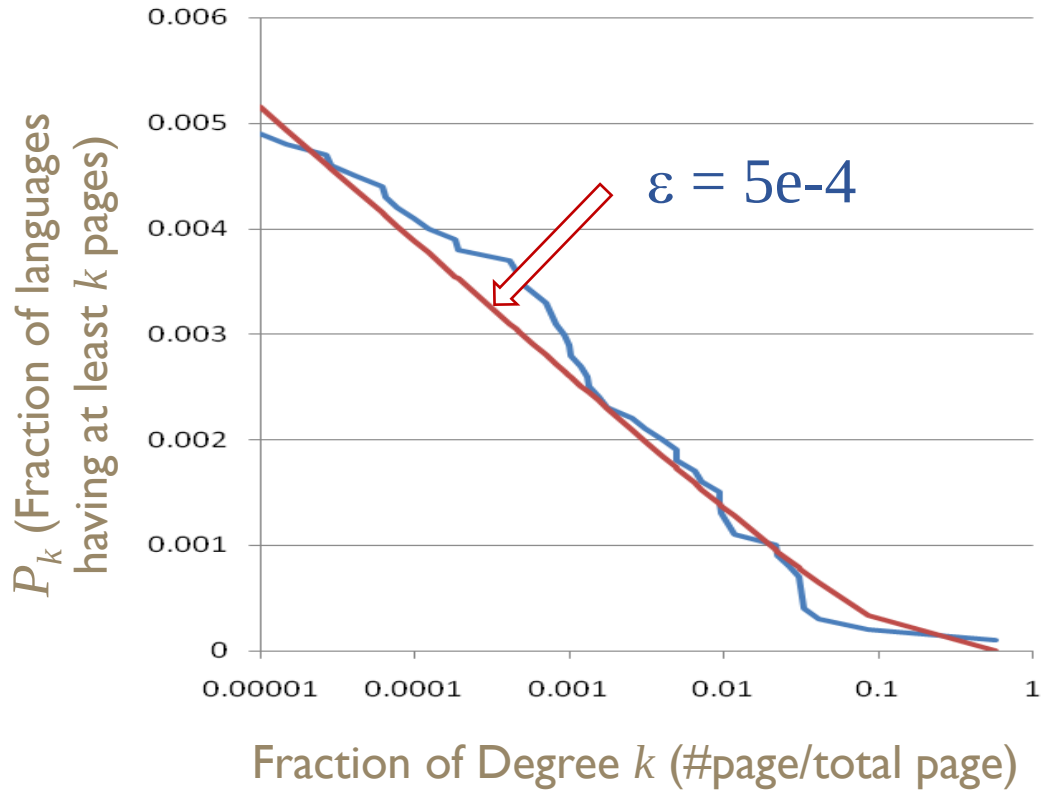
NLP for Social Media

Lecture 6: Processing Multilingual Content

Monojit Choudhury

Microsoft Research Lab, *monojitc@microsoft.com*

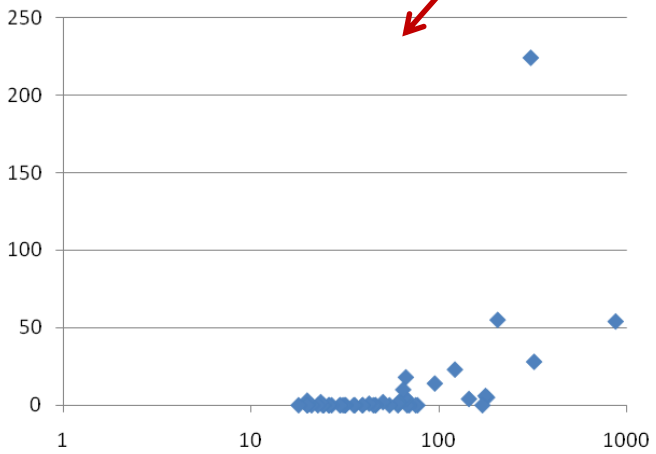
Distribution of Languages over WWW

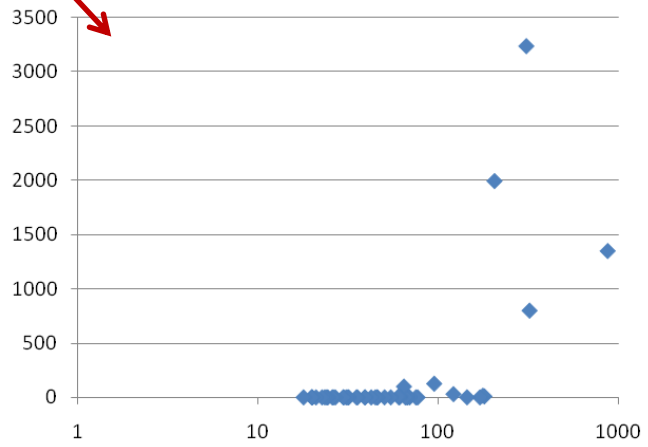


English
German
French
Russian
Japanese
Chinese
Spanish
Italian
Korean
Dutch
Portuguese
Czech
Swedish
Polish
Danish
Catalan
Norwegian
Hungarian
Finnish
Slovak
Turkish

How representative is WWW

Correlations	Speaker	Web	Wiki	LDC-Items
Web	0.29			
Wiki	0.36	0.92		
LDC-Items	0.52	0.94	0.85	
LDC-Words	0.62	0.78	0.71	0.94

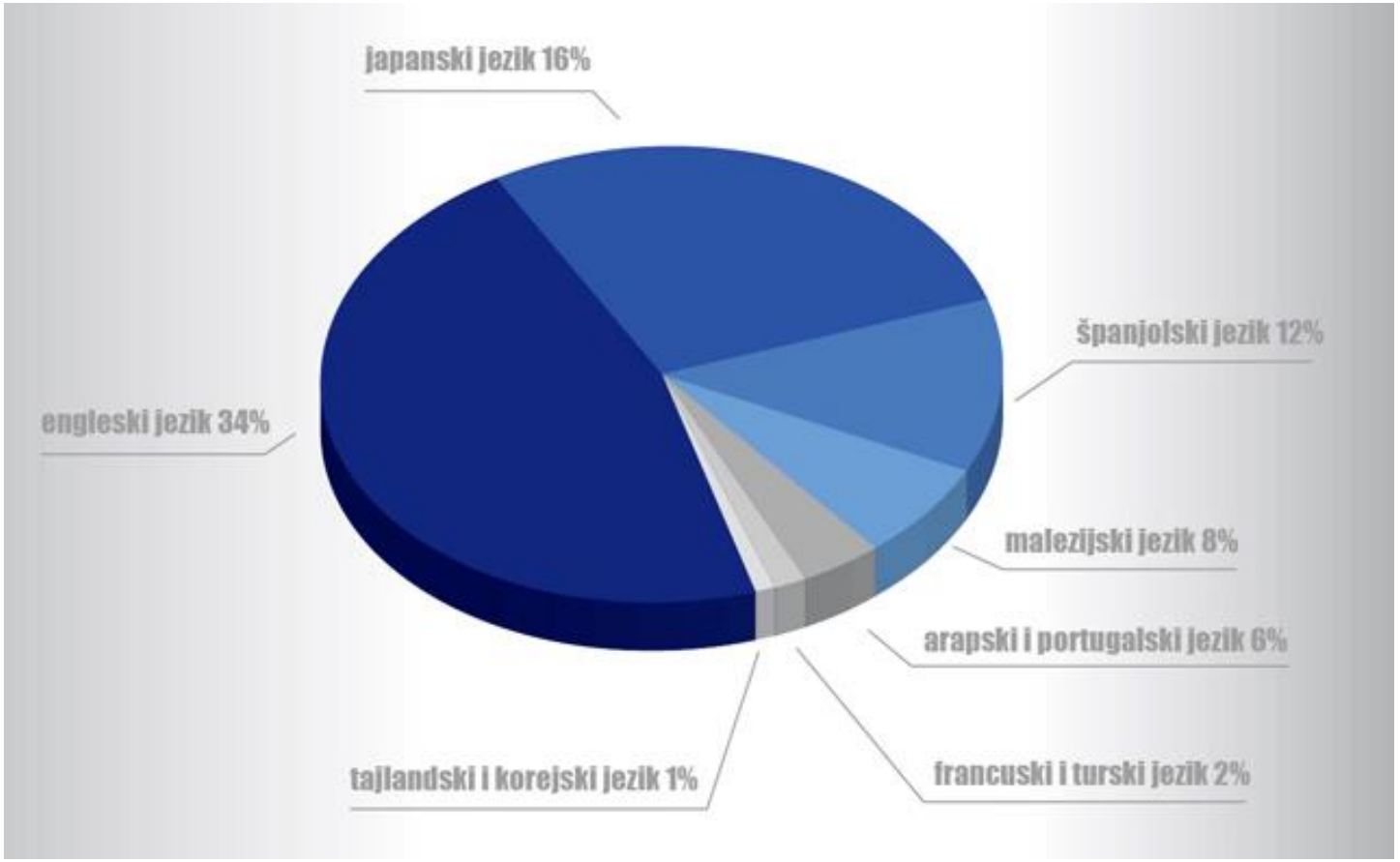




The Outliers!

Speaker	Web	Wiki	LDC-I	LDC-W
Mandarin	English	English	English	English
Spanish	German	German	Arabic	Arabic
English	French	French	Chinese	Chinese
Arabic	Russian	Polish	Spanish	Spanish
Hindi	Japanese	Japanese	Japanese	German
Portuguese	Chinese	Dutch	Korean	French
Bengali	Spanish	Italian	German	Japanese
Russian	Italian	Portuguese	French	Korean
Japanese	Korean	Spanish	Czech	Portuguese
German	Dutch	Swedish	Portuguese	Hindi
Wu	Portuguese	Russian	Hindi	Urdu
Javanese	Czech	Chinese	Tamil	Bengali
Telugu	Swedish	Norwegian	Farsi	Italian
Marathi	Polish	Finnish	Russian	Greek
Vietnamese	Danish	Turkish	Vietnamese	Swedish
Korean	Catalan	Esperanto	Dutch	Norwegian
Tamil	Norwegian	Romanian	Italian	Russian

What about Social Media?



Distribution of Languages on Twitter in 2013

http://www.ciklopea.com/en/blog/consulting/the-distribution-of-languages-on-twitter/531/

Questions to ponder

- What factors determine the distribution of languages on a social network?
- How do we compute or estimate this distribution?
- What technologies, if any, are needed to make an OSN accessible to the speakers of a language?
- What technologies are needed to support and encourage multilingualism on an OSN?
- What can we learn about multilingualism from OSNs?

Questions to ponder

- What factors determine the distribution of languages on a social network?
- How do we compute or estimate this distribution?
- What technologies, if any, are needed to make an OSN accessible to the speakers of a language?
- What technologies are needed to support and encourage multilingualism on an OSN?
- What can we learn about multilingualism from OSNs?

Language Detection

Processing Code-switching

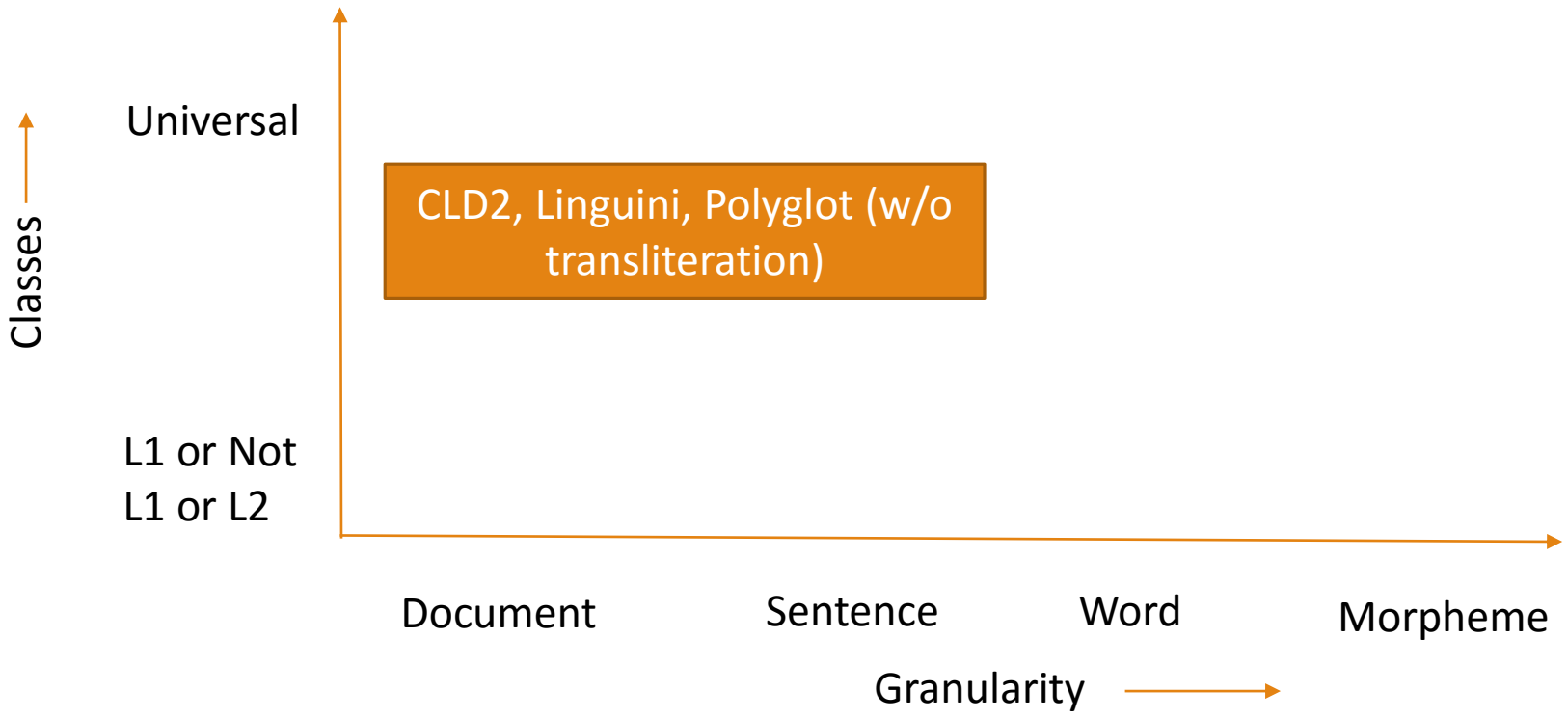
Some interesting stats

Agenda

- Language Detection
- Processing Code-switched text
- Some interesting stats

Scope of Language Identification

Scope of Language Identification



Doc-level Language Detection

Each Document is Monolingual

Most language identification research focuses on language identification for *monolingual* documents (Hughes et al., 2006). In monolingual LangID, the task is to assign each document D a unique language $L_i \in L$.

Language identification for multilingual documents is a multi-label classification task, in which a document can be mapped onto any number of labels from a closed set. In the remainder of this paper, we denote the set of all languages by L. We denote a document D which contains languages L_x and L_y as $D \rightarrow \{L_x, L_y\}$, where $L_x, L_y \in L$.

Documents can be multilingual

Lui, Lao and Baldwin (2014), Transactions of ACL

A brief History of Doc-level LI

LI for Web Documents (for Information Retrieval/Web Search)

1994: William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. Ann Arb[illegible], 48113(2):161–175.

1999: John M Prager. Linguini: Language identification for multilingual documents. In Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference.

2005: P. McNamee. Language identification: ***A solved problem*** suitable for undergraduate instruction. Journal of Computing Sciences in Colleges, 20(3).

2011: Erik Tromp and Mykola Pechenizkiy. Graph-based n-gram language identification on short texts. In Proc. 20th Machine Learning conference of Belgium and The Netherlands, pages 27–34.

2012: Marco Lui and Timothy Baldwin. ***langid.py***: An off-the-shelf language identification tool. In Proceedings of the ACL 2012 System Demonstrations, pages 25–30.

2013: Moises Goldszmidt, Marc Najork, and Stelios Paparizos. Boot-strapping language identifiers for short colloquial postings. In Proc. of ECMLPKDD.

LI for short and noisy text (Twitter & other user generated content)

Doc-level LI: Approaches

How many languages use the Devanagari script?

Unicode Block

- Idea: Different languages use different scripts

English, French German, Spanish Portuguese, Swedish, Vietnamese, Tagalog, Malay, …

Russian, Bulgarian, Belorussian, Abkhasian, Serbian

Unicode blocks and contained scripts

Block range	Block name	Code points[a]	Assigned characters	Scripts[b][c][d][e][f]
U+0000..U+007F	Basic Latin[g]	128	128	Latin (52 characters), Common (76 characters)
U+0080..U+00FF	Latin-1 Supplement[h]	128	128	Latin (64 characters), Common (64 characters)
U+0100..U+017F	Latin Extended-A	128	128	Latin
U+0180..U+024F	Latin Extended-B	208	208	Latin
U+0250..U+02AF	IPA Extensions	96	96	Latin
U+02B0..U+02FF	Spacing Modifier Letters	80	80	Bopomofo (2 characters), Latin (14 characters), Common (64 characters)
U+0300..U+036F	Combining Diacritical Marks	112	112	Inherited
U+0370..U+03FF	Greek and Coptic	144	135	Greek (14 characters), Coptic (117 characters), Common (4 characters)
U+0400..U+04FF	Cyrillic	256	256	Cyrillic (254 characters), Inherited (2 characters)
U+0500..U+052F	Cyrillic Supplement	48	48	Cyrillic
U+0530..U+058F	Armenian	96	89	Armenian (88 characters), Common (1 character)
U+0590..U+05FF	Hebrew	112	87	Hebrew
U+0600..U+06FF	Arabic	256	255	Arabic (226 characters), Common (17 characters), Inherited (12 characters)

Doc-level LI: Approaches

Unicode Block

- Idea: Different languages use different scripts

Dictionary based

- Compute the intersection with each of the language lexicon. Declare the highest matching lexicon as the winner.
- Issues: Resource intensive; coverage; short text

N-gram based techniques

Which of this is Sanskrit?
kshiprata, altakmbil

Character n-gram based word classifiers

Task:

Input: A word *w*

Output: Yes (if w belongs to L1) or No (otherwise)

Features: *character n-grams (n* = 2 to 5)

Classifier: Naïve Bayes*, Max-Ent, SVMs

Data:

- Positive Examples: words of L1
- Negative example: words from other languages

Output: *prob or score* of *w* being L1

kshiprata → $kshiprata$

2: $k, ks, sh, hi, ip, pr, ra, at, ta, a$

3: $ks, ksh, shi, hip, ipr, … ta$

4: $ksh, kshi, ship, …, ata$

5: $kshi, kship, shipr, …, rata$

Prob(kshiprata is Sanskrit) >>
Prob(altakmbil is Sanskrit)

Doc-level LI: Approaches

Unicode Block

- Idea: Different languages use different scripts

Dictionary based

- Compute the intersection with each of the language lexicon. Declare the highest matching lexicon as the winner.
- Issues: Resource intensive; coverage; short text

N-gram based techniques

- Robust, easy to build, can be bootstrapped
- Issues: very short text, very noisy text

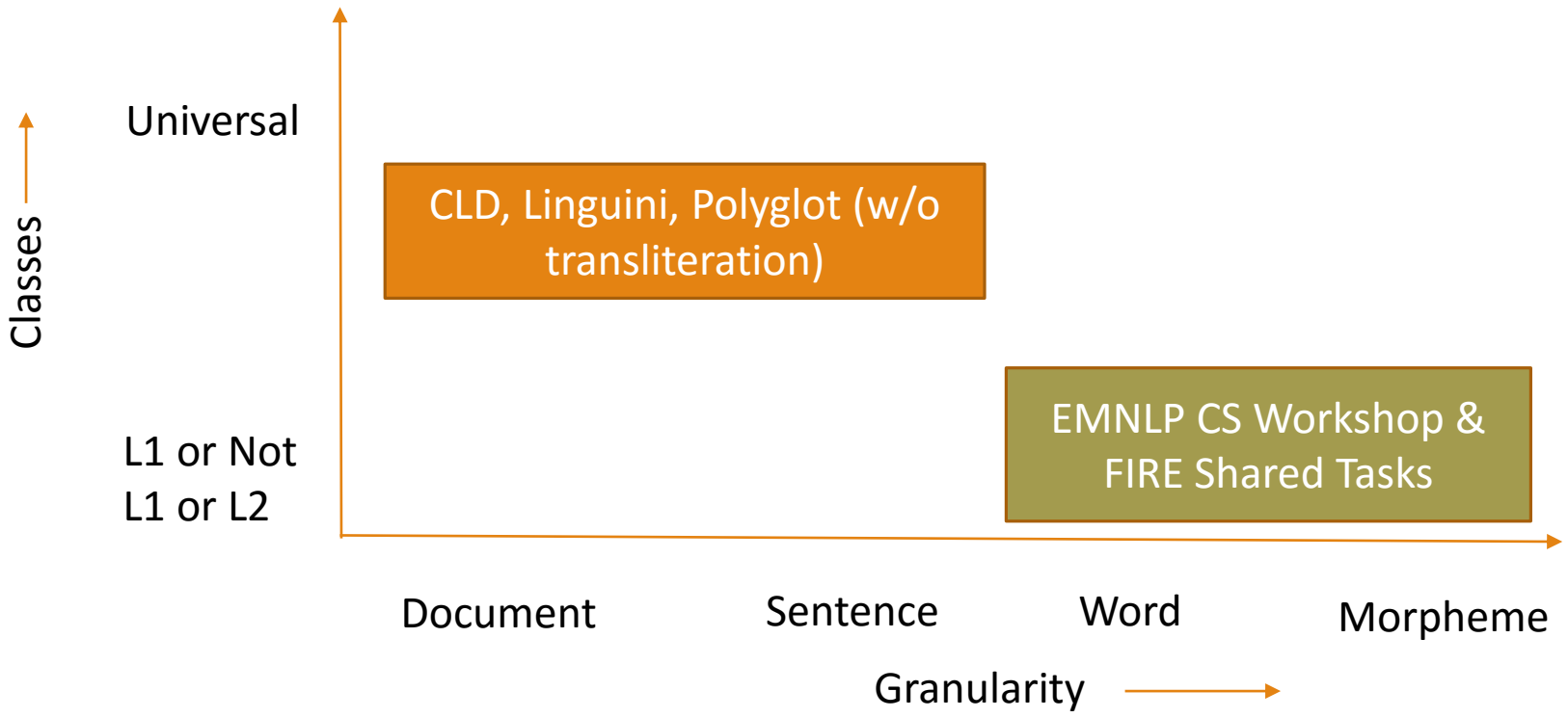
Other Features:

- Meta-data of a webpage
- User Info (in Twitter/social media)

Some off-the-shelf Tools

Tool	Reference	#Lang	Approach	Features	Type
linguini	Prager, 1999		Vector-space model	2-5 Byte n-grams	Multi
polyglot	Lui and Baldwin, 2011/14	44	Generative mixture model	Byte n-grams	Multi
langid.py	Lui and Baldwin, 2012	97	Naïve Bayes Classifier	1,2,3,4 Byte-gram	Multi
CLD2	Google, 2013	83	Naïve Bayes Classifier	character 4-grams	Mono

Scope of Language Identification



Word-level Language Labeling: Problem Definition

Modi	ke	speech	se	India	inspired	ho	gaya	#namo
NE	Hn	En	Hn	NE	En	Hn	Hn	Other
	के		से			हो	गया	

Other Labels:

- Mix: Part L1, part L2 (e.g., *artiston, nachoing*)
- Ambiguous: can be either language (e.g., computer, vote, football)

A brief History of Doc-level LI

LI for Web Documents (for Information Retrieval/Web Search)

1994: William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. Ann Arb[illegible], 48113(2):161–175.

1999: John M Prager. Linguini: Language identification for multilingual documents. In Systems Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference.

2005: P. McNamee. Language identification: ***A solved problem*** suitable for undergraduate instruction. Journal of Computing Sciences in Colleges, 20(3).

2011: Erik Tromp and Mykola Pechenizkiy. Graph-based n-gram language identification on short texts. In Proc. 20th Machine Learning conference of Belgium and The Netherlands, pages 27–34.

2012: Marco Lui and Timothy Baldwin. ***langid.py***: An off-the-shelf language identification tool. In Proceedings of the ACL 2012 System Demonstrations, pages 25–30.

2013: Moises Goldszmidt, Marc Najork, and Stelios Paparizos. Boot-strapping language identifiers for short colloquial postings. In Proc. of ECMLPKDD.

LI for short and noisy text (Twitter & other user generated content)

A Brief History of Word-level Language Labeling

2008: T Solorio and Y. Liu. Parts-of-speech tagging for English-Spanish code-switched text. In Proceedings of the Empirical Methods in natural Language Processing.

2013: Ben King and Steven Abney. Labeling the languages of words in mixed-language documents using weakly supervised methods. In Proceedings of NAACL-HLT, pages 1110–1119.

2013: Rishiraj Saha Roy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. Overview and datasets of FIRE 2013 **track on Transliterated Search**. In FIRE Working Notes.

2014: Monojit Choudhury, Gokul Chittaranjan, Parth Gupta and Amitava Das. Overview FIRE 2014 **track on Transliterated Search**. In FIRE Working Notes.

2014: Thamar Solorio et al. Overview for the **First Shared Task on Language Identification** in Code-Switched Data.

2014: Utsab Barman, Amitava Das, Joachim Wagner and Jennifer Foster. Code Mixing: A Challenge for Language Identification in the Language of Social Media. 1st Workshop on Code-switching, EMNLP'14

Word-level Language Labeling: Problem Definition

Modi	ke	speech	se	India	inspired	ho	gaya	#namo
NE	Hn	En	Hn	NE	En	Hn	Hn	Other
	के		से			हो	गया	

Other Labels:

- Mix: Part L1, part L2 (e.g., *artiston, nachoing*)
- Ambiguous: can be either language (e.g., computer, vote, football)

Modeling as a Structured Prediction Problem

Given $\mathbf{X}$: $X_1 = Modi$, $X_2 = ke$, …,

Output: $\mathbf{Y} = Y_1$ (label for X_1), Y_2 (label for X_2) …

Such that $p(\mathbf{Y}|\mathbf{X})$ is maximized

Hidden Markov Models, Conditional Random Fields,

Features

Training & Test Data

Features

Token-based features	Lexical Features	Context Features
• Capitalization • Script • Special Characters • Character n-gram based classifiers • Word length	• Regular lexicon • Unigram Frequency • Entity Lexicon • Acronym/slang lexicon	• Next 3 tokens • Last 3 tokens • Current token • Previous label (Bigram or B)

Datasets & Metrics

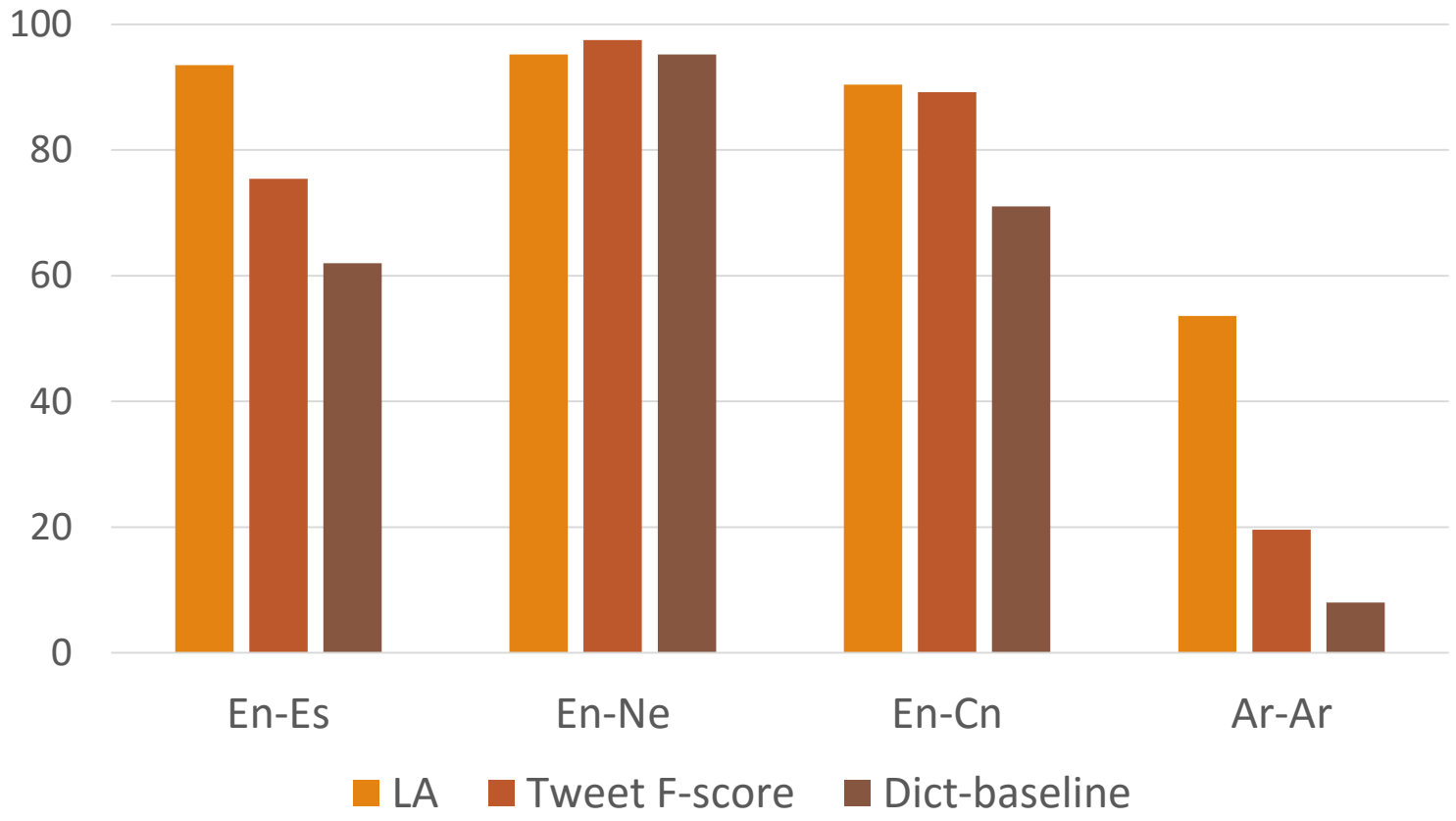
Language-pair	Training	Test	Surprise
MAN-EN	1000	313	n/a
MSA-DA	5,838	2332, 1,777	12,017
NEP-EN	9,993	3,018 (2,874)	1,087
SPA-EN	11,400	3,060 (1,626)	1,102

Metrics:

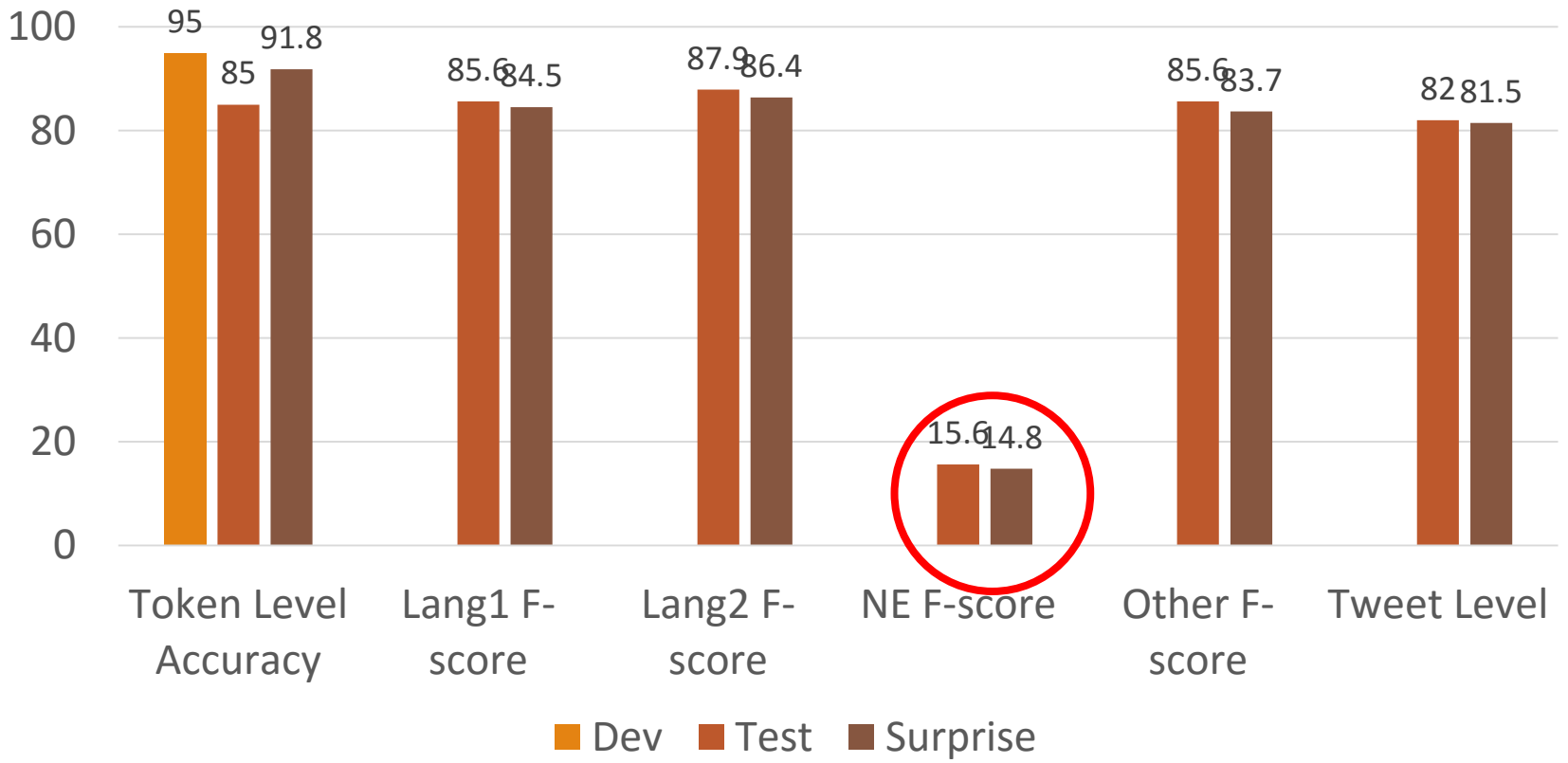
- Word-level labeling accuracy
- Word level Class-wise Precision, Recall and F-score
- Tweet (doc) level accuracy
- Tweet (doc) level CS Precision, Recall and F-score.

Performance

Shared Task in Code-switching Workshop@ EMNLP



Pain points



English-Spanish class wise F-score

Agenda

Language Detection

Processing Code-switched text

Some interesting stats

Did you like Interstellar?

Interstellar es una amazing movie.

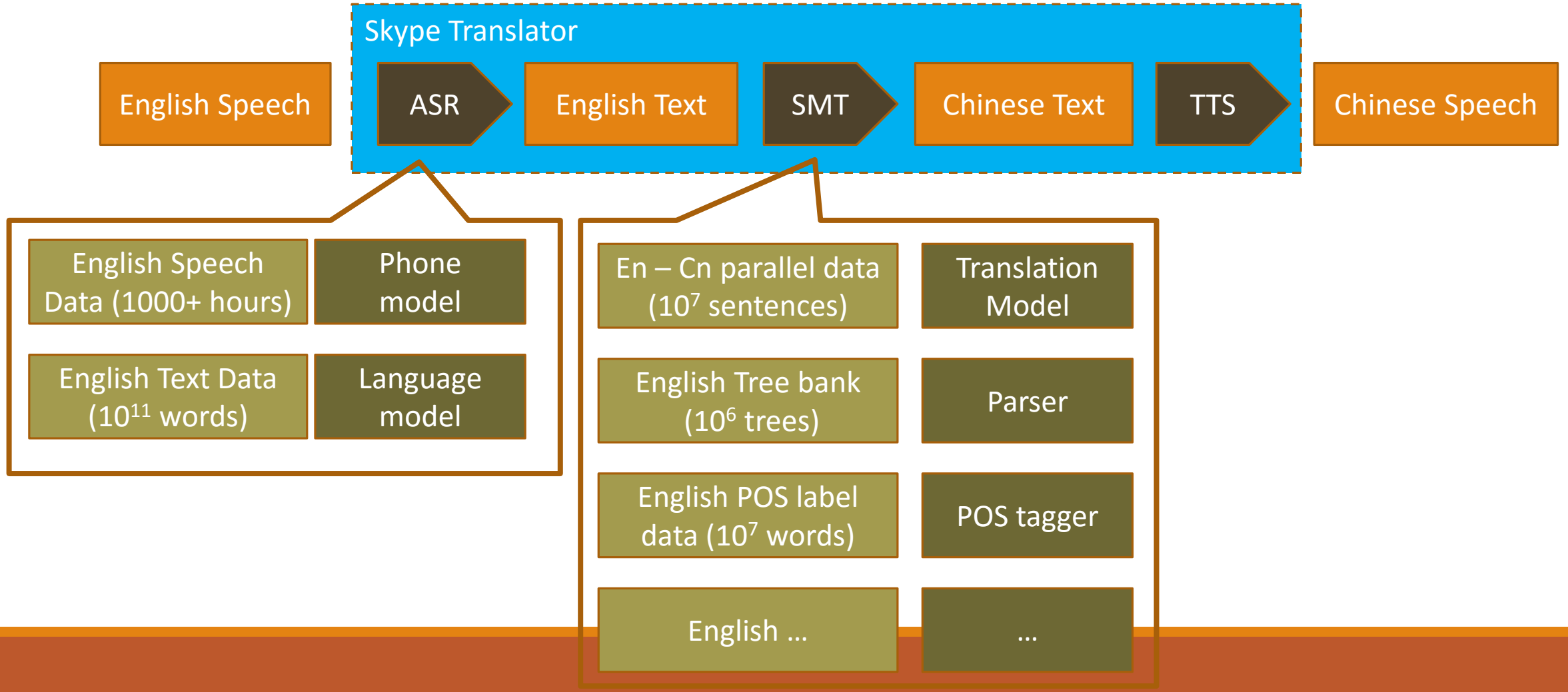
Spanglish

星际 es una 了不起的电影。

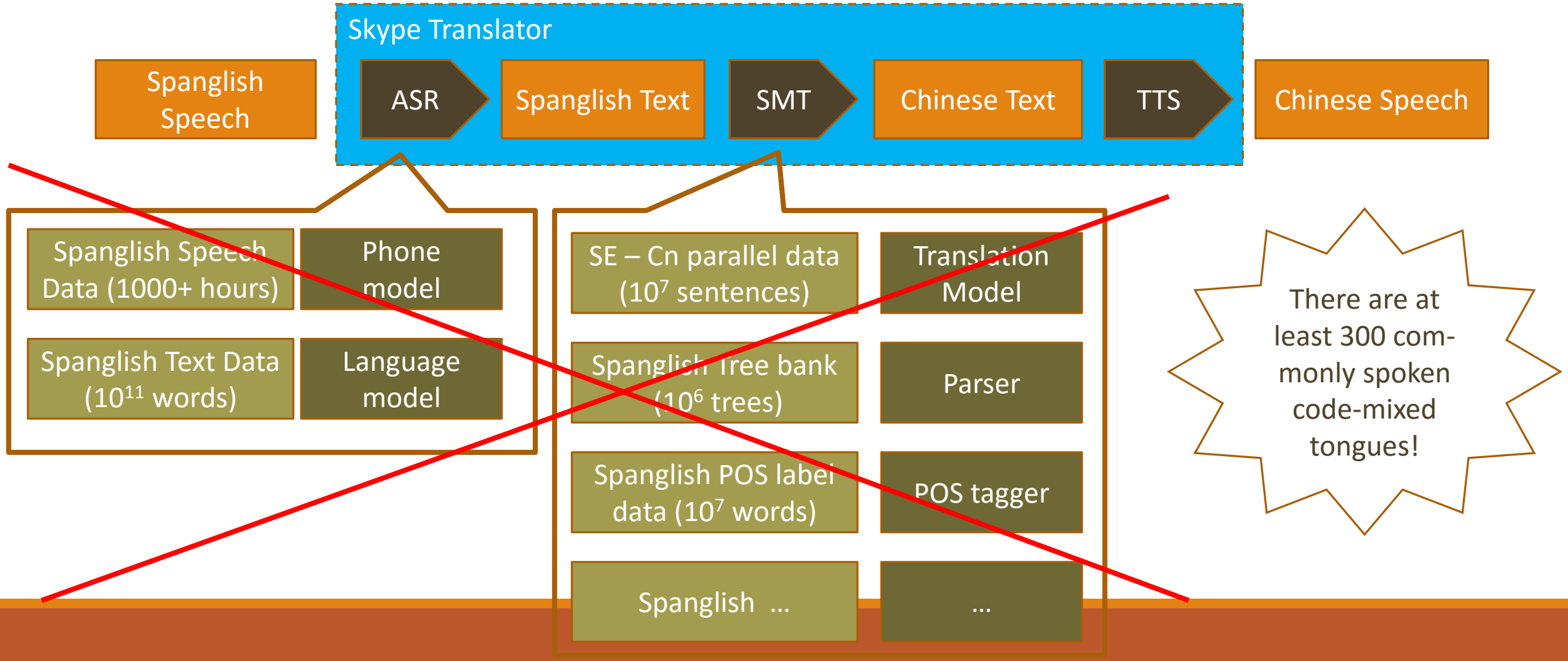
Interstellar 是了不起的电影。

Chinese

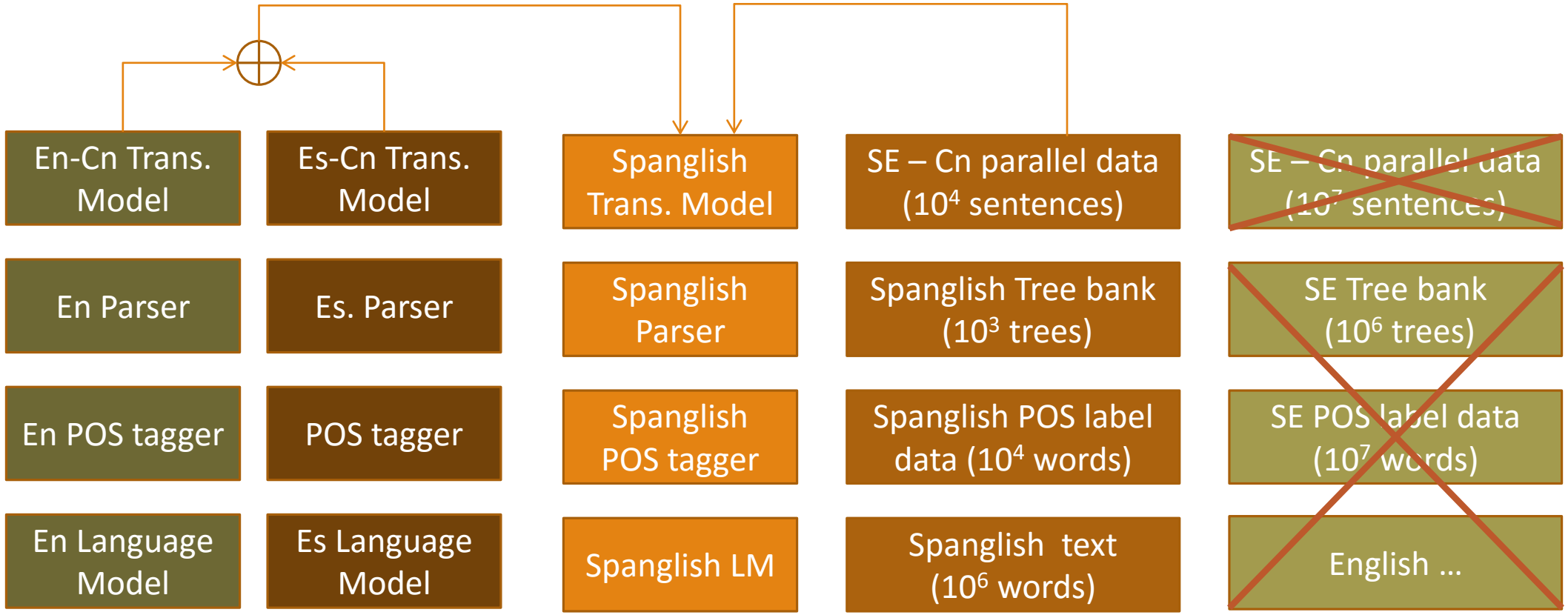
How does Skype Translator work?



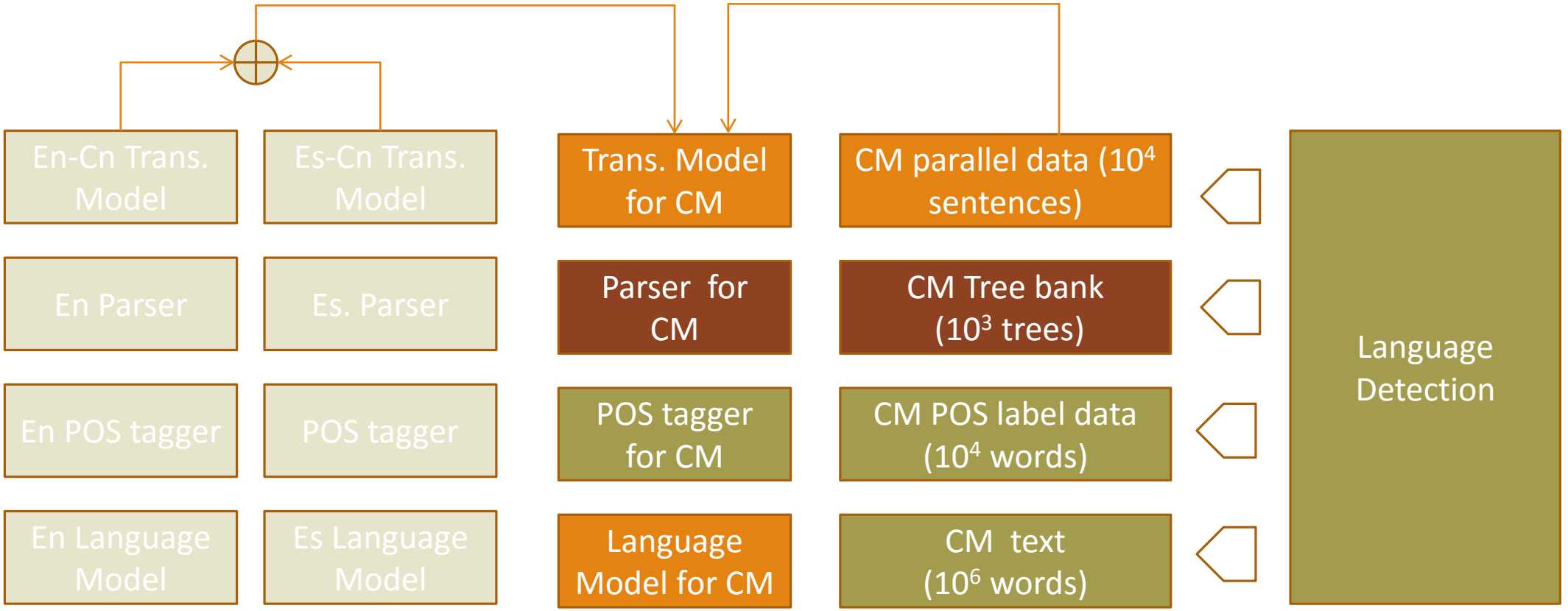
For Skyping in Spanglish…



For Skyping in Spanglish…



State-of-the-art



POS Tagging

Modi	ke	speech	se	India	inspired	ho	gaya	#namo
NE	Hn	En	Hn	NE	En	Hn	Hn	Other
	के		से			हो	गया	
NP	ADP	NN	ADP	NP	VB	VB	VB	X

T Solorio and Y. Liu. 2008. Parts-of-speech tagging for English-Spanish code-switched text. In Proceedings of the Empirical Methods in natural Language Processing.

---

1. Tag the whole sentence using L1 tagger [L1 POS annotated data]
2. Tag the whole sentence using L2 tagger [L2 POS annotated data]
3. Use the L1 tag and L2 tag as features (plus more) and learn to predict the POS tag for CM text [CM annotated data]

En-Es Results: Heuristic based combinations

	Heuristic	Accuracy (%)
1	Spanish Tree Tagger	25.99
	English Tree Tagger	54.59
2	Highest prob tag or English	51.51
	Highest prob tag or Spanish	49.16
3	Prob + special tags + lemmas	64.27
4	Dictionary-based Language Id	**86.03**
	Character 5-grams Language Id	81.46
	Human Language Id	89.72

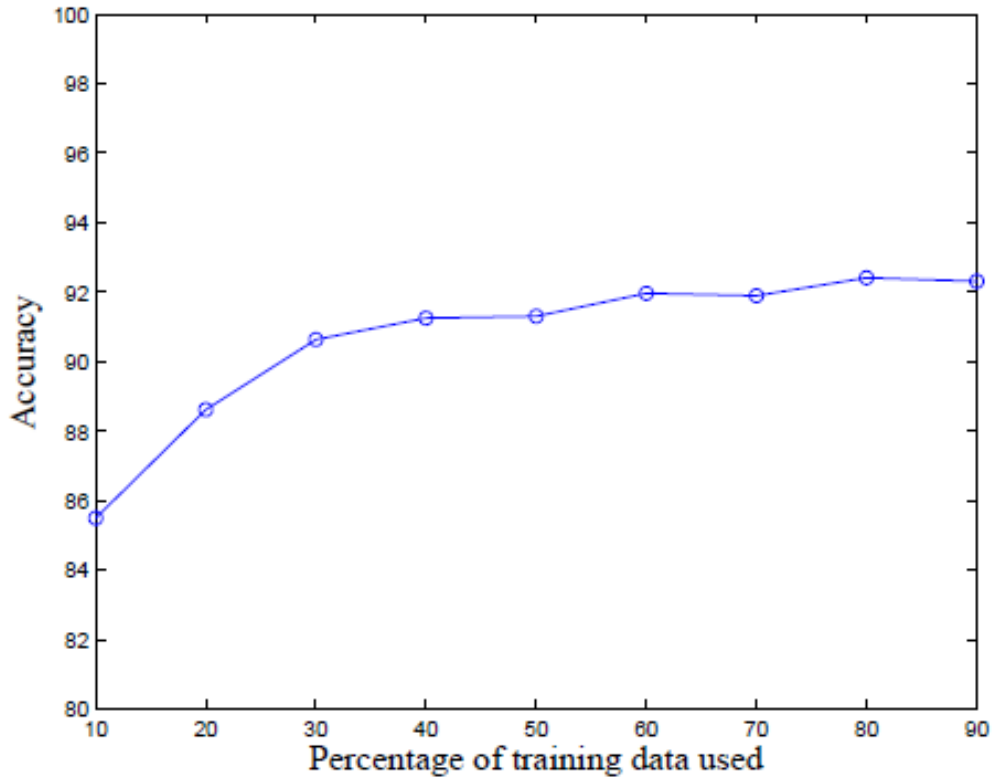
En-Es Results: Machine Learning Techniques

Features

1. The word (word)
2. English POS tag (E_t)
3. English POS tagger lemma (E_l)
4. English POS tagger confidence (E_p)
5. Spanish POS tag (S_t)
6. Spanish POS tagger lemma (S_l)
7. Spanish POS tagger confidence (S_p)

ML Algorithms	Mean Accuracy (%)	Variance
Naive Bayes	88.50	1.9280
SVM	**93.48**	1.2784
Logit Boost	**93.19**	1.4437
J48	91.11	2.1527
Oracle	90.31	-
Language Id	85.80	-

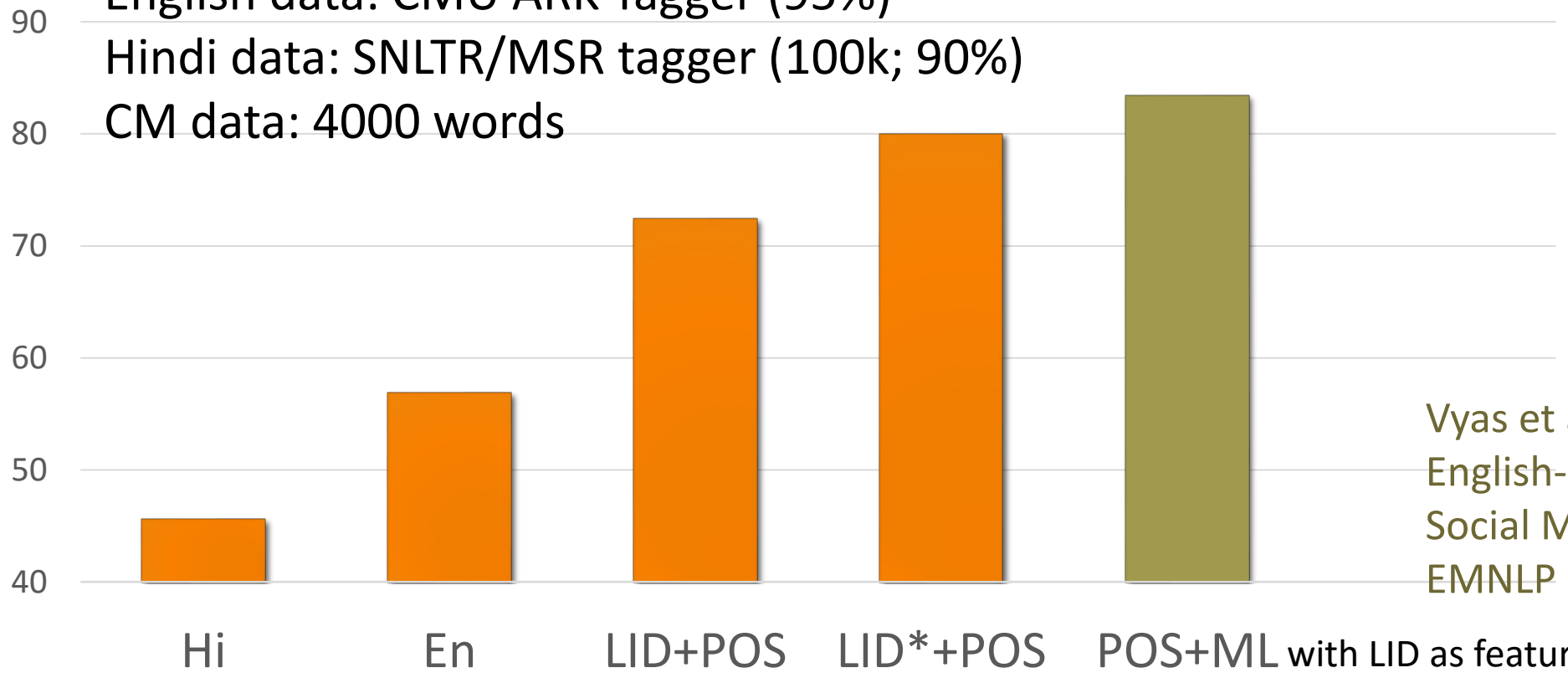
POS-tagged CM data requirement



English data: Penn Treebank (97%)
5 Million words
Spanish data: CRATER
CM data: 8000 words

Some experiments with Hindi

English data: CMU ARK Tagger (95%)
Hindi data: SNLTR/MSR tagger (100k; 90%)
CM data: 4000 words



Vyas et al. POS Tagging of English-Hindi Code-Mixed Social Media Content. EMNLP 2014

Agenda

Language Detection

Processing Code-switched text

Some interesting stats

Script Distribution of FB Posts

Facebook Page	Deva-nagari	Roman	Mixed Script	Other Script
Amitabh Bachhcan	73	3168	112	16
BBC Hindi	56	175	27	0
Narendra Modi	77	2633	84	11
Shahrukh Khan	0	578	23	1

Code-Switching Stats on FB

In the 4 public forums studied:

- All threads are multilingual
- 17.2% of the comments/posts have code-switching or mixing
- 04.2% have code-switching
- 23.7% of Romanized Hindi posts have at least one or more English embeddings
- 7.20% of the English posts have at least one or more Hindi embeddings