Natural Language Processing and Machine Leaning: Synergy or Discord- a Case Study with MT, IR and Sentiment

FIRE 2016
Pushpak Bhattacharyya
IIT Patna and IIT Bombay
pb@cse.iitb.ac.in

9th Dec, 2016

Need for NLP

- Huge amount of language data in electronic form
- Unstructured data (like free flowing text) will grow to 40 zetabytes (1 zettabyte= 10^{21} bytes) by 2020.
- How to make sense of this huge data?

- Example-1: e-commerce companies need to know **sentiment** of online users, sifting through 1 lakh e-opinions per week: needs NLP
- Example-2: **Translation** industry to grow to $37 billion business by 2020

Nature of Machine Learning

- Automatically learning rules and concepts from data

Learning the concept of table.

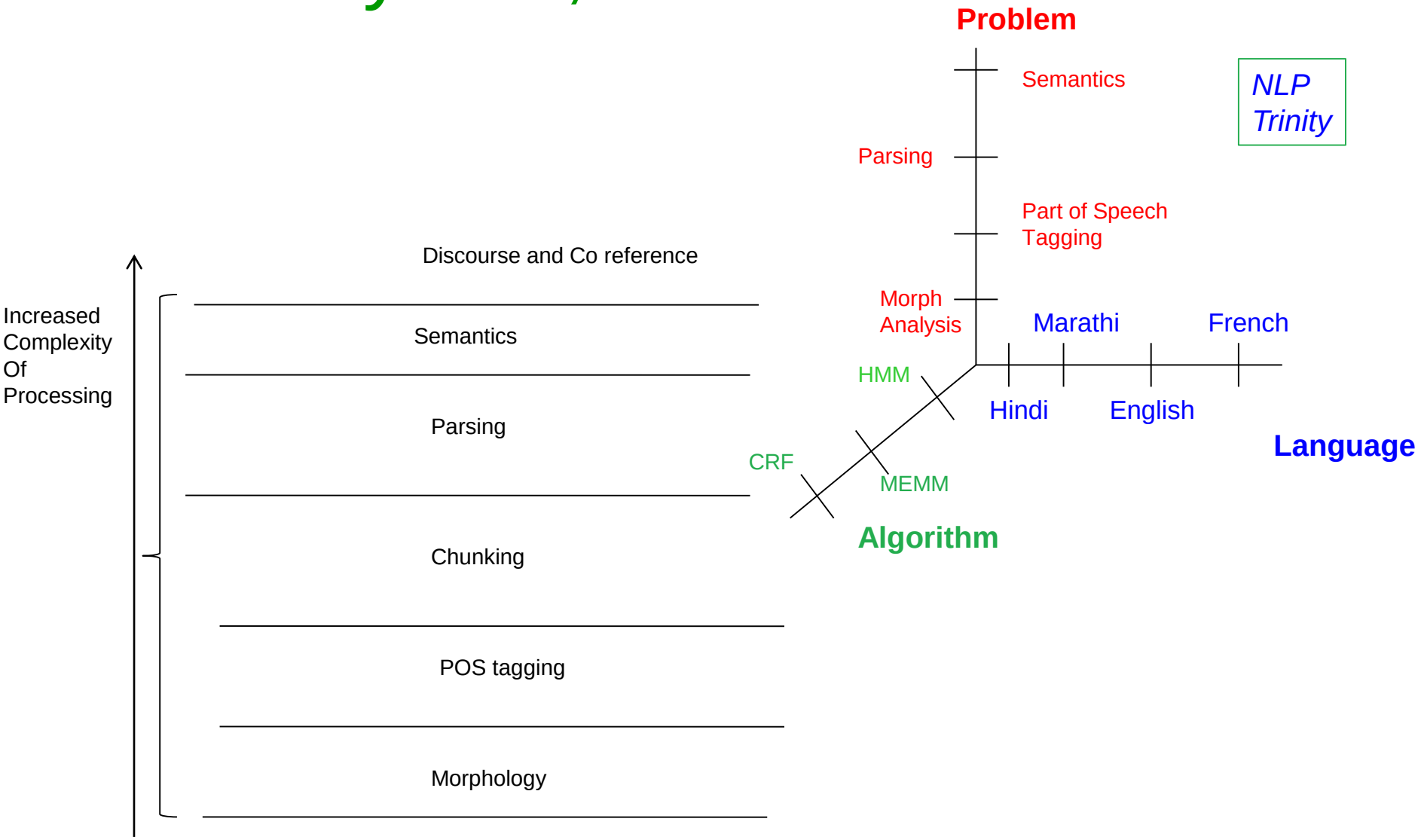
What is “tableness”

Rule: a flat surface with 4 legs (approx.: to be refined gradually)

Why NLP and ML?

- Impossible for humans (single or a team) to makes sense of and analyse humongous text data
- Many processing steps in NLP
- Impossible to give correct-consistent-complete rules covering each and every situation
- Example: Rule: Adjectives preceded Nouns (“blue sky”), but not in French! (“ciel bleu”)

NLP: layered, multidimensional



NLP= Ambiguity Processing

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

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)

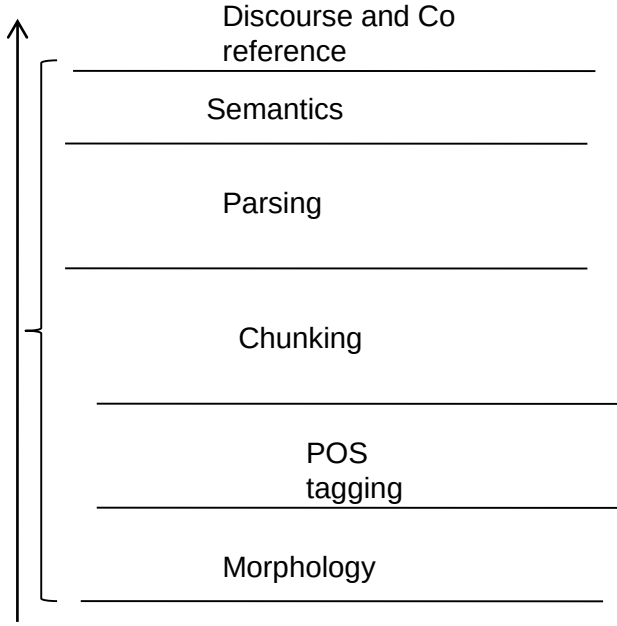
4. After BREXIT (UK's exit from EU), in Mumbai Mirror, and on Tweet: We got Brexit. What's next? Grexit. Departugal. Italeave. Fruckoff. Czechout. Oustria. Finish. Slovakout. Latervia. Byegium

Inter layer interaction

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*

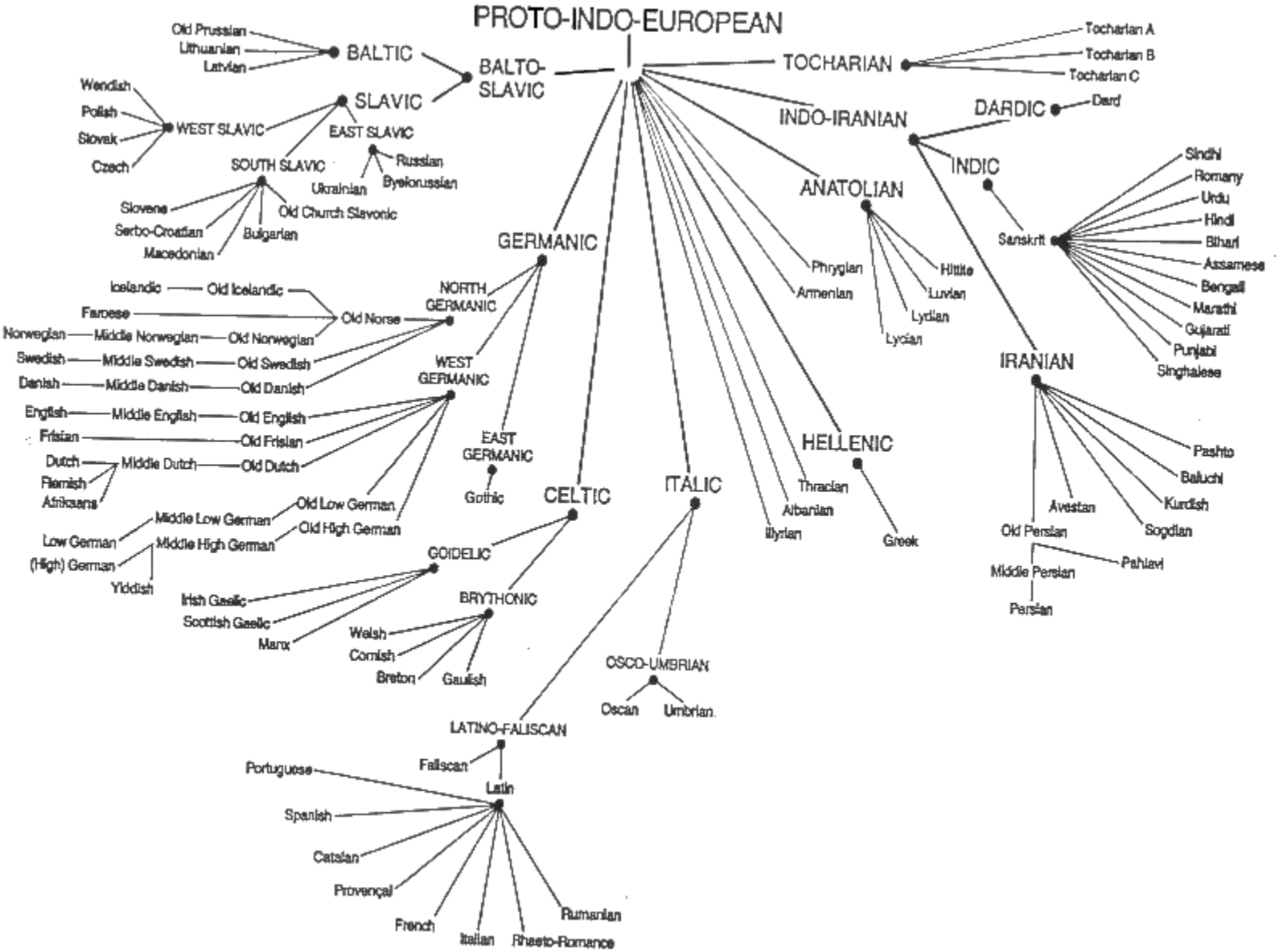
nsubj(saw-2, I-1)
root(ROOT-0, saw-2)
det(boy-4, the-3)
dobj(saw-2, boy-4)
det(telescope-7, a-6)
prep_with(saw-2, telescope-7)
dobj(dropped-10, telescope-7)
nsubj(dropped-10, I-9)
rcmod(telescope-7, dropped-10)
advmod(dropped-10, accidentally-11)

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



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!
  - *Why do we say “Many Thanks” and not “Several Thanks”!*
  - *Impossible to give a rule*
- Rely on machine learning to tease truth out of data; Expectation not always met with☹

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.

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"

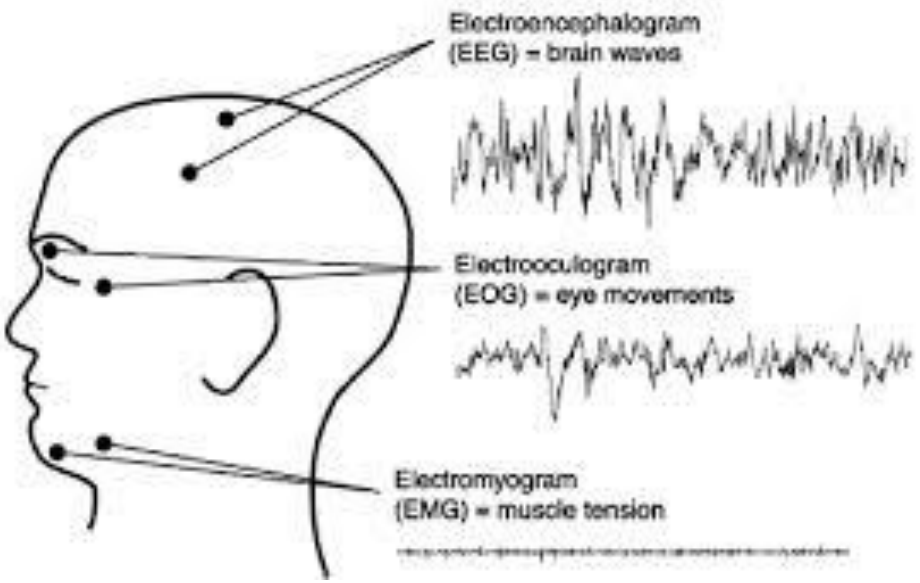
Automatic image labeling (cntd)

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image
			
A person riding a motorcycle on a dirt road.	Two dogs play in the grass.	A skateboarder does a trick on a ramp.	A dog is jumping to catch a frisbee.
			
A group of young people playing a game of frisbee.	Two hockey players are fighting over the puck.	A little girl in a pink hat is blowing bubbles.	A refrigerator filled with lots of food and drinks.
			
A herd of elephants walking across a dry grass field.	A close up of a cat laying on a couch.	A red motorcycle parked on the side of the road.	A yellow school bus parked in a parking lot.

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

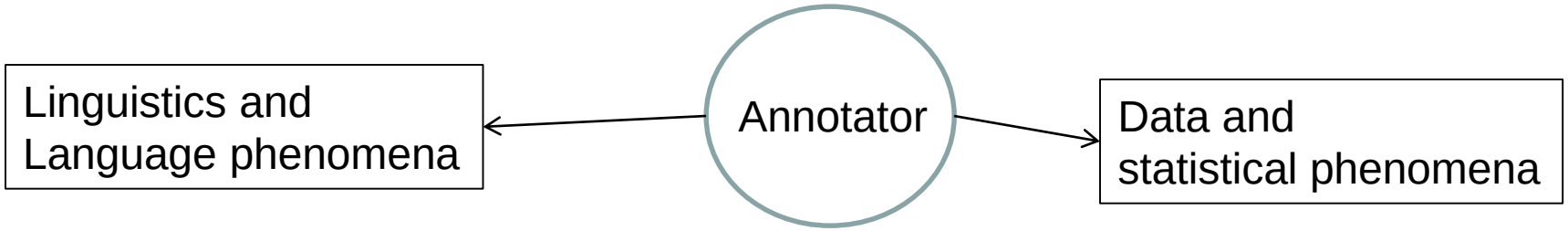
Feature correspondence



"I am hungry now"

Linguistics-Computation Interaction

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



Case Study-1: Machine Translation

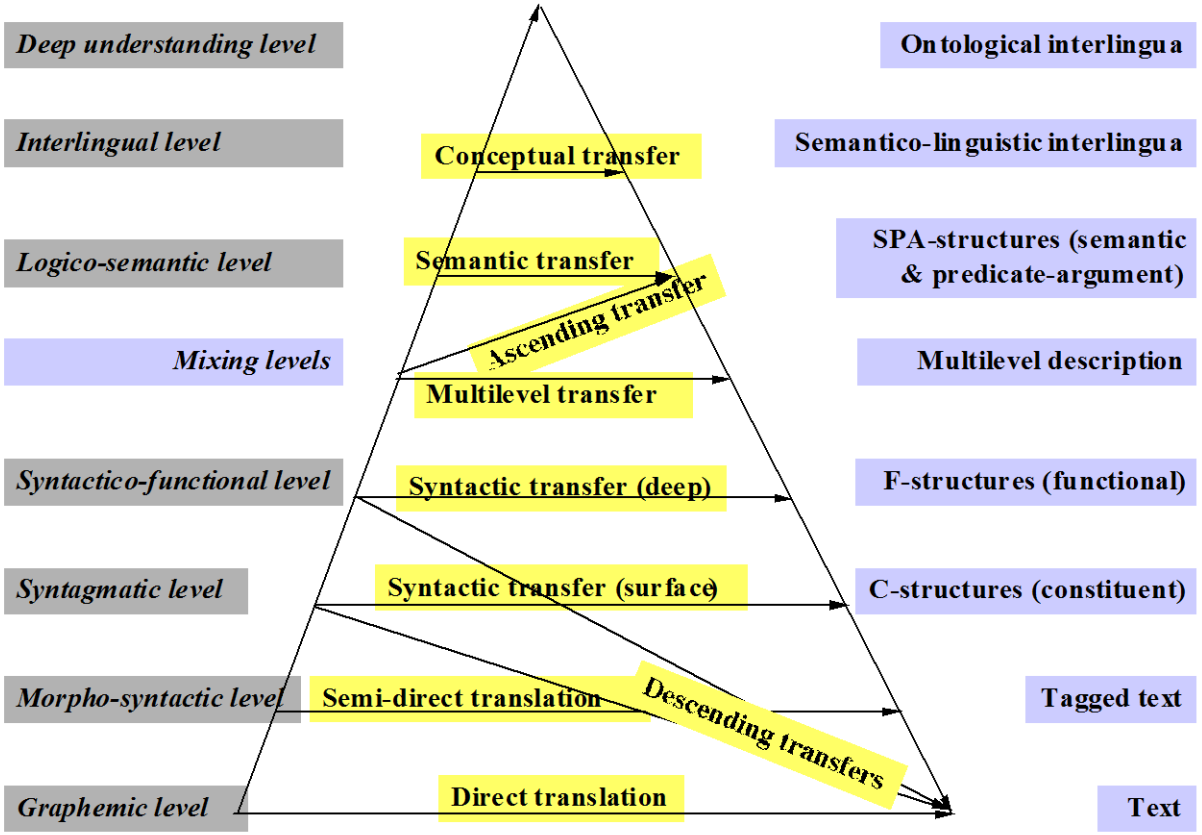
Good Linguistics + Good ML

Pushpak Bhattacharyya, *Machine Translation*, CRC Press, 2015

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.

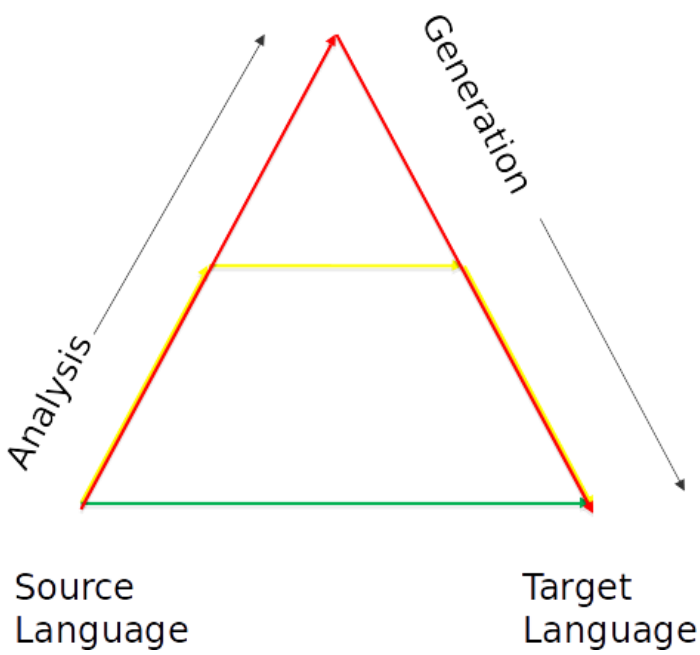
Kinds of MT Systems

(point of entry from source to the target text)



(Vauquois. 1968)

Simplified Vauquois



Interlingua Based Translation

Transfer Based Translation

Direct Translation

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

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

Illustration of difference of RBMT, SMT, EMT

- *Peter has a house*
- *Peter has a brother*
- *This hotel has a museum*

The tricky case of ‘have’ translation

English

- *Peter has a house*
- *Peter has a brother*
- *This hotel has a museum*

Marathi

- पीटरकडे एक घर आहे/ piitar kade ek ghar aahe
- पीटरला एक भाऊ आहे/ piitar laa ek bhaauu aahe
- ह्या हॉटेलमध्ये एक संग्रहालय आहे/ hyaa hotel madhye ek saMgrahaalay aahe

RBMT

If

syntactic subject is animate AND syntactic object is **owned** by subject

Then

"have" should translate to "kade … aahe"

If

syntactic subject is animate AND syntactic object denotes **kinship** with subject

Then

"have" should translate to "laa … aahe"

If

syntactic subject is **inanimate**

Then

"have" should translate to "madhye … aahe"

EBMT

X have Y →

X_kade Y aahe /

X_laa Y aahe /

X_madhye Y aahe

SMT

- *has a house ←→ kade ek ghar aahe*

  <cm> one house has
- *has a car ←→ kade ek gaadii aahe*

  <cm> one car has
- *has a brother ←→ laa ek bhaau aahe*

  <cm> one brother has
- *has a sister ←→ laa ek bahiin aahe*

  <cm> one sister has
- *hotel has ←→ hotel madhye aahe*

  hotel <cm> has
- *hospital has ←→ haspital madhye aahe*

  hospital <cm> has

SMT: new sentence

"This hospital has 100 beds"

- *n*-grams (*n=1, 2, 3, 4, 5*) like the following will be formed:
  - *"This", "hospital",… (unigrams)*
  - *"This hospital", "hospital has", "has 100",… (bigrams)*
  - *"This hospital has", "hospital has 100", … (trigrams)*

  DECODING !!!

Foundation of SMT

- Data driven approach
- Goal is to find out the English sentence *e* given foreign language sentence *f* whose *p(e|f)* is maximum.

$$\tilde{e} = \underset{e \in e^*}{\operatorname{argmax}}\, p(e|f) = \underset{e \in e^*}{\operatorname{argmax}}\, p(f|e)p(e)$$

- Translations are generated on the basis of statistical model
- Parameters are estimated using bilingual parallel corpora

The all important **word alignment**

- The edifice on which the structure of SMT is built
  (Brown et. Al., 1990, 1993; Och and Ney, 1993)
- Word alignment → Phrase alignment (Koehn et al, 2003)
- Word alignment → Tree Alignment (Chiang 2005, 200t; Koehn 2010)
- Alignment at the heart of Factor based SMT too (Koehn and Hoang 2007)

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,

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

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

.....

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

.....

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

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

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

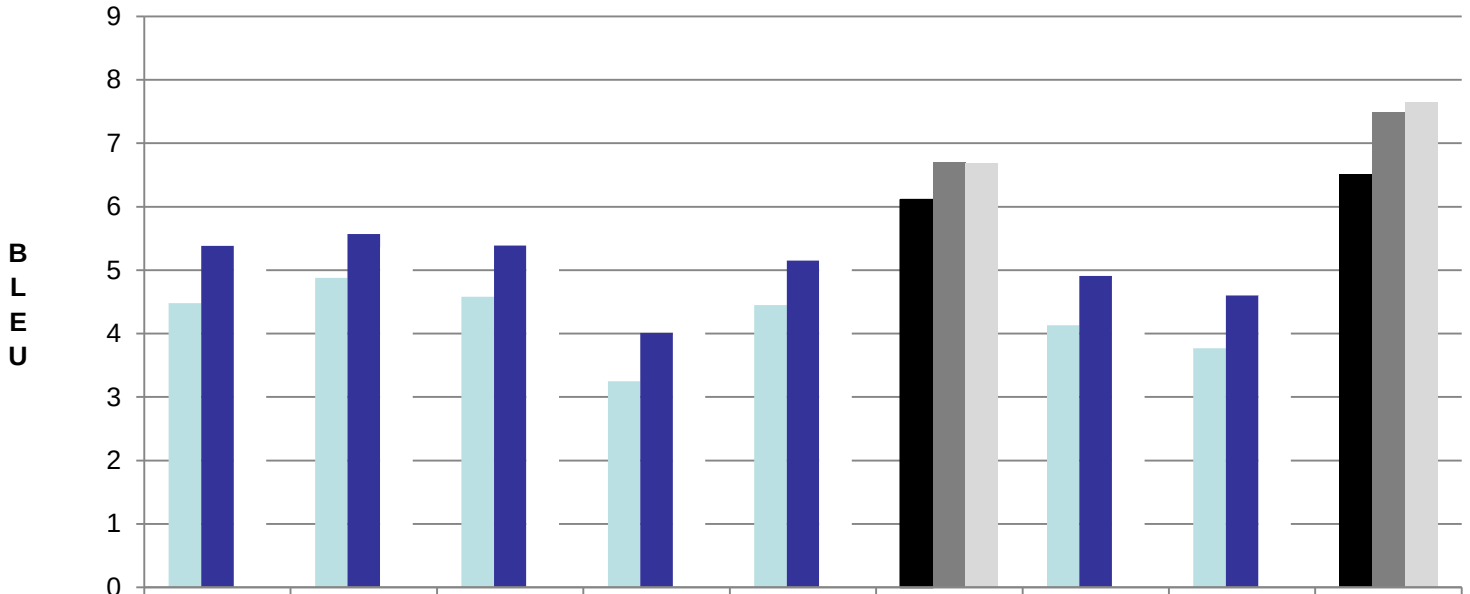
E-step

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

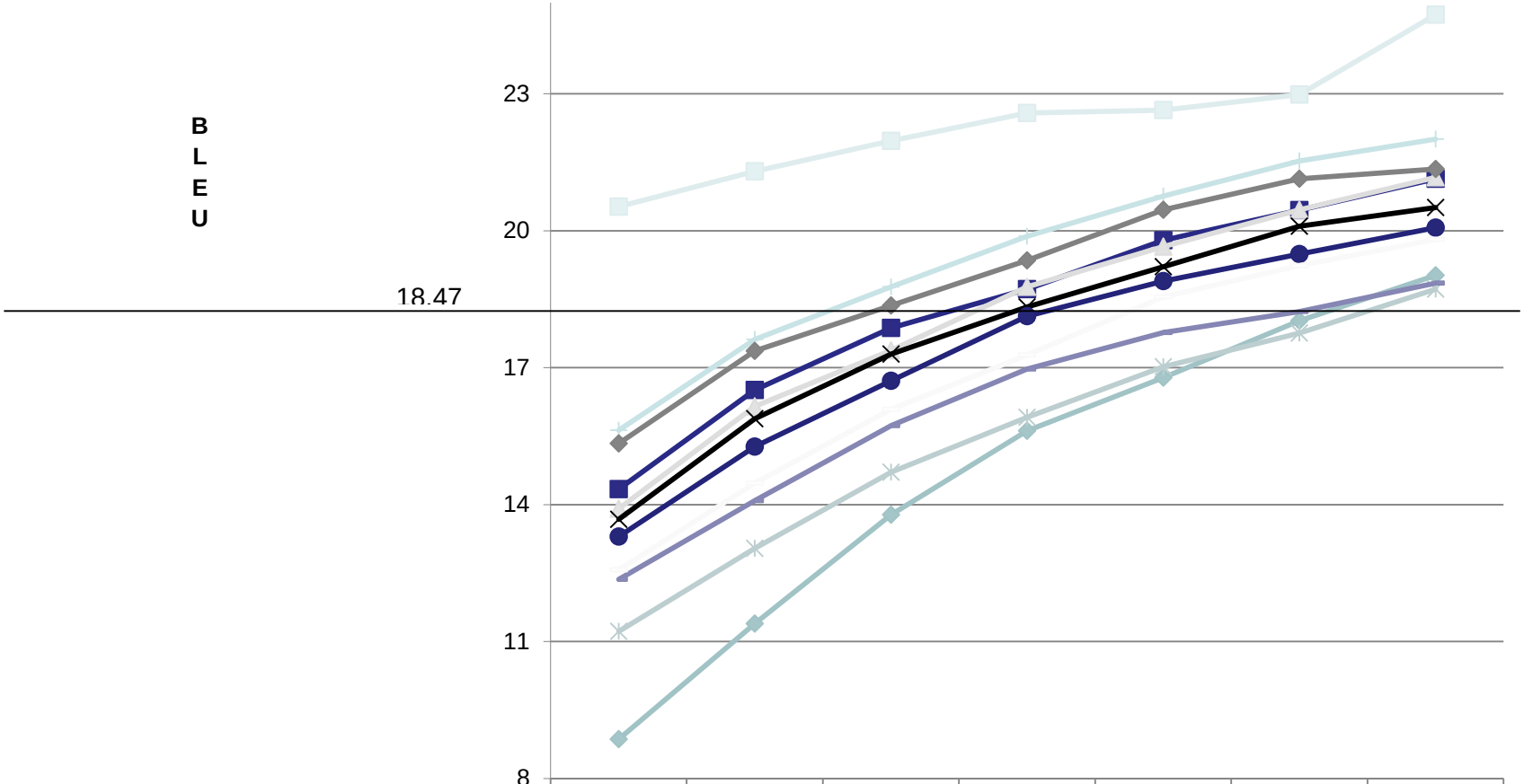
Pivot based MT

Again language property + ML

Pivot for Indian language translation



	Bengali	Gujarati	Konkani	Malayalam	Marathi	Punjabi	Tamil	Telugu	Urdu
p=0.1	4.48	4.88	4.58	3.25	4.45	6.11	4.13	3.77	6.51
p=0.01	5.38	5.57	5.39	4.01	5.15	6.71	4.91	4.6	7.49
p=0.001	5.38	5.36	5.59	4.15	5.2	6.69	4.86	4.59	7.64

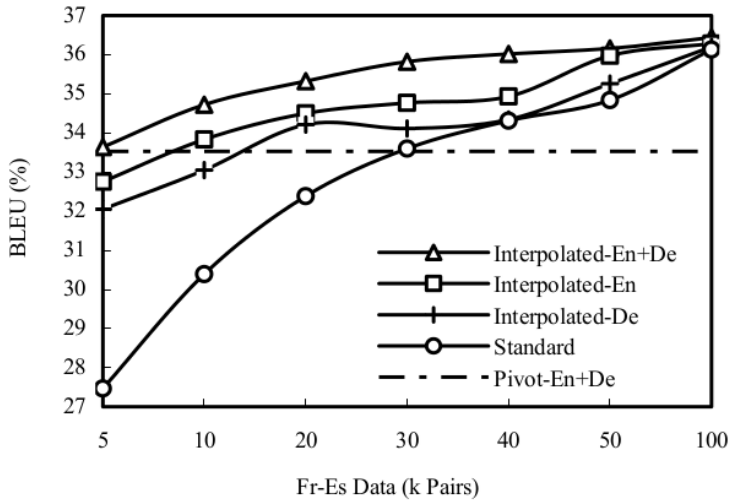


	l=1k	l=2k	l=3k	l=4k	l=5k	l=6k	l=7k
DIRECT_l	8.86	11.39	13.78	15.62	16.78	18.03	19.02
DIRECT_l+BRIDGE_BN	14.34	16.51	17.87	18.72	19.79	20.45	21.14
DIRECT_l+BRIDGE_GU	13.91	16.15	17.38	18.77	19.65	20.46	21.17
DIRECT_l+BRIDGE_KK	13.68	15.88	17.3	18.33	19.21	20.1	20.51
DIRECT_l+BRIDGE_ML	11.22	13.04	14.71	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)



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

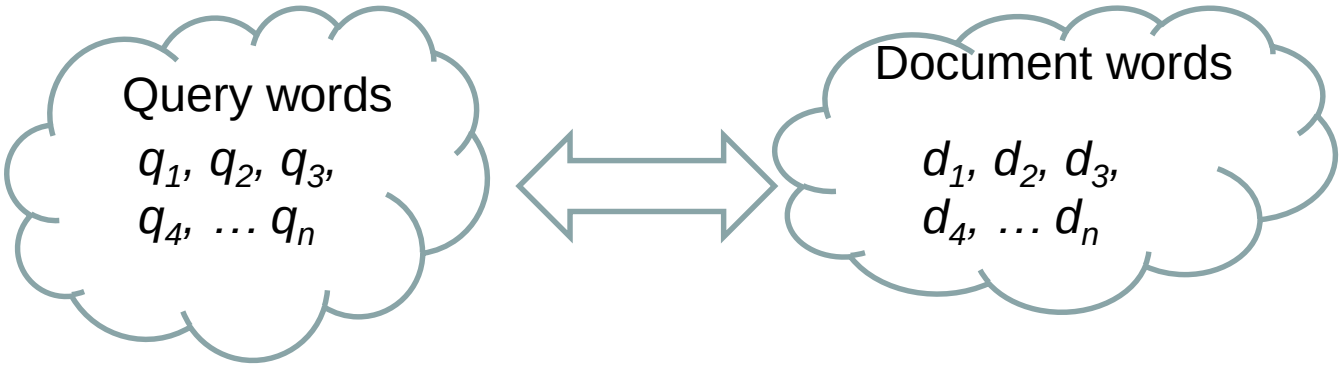
Multilingual Pseudo Relevance Feedback: A way of Query Expansion and Disambiguation

*(Manoj Chinnakotla, Karthik Raman and Pushpak Bhattacharyya, Multilingual PRF: English Lends a Helping Hand, **SIGIR 2010**, Geneva, Switzerland, July, 2010.)*

Manoj Chinnakotla, Karthik Raman and Pushpak Bhattacharyya, *Multilingual Relevance Feedback: One Language Can Help Another*, Conference of Association of Computational Linguistics (**ACL 2010**), Uppsala, Sweden, July 2010.

Arjun Atreya, Ashish Kankaria, Pushpak Bhattacharyya and Ganesh Ramakrishnan *Query Expansion in Resource Scarce Languages: A Multilingual Framework Utilizing Document Structure*, **TALLIP** (Transactions on Asian and Low-resource Language Processing), 2016.

Ranking: computing divergence

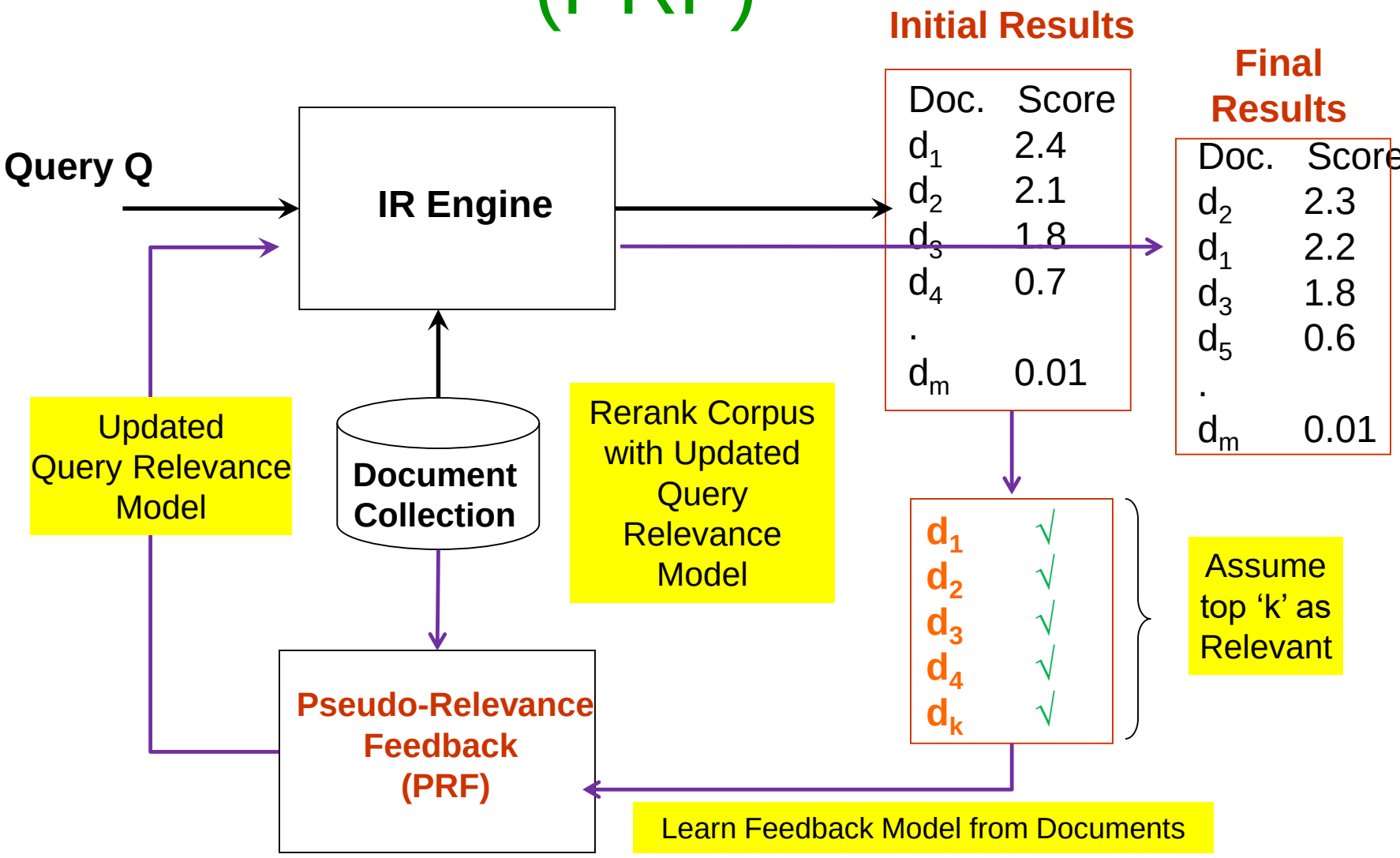


Ranking Function – KL Divergence

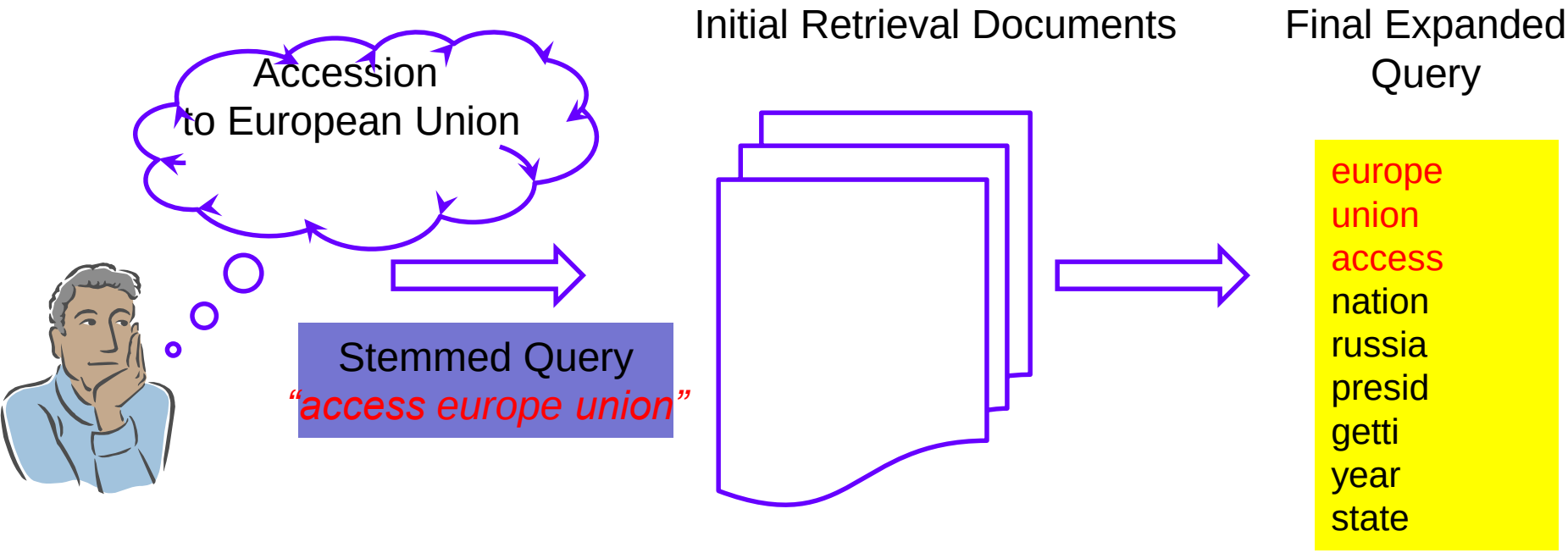
$$\text{Score(D)} = KL(\Theta_R, D)$$

$$\equiv -\sum_w \overbrace{P(w|\Theta_R)}^{\text{Importance of term in Query}} \times \underbrace{\log P(w|D)}_{\text{Importance of term in Document}}$$

Pseudo-Relevance Feedback (PRF)

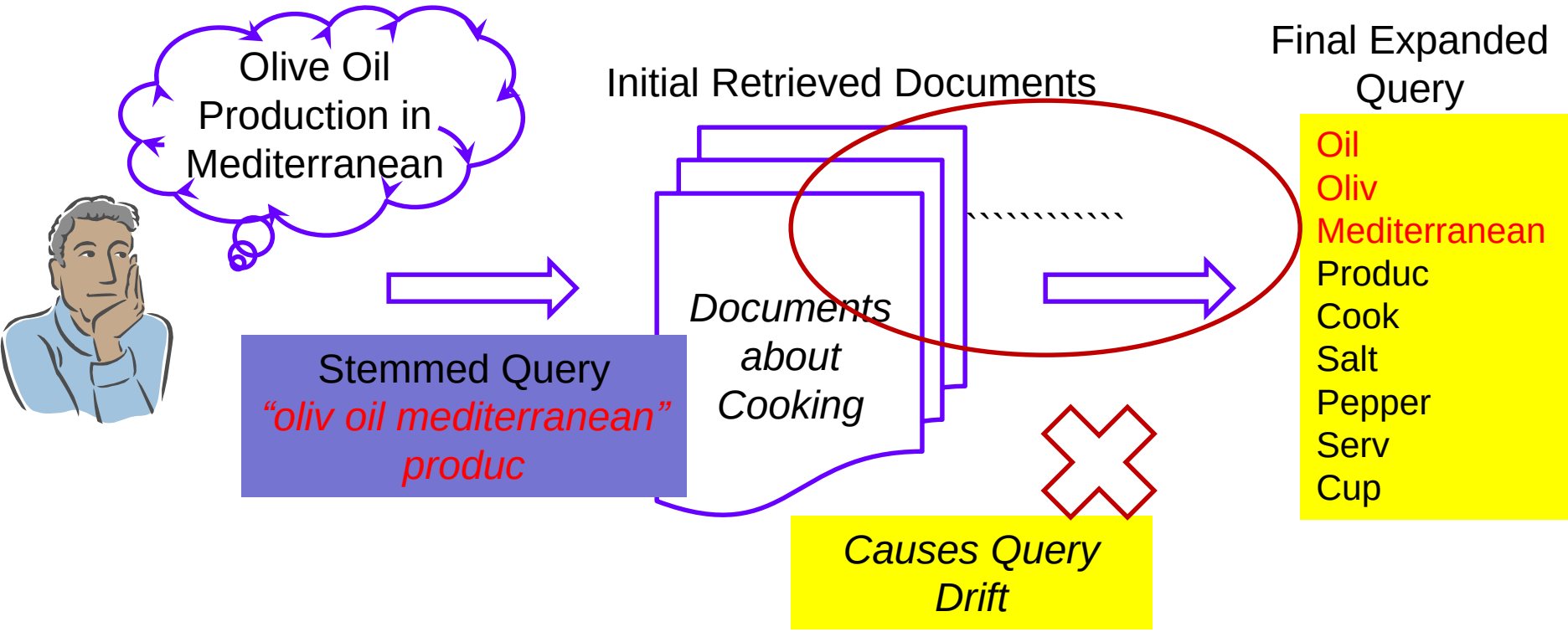


Misses related words



Relevant documents with terms like *"Membership"*, *"Member"*, *"Country"* *not ranked high enough*

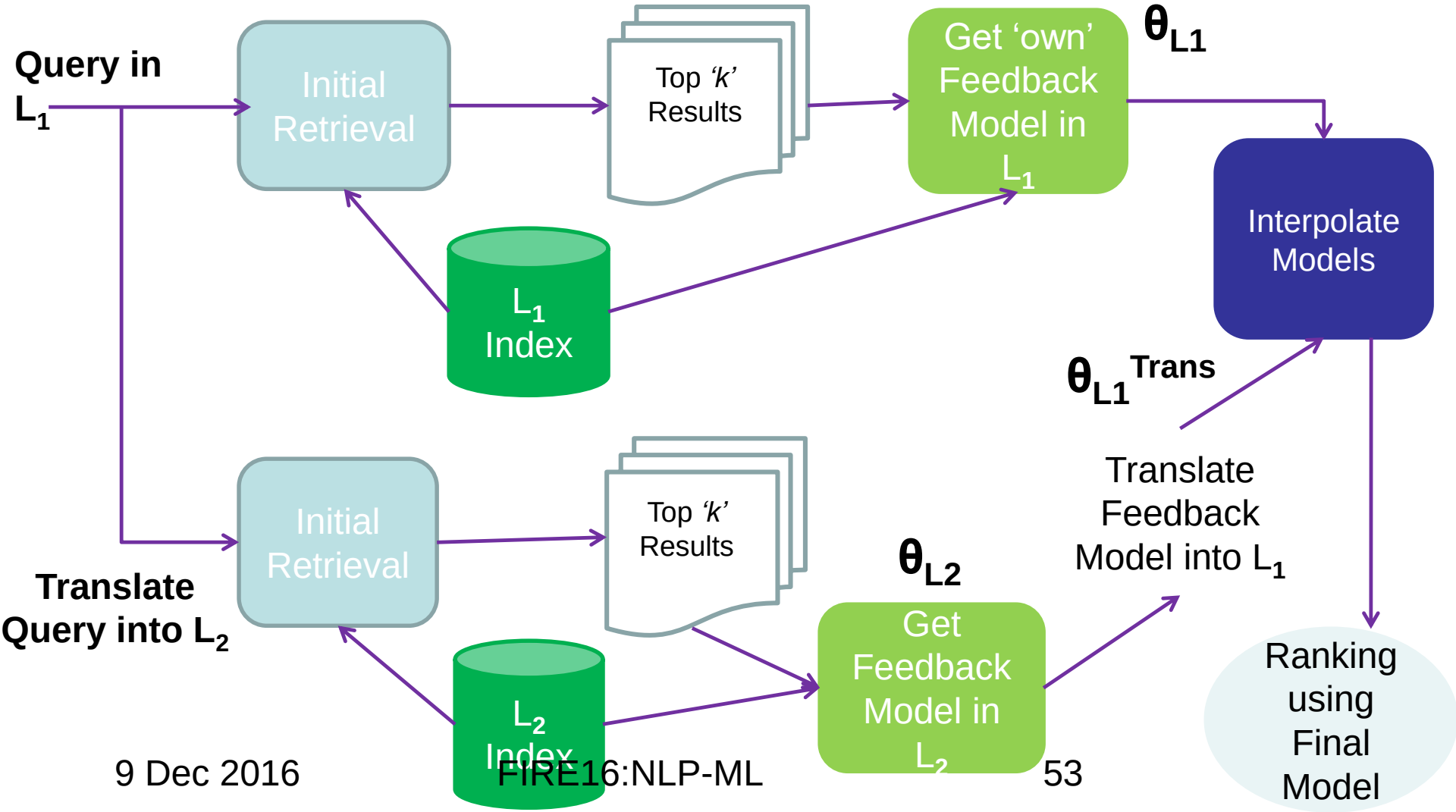
Lack of Robustness



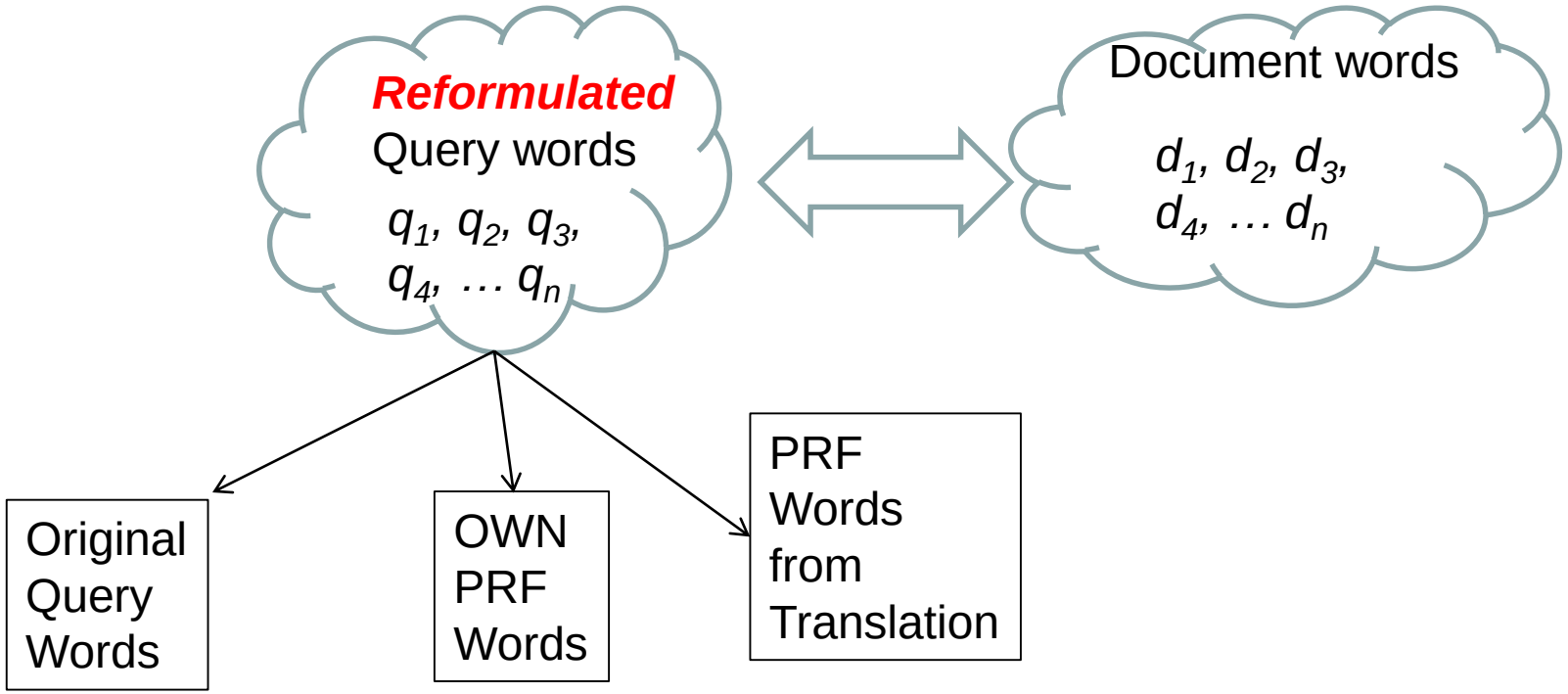
Harness Multilinguality

- Use Assisting Language
- An attractive proposition for languages that have poor monolingual performance due to
  - Resource constraints like inadequate coverage
  - Morphological complexity

Multilingual PRF: System Flow



KLD with Augmented Query

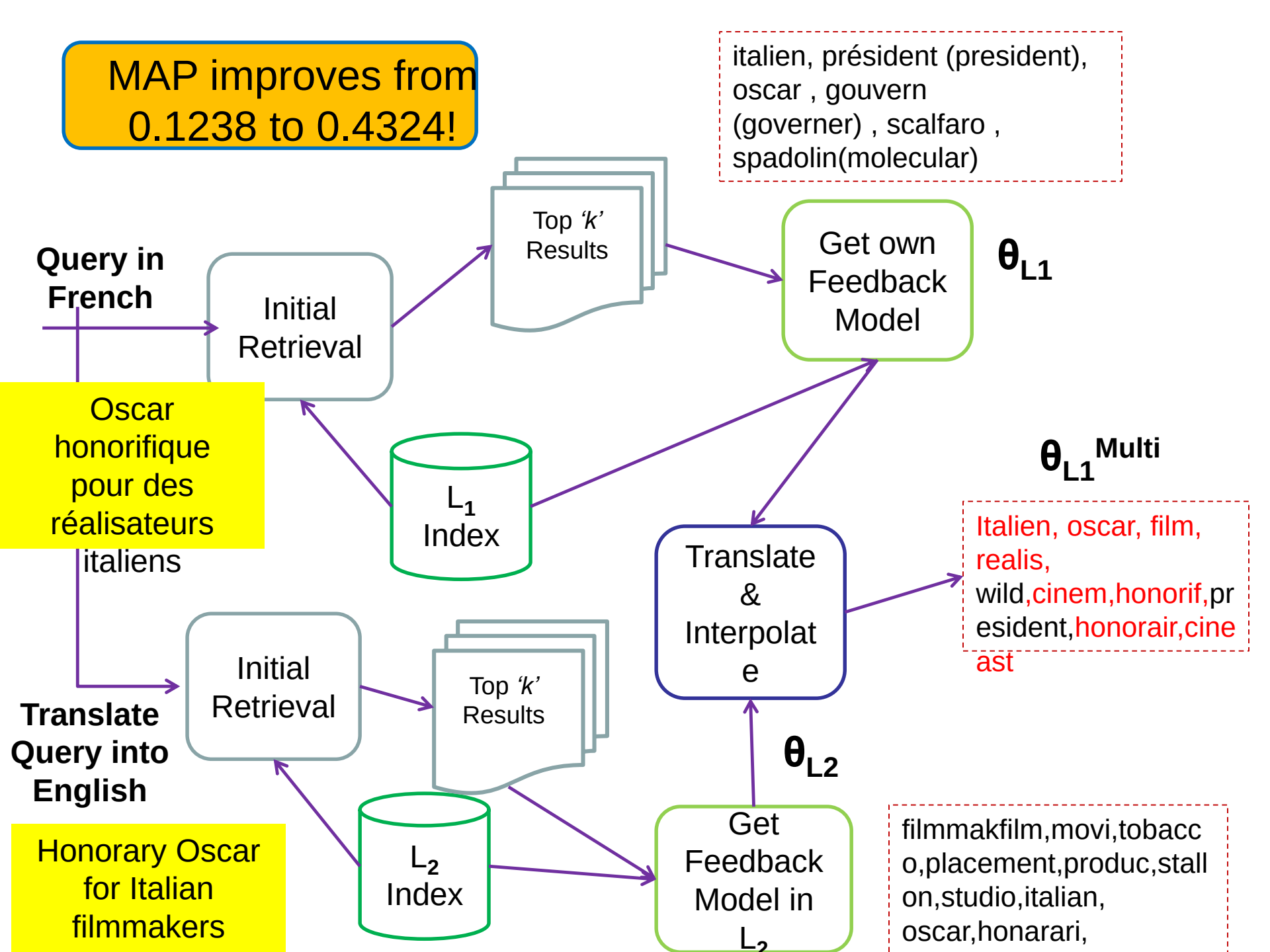


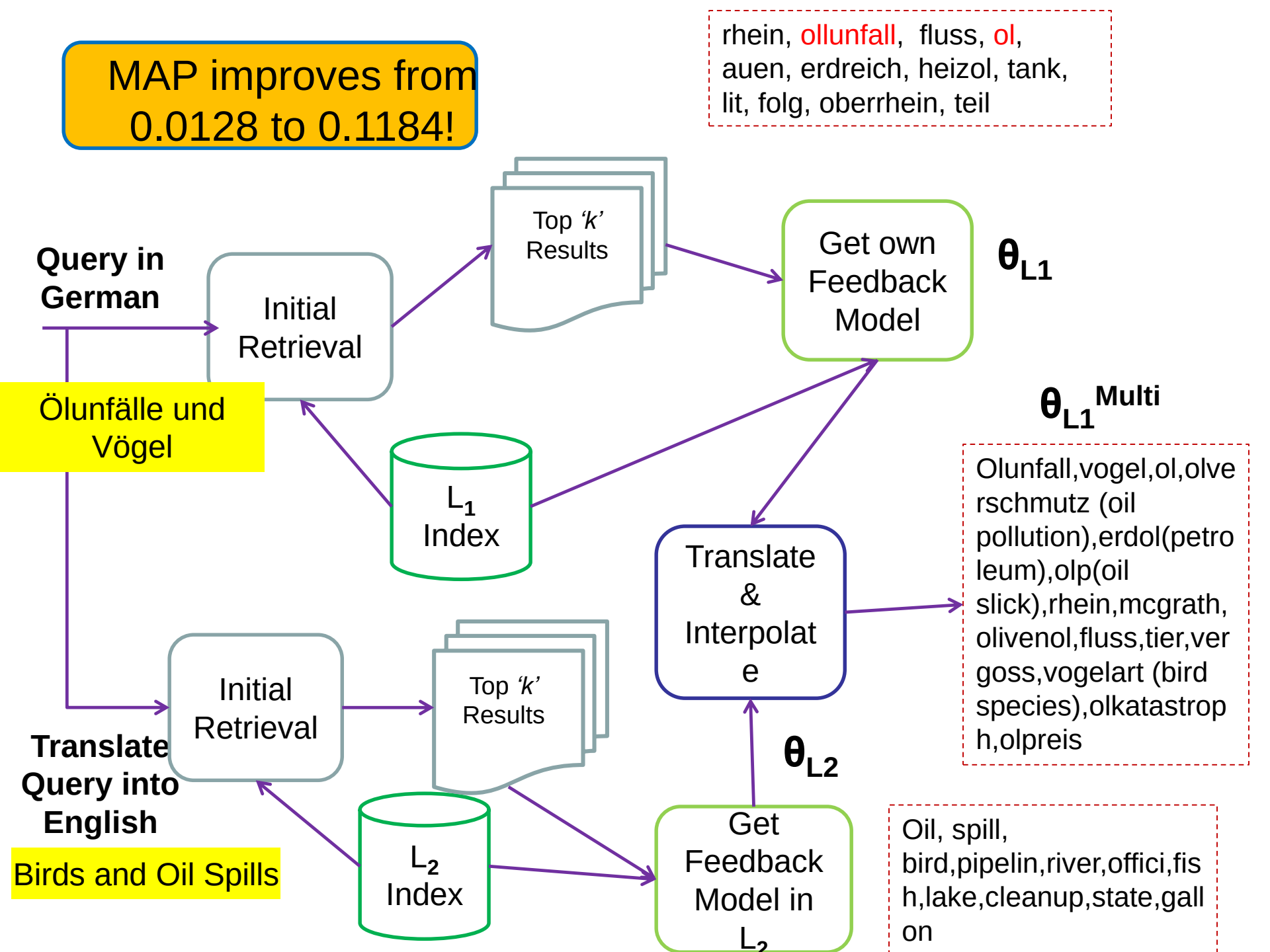
English Lends a Helping Hand!

- English used as assisting language
  - Good monolingual performance
  - Ease of processing
- MultiPRF consistently and significantly outperforms monolingual PRF baseline

Experimental Setup

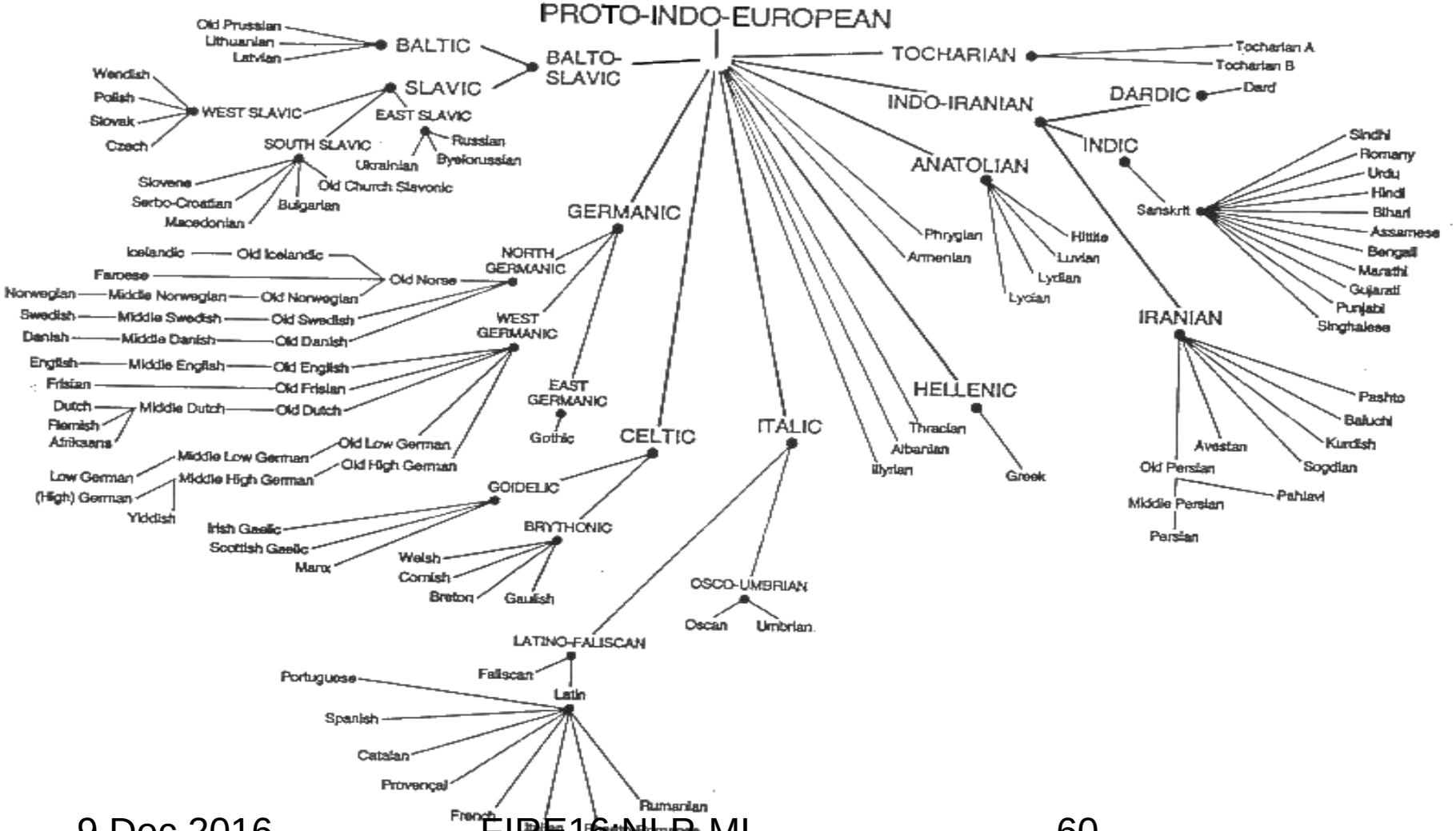
- English chosen as assisting language
- CLEF Standard Dataset for Evaluation
  - Four widely differing source languages uses
    - French (Romance Family), German(West Germanic)
    - Finnish (Baltic-Finnic), Hungarian (Uralic-Ugric)
  - On more than 600 topics (only Title field)
- Use *Google Translate* for Query Translation


MAP improves from 0.1238 to 0.4324!
italien, président (president), oscar , gouvern (governer) , scalfaro , spadolin(molecular)
Query in French
Initial Retrieval
Top 'k' Results
Get own Feedback Model
θL1
Oscar honorifique pour des réalisateurs italiens
L1 Index
θL1Multi
Translate & Interpolate
Italien, oscar, film, realis, wild,cinem,honorif,president,honorair,cineast
Translate Query into English
Initial Retrieval
Top 'k' Results
θL2
L2 Index
Get Feedback Model in L2
filmmakfilm,movi,tobacco,placement,produc,stallon,studio,italian, oscar,honarari,
Honorary Oscar for Italian filmmakers


MAP improves from 0.0128 to 0.1184!
rhein, ollunfall, fluss, ol, auen, erdreich, heizol, tank, lit, folg, oberrhein, teil
Query in German
Initial Retrieval
Top 'k' Results
Get own Feedback Model
θL1
Ölunfälle und Vögel
L1 Index
θL1Multi
Olunfall,vogel,ol,olverschmutz (oil pollution),erdol(petroleum),olp(oil slick),rhein,mcgrath,olivenol,fluss,tier,vergoss,vogelart (bird species),olkatastroph,olpreis
Translate & Interpolate
Translate Query into English
Initial Retrieval
Top 'k' Results
θL2
L2 Index
Get Feedback Model in L2
Oil, spill, bird,pipelin,river,offici,fish,lake,cleanup,state,gallon
Birds and Oil Spills

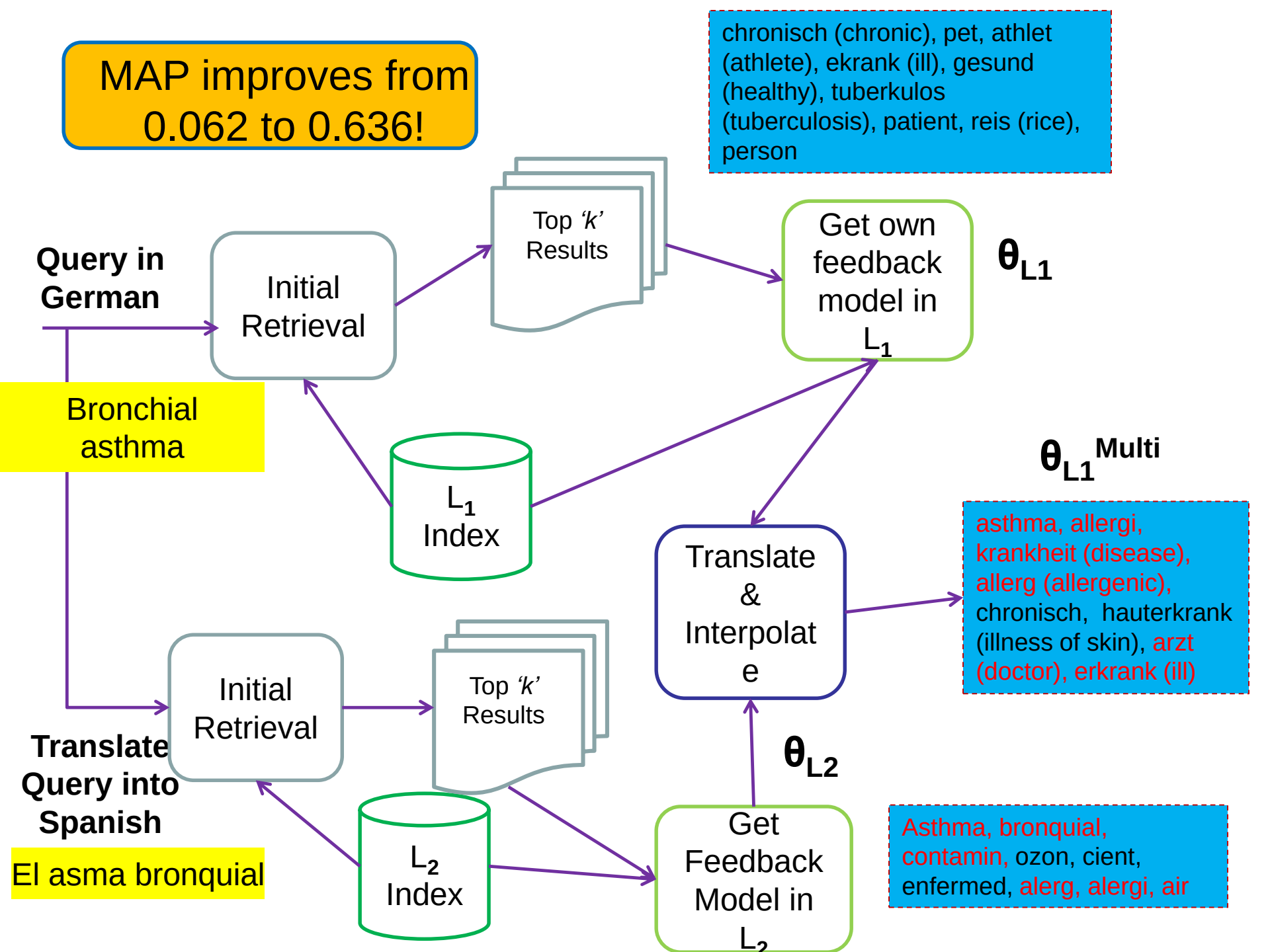
Can languages other than English help?

Language Typology



MultiPRF with Non-English Assisting Languages

Collection	Assist. Lang	P@5 MBF	P@5 MultiPRF	P@5 % Impr.	P@10 MBF	P@10 MultiPRF	P@10 % Impr.	MAP MBF	MAP MultiPRF	MAP % Impr.	GMAP MBF	GMAP MultiPRF	GMAP % Impr.
FR-00	EN	0.4690	0.5241	**11.76‡**	0.4000	0.4000	0.00	0.4220	0.4393	4.10	0.2961	0.3413	**15.27**
	ES		0.5034	**7.35‡**		0.4103	2.59		0.4418	4.69		0.3382	**14.22**
	NL		0.5034	7.35		0.4103	2.59		0.4451	5.47		0.3445	**16.34**
FR-01+02	EN	0.4636	0.4818	3.92	0.4068	0.4386	**7.82‡**	0.4342	0.4535	**4.43‡**	0.2395	0.2721	**13.61**
	ES		0.4977	**7.35‡**		0.4363	**7.26‡**		0.4416	1.70		0.2349	-1.92
	NL		0.4818	3.92		0.4409	**8.38‡**		0.4375	0.76		0.2534	**5.80**
FR-03+05	EN	0.4545	0.4768	**4.89‡**	0.4040	0.4202	**4‡**	0.3529	0.3694	**4.67‡**	0.1324	0.1411	**6.57**
	ES		0.4727	4.00		0.4080	1.00		0.3582	1.50		0.1325	0.07
	NL		0.4525	-0.44		0.4010	-0.75		0.3513	0.45		0.1319	-0.38
FR-06	EN	0.4917	0.5083	3.39	0.4625	0.4729	2.25	0.3837	0.4104	6.97	0.2174	0.2810	**29.25**
	ES		0.5083	3.39		0.4687	1.35		0.3918	2.12		0.2617	**20.38**
	NL		0.5083	3.39		0.4646	0.45		0.3864	0.71		0.2266	**4.23**
DE-00	EN	0.2303	0.3212	**39.47‡**	0.2394	0.2939	**22.78‡**	0.2158	0.2273	5.31	0.0023	0.0191	**730.43**
	ES		0.3212	**39.47‡**		0.2818	**17.71‡**		0.2376	10.09		0.0123	**434.78**
	NL		0.3151	**36.82‡**		0.2818	**17.71‡**		0.2331	8.00		0.0122	**430.43**
DE-01+02	EN	0.5341	0.6000	**12.34‡**	0.4864	0.5318	**9.35‡**	0.4229	0.4576	**8.2‡**	0.1765	0.2721	**9.19**
	ES		0.5682	**6.39‡**		0.5091	**4.67‡**		0.4459	5.43		0.2309	**30.82**
	NL		0.5773	**8.09‡**		0.5114	**5.15‡**		0.4498	**6.35‡**		0.2355	**33.43**
DE-03	EN	0.5098	0.5412	6.15	0.4784	0.4980	4.10	0.4274	0.4355	1.91	0.1243	0.1771	**42.48**
	ES		0.5647	**10.77‡**		0.4980	4.10		0.4568	**6.89‡**		0.1645	**32.34**
	NL		0.5529	**8.45‡**		0.4941	3.27		0.4347	1.72		0.1490	**19.87**
FI-02+03+04	EN	0.3782	0.4034	**6.67‡**	0.3059	0.3319	**8.52‡**	0.3966	0.4246	**7.06‡**	0.1344	0.2272	**69.05**
	ES		0.3879	2.58		0.3267	6.81		0.3881	-2.15		0.1755	**30.58**
	NL		0.3948	4.40		0.3301	7.92		0.4077	2.79		0.1839	**36.83**


MAP improves from 0.062 to 0.636!
chronisch (chronic), pet, athlet (athlete), ekrank (ill), gesund (healthy), tuberkulos (tuberculosis), patient, reis (rice), person
Query in German
Initial Retrieval
Top 'k' Results
Get own feedback model in L_1
θ_{L1}
Bronchial asthma
L_1 Index
θ_{L1}^{Multi}
asthma, allergi, krankheit (disease), allerg (allergenic), chronisch, hauterkrank (illness of skin), arzt (doctor), erkrank (ill)
Translate & Interpolate
Translate Query into Spanish
Initial Retrieval
Top 'k' Results
θ_{L2}
El asma bronquial
L_2 Index
Get Feedback Model in L_2
Asthma, bronquial, contamin, ozon, cient, enfermed, alerg, alergi, air

Results

Source \ Assisting →	English	German	Dutch	Spanish	French	Finnish
English	-	3	4	1	2	5
German	1	-	3	2	4	5
Dutch	1	2	-	4	3	5
Spanish	4	2	3	-	1	5
French	2	3	4	1	-	5
Finnish	1	5	3	2	4	-
Avg. Posn. as Assisting	1.80	3.00	3.40	2.40	2.40	5.00

Dependence on Monolingual Performance

Assisting → / Source ↓	English	German	Dutch	Spanish	French	Finnish
English	-	3	4	1	2	5
German	1	-	3	2	4	5
Dutch	1	2	-	4	3	5
Spanish	4	2	3	-	1	5
French	2	3	4	1	-	5
Finnish	1	5	3	2	4	-
Avg. Posn. as Assisting	1.80	3.00	3.40	2.40	2.40	5.00

Monolingual MAP	0.4495	0.4033	0.4153	0.4805	0.4356	0.3578
Rank	2	5	4	1	3	6

More than one assisting language

- Tried parallel composition for two assisting languages
- Uniform interpolation weights used
- Exhaustively tried all 60 combinations
- Improvements reported over best performing PRF of L_1 or L_2

Source Language	Assisting Language Pairs with Improvement >3%
English	FR-DE (4.5%), FR-ES (4.8%), DE-NL (+3.1%)
French	EN-DE (4.1%), DE-ES (3.4%), NL-FI (4.8%)
German	None
Spanish	None
Dutch	EN-DE (3.9%), DE-FR (4.1%), FR-ES (3.8%), DE-ES (3.9%)
Finnish	EN-ES (3.2%), FR-DE (4.6%), FR-ES (6.4%), DE-ES (11.2%), DE-NL (4.4%), ES-NL (5.9%)
Total - 16	EN – 3 Pairs; FR – 6 Pairs; DE – 10 Pairs; ES - 8 Pairs; NL – 4 Pairs; FI – 1 Pair

Structure aware feedback terms

(Atreya et. al, IJCNLP 2013)

- Title and conclusion are high importance regions
- In Wikipedia documents, get PRF terms from: title, body, infobox and categories

	NORF	*PRF*	*StructPRF*
English	0.1758	0.2022 (+15%)	0.2189 (+24.5%)
Spanish	0.0433	0.1352 (+212%)	0.1778 (+310%)
Finnish	0.1532	0.2477 (+61.6%)	0.2517 (+64.3%)
Hindi	0.2321	0.2364 (+1.8%)	0.2529 (+9%)

← MAP improvement

	English	Spanish	Finnish	Hindi
NoTitle	0.1953(-11%)	0.1179(-33%)	0.1914(-23%)	0.2086(-17%)
NoBody	0.2059(-6%)	0.1383(-22%)	0.2333(-8%)	0.2185(-13%)
NoCategories	0.2172(-0.7%)	0.1436(-19%)	0.2358(-7%)	0.2209(-12%)
NoInfobox	0.2178(-0.5%)	0.1467(-17%)	0.2449(-3%)	0.2234(-11%)

↑ Ablation results

Cooperative Word Sense Disambiguation

Niladri Dash, Pushpak Bhattacharyya, Jyoti Pawar (eds.), *Wordnets of Indian Languages*, Springer, ISBN 978-981-10-1909-8, 2016.

Mitesh Khapra, Salil Joshi and Pushpak Bhattacharyya, *It takes two to Tango: A Bilingual Unsupervised Approach for Estimating Sense Distributions using Expectation Maximization*, 5th International Joint Conference on Natural Language Processing (**IJCNLP 2011**), Chiang Mai, Thailand, November 2011.

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

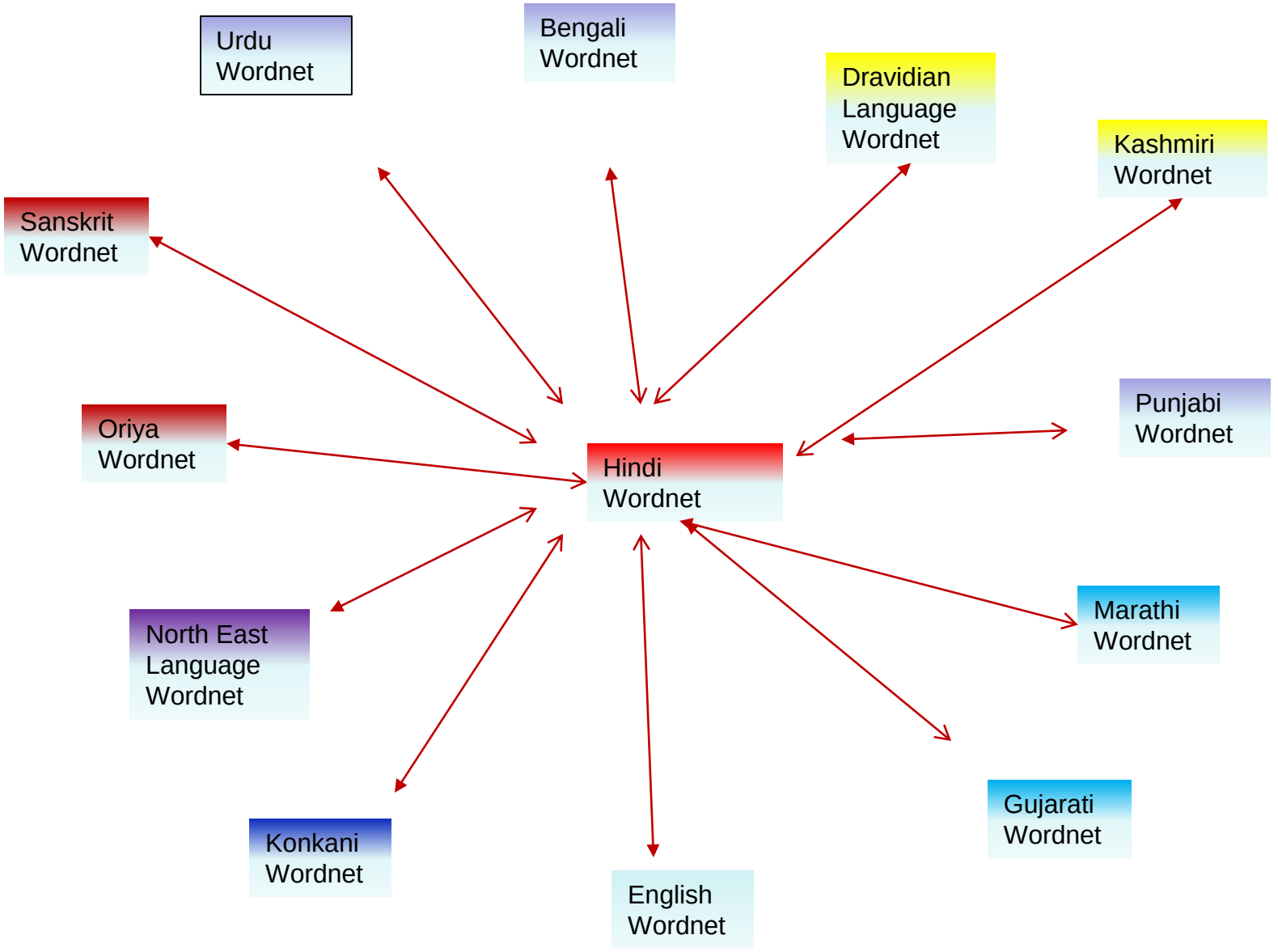
Example: "*operation*" (from Princeton Wordnet)

- **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**



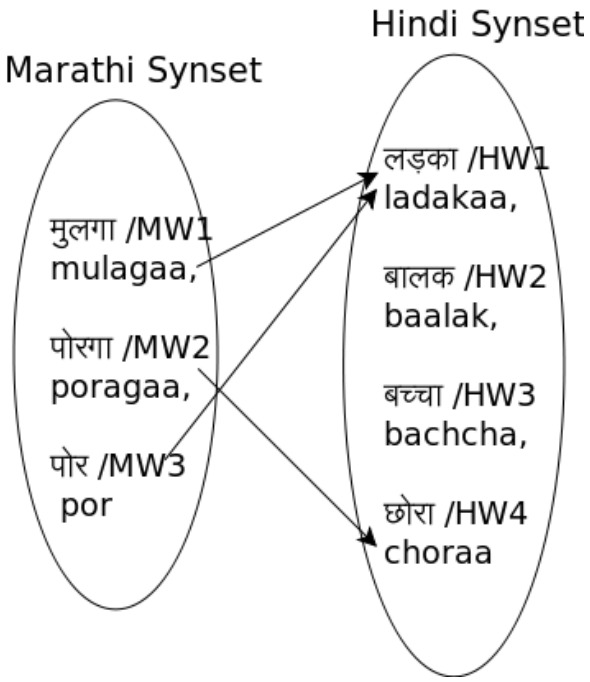
Synset Based Multilingual Dictionary

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

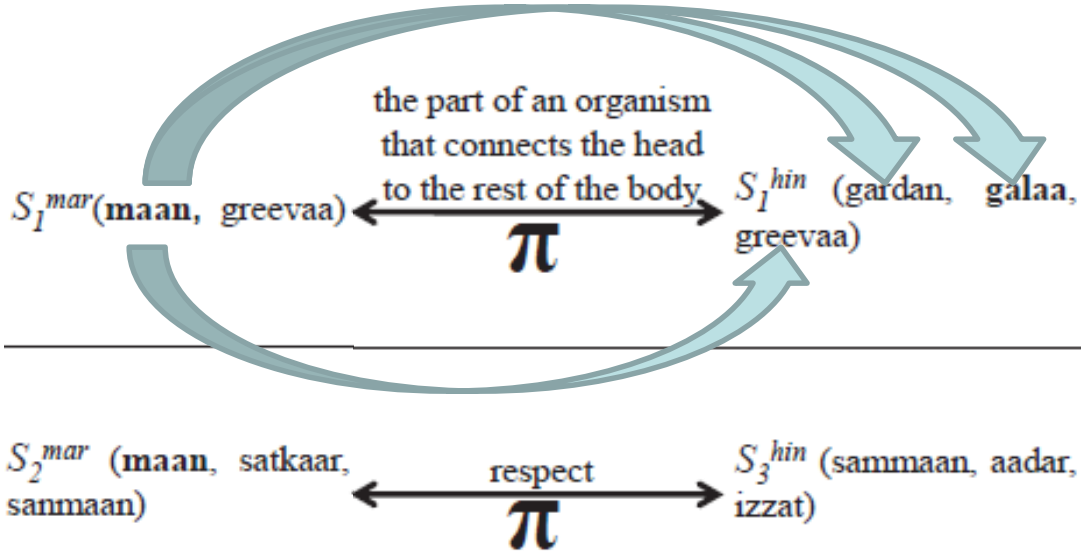
Resources for WSD- wordnet and corpora: 5 scenarios

	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

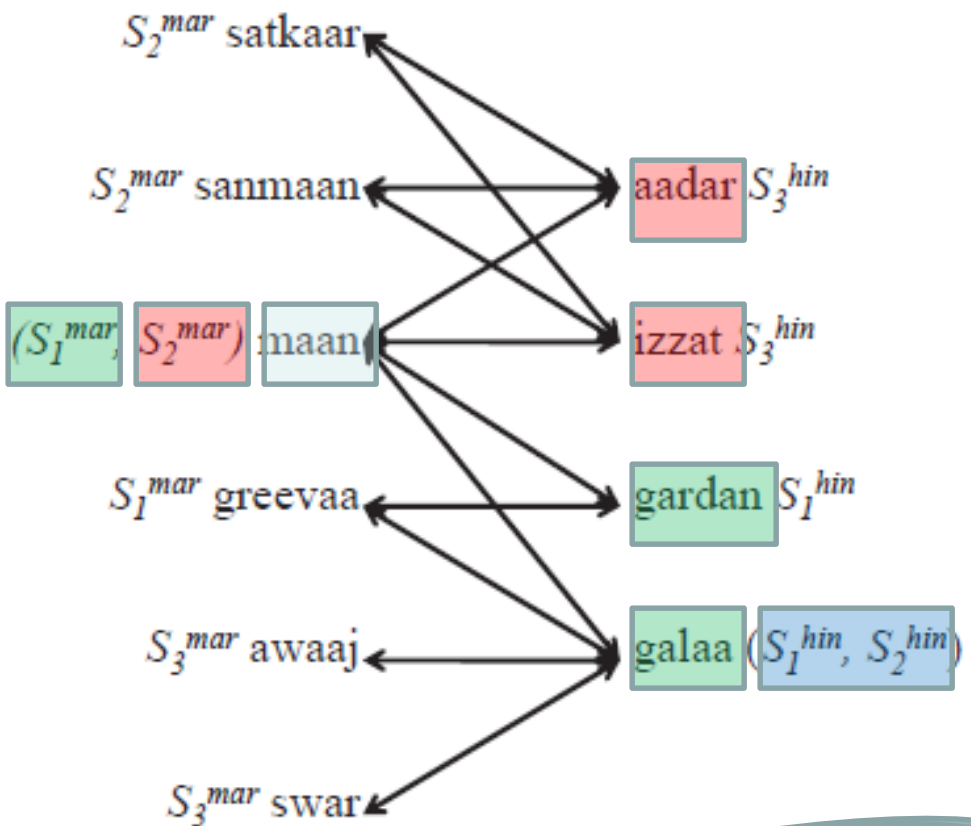


If sense tagged Marathi corpus were available, we could have estimated

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

But such a corpus is not available

EM for estimating sense distributions



E-Step

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

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

Algorithms	Tourism			Health		
	P%	R%	F%	P%	R%	F%
MCL	73.36	68.83	71.02	75.86	66.6	70.93
PCL	68.57	67.93	68.25	65.75	64.53	65.14
IWSD-Self	78.36	77.77	78.07	78.15	75.91	77.01
WFS	57.15	57.15	57.15	55.55	55.55	55.55
PPR	51.49	51.49	51.49	48.32	48.32	48.32
Unsup	9.01	9.01	9.01	9.72	9.72	9.72

Our values

Manual Cross Linkages

Probabilistic Cross Linkages

Skyline - self training data is available

Wordnet first sense baseline

S-O-T-A Knowledge Based Approach

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*

Sarcasm Detection Using Semantic incongruity

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.

Also covered in: How Vector Space Mathematics Helps Machines Spot Sarcasm, MIT Technology Review, 13th October, 2016.

www.cfilt.iitb.ac.in/sarcasmsuite/

Sarcasm

Sarcasm is defined as ‘the use of irony to mock or convey contempt’

(Source: Oxford Dictionary)

I had a great time waiting for you in the sun for two hours.

Three components of sarcasm:

(a) Ironic language (implied meaning different from surface meaning),

(b) Negative sentiment,

(c) Presence of a target

Motivation for Computational Sarcasm

	Precision (Sarc)	Precision (Non-sarc)
Conversation Transcripts		
MeaningCloud	20.14	49.41
NLTK (Bird, 2006)	38.86	81
Tweets		
MeaningCloud	17.58	50.13
NLTK (Bird, 2006)	35.17	69

A challenge to dialogue agents

Human: You are fast like a snail

ALICE (Wallace, 2009): Thank you for telling me I am fast like a snai

Capture Incongruity

Some incongruity may occur without the presence of sentiment words

This can be captured using word embedding-based features, **in addition to other features**

"A man needs a woman like a fish needs bicycle."

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

Word embedding-based features

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

Experiment Setup

- Dataset: 3629 Book snippets (759 sarcastic) downloaded from GoodReads website
- Labelled by users with tags
- 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

	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)

Results (2/2)

	Word2Vec	LSA	GloVe	Dep. Wt.
+S	0.835	0.86	0.918	**0.978**
+WS	**1.411**	0.255	0.192	1.372
+S+WS	**1.182**	0.24	0.845	0.795

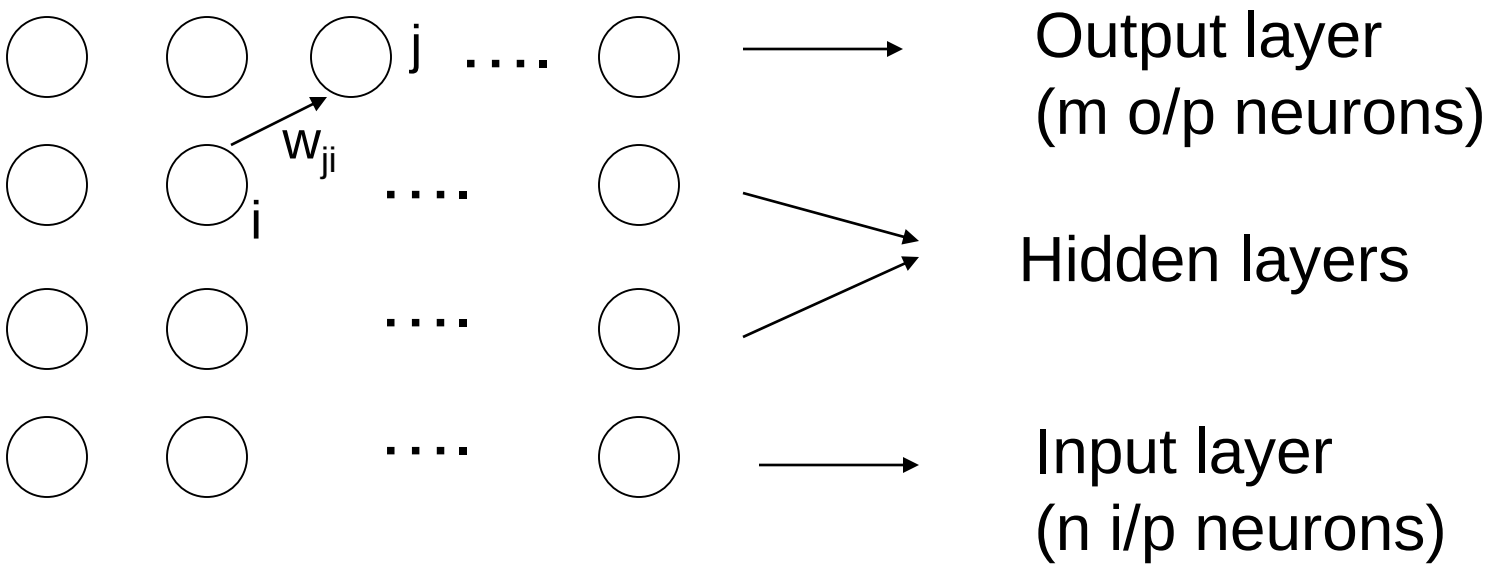
Table 4: Average gain in F-Scores obtained by using intersection of the four word embeddings, for three word embedding feature-types, augmented to four prior works; Dep. Wt. indicates vectors learned from dependency-based weights

Word Embedding	Average F-score Gain
LSA	0.452
Glove	0.651
Dependency	1.048
Word2Vec	1.143

Table 5: Average gain in F-scores for the four types of word embeddings; These values are computed for a subset of these embeddings consisting of words common to all four

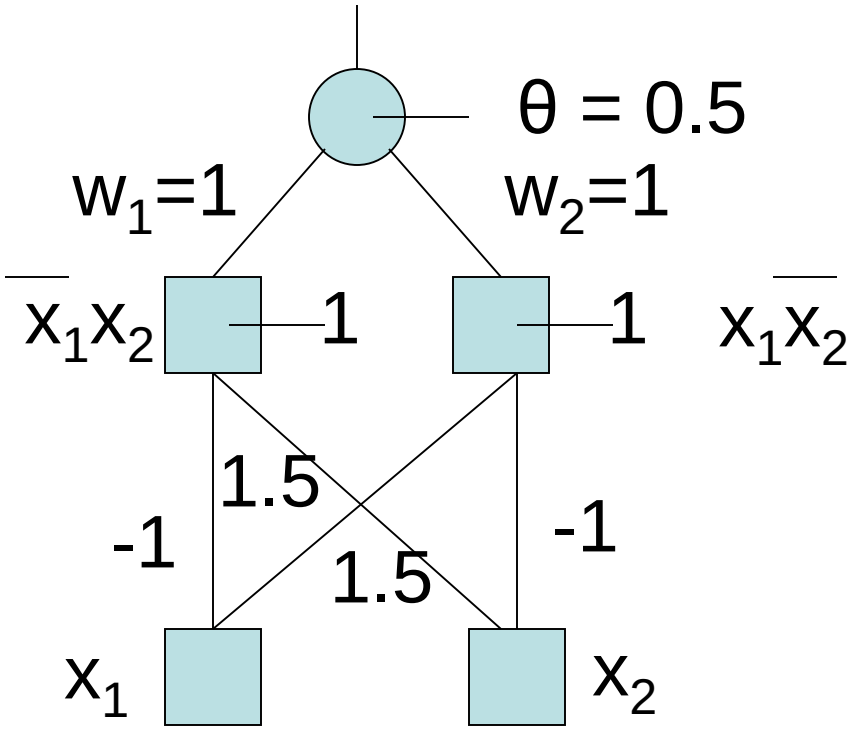
NLP and Deep Neural Nets

Deep neural net

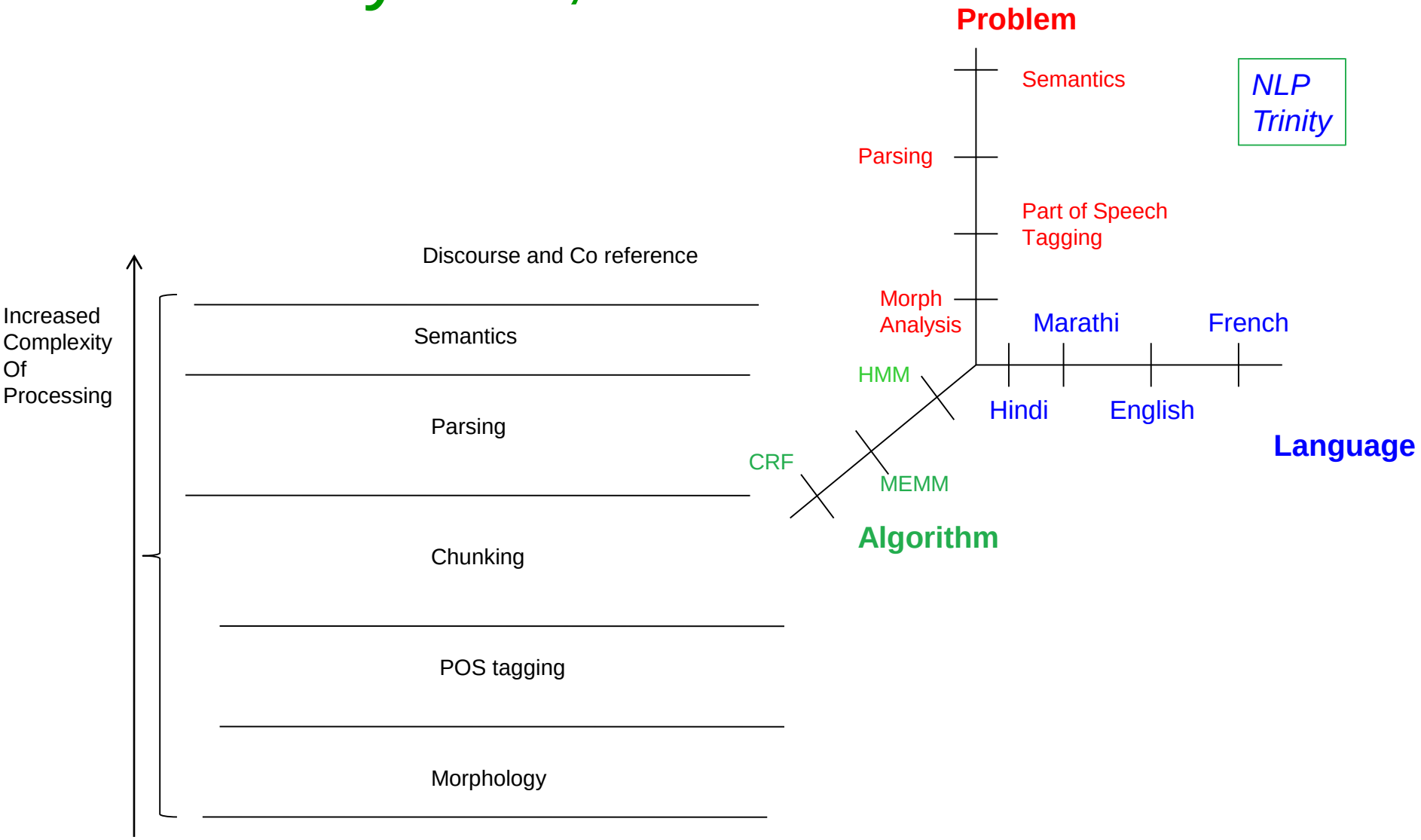


- NLP pipeline ←→ NN layers
- Discover bigger structures bottom up, starting from character?
- Words, POS, Parse, Sentence, Discourse?

Example- XOR: automatic discovery of computation (features)



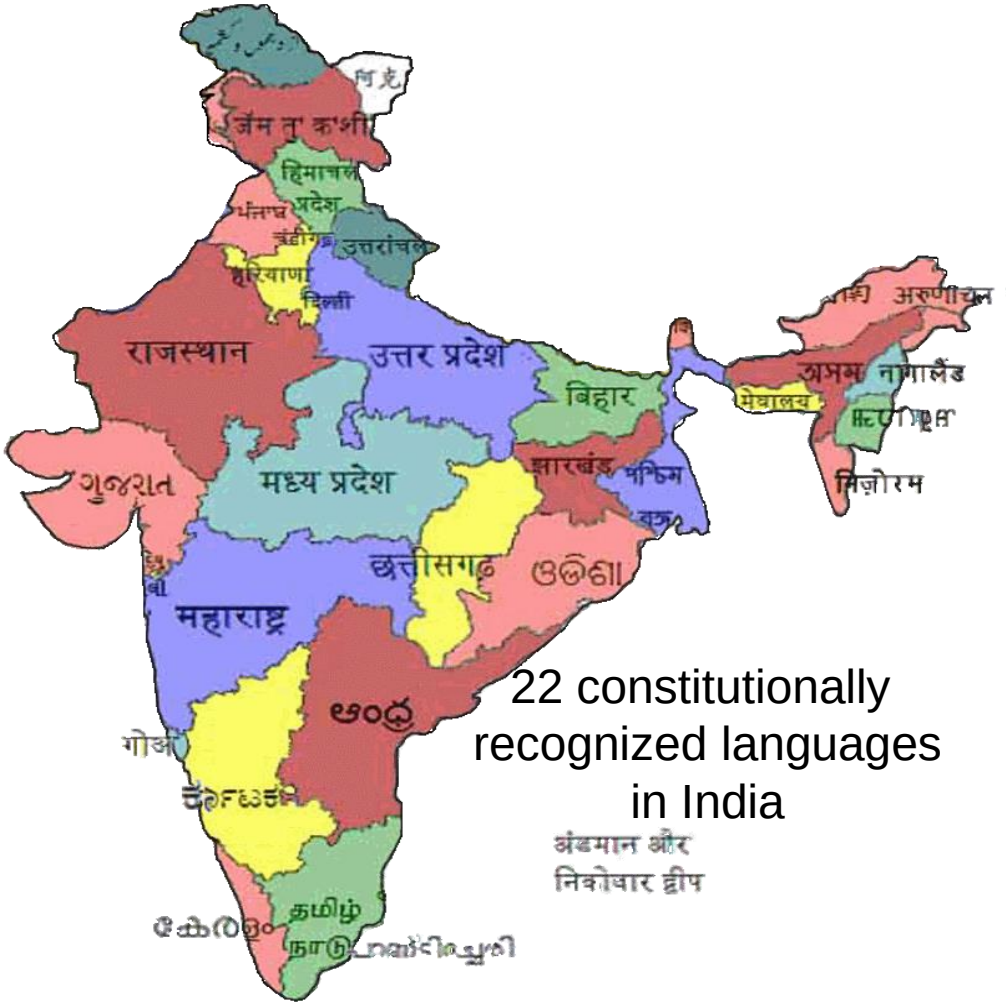
NLP: layered, multidimensional



DL yet to prove itself for text

- NMT a particular instance of solving mapping problems by neural networks
- Spectacular success in speech and vision (as high as 50% reduction in error rate)

a multilingual world, A Multilingual country



22 constitutionally recognized languages in India

>1500 languages spoken in India

First 10 spoken languages (by population)

Rank	Language	Native speakers in millions 2007 (2010)	Fraction of world population (2007)
1	**Mandarin** (entire branch)	935 (955)	14.1%
2	**Spanish**	390 (405)	5.85%
3	**English**	365 (360)	5.52%
4	**Hindi** [Note 1]	295 (310)	4.46%
5	**Arabic**	280 (295)	4.23%
6	**Portuguese**	205 (215)	3.08%
7	**Bengali**	200 (205)	3.05%
8	**Russian**	160 (155)	2.42%
9	**Japanese**	125 (125)	1.92%
10	**Punjabi**	95 (100)	1.44%

Summary

- NLP=ambiguity processing
  - Hence becomes a classification problem
- Alignment in MT: predominantly ML; but cannot do without linguistics when dealing with rich morphology
- Word sense disambiguation using E-M algorithm

- Sarcasm (difficult sentiment analysis problem)
  - Good NLP (incongruity) + good ML

Conclusions

- Huge volume of text data needs automation- NLP and ML
- Both Linguistics and Computation needed: **Linguistics is the eye, Computation the body**
- 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?