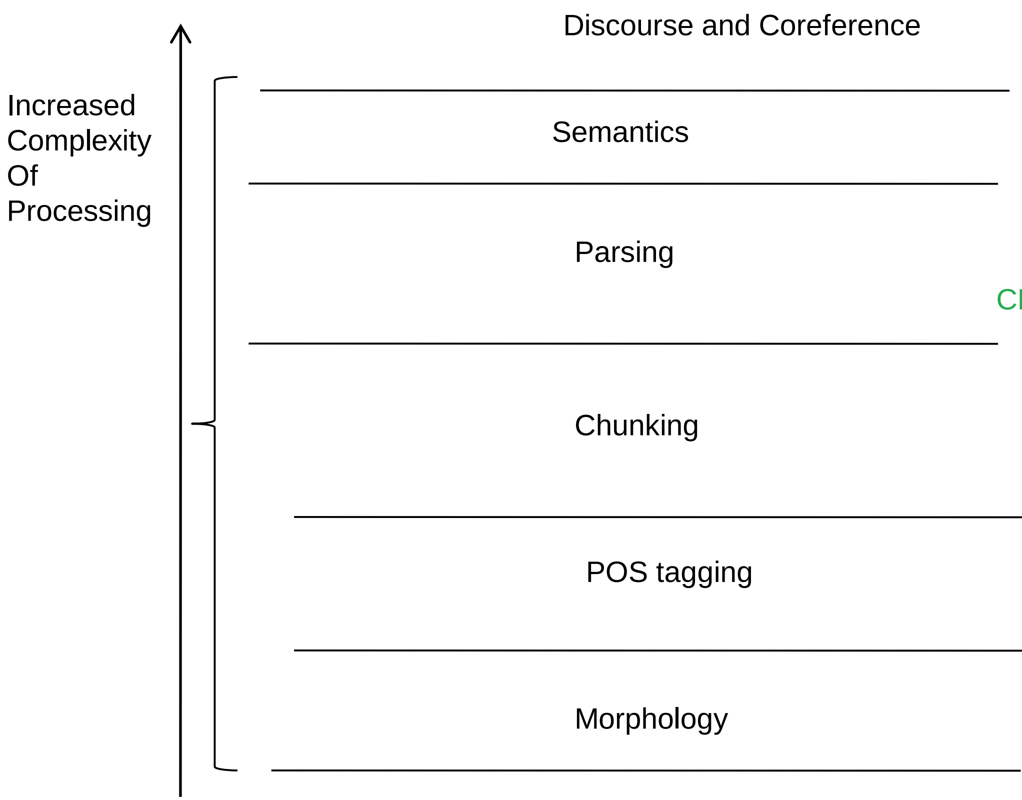
(21st Sept, 2016)

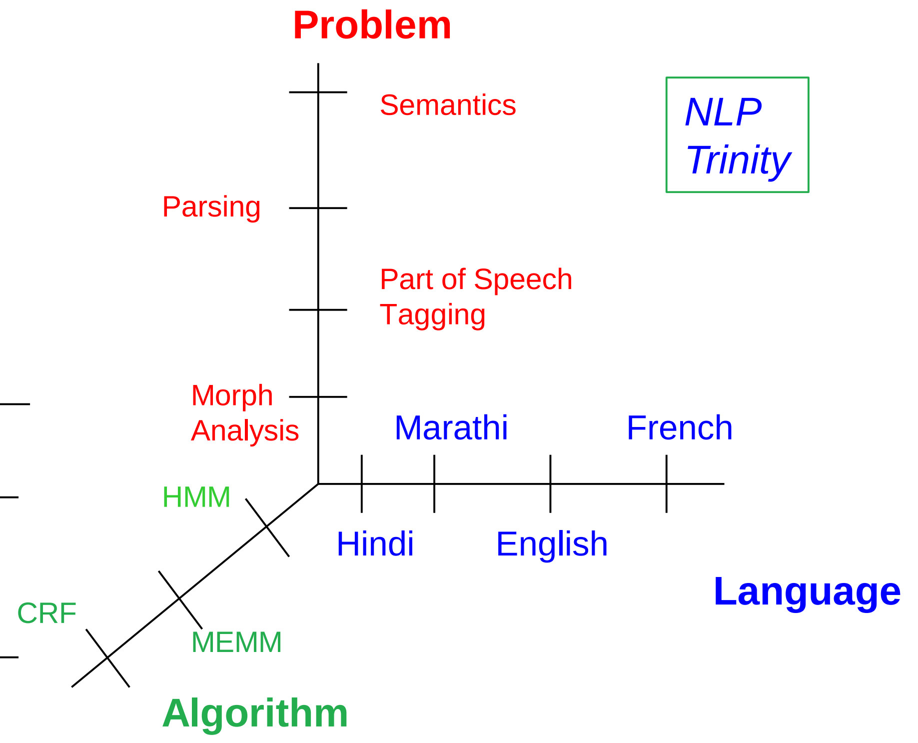
Roadmap

- Perspective
- Power of Data
- Some “lower level” NLP tasks
- Alignment in MT
- Annotation
- Cooperative WSD
- Sarcasm
- Conclusions

Perspective

NLP: a useful view





Why is NLP hard?

Ambiguity

- Lexical Ambiguity
- Structural Ambiguity
- Semantic Ambiguity
- Pragmatic Ambiguity

Examples

1. (ellipsis) Amsterdam airport: “Baby Changing Room”

2. (Attachment/grouping) Public demand changes (credit for the phrase: Jayant Haritsa):

   (a) Public demand changes, but does any body listen to them?

   (b) Public demand changes, and we companies have to adapt to such changes.

   (c) Public demand changes have pushed many companies out of business

3. (Attachment) *Ishant ruled out of first test with Chickengunia* (ToI: 21/9/16)

3. (Pragmatics-1) The use of shin bone is to locate furniture in a dark room

4. (Pragmatics-2) Blood flows on streets of Dhaka on Eid after animal sacrifice

New words and terms (people are very creative!!)

1. ROFL: rolling on the floor laughing; *LOL*: laugh out loud

2. facebook: to use facebook; *google*: to search

3. communifake: faking to talk on mobile; *Obamacare*: medical care system introduced through the mediation of President Obama (portmanteau words)

4. After BREXIT (UK's exit from EU), in Mumbai Mirror, and on Tweet: We got Brexit. What's next? Grexit. Departugal. Italeave. Fruckoff. Czechout. Oustria. Finish. Slovakout. Latervia. Byegium

Example: Humour

1. (for a student of mine)

Student: my thesis is on unsupervised WSD

Prof. Sivakumar: But I thought Pushpak is supervising your thesis!

2. (ToI, 11/4/15)

If money does not grow on trees, why do banks have branches?

3. (ToI 2/3/15)

Q: Have you heard of the kidnapping in the school?

A: no, he got up

NLP: compulsory Inter layer interaction (1/2)

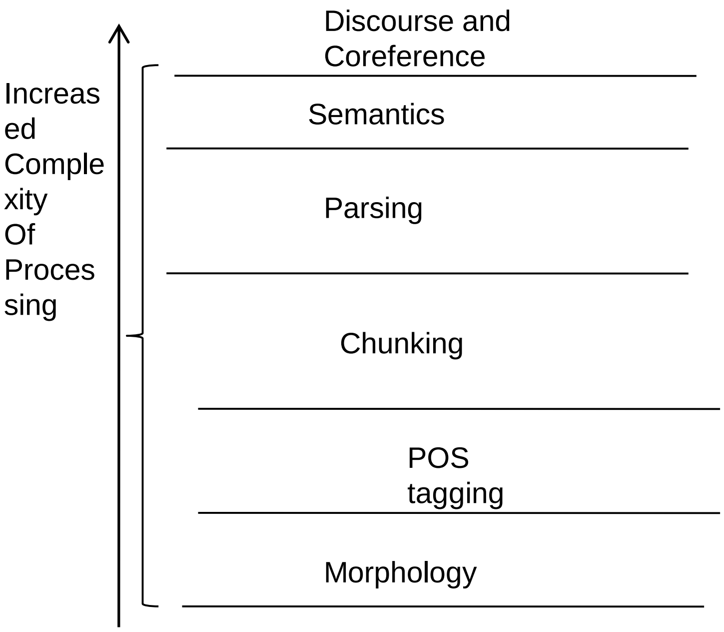
Text-1: “*I saw the boy with a telescope which he dropped accidentally*”
Text-2: “*I saw the boy with a telescope which I dropped accidentally*

Text-1:

```
(S
  (NP (PRP I))
  (VP
       (VBD saw)
       (NP (DT the) (NN boy))
       (PP (IN with) (NP (NP (DT a) (NN telescope))
            (SBAR (WHNP (WDT which)) (S (NP (PRP I))
               (VP (VBD dropped)
              (ADVP (RB accidentally)))))))) (. .)))
```

Text-2:

```
(S
  (NP (PRP I))
  (VP
       (VBD saw)
       (NP (DT the) (NN boy))
       (PP (IN with) (NP (NP (DT a) (NN telescope))
            (SBAR (WHNP (WDT which)) (S (NP (PRP he))
               (VP (VBD dropped) (ADVP (RB accidentally)))))))) (. .)))
```



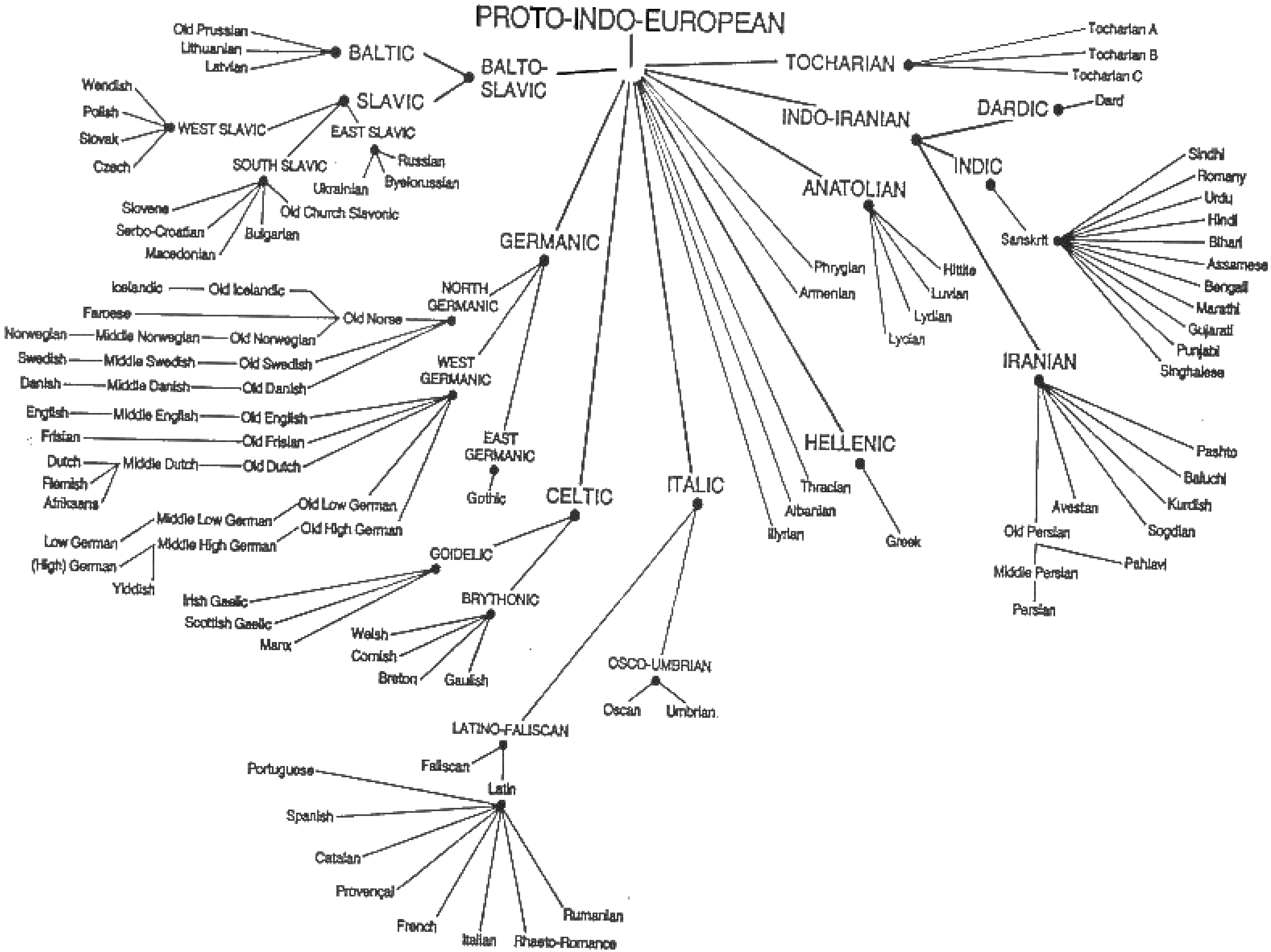
Inter layer interaction (2/2)

Text-1: "*I saw the boy with a telescope which he dropped accidentally*"
Text-2: "*I saw the boy with a telescope which I dropped accidentally*

nsubj(saw-2, I-1)
root(ROOT-0, saw-2)
det(boy-4, the-3)
dobj(saw-2, boy-4)
det(telescope-7, a-6)
prep_with(saw-2, telescope-7)
dobj(dropped-10, telescope-7)
nsubj(dropped-10, I-9)
rcmod(telescope-7, dropped-10)
advmod(dropped-10, accidentally-11)

nsubj(saw-2, I-1)
root(ROOT-0, saw-2)
det(boy-4, the-3)
dobj(saw-2, boy-4)
det(telescope-7, a-6)
prep_with(saw-2, telescope-7)
dobj(dropped-10, telescope-7)
nsubj(dropped-10, he-9)
rcmod(telescope-7, dropped-10)
advmod(dropped-10, accidentally-11)

NLP: deal with multilinguality
Language Typology



Languages differ in expressing thoughts: Agglutination

- Finnish*:* “istahtaisinkohan”
- English: "I wonder if I should sit down for a while"

Analysis:

- ist + "sit", verb stem
- ahta + verb derivation morpheme, "to do something for a while"
- isi + conditional affix
- n + 1st person singular suffix
- ko + question particle
- han a particle for things like reminder (with declaratives) or "softening" (with questions and imperatives)

Consider Malayalam → Hindi translation

Source

കുറച്ച് ശാസ്ത്രജ്ഞരർ പറയുന്നു നമ്മുടെ മനസ്സിൽ ഉണ്ടാകുന്ന ചിന്തകളാണ് സ്വപ്നമായി കാണുന്നതെന്ന് .

kuRacc shAstrajJNar paRayunnu nammuT.e manassila uNTAkunna cintakaLAN svapnamAyi kANunnat.enn .

Some scientists say our mind+in happening thoughts dream+become see

Some scientists opine that whatever we see in dreams are thoughts encased in our unconscious mind .

Word-level Translation output

कुछ वैज्ञानिकों ने कहा कि हमारे मन में होने वाले ചിന്തകളാണ് സ്വപ്നമായി കാണുന്നതെന്ന് है ।

Morpheme-level output

कुछ वैज्ञानिकों ने कहा जाता है कि हमारे मन में होने वाले चिंता होते हैं , स्वप्न रूप से देख सकते हैं ।

So far we have meaningful units of text.

But, we needs lot of data to achieve good vocabulary coverage and probability estimates

Use character as basic unit

कुछ शास्त्र में ने कहा हमारे मन मस्सों वाली चिंता स्वप्न माना जाता है ।

That’s looks like a good start, given we have no linguistic knowledge

Though, we essentially threw away the notion of a word !

The basic units don’t convey any meaning !

Can we do better?

Let’s try something better

First segment the character stream into *akshar*

ie. Consonant-vowel+ combinations

वैज्ञानिकों → वै ज्ञा नि कों

Why?

- Character vocabulary very small, ambiguous translations
- Syllable as a basic unit of speech

Translation output

कुछ वैज्ञानिकों का कहना है कि हमारे मन में होने वाले चिंताओं स्वप्न से देख लेते हैं ।

We get even better results !

But, these basic units aren’t meaningful either !!

This works for many language pairs

(Kunchukuttan & Bhattacharyya, 2016)

Source	Target	Word	Morph	Character	Orth-Syllable
bn	hi	31.23	32.17	27.95	**33.46**
kK	mr	21.39	22.81	19.83	**23.53**
ml	ta	6.52	7.61	4.50	**7.86**
hi	ml	8.49	9.23	6.28	**10.45**
ml	hi	15.23	17.08	12.33	**18.50**
pa	hi	68.96	71.29	71.26	**72.51**
te	ml	6.62	7.86	6.00	**8.51**

So, what's happening?

Anoop Kunchukuttan, Pushpak Bhattacharyya. *Orthographic Syllable as basic unit for SMT between Related Languages*. EMNLP. 2016.

Language Similarity

കുറച്ച് ശാസ്ത്രജ്ഞരർ പറയുന്നു നമ്മുടെ മനസ്സിൽ ഉണ്ടാകുന്ന ചിന്തകളാണ് സ്വപ്നമായി കാണുന്നതെന്ന് .

kuRacc shAstrajJNar paRayunnu nammuT.e manassil uNTAkunna cintakaLAN svapnamAyi kANunnat.enn .

कुछ वैज्ञानिकों का कहना है कि हमारे मन में होने वाले विचार सपने बनकर देखते है

These language pairs exhibit the following properties

Lexical Similarity: Cognates, loan-words, lateral borrowings

Structural Correspondence: Similar word order and parse structures

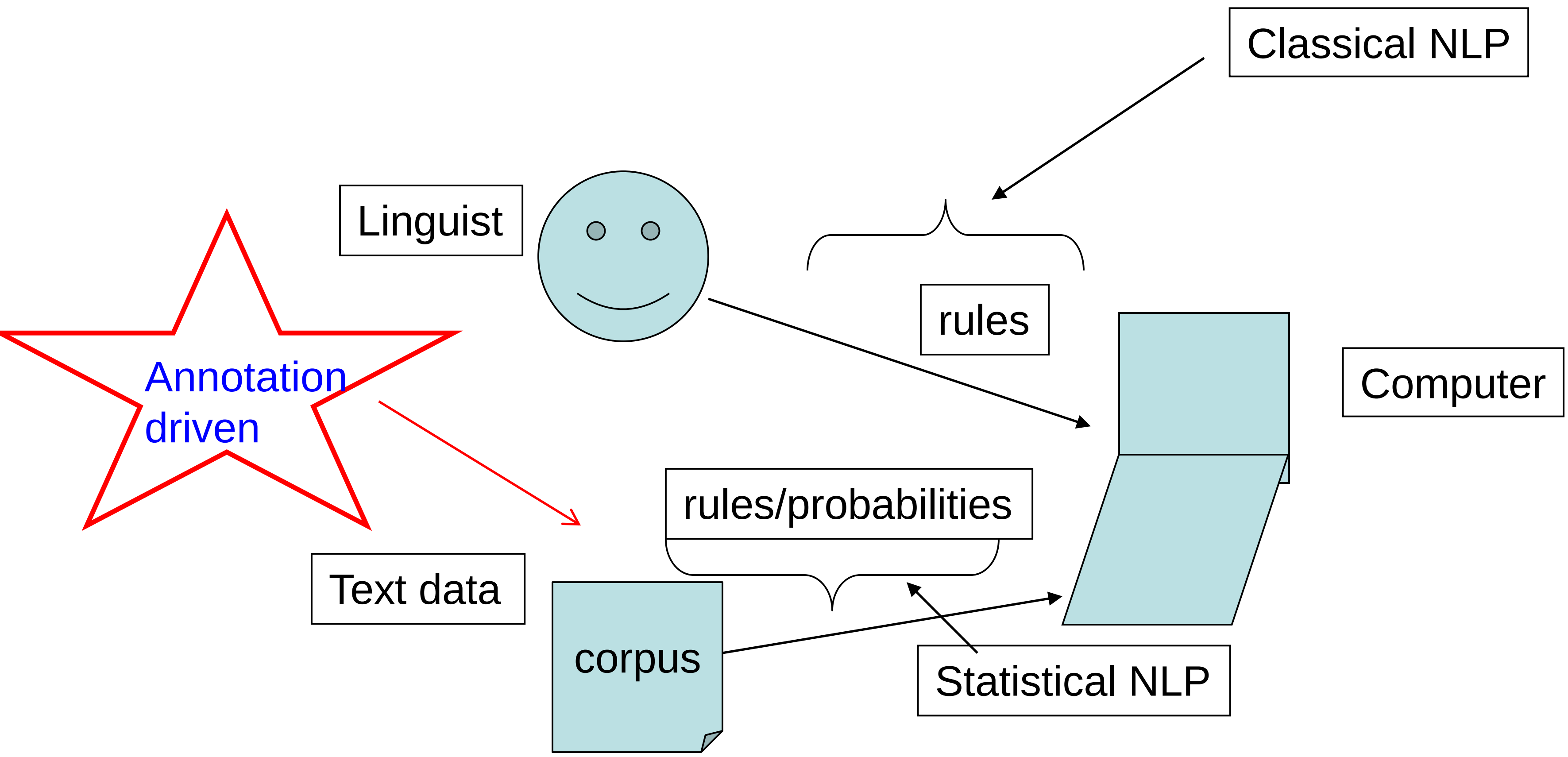
Morphological Isomorphism: Correspondence between suffixes/post-positions in language pairs

Implicit use of linguistic knowledge

- This technique worked because the properties of lexical similarity, structural correspondence and morphological isomorphism hold between related languages
- A linguistic understanding is needed to understand the applicability and viability of NLP techniques
- Many SMT techniques which claim language independence use implicit linguistic knowledge (Bender, 2011)
  - Classical methods of POS tagging and n-gram modelling assume simple morphology and rigid word-order

Emily Bender *On achieving and evaluating language-independence in NLP*. Linguistic Issues in Language Technology. 2011.

Two approaches to NLP: Knowledge Based and ML based



Rules: when and when not

- When the phenomenon is understood AND expressed, rules are the way to go
- "Do not learn when you know!!"
- When the phenomenon "seems arbitrary" at the current state of knowledge, DATA is the only handle!
- Rely on machine learning to tease truth out of data
- Expectation not always met with☹

Why is probability important for NLP

Choose amongst competing options

Impact of probability: Language modeling

Probabilities computed in the context of corpora

1. P(“The sun rises in the east”)
2. P(“The sun rise in the east”)
   - Less probable because of grammatical mistake.
3. P(The svn rises in the east)
   - Less probable because of lexical mistake.
4. P(The sun rises in the west)
   - Less probable because of semantic mistake.

Empiricism vs. Rationalism

- Ken Church, “A Pendulum Swung too Far”, LILT, 2011
  - Availability of huge amount of data: what to do with it?
  - 1950s: Empiricism (Shannon, Skinner, Firth, Harris)
  - 1970s: Rationalism (Chomsky, Minsky)
  - 1990s: Empiricism (IBM Speech Group, AT & T)
  - 2010s: Return of Rationalism?

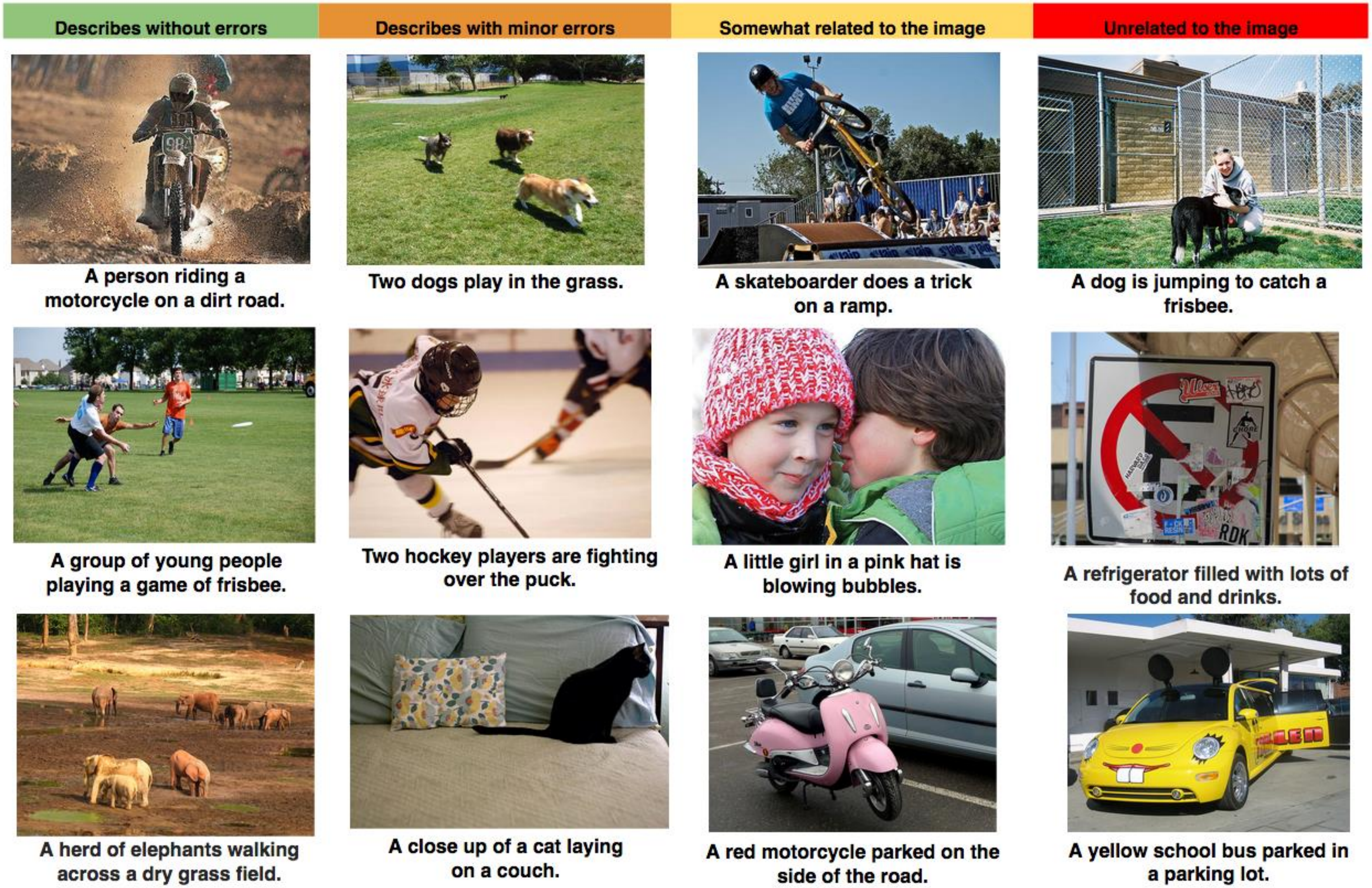
Resource generation will play a vital role in this revival of rationalism

Power of Data

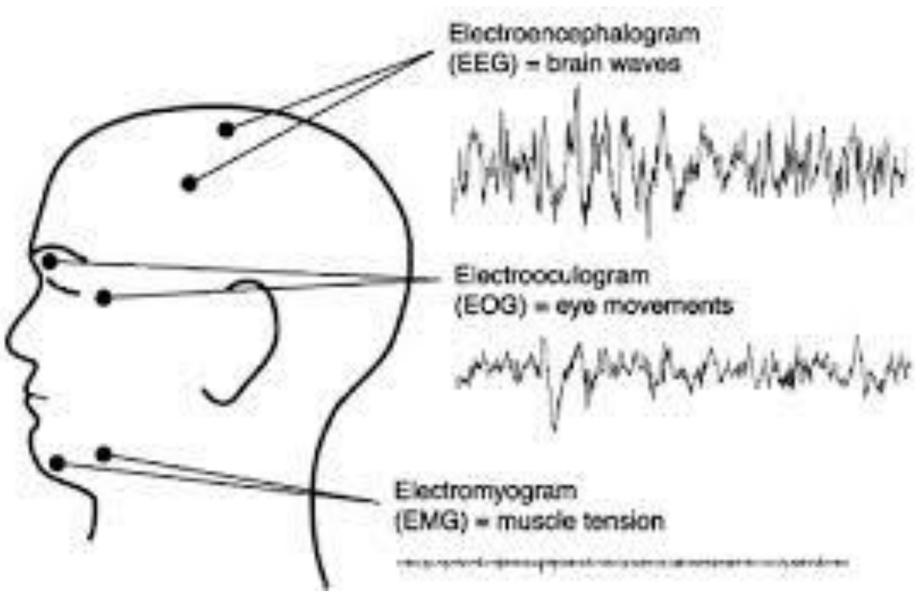
Automatic image labeling
(Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, 2014)

Automatically captioned: "Two pizzas sitting on top of a stove top oven"

Automatic image labeling (cntd)



Thought Reader!



"I am hungry now"

Main methodology

- Object A: extract parts and features
- Object B which is in correspondence with A: extract parts and features
- LEARN mappings of these features and parts
- Use in NEW situations: called DECODING

Some foundational NLP tasks

Part of Speech Tagging

- POS Tagging: attaches to each word in a sentence a part of speech tag from a given set of tags called the **Tag-Set**
- Standard Tag-set : Penn Treebank (for English).

Example

– “_“ The_DT mechanisms_NNS that_WDT make_VBP traditional_JJ hardware_NN are_VBP really_RB being_VBG obsoleted_VBN by_IN microprocessor-based_JJ machines_NNS ,_, ”_” said_VBD Mr._NNP Benton_NNP ._.

Where does POS tagging fit in

Increased Complexity Of Processing

Discourse and Corefernce

Semantics Extraction

Parsing

Chunking

POS tagging

Morphology

Penn tag set

CC	Coord Conjuncn	*and,but,or*	NN	Noun, sing. or mass	*dog*
CD	Cardinal number	*one,two*	NNS	Noun, plural	*dogs*
DT	Determiner	*the,some*	NNP	Proper noun, sing.	*Edinburgh*
EX	Existential there	*there*	NNPS	Proper noun, plural	*Orkneys*
FW	Foreign Word	*mon dieu*	PDT	Predeterminer	*all, both*
IN	Preposition	*of,in,by*	POS	Possessive ending	*'s*
JJ	Adjective	*big*	PP	Personal pronoun	*I,you,she*
JJR	Adj., comparative	*bigger*	PP$	Possessive pronoun	*my,one's*
JJS	Adj., superlative	*biggest*	RB	Adverb	*quickly*
LS	List item marker	*1,One*	RBR	Adverb, comparative	*faster*
MD	Modal	*can,should*	RBS	Adverb, superlative	*fastest*

Penn Tagset cntd.

VB	Verb, base form subsumes imperatives, infinitives and subjunctives
VBD	Verb, past tense includes the conditional form of the verb to be
VBG	Verb, gerund or persent participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
TO	*to*

Language Phenomena

<u>To</u>

1. *I want to dance*
2. *I went to dance*
3. *I went to dance parties*

<u>NNS & VBZ</u>

1. Most English nouns can act as verbs
2. Noun plurals have the same form as 3p1n verbs

Christopher D. Manning. 2011. Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing, 12th International Conference, CICLing 2011, Proceedings, Part I*. Lecture Notes in Computer Science 6608, pp. 171--189.

Indian Language Tag set: Noun

Sl. No	Category			Label	Annotation Convention**	Examples
	Top level	Subtype (level 1)	Subtype (level 2)			
1	**Noun**			**N**	**N**	ladakaa, raajaa, kitaaba
1.1		Common		NN	N__NN	kitaaba, kalama, cashmaa
1.2		Proper		NNP	N__NNP	Mohan, ravi, rashmi
1.4		Nloc		NST	N__NST	Uupara, niice, aage,

Argmax computation (1/2)

Best tag sequence

= T*

= argmax P(T|W)

= argmax P(T)P(W|T) (by Baye's Theorem)

$P(T) = P(t_0=\text{^}\ t_1 t_2 \ldots t_{n+1}=.)$

$= P(t_0)P(t_1|t_0)P(t_2|t_1 t_0)P(t_3|t_2 t_1 t_0) \ldots P(t_n|t_{n-1} t_{n-2} \ldots t_0)P(t_{n+1}|t_n t_{n-1} \ldots t_0)$

$= P(t_0)P(t_1|t_0)P(t_2|t_1) \ldots P(t_n|t_{n-1})P(t_{n+1}|t_n)$

$= \prod_{i=0}^{N+1} P(t_i|t_{i-1})$ Bigram Assumption

Argmax computation (2/2)

$P(W|T) = P(w_0|t_0\text{-}t_{n+1})P(w_1|w_0t_0\text{-}t_{n+1})P(w_2|w_1w_0t_0\text{-}t_{n+1}) \ldots$
$P(w_n|w_0\text{-}w_{n-1}t_0\text{-}t_{n+1})P(w_{n+1}|w_0\text{-}w_nt_0\text{-}t_{n+1})$

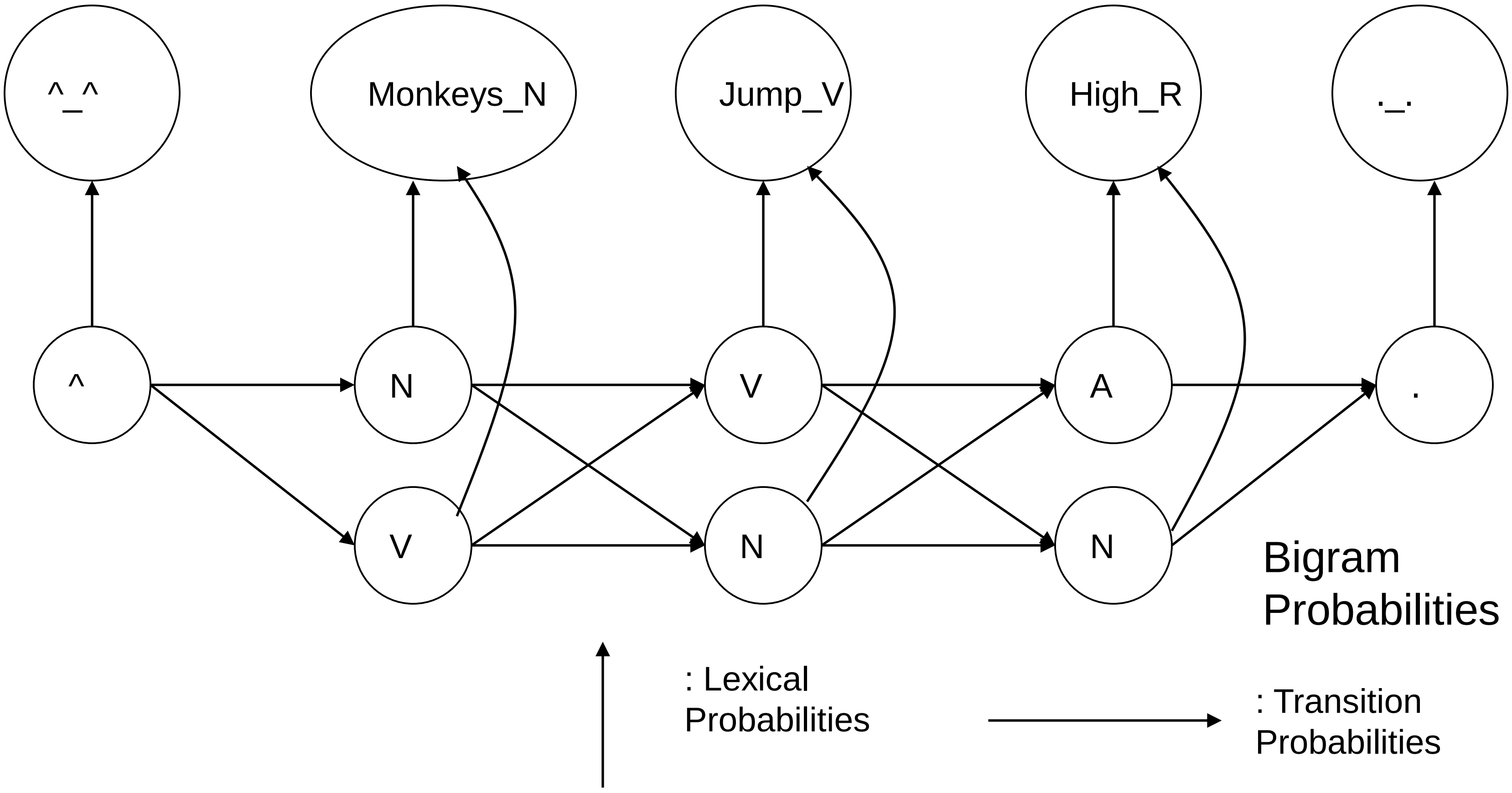
Assumption: A word is determined completely by its tag. This is inspired by speech recognition

$= P(w_o|t_o)P(w_1|t_1) \ldots P(w_{n+1}|t_{n+1})$

$= \prod_{i=0}^{n+1} P(w_i|t_i)$

$= \prod_{i=1}^{n+1} P(w_i|t_i)$ (Lexical Probability Assumption)

Generative Model



Machine Translation and Machine Learning

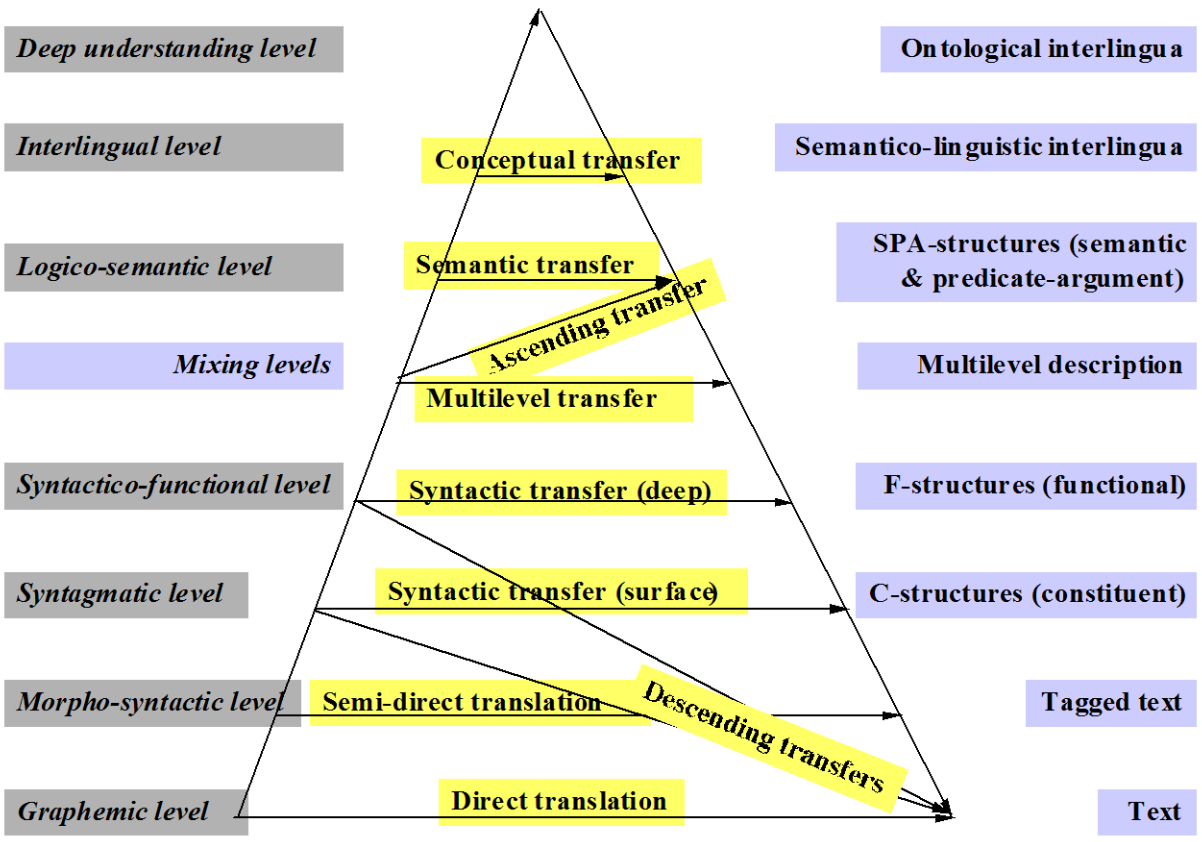
Why is MT difficult: Language Divergence

- Languages have different ways of expressing meaning
  - Lexico-Semantic Divergence
  - Structural Divergence

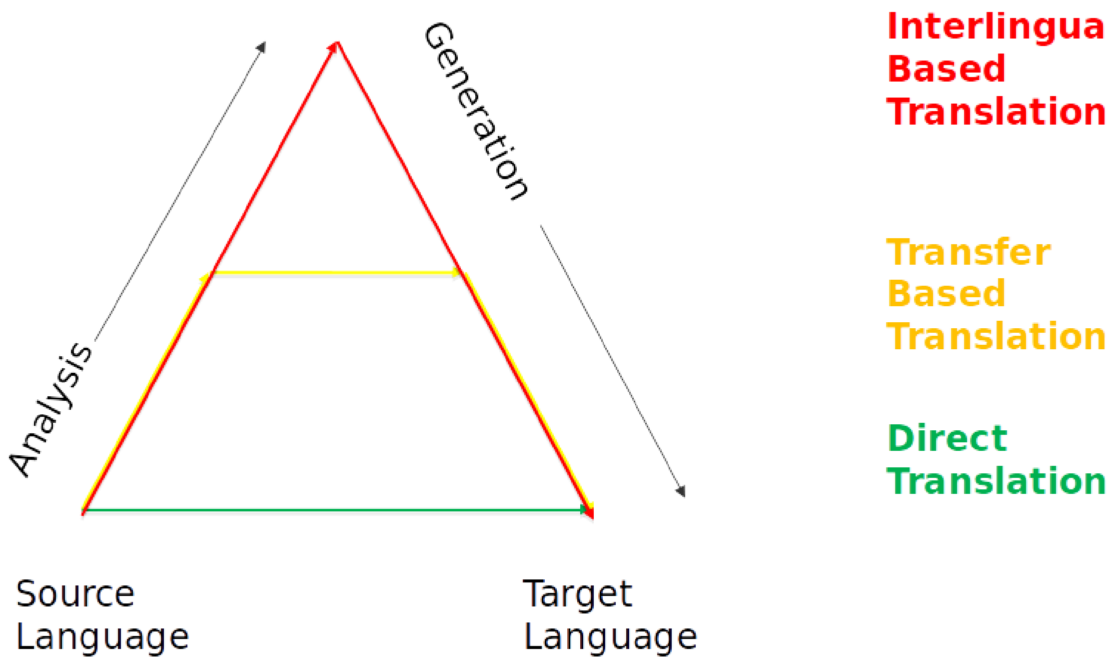
Our work on English-IL Language Divergence with illustrations from Hindi
(Dave, Parikh, Bhattacharyya, Journal of MT, 2002)

Kinds of MT Systems

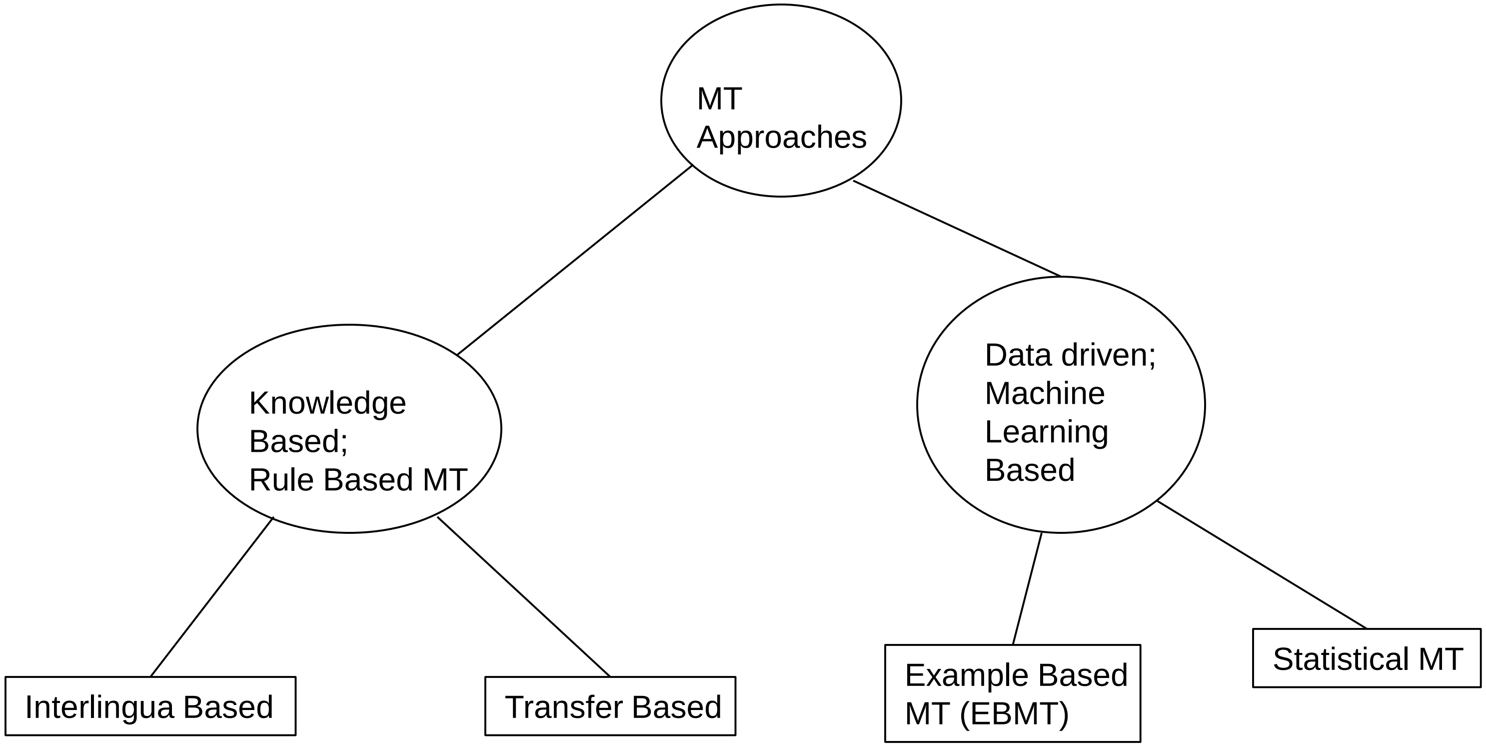
(point of entry from source to the target text)



Simplified Vauquois



Taxonomy of MT systems



RBMT-EBMT-SMT spectrum: knowledge (rules) intensive to data (learning) intensive

RBMT ---------- EBMT ---------- SMT

Can and should choose level of transfer

- **राजा को नमन करो** (Hindi; Indo Aryan)

  raajaa ko naman karo

  HG: king to obeisance do

  Give obeisance to the king (English; Indo-Aryan)

- **राजाला नमन करा** (Marathi; Indo Aryan)

  raajaalaa naman karaa

  king_to obeisance do

- **அரசரை வணங்கு** (Tamil; Dravidian)

  aracarai vanaNku

  king_to obeisance_do

- নিংথৌবু খইরম্মু (Manipuri; Tibeto Burman)

  niNgthoubu khoirammu

  king_to obeisance do

transfer amongst different language families

Language	Inflected Verb/Inflected verb complex	Inflected Noun/Inflected Noun chunk
English	give obeisance	To the king
Hindi	naman karo	raajaa ko
Marathi	naman karaa	raajaalaa
Tamil	vanaNku	aracarai
Manipuri	Khoirammu	niNgthoubu

Data driven translation: Czeck-English data

• [nesu]	“I carry”
• [ponese]	“He will carry”
• [nese]	“He carries”
• [nesou]	“They carry”
• [yedu]	“I drive”
• [plavou]	“They swim”

To translate …

- I will carry.
- They drive.
- He swims.
- They will drive.

Hindi-English data

• [DhotA huM]	“I carry”
• [DhoegA]	“He will carry”
• [DhotA hAi]	“He carries”
• [Dhote hAi]	“They carry”
• [chalAtA huM]	“I drive”
• [tErte hEM]	“They swim”

Bangla-English data

• [bai]	“I carry”
• [baibe]	“He will carry”
• [bay]	“He carries”
• [bay]	“They carry”
• [chAlAi]	“I drive”
• [sAMtrAy]	“They swim”

Word alignment as the crux of Statistical Machine Translation

English	French
(1) three rabbits	(1) trois lapins
a b	w x
(2) rabbits of Grenoble	(2) lapins de Grenoble
b c d	x y z

Initial Probabilities:
each cell denotes *t(a←→ w), t(a←→ x) etc.*

	a	b	c	d
w	1/4	1/4	1/4	1/4
x	1/4	1/4	1/4	1/4
y	1/4	1/4	1/4	1/4
z	1/4	1/4	1/4	1/4

“counts”

a b* ←→ *w x	a	b	c	d
w	1/2	1/2	0	0
x	1/2	1/2	0	0
y	0	0	0	0
z	0	0	0	0

b c d* ←→ *x y z	a	b	c	d
w	0	0	0	0
x	0	1/3	1/3	1/3
y	0	1/3	1/3	1/3
z	0	1/3	1/3	1/3

Revised probabilities table

	a	b	c	d
w	1/2	1/4	0	0
x	1/2	5/12	1/3	1/3
y	0	1/6	1/3	1/3
z	0	1/6	1/3	1/3

“revised counts”

a b ←→ ***w x***	a	b	c	d
w	1/2	3/8	0	0
x	1/2	5/8	0	0
y	0	0	0	0
z	0	0	0	0

b c d ←→ ***x y z***	a	b	c	d
w	0	0	0	0
x	0	5/9	1/3	1/3
y	0	2/9	1/3	1/3
z	0	2/9	1/3	1/3

Re-Revised probabilities table

	a	b	c	d
w	1/2	3/16	0	0
x	1/2	**85/144**	1/3	1/3
y	0	1/9	1/3	1/3
z	0	1/9	1/3	1/3

Continue until convergence; notice that (b,x) binding gets progressively stronger; b=rabbits, x=lapins

Derivation: Key Notations

English vocabulary : V_E
French vocabulary : V_F
No. of observations / sentence pairs : S
Data D which consists of S observations looks like,

$$e^1{}_1, e^1{}_2, \ldots, e^1{}_{l^1} \Leftrightarrow f^1{}_1, f^1{}_2, \ldots, f^1{}_{m^1}$$

$$e^2{}_1, e^2{}_2, \ldots, e^2{}_{l^2} \Leftrightarrow f^2{}_1, f^2{}_2, \ldots, f^2{}_{m^2}$$

.....

$$e^s{}_1, e^s{}_2, \ldots, e^s{}_{l^s} \Leftrightarrow f^s{}_1, f^s{}_2, \ldots, f^s{}_{m^s}$$

.....

$$e^S{}_1, e^S{}_2, \ldots, e^S{}_{l^S} \Leftrightarrow f^S{}_1, f^S{}_2, \ldots, f^S{}_{m^S}$$

No. words on English side in s^{th} sentence : l^s
No. words on French side in s^{th} sentence : m^s
$index_E(e^s{}_p)$ =Index of English word $e^s{}_p$in English vocabulary/dictionary
$index_F(f^s{}_q)$ =Index of French word $f^s{}_q$in French vocabulary/dictionary

(Thanks to Sachin Pawar for helping with the maths formulae processing)

Modeling: Hidden variables and parameters

Hidden Variables (Z) :
Total no. of hidden variables $= \sum_{s=1}^{S} l^s \, m^s$ where each hidden variable is as follows:
$z_{pq}^{s} = 1$, if in s^{th} sentence, p^{th} English word is mapped to q^{th} French word.
$z_{pq}^{s} = 0$, otherwise

Parameters (*Θ*) :
Total no. of parameters $= |V_E| \times |V_F|$, where each parameter is as follows:
$P_{i,j} =$ Probability that i^{th} word in English vocabulary is mapped to j^{th} word in French vocabulary

Likelihoods

Data Likelihood *L(D; Θ)* :

$$L(D;\Theta) = \prod_{s=1}^{S}\prod_{p=1}^{l^s}\prod_{q=1}^{m^s}\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)^{z_{pq}^s}$$

Data Log-Likelihood LL(D; Θ) :

$$LL(D;\Theta) = \sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} z_{pq}^s \log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)$$

Expected value of Data Log-Likelihood E(LL(D; Θ)) :

$$E(LL(D;\Theta)) = \sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s} E(z_{pq}^s)\log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right)$$

Constraint and Lagrangian

$$\sum_{j=1}^{|V_F|} P_{i,j} = 1 \ , \forall i$$

$$\sum_{s=1}^{S} \sum_{p=1}^{l^s} \sum_{q=1}^{m^s} E(z_{pq}^s) \, log\left(P_{index_E(e_p^s),index_F(f_q^s)}\right) - \sum_{i=1}^{|V_E|} \lambda_i \left(\sum_{j=1}^{|V_F|} P_{i,j} - 1 \right)$$

Differentiating wrt P_{ij}

$$\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\left(\frac{E(z_{pq}^s)}{P_{i,j}}\right)-\lambda_i=0$$

$$P_{i,j}=\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\,E(z_{pq}^s)$$

$$\sum_{j=1}^{|V_F|}P_{i,j}=1=\sum_{j=1}^{|V_F|}\frac{1}{\lambda_i}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}\,E(z_{pq}^s)$$

Final E and M steps

M-step

$$P_{i,j} = \frac{\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}E(z_{pq}^s)}{\sum_{j=1}^{|V_F|}\sum_{s=1}^{S}\sum_{p=1}^{l^s}\sum_{q=1}^{m^s}\delta_{index_E(e_p^s),i}\,\delta_{index_F(f_q^s),j}E(z_{pq}^s)}, \forall i,j$$

E-step

$$E(z_{pq}^s) = \frac{P_{index_E(e_p^s),index_F(f_q^s)}}{\sum_{q'=1}^{m^s}P_{index_E(e_p^s),index_F(f_{q'}^s)}}, \forall s,p,q$$

A recent study

PAN Indian SMT

(Anoop K, Abhijit Mishra, Pushpak Bhattacharyya, LREC 2014)

Natural Partitioning of SMT systems

	hi	ur	pa	bn	gu	mr	kK	ta	te	ml	en
hi		61.28	68.21	34.96	51.31	39.12	37.81	14.43	21.38	10.98	29.23
ur	61.42		52.02	29.59	39.00	27.57	28.29	11.95	16.61	8.65	22.46
pa	73.31	56.00		29.89	43.85	30.87	30.72	10.75	18.81	9.11	23.83
bn	37.69	32.08	31.38		28.14	22.09	23.47	10.94	13.40	8.10	18.76
gu	55.66	44.12	45.14	28.50		32.06	30.48	12.57	17.22	8.01	19.78
mr	45.11	32.60	33.28	23.73	32.42		27.81	10.74	12.89	7.65	17.62
kK	41.92	34.00	34.31	24.59	31.07	27.52		10.36	14.80	7.89	17.07
ta	20.48	18.12	15.57	13.21	16.53	11.60	11.87		8.48	6.31	11.79
te	28.88	25.07	25.56	16.57	20.96	14.94	17.27	8.68		6.68	12.34
ml	14.74	13.39	12.97	10.67	9.76	8.39	9.18	5.90	5.94		8.61
en	28.94	22.96	22.33	15.33	15.44	12.11	13.66	6.43	6.55	4.65	

Baseline PBSMT - % BLEU scores (S1)

- **Clear partitioning of translation pairs by language family pairs**, based on translation accuracy.
  - Shared characteristics within language families make translation simpler
  - Divergences among language families make translation difficult

Using Bridge to mitigate resource scarcity L1→bridge→L2 *(Wu and Wang 2009)*

- Resource rich and resource poor language pairs
- Question-1: How about translating through a 'bridge'?
- Question-2: how to choose the bridge?

Mathematical preliminaries

$$e_{best} = \arg\max_{\mathbf{e}} p(\mathbf{e}|\mathbf{f})$$
$$= \arg\max_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p_{LM}(\mathbf{e})$$

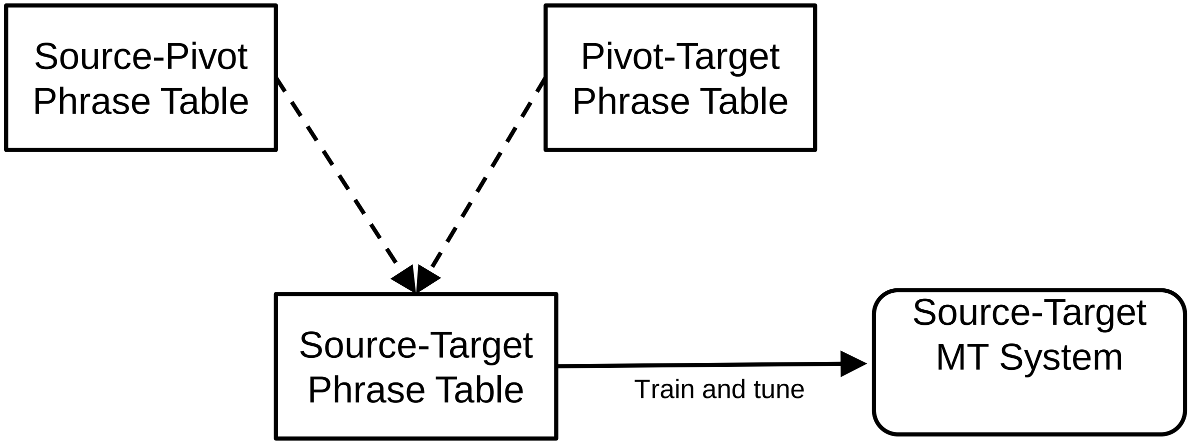
Where p(***f***|***e***) is given by:

$$p(\mathbf{f}|\mathbf{e}) = p\left(\bar{f}^I \middle| \bar{e}^I\right) = \prod_{i=1}^{I} \emptyset\left(\bar{f}_i \middle| \bar{e}_i\right) d(a_i - b_{i-1}) p_w\left(\bar{f}_i \middle| \bar{e}_i, a\right)^{\gamma}$$

$$\emptyset\left(\bar{f}_i \middle| \bar{e}_i\right) = \sum_{\bar{p}_i} \emptyset\left(\bar{f}_i \middle| \bar{p}_i\right) \emptyset\left(\bar{p}_i \middle| \bar{e}_i\right)$$

$$p_w\left(\bar{f}_i \middle| \bar{e}_i, a\right) = \prod_{l=1}^{n} \frac{1}{|m/(l,m) \in a|} \sum_{\forall (l,m) \in a} w(f_l/e_l)$$

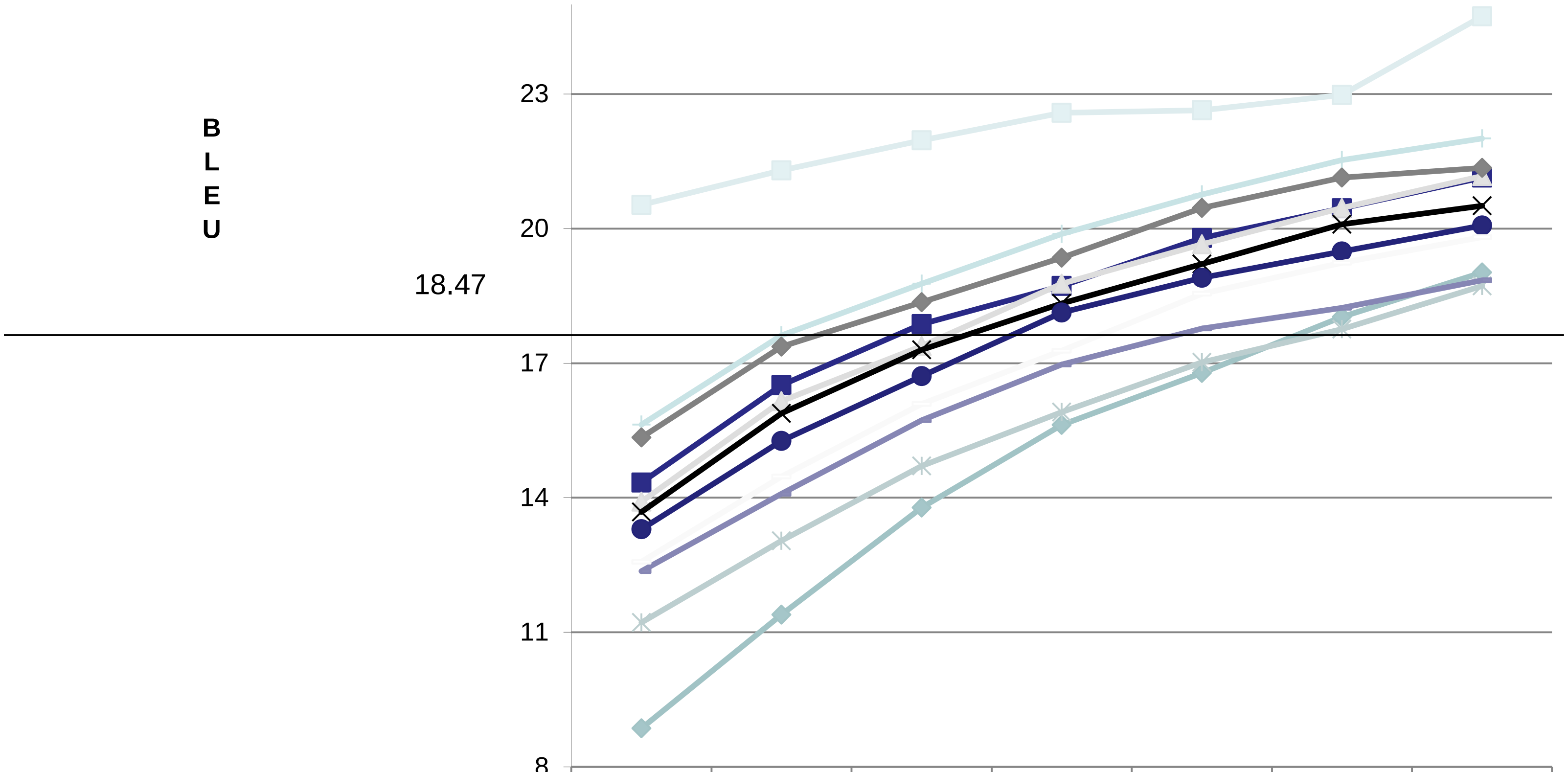
Triangulation approach



- Important to induce language dependent components such as phrase translation probability and lexical weight

English-Hindi SMT: Resource Details

Segment	#Sentences	#Unique Words
Training	46277	39452 (en), 41418 (hi)
Tuning	500	2623 (en), 2816 (hi)
Test	2000	6722 (en), 7102 (hi)
Monolingual	1538429	558206

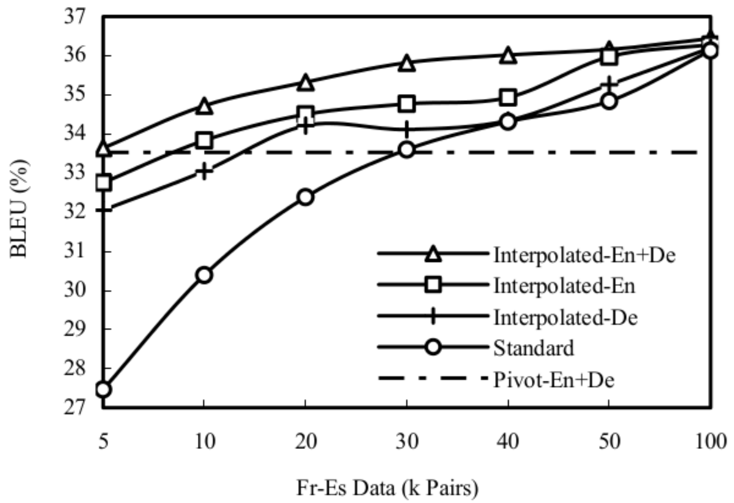


	l=1k	l=2k	l=3k	l=4k	l=5k	l=6k	l=7k
DIRECT_l	8.86	11.39	13.78	15.62	16.78	18.03	19.02
DIRECT_l+BRIDGE_BN	14.34	16.51	17.87	18.72	19.79	20.45	21.14
DIRECT_l+BRIDGE_GU	13.91	16.15	17.38	18.77	19.65	20.46	21.17
DIRECT_l+BRIDGE_KK	13.68	15.88	17.3	18.33	19.21	20.1	20.51
DIRECT_l+BRIDGE_ML	11.22	13.04	14.71	15.91	17.02	17.76	18.72
DIRECT_l+BRIDGE_MA	13.3	15.27	16.71	18.13	18.9	19.49	20.07
DIRECT_l+BRIDGE_PU	15.63	17.62	18.77	19.88	20.76	21.53	22.01
DIRECT_l+BRIDGE_TA	12.36	14.09	15.73	16.97	17.77	18.23	18.85
DIRECT_l+BRIDGE_TE	12.57	14.47	16.09	17.28	18.55	19.24	19.81
DIRECT_l+BRIDGE_UR	15.34	17.37	18.36	19.35	20.46	21.14	21.35
DIRECT_l+BRIDGE_PU_UR	20.53	21.3	21.97	22.58	22.64	22.98	24.73

Effect of Multiple Pivots

Fr-Es translation using 2 pivots

Source: Wu & Wang (2007)



- Raj Dabre, Fabien Cromiere, Sadao Kurohash and Pushpak Bhattacharyya, Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages, **NAACL 2015**, Denver, Colorado, USA, May 31 - June 5, 2015.

Hi-Ja translation using 7 pivots

Source: Dabre et al (2015)

System	**Ja→Hi**	**Hi→Ja**
Direct	33.86	37.47
Direct+best pivot	35.74 (es)	39.49 (ko)
Direct+Best-3 pivots	38.22	41.09
Direct+All 7 pivots	38.42	40.09

Annotation

Definition

(*Eduard Hovy, ACL 2010, tutorial on annotation)*

- Annotation (‘tagging’) is the process of adding new information into raw data by human annotators.

- Typical annotation steps:
  - Decide which fragment of the data to annotate
  - Add to that fragment a specific bit of information
  - chosen from a fixed set of options

Example of annotation: sense marking

एक_4187 नए शोध_1138 के अनुसार_3123 जिन लोगों_1189 का सामाजिक_43540 जीवन_125623 व्यस्त_48029 होता है उनके दिमाग_16168 के एक_4187
हिस्से_120425 में अधिक_42403 जगह_113368 होती है।

(According to a new research, those people who have a busy social life, have larger space in a part of their brain).

नेचर न्यूरोसाइंस में छपे एक_4187 शोध_1138 के अनुसार_3123 कई_4118 लोगों_1189 के दिमाग_16168 के स्कैन से पता_11431 चला कि दिमाग_16168 का एक_4187 हिस्सा_120425 एमिगडाला सामाजिक_43540 व्यस्तताओं_1438 के साथ_328602 सामंजस्य_166
के लिए थोड़ा_38861 बढ़_25368 जाता है। यह शोध_1138 58 लोगों_1189 पर किया गया जिसमें उनकी उम्र_13159 और दिमाग_16168 की साइज़ के आँकड़े_128065
लिए गए। अमरीकी_413405 टीम_14077 ने पाया_227806 कि जिन लोगों_1189 की सोशल नेटवर्किंग अधिक_42403 है उनके दिमाग_16168 का एमिगडाला
वाला हिस्सा_120425 बाकी_130137 लोगों_1189 की तुलना_में_38220 अधिक_42403 बड़ा_426602 है। दिमाग_16168 का एमिगडाला वाला हिस्सा_120425
भावनाओं_1912 और मानसिक_42151 स्थिति_1652 से जुड़ा हुआ माना_212436 जाता है।

Ambiguity of लोगों (People)

- **लोग**, **जन**, **लोक**, **जनमानस**, **पब्लिक** - एक से अधिक व्यक्ति *"लोगों के हित में काम करना चाहिए"*
  - **(English synset) multitude, masses, mass, hoi_polloi, people, the_great_unwashed - the common people generally *"separate the warriors from the mass" "power to the people"***
- **दुनिया**, **दुनियाँ**, **संसार**, **विश्व**, **जगत**, **जहाँ**, **जहान**, **ज़माना**, **जमाना**, **लोक**, **दुनियावाले**, **दुनियाँवाले**, **लोग** - संसार में रहने वाले लोग *"महात्मा गाँधी का सम्मान पूरी दुनिया करती है / मैं इस दुनिया की परवाह नहीं करता / आज की दुनिया पैसे के पीछे भाग रही है"*
  - **(English synset) populace, public, world - people in general considered as a whole *"he is a hero in the eyes of the public"***

Structural annotation

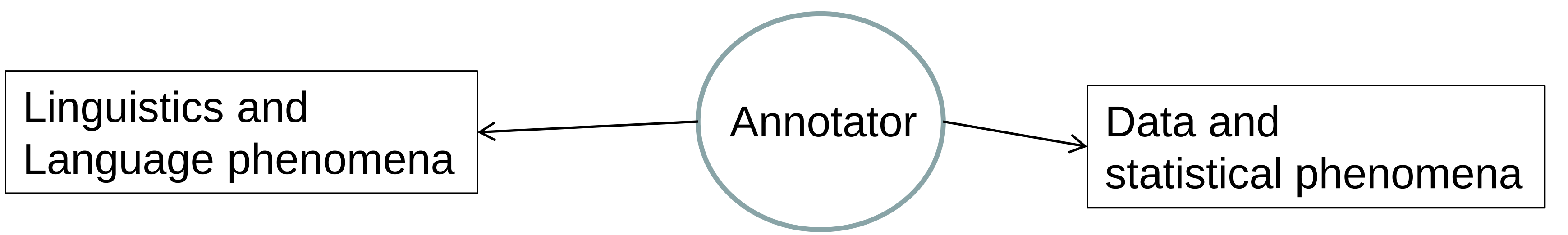
Raw Text: "My dog also likes eating sausage."

```
(ROOT
      (S
         (NP
            (PRP$ My) (NN dog))
         (ADVP (RB also))
         (VP (VBZ likes)
            (S (VP (VBG eating)
               (NP (NN sausage))))) (. .)))
```

poss(dog-2, My-1)
nsubj(likes-4, dog-2)
advmod(likes-4, also-3)
root(ROOT-0, likes-4)
xcomp(likes-4, eating-5)
dobj(eating-5, sausage-6)

Good annotators and good annotation designers are rare to find

- An annotator has to understand BOTH language phenomena and the data
- An annotation designer has to understand BOTH linguistics and statistics!



Scale of effort involved in annotation (1/2)

- Penn Treebank
  - Total effort: *8 million words, 20-25 man years (5 persons for 4-5 years)*
- Ontonotes: Annotate 300K words per year (*1 person per year*)
  - news, conversational telephone speech, weblogs, usenet newsgroups, broadcast, talk shows,
  - with structural information (syntax and predicate argument structure) and shallow semantics (word sense linked to an ontology and coreference)
  - in English, Chinese, and Arabic
- Prague Discourse Treebank (Czeck): 500,000 words, 20-25 man years (*4-5 persons for 5 years*)

Scale of effort in annotation (2/2)

Sense marked corpora created at IIT Bombay

- http://www.cfilt.iitb.ac.in/wsd/annotated_corpus
- English: Tourism (~170000), Health (~150000)
- Hindi: Tourism (~170000), Health (~80000)
- Marathi: Tourism (~120000), Health (~50000)
  - 6 man years for each <L,D> combination (3 persons for 2 years)

Serious world wide effort on leveraging multiliguality

- Greg Durrett, Adam Pauls, and Dan Klein, *Syntactic Transfer Using Bilingual Lexicon*, EMNLP-CoNLL, 2012
- Balamurali A.R., Aditya Joshi and Pushpak Bhattacharyya, *Cross-Lingual Sentiment Analysis for Indian Languages using Wordent Synsets*, COLING 2012
- Dipanjan Das and Slav Petrov, *Unsupervised Part of Speech Tagging with Bilingual Graph-Based Projections,* ACL, 2011
- Benjamin Snyder, Tahira Naseem, and Regina Barzilay, *Unsupervised multilingual grammar induction*, ACL-IJCNLP, 2009

Cooperative Word Sense Disambiguation

Definition: WSD

- Given a context:
  - Get "meaning"s of
    - *a set of words (targetted wsd)*
    - or all words (*all words wsd*)
- The "Meaning" is usually given by the id of senses in a sense repository
  - usually the wordnet

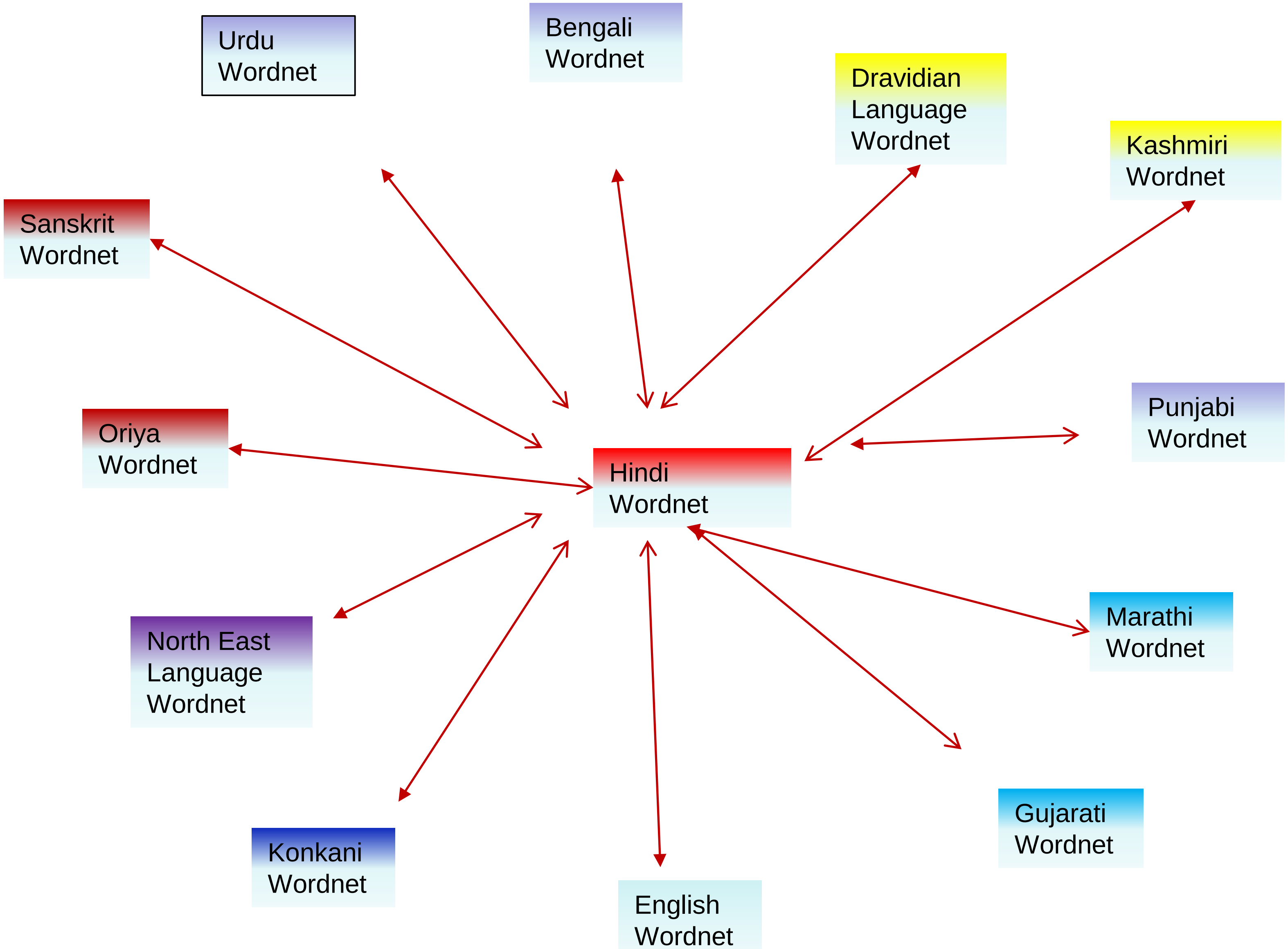
Example: “*operation*” (from Princeton Wordnet)

- **Operation**, surgery, surgical operation, surgical procedure, surgical process -- (a medical procedure involving an incision with instruments; performed to repair damage or arrest disease in a living body; "they will schedule the operation as soon as an operating room is available"; "he died while undergoing surgery") TOPIC->(noun) surgery#1

- **Operation**, military operation -- (activity by a military or naval force (as a maneuver or campaign); "it was a joint operation of the navy and air force") TOPIC->(noun) military#1, armed forces#1, armed services#1, military machine#1, war machine#1

- mathematical process, mathematical **operation**, **operation** -- ((mathematics) calculation by mathematical methods; "the problems at the end of the chapter demonstrated the mathematical processes involved in the derivation"; "they were learning the basic operations of arithmetic") TOPIC->(noun) mathematics#1, math#1, maths#1

WSD for ALL Indian languages:
Critical resource: I**NDOWORDNET**



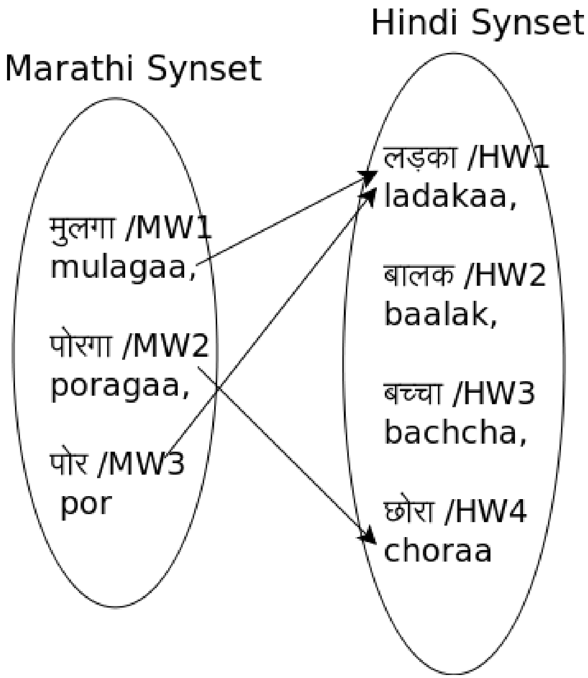
Synset Based Multilingual Dictionary

Concepts	L1 (English)	L2 (Hindi)	L3 (Marathi)
04321: a youthful male person	{malechild, boy}	{लड़का (*ladkaa*), बालक (*baalak*), बच्चा (*bachchaa*)}	{मुलगा (*mulgaa*), पोरगा (*porgaa*), पोर (*por*)}

A sample entry from the *MultiDict*

- Expansion approach for creating wordnets [Mohanty et. al., 2008]
- Instead of creating from scratch link to the synsets of existing wordnet
- Relations get borrowed from existing wordnet

Cross Linkages Between Synset Members



- Captures native speakers intuition
- Wherever the word *ladkaa* appears in Hindi one would expect to see the word *mulgaa* in Marathi
- A few wordnet pairs do not have explicit word linkages within synset, in which case one assumes every word is linked all words on the other side

Resources for WSD- wordnet and corpora: 5 scenarios

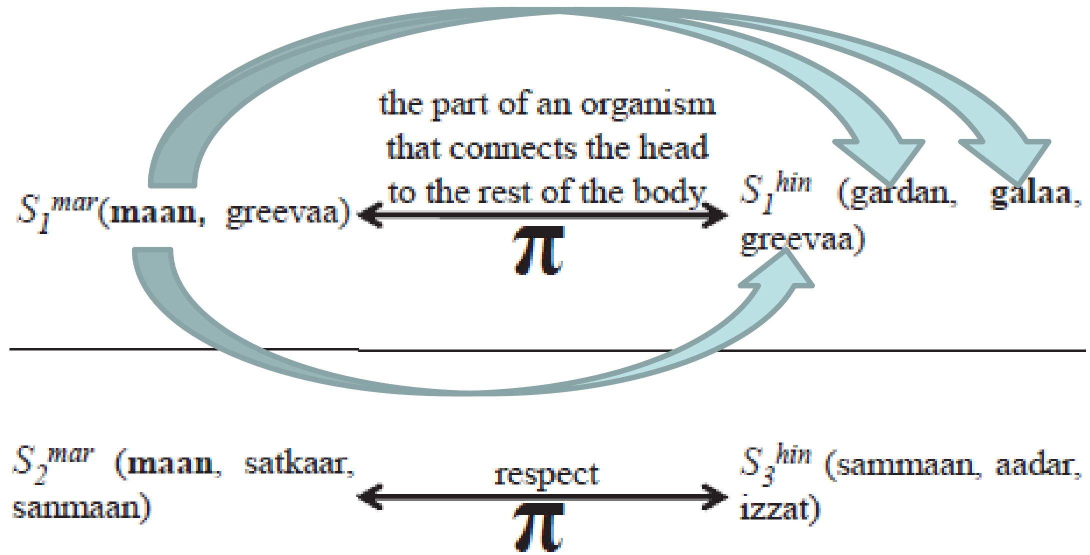
	Annotated Corpus in L1	*Aligned Wordnets*	*Annotated Corpus in L2*
Scenario 1	✓	✓	✗
Scenario 2	✓	✓	✗
Scenario 3	✓	✓	***Varies***
Scenario 4	✗	✓	✗
Scenario 5	***Seed***	✓	***Seed***

Unsupervised WSD

(No annotated corpora)

Khapra, Joshi and Bhattacharyya, IJCNLP 2011

ESTIMATING SENSE DISTRIBUTIONS

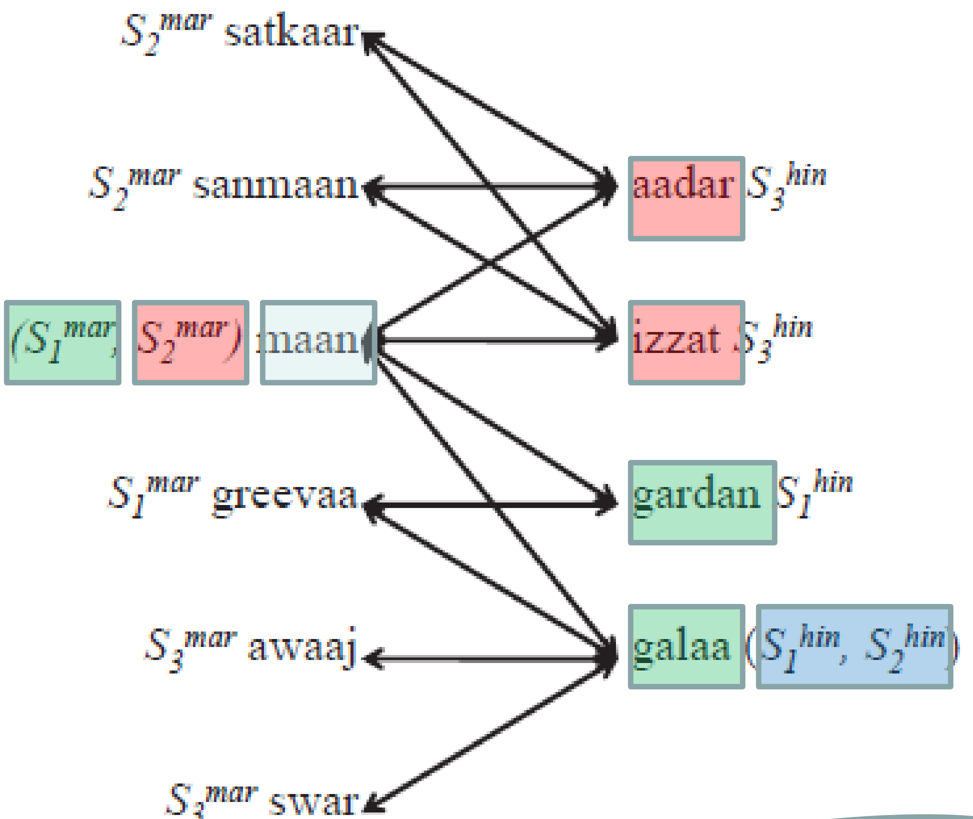


If sense tagged Marathi corpus were available, we could have estimated

$$P(S_1^{mar}|maan) = \frac{\#(S_1^{mar}, maan)}{\#(S_1^{mar}, maan) + \#(S_2^{mar}, maan)}$$

But such a corpus is not available

EM for estimating sense distributions



E-Step

$$P(S_1^{mar}|maan) = \frac{\#(gardan) + \ldots \cdot \#(gala)}{\#(gardan) + \ldots \cdot \#(gala) + \ldots) \cdot \#(aadar) + \ldots \cdot \#(izzat)}$$

M-Step

$$P(S_1^{hin}|gala) = \frac{P(S_1^{mar}|maan) \cdot \#(maan) + P(S_1^{mar}|greeva) \cdot \#(greeva)}{P(S_1^{mar}|maan) \cdot \#(maan) + P(S_1^{mar}|greeva) \cdot \#(greeva) + P(S_3^{mar}|awaaj) \cdot \#(awaaj) + P(S_3^{mar}|swar) \cdot \#(swar)}$$

Results & Discussions

Our values

Algorithms	Tourism			Health			
	P%	R%	F%	P%	R%	F%	
MCL	73.36	68.83	71.02	75.86	66.6	70.93	Manual Cross Linkages
PCL	68.57	67.93	68.25	65.75	64.53	65.14	Probabilistic Cross Linkages
IWSD-Self	78.36	77.77	78.07	78.15	75.91	77.01	Skyline - self training data is available
WFS	57.15	57.15	57.15	55.55	55.55	55.55	Wordnet first sense baseline
PPR	51.49	51.49	51.49	48.32	48.32	48.32	S-O-T-A Knowledge Based Approach
Unsup	9.01	9.01	9.01	9.72	9.72	9.72	S-O-T-A Unsupervised Approach

- Performance of projection using manual cross linkages is within 7% of Self-Training
- Performance of projection using probabilistic cross linkages is within 10-12% of Self-Training – remarkable since no additional cost incurred in target language
- Both MCL and PCL give 10-14% improvement over Wordnet First Sense Baseline
- *Not prudent to stick to knowledge based and unsupervised approaches – they come nowhere close to MCL or PCL*

Harnessing Context Incongruity for Sarcasm Detection

1. Aditya Joshi, Vinita Sharma, Pushpak Bhattacharyya*, Harnessing Context Incongruity for Sarcasm Detection*, ACL 2015

2. Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya and Mark Carman, *Are Word Embedding-based Features Useful for Sarcasm Detection?*, EMNLP 2016

Goal

The relationship between context incongruity and sarcasm has been studied in linguistics.

We present a statistical system that harnesses context incongruity as a basis for sarcasm detection in the form of **two kinds of incongruity features: explicit and implicit**.

Context Incongruity

- Incongruity is defined as *‘the state of being not in agreement, as with principles’.*

- Ivanko and Pexman (2003) state that the sarcasm processing time (time taken by humans to understand sarcasm) depends on the **degree of context incongruity** between the statement and the context.

Two kinds of incongruity

- **Explicit incongruity**
  - Overtly expressed through sentiment words of both polarities
  - Contribute to almost 11% of sarcasm instances

    'I <u>love</u> being <u>ignored</u>'
- **Implicit incongruity**
  - Covertly expressed through phrases of implied sentiment

    'I <u>love</u> this paper so much that I <u>made a doggy bag out of</u> it'

Feature Set

Lexical	
Unigrams	Unigrams in the training corpus
Pragmatic	
Capitalization	Numeric feature indicating presence of capital letters
Emoticons & laughter expressions	Numeric feature indicating presence of emoticons and 'lol's
Punctuation marks	Numeric feature indicating presence of punctuation marks
Implicit Incongruity (Based on Riloff et al	
Implicit Sentiment Phrases	Boolean feature indicating phrases extracted from the implicit phrase extraction step
Explicit Incongruity (Based on Ramteke et al	
#Explicit incongruity	Number of times a word is followed by a word of opposite polarity
Largest positive /negative subsequence	Length of largest series of words with polarity unchanged
#Positive words	Number of positive words
#Negative words	Number of negative words
Lexical Polarity	Polarity of a tweet based on words present

Datasets

Name	Text-form	Method of labeling	Statistics
Tweet-A	Tweets	Using sarcasm-based hashtags as labels	5208 total, 4170 sarcastic
Tweet-B	Tweets	Manually labeled (Given by Riloff et al(2013))	2278 total, 506 sarcastic
Discussion-A	Discussion forum posts (IAC Corpus)	Manually labeled (Given by Walker et al (2012))	1502 total, 752 sarcastic

Results

Features	P	R	F
Original Algorithm by Riloff et al. (2013)			
Ordered	0.774	0.098	0.173
Unordered	0.799	0.337	0.474
Our system			
Lexical (**Baseline**)	0.820	0.867	0.842
Lexical+Implicit	0.822	0.887	0.853
Lexical+Explicit	0.807	0.985	0.8871
All features	0.814	0.976	**0.8876**

Tweet-A

Approach	P	R	F
Riloff et al. (2013) (**best reported**)	0.62	0.44	0.51
Maynard and Greenwood (2014)	0.46	0.38	0.41
Our system (all features)	**0.77**	**0.51**	**0.61**

Tweet-B

Features	P	R	F
Lexical (**Baseline**)	0.645	0.508	0.568
Lexical+Explicit	0.698	0.391	0.488
Lexical+Implicit	0.513	0.762	0.581
All features	0.489	0.924	**0.640**

Discussion-A

When explicit incongruity is absent

A <u>woman</u> needs a <u>man</u> like a <u>fish</u> needs <u>bicycle</u>

Word2Vec similarity(man,woman) = 0.766
Word2Vec similarity(fish, bicycle) = 0.131

Can word embedding-based features when augmented to features reported in prior work improve the performance of sarcasm detection?

Word embedding-based features

(Stop words removed)

Unweighted similarity features (S):
For every word and word pair,
1) Maximum score of most similar word pair
2) Minimum score of most similar word pair
3) Maximum score of most dissimilar word pair
4) Minimum score of most dissimilar word pair

Distance-weighted similarity features (WS): 4 S features weighted by linear distance between the two words

Both (S+WS): 8 features

	man	**woman**	**fish**	**needs**	**bicycle**
man	-	**0.766**	**0.151**	**0.078**	**0.229**
woman	0.766	-	0.084	0.060	0.229
fish	0.151	0.084	-	0.022	0.130
needs	0.078	0.060	0.022	-	0.060
bicycle	0.229	0.229	0.130	0.060	-

Experiment setup

- **Dataset:** 3629 Book snippets (759 sarcastic) downloaded from GoodReads website. Labeled by users with tags. We download ones with ‘sarcasm’ as sarcastic, ones with ‘philosophy’ as non-sarcastic
- Five-fold cross-validation
- **Classifier:** SVM-Perf optimised for F-score
- **Configurations**:
  - Four prior works (augmented with our sets of features)
  - Four implementations of word embeddings (Word2Vec, LSA, GloVe, Dependency weights-based)

Results (1/2)

Features	**P**	**R**	**F**
	Baseline		
Unigrams	67.2	78.8	72.53
S	64.6	75.2	69.49
WS	67.6	51.2	58.26
Both	67	52.8	59.05

- **Observation**: Only word embedding-based features will not suffice. ‘Augmentation’ to other known useful features necessary

Results (2/2)

	LSA			GloVe			Dependency Weights			Word2Vec		
	P	R	F	P	R	F	P	R	F	P	R	F
L	73	79	75.8	73	79	75.8	73	79	75.8	73	79	75.8
+S	81.8	78.2	**79.95**	81.8	79.2	**80.47**	81.8	78.8	80.27	80.4	80	**80.2**
+WS	76.2	79.8	77.9	76.2	79.6	77.86	81.4	80.8	81.09	80.8	78.6	79.68
+S+WS	77.6	79.8	78.68	74	79.4	76.60	82	80.4	**81.19**	81.6	78.2	79.86
G	84.8	73.8	78.91	84.8	73.8	78.91	84.8	73.8	**78.91**	84.8	73.8	**78.91**
+S	84.2	74.4	**79**	84	72.6	77.8	84.4	72	77.7	84	72.8	78
+WS	84.4	73.6	78.63	84	75.2	**79.35**	84.4	72.6	78.05	83.8	70.2	76.4
+S+WS	84.2	73.6	78.54	84	74	78.68	84.2	72.2	77.73	84	72.8	78
B	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61
+S	78.2	75.6	**76.87**	80.4	76.2	**78.24**	81.2	74.6	**77.76**	81.4	72.6	76.74
+WS	75.8	77.2	76.49	76.6	77	76.79	76.2	76.4	76.29	81.6	73.4	77.28
+S+WS	74.8	77.4	76.07	76.2	78.2	77.18	75.6	78.8	77.16	81	75.4	**78.09**
J	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43
+S	84.8	73.8	78.91	85.6	74.8	79.83	85.4	74.4	79.52	85.4	74.6	**79.63**
+WS	85.6	75.2	**80.06**	85.4	72.6	78.48	85.4	73.4	78.94	85.6	73.4	79.03
+S+WS	84.8	73.6	78.8	85.8	75.4	**80.26**	85.6	74.4	**79.6**	85.2	73.2	78.74

Table 3: Performance obtained on augmenting word embedding features to features from four prior works, for four word embeddings; L: Liebrecht et al. (2013), G: González-Ibánez et al. (2011a), B: Buschmeier et al. (2014) , J: Joshi et al. (2015)

- **Observation**: Using word embedding-based features improves sarcasm detection, for multiple word embedding types and feature sets

Multiword Expressions

About half the lexical items in most languages are multiwords!

Multi-Word Expressions (MWE)

- Necessary Condition
  - Word sequence separated by space/delimiter
- Sufficient Conditions
  - Non-compositionality of meaning
  - Fixity of expression
    - In lexical items
    - In structure and order

Examples – Necessary condition

- Non-MWE example:
  - Marathi: सरकार हक्काबक्का झाले
  - Roman: sarakAra HakkAbakkA JZAle
  - Meaning: government was surprised
- MWE example:
  - Hindi: गरीब नवाज़
  - Roman: garIba navAjZa
  - Meaning: who nourishes poor

Examples - Sufficient conditions (Non-compositionality of meaning)

- Konkani: पोटांत चाबता
- Roman: poTAMta cAbatA
- Meaning: to feel jealous

- Telugu: చెట్టు కింది ప్లీడరు
- Roman: ceVttu kiMXa pLIdaru
- Meaning: an idle person

- Bangla: মাটির মানুষ
- Roman: mAtira mAnuSa
- Meaning: a simple person/son of the soil

Examples – Sufficient conditions (Fixity of expression)

In lexical items

- Hindi
  - usane muJe gAlI dI
  - *usane muJe galI pradAna kI
- Bangla
  - jabajjIbana karadaMda
  - *jIbanabhara karadaMda
  - *jabajjIbana jela
- English (1)
  - life imprisonment
  - *lifelong imprisonment
- English (2)
  - Many thanks
  - *Plenty of thanks

Examples – Sufficient conditions (In structure and order)

- English example
  - kicked the bucket (died)
  - the bucket was kicked
    (not passivizable in the sense of dying)
- Hindi example
  - उम्र क़ैद
  - umra kEda (life imprisonment)
  - umra bhara kEda

MW task (NLP + ML)

NLP

ML

	String + Morph	*POS*	*POS+ WN*	*POS + List*	*Chun k-ing*	*Parsing*
Rules	Onomaetopic Redupli-cation *(tik tik, chham chham)*	Non-Onomaetopic Redupli-cation *(ghar ghar)*	Non-redup (Syn, Anto, Hypo) *(raat din, dhan doulat)*			Non-contiguous something
Statistical		Colloctions or fixed expressions *(many thanks)*		Conjunct verb (verbalizer list), Compund verb (verctor verb list) *(salaha dena, has uthama)*		Non-contiguous Complex Predicate

Idioms will be list morph + look up

Summary

- POS tagging: done by ML predominantly
- Alignment in MT: predominantly ML; but cannot do without linguistics when facing rich morphology
- Co-operative WSD
  - Good linguistics (high quality linked wordnets) + Good ML (novel EM formulation)
- Sarcasm (difficult sentiment analysis problem)
  - Good NLP (incongruity) + good ML (string kernels?)
- MWE processing: FIXITY or colocation: ML is the only way; no apparent *reason* for fixity.

Conclusions

- Both Linguistics and Computation needed: **Linguistics is the eye, Computation the body**

- It is possible to leverage the resources created for one language in another

- Language phenomenon → Formalization → Hypothesis formation → Experimentation → Interpretation (Natural Science like flavor)

- Theory=Linguistics+NLP, Technique=ML

URLS

(publications) http://www.cse.iitb.ac.in/~pb

(resources) http://www.cfilt.iitb.ac.in

Thank you

Questions?

Word embedding-based features for sarcasm detection

(To appear in EMNLP 2016)

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya and Mark Carman,
Are Word Embedding-based Features Useful for Sarcasm Detection?, EMNLP 2016, Austin, Texas, USA, November 1-5, 2016.

Introduction

- Sarcasm detection is the task of predicting whether a given piece of text is sarcastic
- The ruling paradigm in sarcasm detection research is to design features that incorporate contextual information to understand context incongruity that lies at the heart of sarcasm
- 'I love being ignored' : Incorporating context incongruity using sentiment flips
- What happens in case of sentences with few or no sentiment words?

Motivation

A <u>woman</u> needs a <u>man</u> like a <u>fish</u> needs <u>bicycle</u>

Word2Vec similarity(man,woman) = 0.766
Word2Vec similarity(fish, bicycle) = 0.131

Can word embedding-based features when augmented to features reported in prior work improve the performance of sarcasm detection?

Word embedding based features

(Stop words removed)

Unweighted similarity features (S):
For every word and word pair,
1) Maximum score of most similar word pair
2) Minimum score of most similar word pair
3) Maximum score of most dissimilar word pair
4) Minimum score of most dissimilar word pair

Distance-weighted similarity features (WS): 4 S features weighted by linear distance between the two words

Both (S+WS): 8 features

	man	woman	fish	needs	bicycle
man	-	**0.766**	**0.151**	**0.078**	**0.229**
woman	0.766	-	0.084	0.060	0.229
fish	0.151	0.084	-	0.022	0.130
needs	0.078	0.060	0.022	-	0.060
bicycle	0.229	0.229	0.130	0.060	-

Experiment setup

- **Dataset:** 3629 Book snippets (759 sarcastic) downloaded from GoodReads website. Labeled by users with tags. We download ones with ‘sarcasm’ as sarcastic, ones with ‘philosophy’ as non-sarcastic
- Five-fold cross-validation
- **Classifier:** SVM-Perf optimised for F-score
- **Configurations**:
  - Four prior works (augmented with our sets of features)
  - Four implementations of word embeddings (Word2Vec, LSA, GloVe, Dependency weights-based)

Results (1/2)

Features	P	R	F
	Baseline		
Unigrams	67.2	78.8	72.53
S	64.6	75.2	69.49
WS	67.6	51.2	58.26
Both	67	52.8	59.05

- **Observation**: Only word embedding-based features will not suffice. ‘Augmentation’ to other known useful features necessary

Results (2/2)

	LSA			GloVe			Dependency Weights			Word2Vec		
	P	R	F	P	R	F	P	R	F	P	R	F
L	73	79	75.8	73	79	75.8	73	79	75.8	73	79	75.8
+S	81.8	78.2	**79.95**	81.8	79.2	**80.47**	81.8	78.8	80.27	80.4	80	**80.2**
+WS	76.2	79.8	77.9	76.2	79.6	77.86	81.4	80.8	81.09	80.8	78.6	79.68
+S+WS	77.6	79.8	78.68	74	79.4	76.60	82	80.4	**81.19**	81.6	78.2	79.86
G	84.8	73.8	78.91	84.8	73.8	78.91	84.8	73.8	**78.91**	84.8	73.8	**78.91**
+S	84.2	74.4	**79**	84	72.6	77.8	84.4	72	77.7	84	72.8	78
+WS	84.4	73.6	78.63	84	75.2	**79.35**	84.4	72.6	78.05	83.8	70.2	76.4
+S+WS	84.2	73.6	78.54	84	74	78.68	84.2	72.2	77.73	84	72.8	78
B	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61
+S	78.2	75.6	**76.87**	80.4	76.2	**78.24**	81.2	74.6	**77.76**	81.4	72.6	76.74
+WS	75.8	77.2	76.49	76.6	77	76.79	76.2	76.4	76.29	81.6	73.4	77.28
+S+WS	74.8	77.4	76.07	76.2	78.2	77.18	75.6	78.8	77.16	81	75.4	**78.09**
J	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43
+S	84.8	73.8	78.91	85.6	74.8	79.83	85.4	74.4	79.52	85.4	74.6	**79.63**
+WS	85.6	75.2	**80.06**	85.4	72.6	78.48	85.4	73.4	78.94	85.6	73.4	79.03
+S+WS	84.8	73.6	78.8	85.8	75.4	**80.26**	85.6	74.4	**79.6**	85.2	73.2	78.74

Table 3: Performance obtained on augmenting word embedding features to features from four prior works, for four word embeddings; L: Liebrecht et al. (2013), G: González-Ibánez et al. (2011a), B: Buschmeier et al. (2014) , J: Joshi et al. (2015)

- **Observation**: Using word embedding-based features improves sarcasm detection, for multiple word embedding types and feature sets

Conclusion

- Word embeddings can be used to design novel features for sarcasm detection
- Word embeddings do not operate well on their own as features
- When combined with past feature sets (based on punctuations, sentiment flips, affective lexicons, etc.), these word embedding-based features result in improved performance
- The performance is highest when Word2Vec embeddings are used (Several reasons: Large training corpus, Domain similarity, etc.)

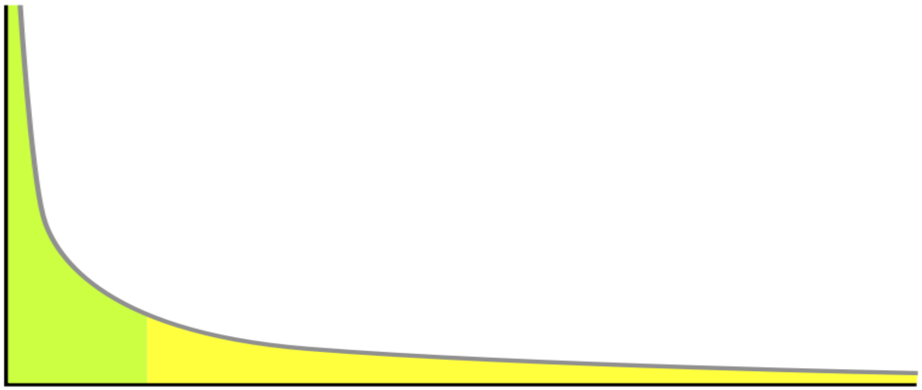
Goal of NLP

- Science Goal
  - Understand human language behaviour

- Engineering goal
  - Unstructured Text → Structured Data

No “democracy”: Tail phenomenon and Language phenomenon

- Long tail Phenomenon: Probability is very low but not zero over a large number of phenomena.



- Language Phenomenon:
  - “people” which is predominantly tagged as “Noun” displays a long tail behaviour.
  - “laugh” is predominantly tagged as “Verb”.

Word embedding-based features for sarcasm detection

(To appear in EMNLP 2016)

Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya and Mark Carman, Are Word Embedding-based Features Useful for Sarcasm Detection?, EMNLP 2016, Austin, Texas, USA, November 1-5, 2016.

Introduction

- Sarcasm detection is the task of predicting whether a given piece of text is sarcastic
- The ruling paradigm in sarcasm detection research is to design features that incorporate contextual information to understand context incongruity that lies at the heart of sarcasm
- ‘I love being ignored’ : Incorporating context incongruity using sentiment flips
- What happens in case of sentences with few or no sentiment words?

Motivation

A <u>woman</u> needs a <u>man</u> like a <u>fish</u> needs <u>bicycle</u>

Word2Vec similarity(man,woman) = 0.766
Word2Vec similarity(fish, bicycle) = 0.131

Can word embedding-based features when augmented to features reported in prior work improve the performance of sarcasm detection?

Word embedding based features

(Stop words removed)

Unweighted similarity features (S): For every word and word pair,

1) Maximum score of most similar word pair
2) Minimum score of most similar word pair
3) Maximum score of most dissimilar word pair
4) Minimum score of most dissimilar word pair

	man	**woman**	**fish**	**needs**	**bicycle**
man	-	**0.766**	**0.151**	**0.078**	**0.229**
woman	0.766	-	0.084	0.060	0.229
fish	0.151	0.084	-	0.022	0.130
needs	0.078	0.060	0.022	-	0.060
bicycle	0.229	0.229	0.130	0.060	-

Experiment setup

- **Dataset:** 3629 Book snippets (759 sarcastic) downloaded from GoodReads website. Labeled by users with tags. We download ones with 'sarcasm' as sarcastic, ones with 'philosophy' as non-sarcastic
- Five-fold cross-validation
- **Classifier:** SVM-Perf optimised for F-score
- **Configurations**:
  - Four prior works (augmented with our sets of features)
  - Four implementations of word embeddings (Word2Vec, LSA, GloVe, Dependency weights-based)

Results (1/2)

Features	**P**	**R**	**F**
	Baseline		
Unigrams	67.2	78.8	72.53
S	64.6	75.2	69.49
WS	67.6	51.2	58.26
Both	67	52.8	59.05

- **Observation**: Only word embedding-based features will not suffice. 'Augmentation' to other known useful features necessary

Results (2/2)

	LSA			GloVe			Dependency Weights			Word2Vec		
	P	R	F	P	R	F	P	R	F	P	R	F
L	73	79	75.8	73	79	75.8	73	79	75.8	73	79	75.8
+S	81.8	78.2	**79.95**	81.8	79.2	**80.47**	81.8	78.8	80.27	80.4	80	**80.2**
+WS	76.2	79.8	77.9	76.2	79.6	77.86	81.4	80.8	81.09	80.8	78.6	79.68
+S+WS	77.6	79.8	78.68	74	79.4	76.60	82	80.4	**81.19**	81.6	78.2	79.86
G	84.8	73.8	78.91	84.8	73.8	78.91	84.8	73.8	**78.91**	84.8	73.8	**78.91**
+S	84.2	74.4	**79**	84	72.6	77.8	84.4	72	77.7	84	72.8	78
+WS	84.4	73.6	78.63	84	75.2	**79.35**	84.4	72.6	78.05	83.8	70.2	76.4
+S+WS	84.2	73.6	78.54	84	74	78.68	84.2	72.2	77.73	84	72.8	78
B	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61	81.6	72.2	76.61
+S	78.2	75.6	**76.87**	80.4	76.2	**78.24**	81.2	74.6	**77.76**	81.4	72.6	76.74
+WS	75.8	77.2	76.49	76.6	77	76.79	76.2	76.4	76.29	81.6	73.4	77.28
+S+WS	74.8	77.4	76.07	76.2	78.2	77.18	75.6	78.8	77.16	81	75.4	**78.09**
J	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43	85.2	74.4	79.43
+S	84.8	73.8	78.91	85.6	74.8	79.83	85.4	74.4	79.52	85.4	74.6	**79.63**
+WS	85.6	75.2	**80.06**	85.4	72.6	78.48	85.4	73.4	78.94	85.6	73.4	79.03
+S+WS	84.8	73.6	78.8	85.8	75.4	**80.26**	85.6	74.4	**79.6**	85.2	73.2	78.74

ble 3: Performance obtained on augmenting word embedding features to features from four prior works, for four word embeddings; L: Liebrecht l. (2013), G: González-Ibánez et al. (2011a), B: Buschmeier et al. (2014) , J: Joshi et al. (2015)

- **Observation**: Using word embedding-based features improves sarcasm detection, for multiple word embedding types and feature sets

Conclusion

- Word embeddings can be used to design novel features for sarcasm detection

- Word embeddings do not operate well on their own as features

- When combined with past feature sets (based on punctuations, sentiment flips, affective lexicons, etc.), these word embedding-based features result in improved performance

- The performance is highest when Word2Vec embeddings are used (Several reasons: Large training corpus, Domain similarity, etc.)