Google

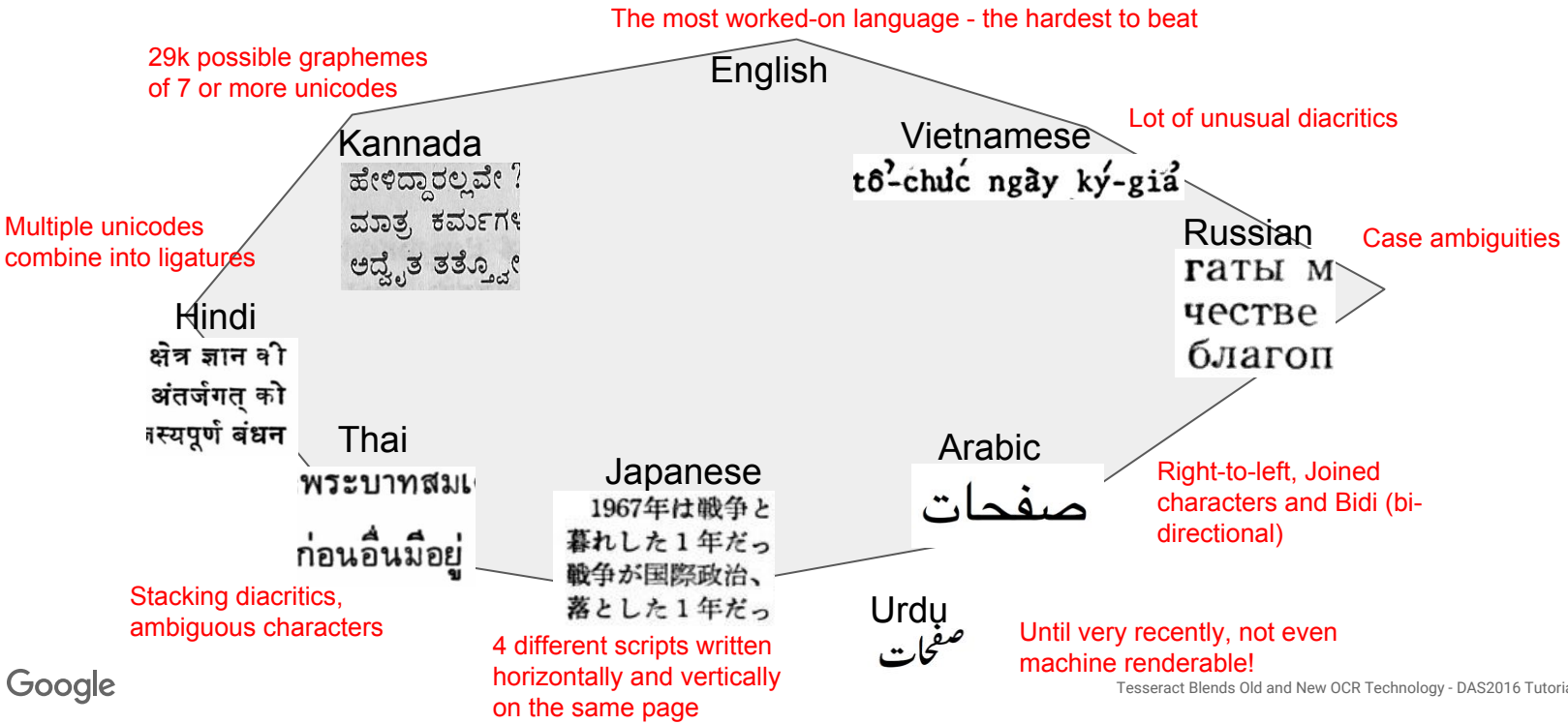
7. Building a Multilingual OCR Engine

Training LSTM networks on 100 languages and test results

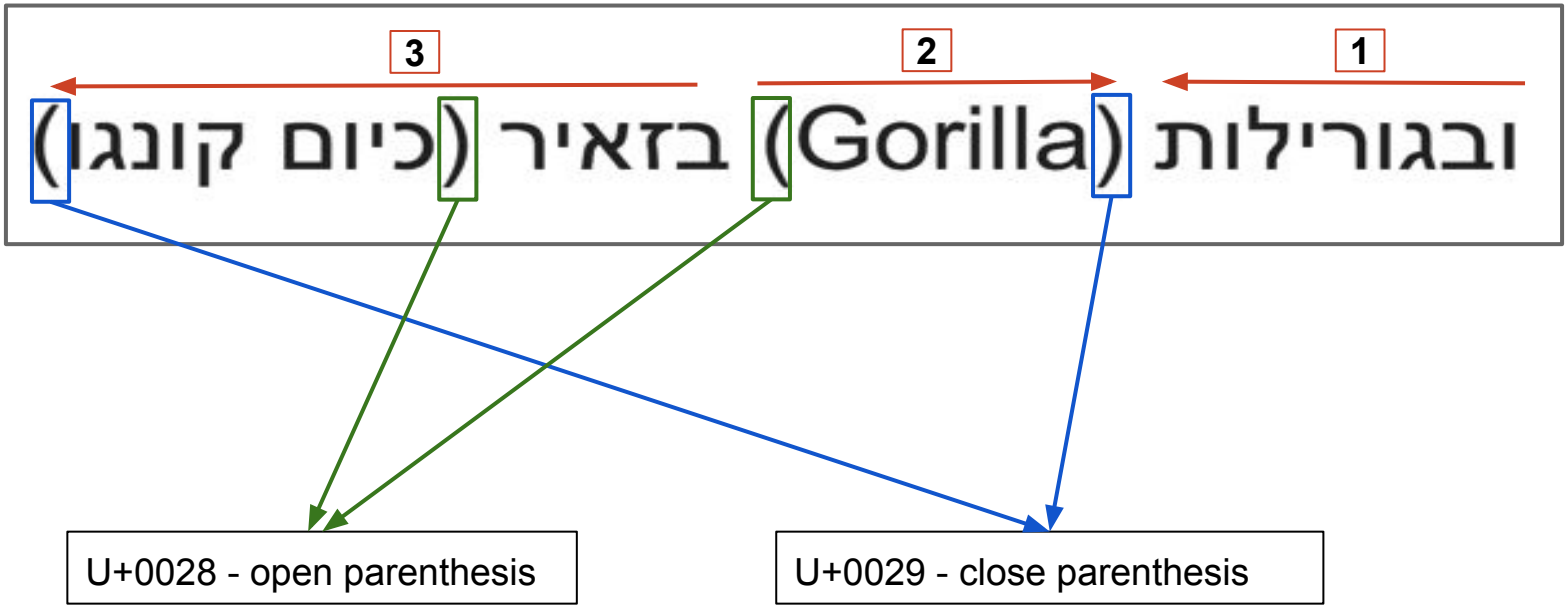
Ray Smith, Google Inc.

Internationalization: The Convex Hull of Languages

If you want to develop multilingual OCR, work on these first...



Bidirectional issues



What's a "character" in Devanagari?

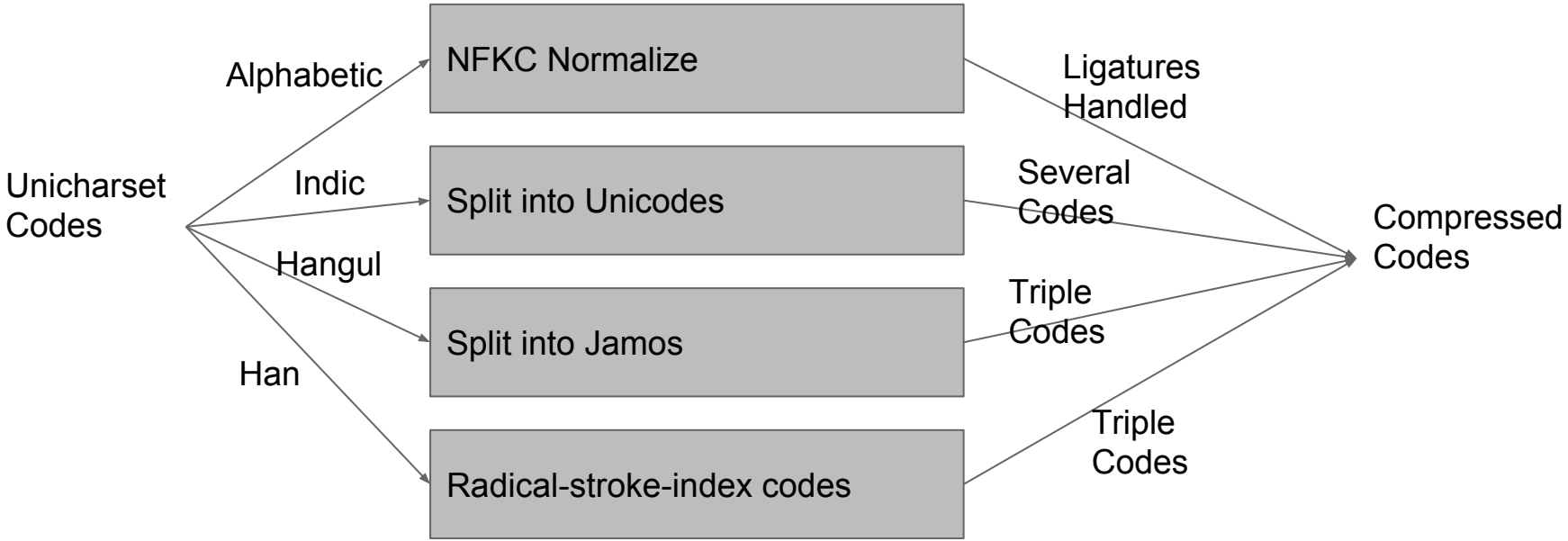
Result	Unicode	Transliteration
र	0930	ra
र्	0930 094d	r
र्द	0930 094d 0926	rda
र्द्	0930 094d 0926 094d	rd
र्द्व	0930 094d 0926 094d 0935	rdva
र्द्वि	0930 094d 0926 094d 0935 093f	rdvi
र्द्विक	0930 094d 0926 094d 0935 093f 0915	rdvika

What's a "character" in Kannada?

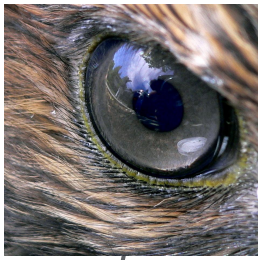
Result	Unicode	Transliteration
ರ	0cb0	ra
ದ	0ca6	da
ವ	0cb5	va
ರ್	0cb0 0ccd	r
ರ್ದ	0cb0 0ccd 0ca6	rda
ರ್ದ್	0cb0 0ccd 0ca6 0ccd	rd
ರ್ದ್ವ ದ್ವ್	0cb0 0ccd 0ca6 0ccd 0cb5	rdva
ರ್ದ್ವ್ ದ್ವ್	0cb0 0ccd 0ca6 0ccd 0cb5 0ccd	rdv
ರ್ದ್ವ್ಕ ದ್ವ್ಕ್	0cb0 0ccd 0ca6 0ccd 0cb5 0ccd 0c95	rdvka
ರ್ದ್ವ್ಕಿ ದ್ವ್ಕಿ	0cb0 0ccd 0ca6 0ccd 0cb5 0ccd 0c95 0cbf	rdvki

Universal Character/Grapheme Encoding/Compression

Extension to Tesseract's UNICHARSET to make the output Softmax smaller



Training International OCR Engines



Tesseract Training

Manually generated Files

Eng

Web Crawl Repository

Language ID Map-Reduce

Eng

Dirty Language Corpora

Text Filtration

Cleaned Language Corpora

Eng

OCR Engine Training

OCR Shape Files

Eng

Realistic Text Rendering

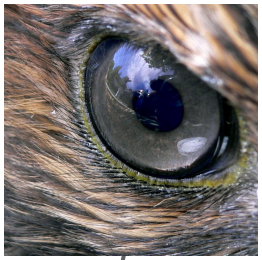
32 fonts

Language Model Generation

Language Model Files

Eng

Training International OCR Engines



T-LSTM Training

Manually generated Files

Eng

Web Crawl Repository

Language ID Map-Reduce

Eng

Dirty Language Corpora

Text Filtration

Cleaned Language Corpora

Eng

OCR Engine Training x500

OCR Shape Files

Eng

x100

5000 fonts

Realistic Text Rendering x500

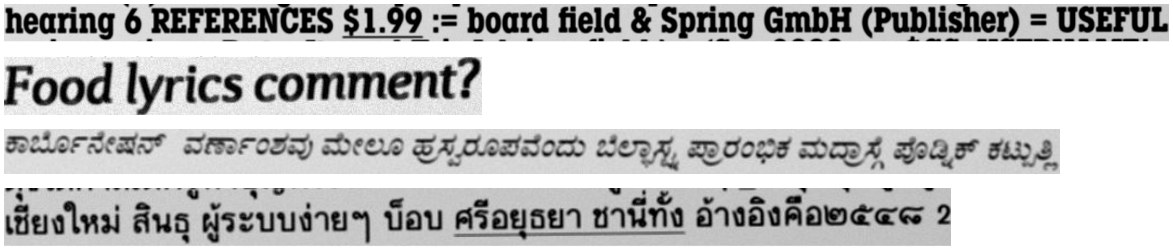
Language Model Generation

Language Model Files

Eng

Photo By Steve Jurvetson (http://www.flickr.com/photos/jurvetson/162116759) [CC BY 2.0 (http://creativecommons.org/licenses/by/2.0)], via Wikimedia Commons
https://commons.wikimedia.org/wiki/File%3AHawk_eye.jpg

Training



- Synthetic Training data:
  With bounding boxes
  Instead of CTC
- About 500k lines per language
- Random book-like degradation
- (Almost) Same network specification for each language:
- Convergence in 3-5 days or more

```
[G2,0C2,2FT16P3,3LQ1,64L1,128RtL1,128LS1,256]
[G2,0C2,2FT16P3,3LQ1,64L1,128RtL1,128LS1,512]
```

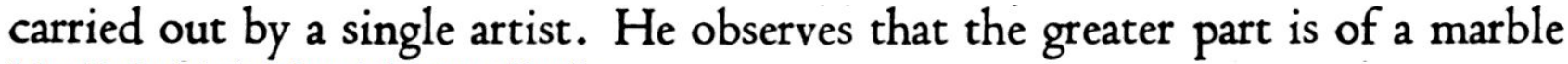
Testing

Testset from Google Books:

- Single Lines cut from older books.
- Hand typed. Accuracy far from perfect.
- 1000 lines * ~50 languages.

Caveat: Does not allow Tesseract to adapt to a whole page. (T-LSTM doesn't adapt)

Example:

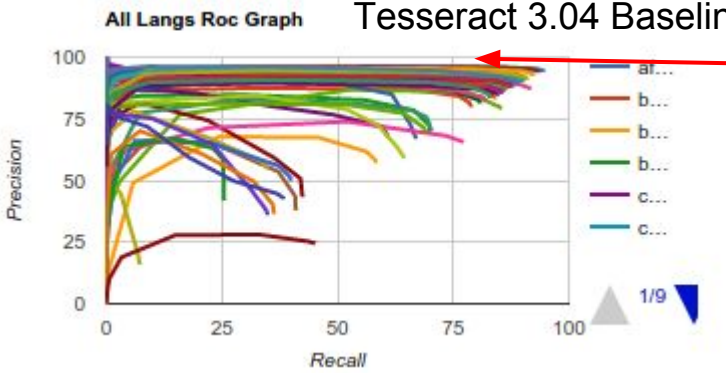


Truth: carried out by a single artist. He observes that the greatest part of the marble
OCR: carried out by a single artist. He observes that the greater part is of a marble
Confs: 0.88 0.93 0.93 0.93 0.93 0.93 0.92 0.93 0.93 0.93 0.93 0.93 0.93 0.93 0.65 0.65
Diff: carried out by a single artist. He observes that the greater part ▪ is of a marble
Recall Errors = 2
Precision Errors = 2

Overall effect on 51 Languages

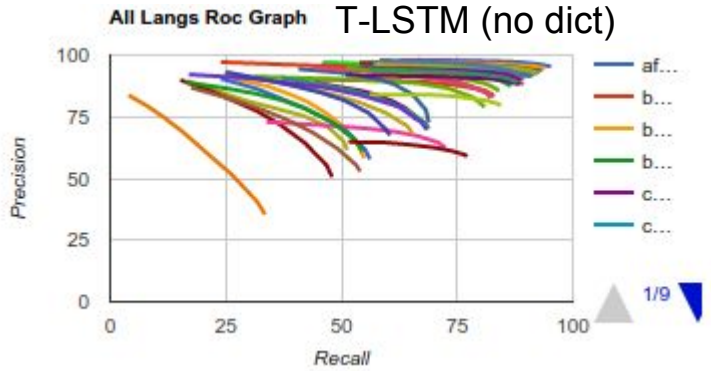
Impossible to resolve individual language results, but overall feel is improved

Tesseract 3.04 Baseline

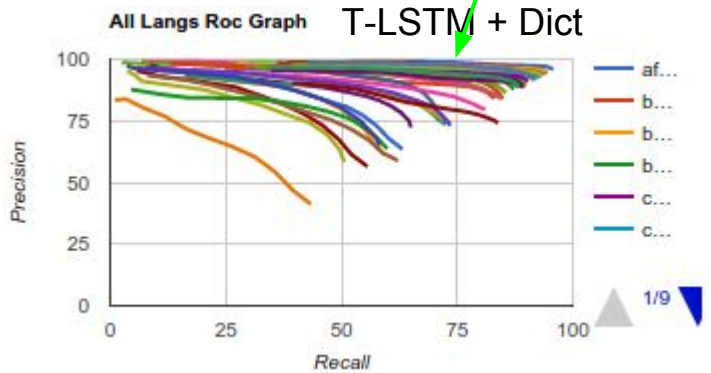


Notice this annoying precision ceiling is completely gone in the new version

T-LSTM (no dict)



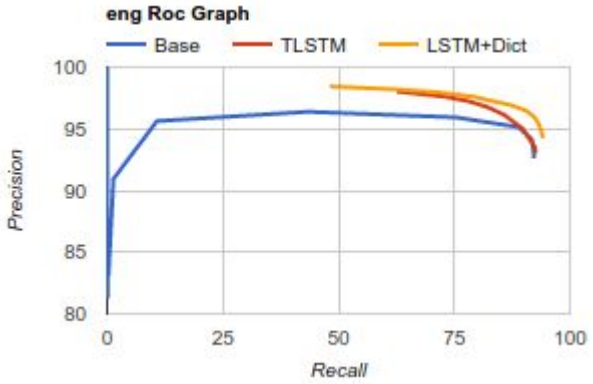
T-LSTM + Dict

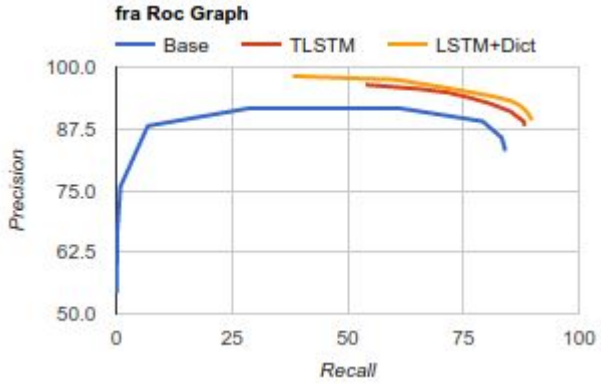


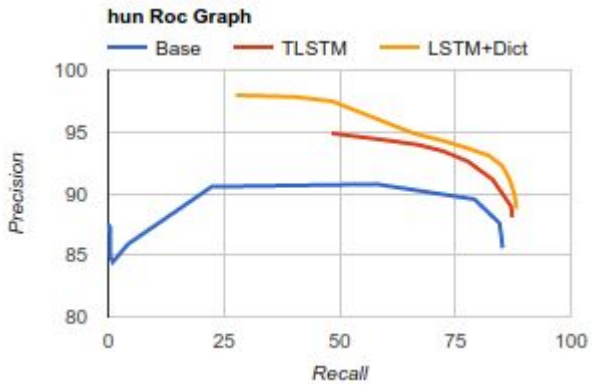
Accuracy Results: Selection of Latin Languages

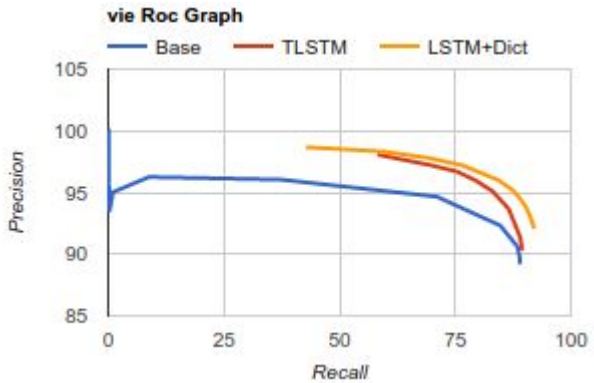
Lang	Truth Words	Char Error Rates			%Change	Word Error Rates			%Change
		Base	T-LSTM	LSTM+Dict		Base	T-LSTM	LSTM+Dict	
Czech	19319	3.08	1.86	1.81	-41.23	12.69	8.26	7.58	-40.27
English	21543	2.51	1.95	1.76	-29.88	7.58	7.18	5.77	-23.88
French	20746	5.98	3.1	2.98	-50.17	16.63	11.82	10.47	-37.04
Hungarian	17977	3.75	2.95	2.86	-23.73	14.61	12.26	11.45	-21.63
Indonesian	16616	2.77	1.96	2.26	-18.41	9.45	8.64	7.51	-20.53
Dutch	18878	6.23	6.65	5.97	-4.17	16.38	17.55	15.75	-3.85
Norwegian	18129	2.19	1.74	1.81	-17.35	7.73	5.87	5.7	-26.26
Portuguese	17726	2.91	1.9	1.87	-35.74	10.09	7.46	6.75	-33.10
Spanish	20333	4.28	2.5	2.31	-46.03	11.63	8.55	7.45	-35.94
Vietnamese	22472	4.04	3.55	2.61	-35.40	10.89	10.05	7.93	-27.18

Accuracy Results: Latin Langs ROC curves





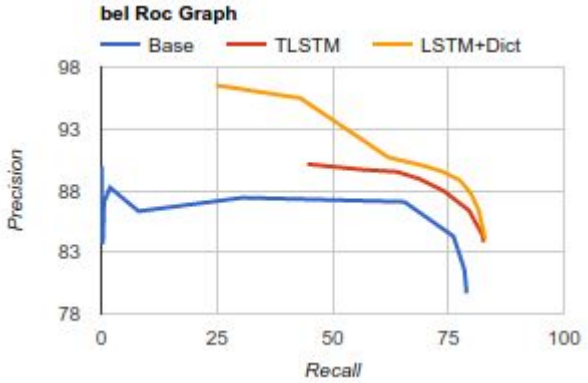


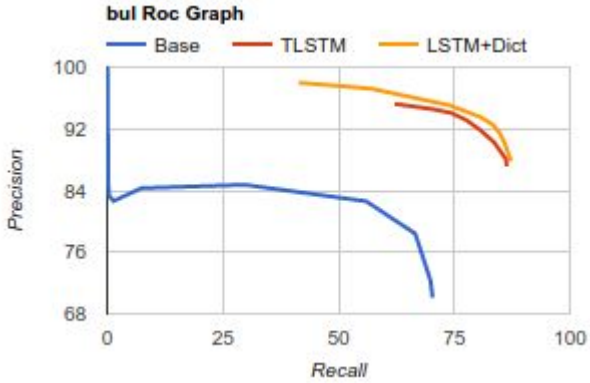


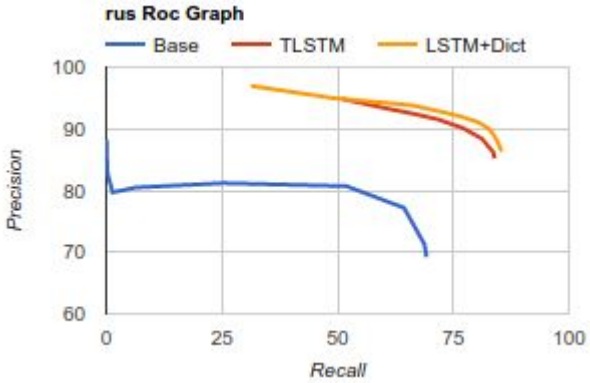
Accuracy Results: Cyrillic Languages

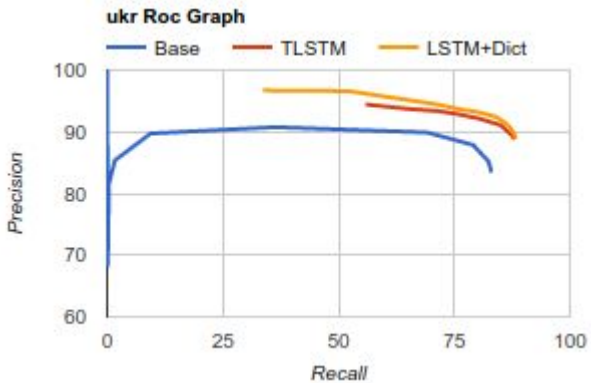
Lang	Truth Words	Char Error Rates			%Change	Word Error Rates			%Change
		Base	T-LSTM	LSTM+Dict		Base	T-LSTM	LSTM+Dict	
Belarusian	11697	7.72	4.25	4.60	-40.41	20.63	16.66	16.38	-20.60
Bulgarian	18457	19.68	3.78	3.85	-80.44	29.88	13.18	12.4	-58.50
Macedonian	17117	11.22	3.18	3.08	-72.55	18.92	11.69	10.61	-43.92
Russian	14993	19.06	4	4.10	-78.49	30.83	15.34	14.04	-54.46
Ukrainian	17123	6.63	2.77	3.27	-50.68	16.78	11.6	11.34	-32.42

Accuracy Results (Cyrillic Langs)





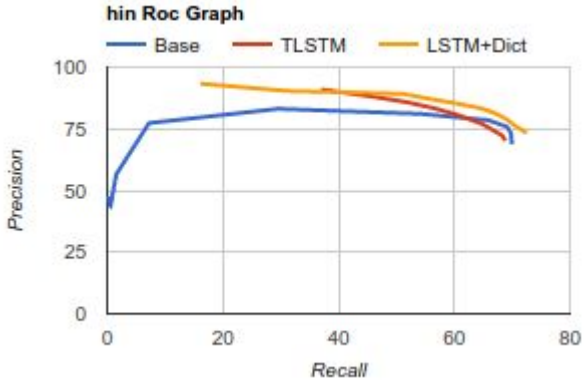


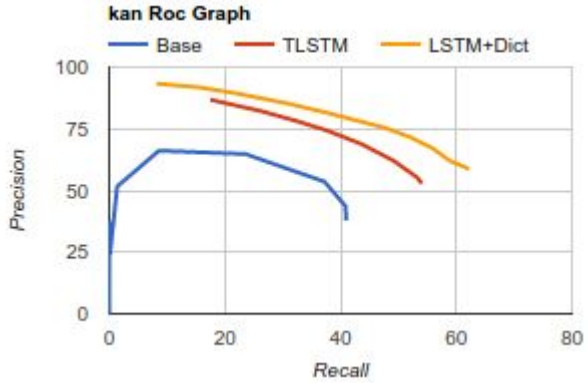


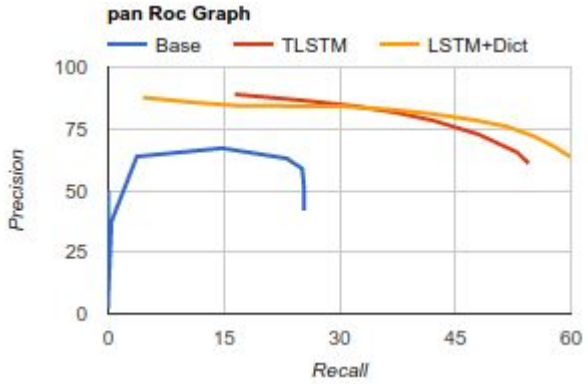
Accuracy Results (Indic Languages)

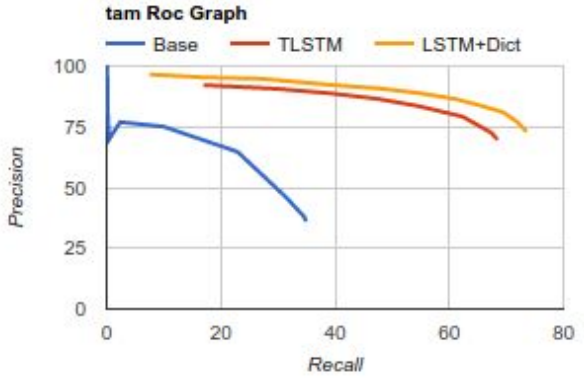
Lang	Truth Words	Char Error Rates			%Change	Word Error Rates			%Change
		Base	T-LSTM	LSTM+Dict		Base	T-LSTM	LSTM+Dict	
Bengali	16865	19.59	22.33	19.25	-1.74	42.79	42.23	39.13	-8.55
Gujarati	19874	45.7	22.38	18.35	-59.85	56.46	49.46	43.62	-22.74
Hindi	27539	14.05	13.56	11.68	-16.87	30.92	30.19	26.92	-12.94
Kannada	13673	31.55	13.57	12.12	-61.58	63.2	47.11	40.71	-35.59
Marathi	21486	29.04	12.07	9.18	-68.39	47.19	32.79	25.92	-45.07
Nepalese	20606	29.56	18.38	15.63	-47.12	50.19	42.39	36.59	-27.10
Panjabi	27651	46.23	19.89	16.1	-65.17	54.89	40.18	37.15	-32.32
Tamil	10033	26.1	10.09	9.3	-64.37	63.58	30.65	26.78	-57.88
Telugu	14133	31.47	31.51	27.8	-11.66	63.63	63.79	59.36	-6.71

Accuracy Results (Indic Langs)





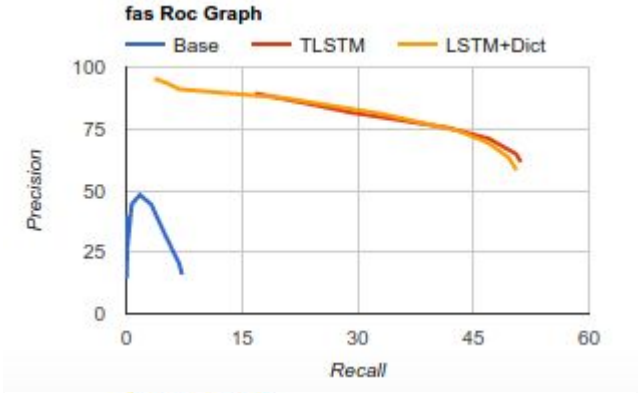


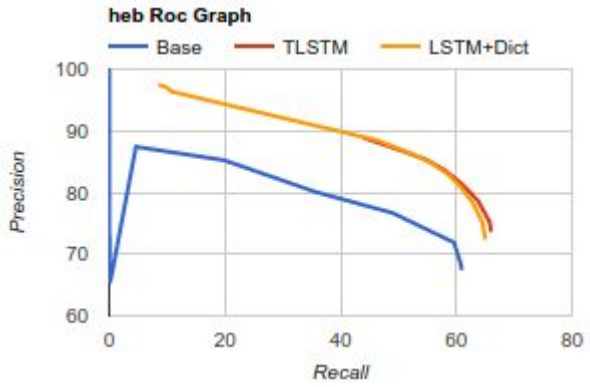


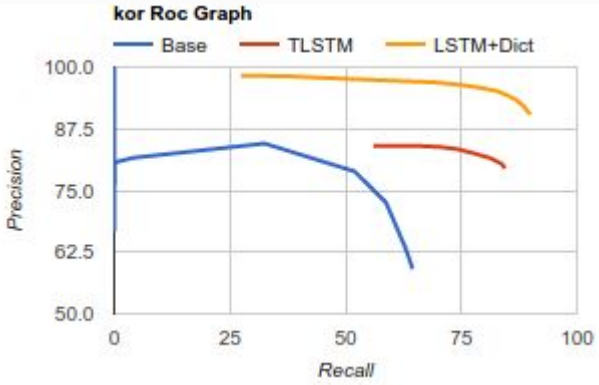
Accuracy Results (Other Languages)

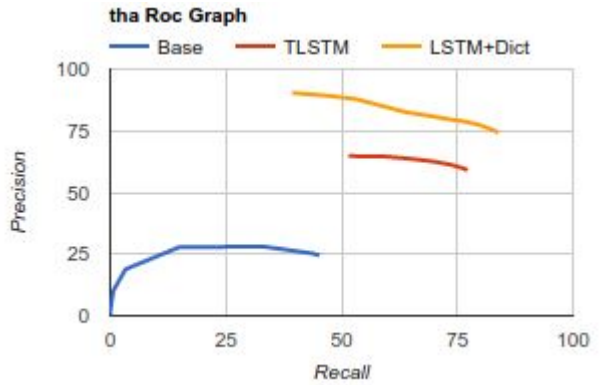
Lang	Truth Words	Char Error Rates			%Change	Word Error Rates			%Change
		Base	T-LSTM	LSTM+Dict		Base	T-LSTM	LSTM+Dict	
Hebrew	21919	12.29	8.08	8.36	-31.98	34.28	28.81	29.91	-12.75
Thai	32173	60.52	12.32	7.73	-87.23	96.91	38.46	22.70	-76.58
Yiddish	21674	20.76	10.04	10.62	-48.84	56.77	34.32	36.53	-35.65
Farsi	23691	74.87	16.53	16.84	-77.51	65.44	40.43	42.80	-34.60
Chinese(S)	31045	8.25	7.25	6.42	-22.18	10.82	11.41	9.32	-13.86
Chinese(T)	28700	13.22	12.83	10.35	-21.71	18.56	20.28	15.78	-14.98
Japanese	29574	18.65	16.97	11.53	-38.18	31.66	35.49	19.94	-37.02
Korean	25687	31.19	9.62	6.67	-78.61	40.09	18.69	9.82	-75.51

Accuracy Results (Other Langs)









Conclusion

- Tesseract is surprisingly general
- T-LSTM is much better almost across the board
  - Some Language model integration issues remain
- Some languages remain difficult, but
- Neural networks **are** taking over!

Thanks for Listening!

Questions?