

Language identification of names with SVMs



ADITYA BHARGAVA AND GRZEGORZ KONDRAK
UNIVERSITY OF ALBERTA

NAACL-HLT 2010
JUNE 3, 2010

Outline







- Introduction: task definition & motivation
- Previous work: character language models
- Using SVMs
- Intrinsic evaluation
 - SVMs outperform language models
- Applying language identification to machine transliteration
 - Training separate models
- Conclusion & future work

Task definition



- Given a name, what is its language?
- Same script (no diacritics)

Beckham		English
Brillault		French
Velazquez		Spanish
Friesenbichler		German

Motivation



- Improving letter-to-phoneme performance (Font Llitjós and Black, 2001)
- Improving machine transliteration performance (Huang, 2005)
- Adjusting for different semantic transliteration rules between languages (Li et al., 2007)

Previous approaches



- Character language models (Cavnar and Trenkle, 1994)
 - Construct models for each language, then choose the language with the most similar model to the test data
 - **99.5%** accuracy **given >300 characters** & 14 languages
- Given 50 bytes (and 17 languages), language models give only **90.2%** (Kruengkrai et al., 2005)
- Between 13 languages, average F1 on last names is **50%**; full names gives **60%** (Konstantopoulos, 2007)
- Easier with more dissimilar languages: English vs. Chinese vs. Japanese (same script) gives **94.8%** (Li et al., 2007)

Using SVMs



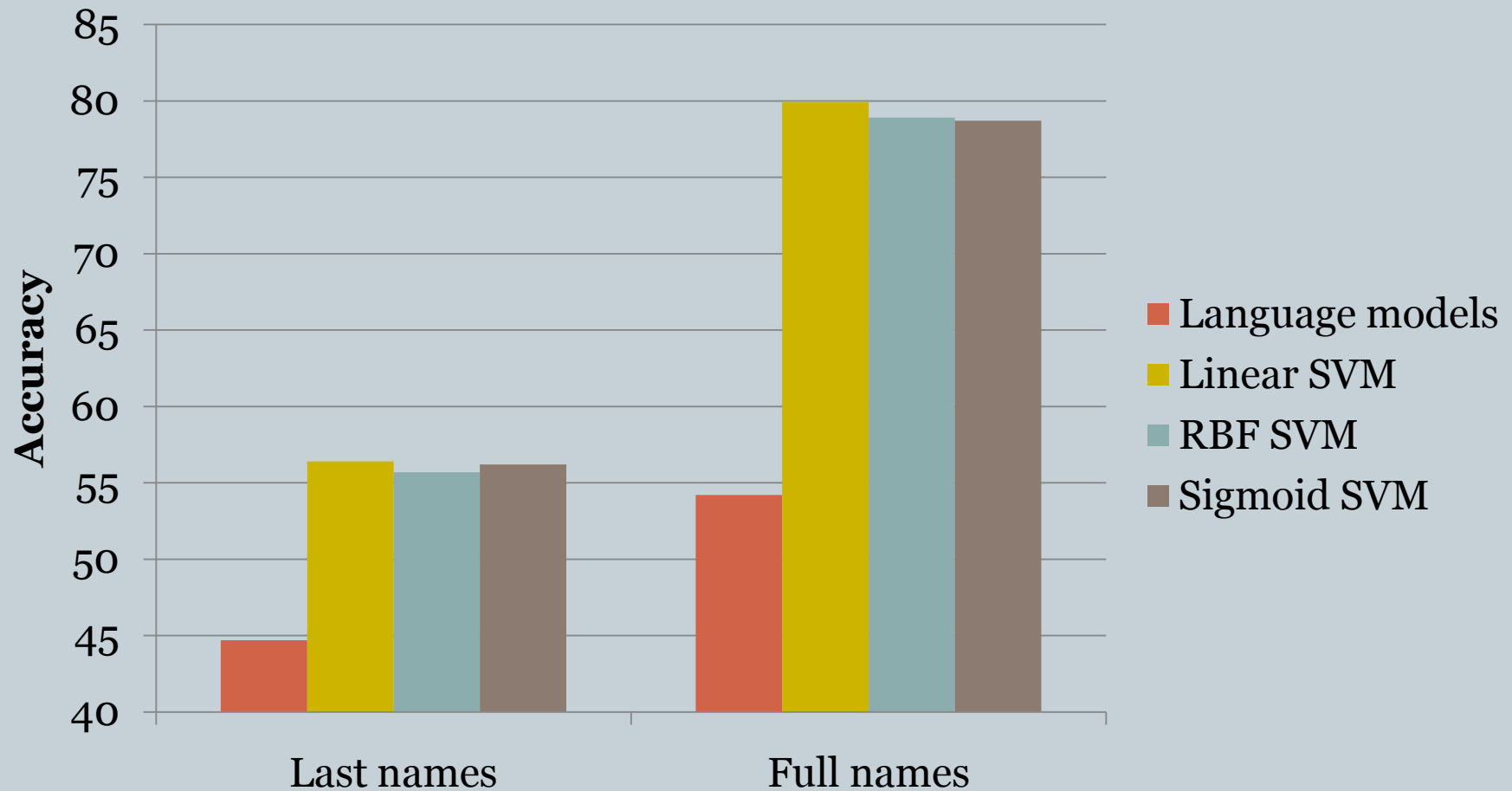
- **Features**
 - Substrings (n-grams) of length n for $n=1$ to 5
 - ✦ Include special characters at the beginning and the end to account for prefixes and suffixes
 - Length of string
- **Kernels**
 - Linear, sigmoid, RBF
 - Other kernels (polynomial, string kernels) did not work well

Evaluation: Transfermarkt corpus



- European national soccer player names (Konstantopoulos, 2007) from 13 national languages
 - ~15k full names (average length 14.8 characters)
 - ~12k last names (average length 7.8 characters)
- Noisy data
 - e.g. Dario Dakovic born in Bosnia but plays for Austria, so annotated as German

Evaluation: Transfermarkt corpus



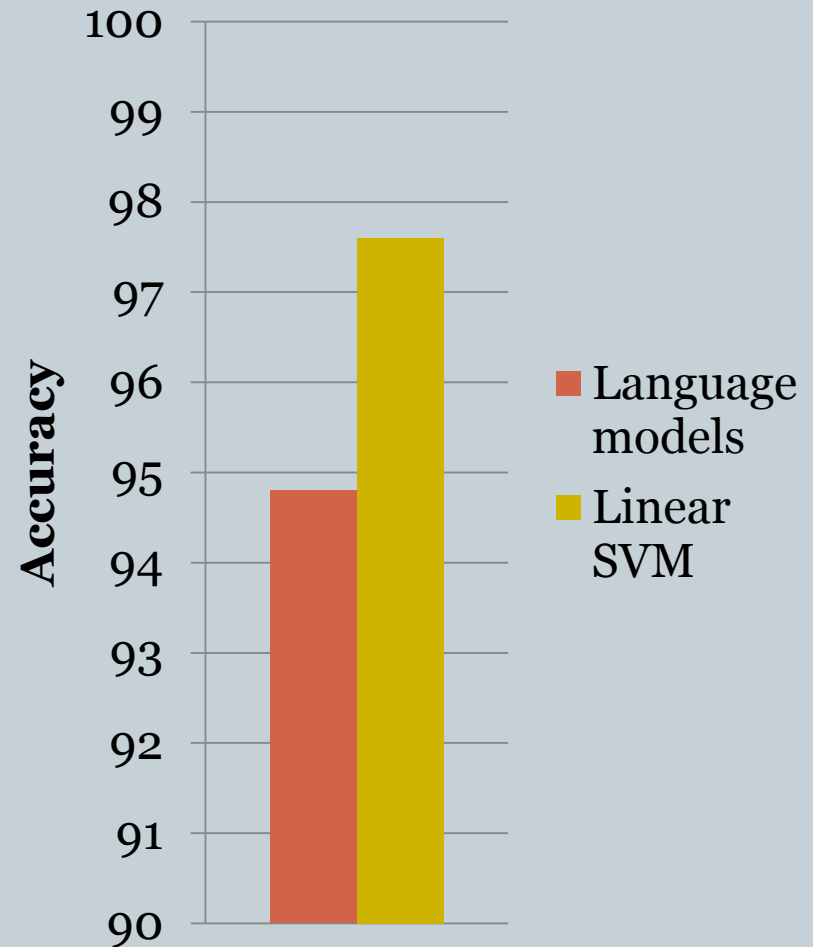
Evaluation: Transfermarkt corpus



	cs	da	de	en	es	fr	it	nl	no	pl	pt	se	yu	Recall
cs	19	0	15	4	1	3	1	0	0	4	2	1	7	0.33
da	0	27	15	2	0	3	1	1	9	0	0	1	0	0.46
de	4	2	183	12	2	11	2	12	5	10	2	2	9	0.72
en	0	1	20	69	1	12	2	2	1	2	1	0	0	0.62
es	2	0	9	4	25	7	23	0	0	1	9	0	2	0.31
fr	0	0	17	10	5	41	13	1	1	1	4	0	2	0.43
it	1	0	6	2	10	5	84	0	0	2	2	0	1	0.74
nl	1	3	19	9	3	9	1	36	1	2	1	0	0	0.42
no	1	7	9	1	1	3	1	3	17	1	0	2	1	0.36
pl	2	0	13	2	3	3	1	2	1	63	0	0	3	0.68
pt	1	0	4	4	8	7	8	1	0	1	8	0	1	0.19
se	2	0	14	0	1	2	1	2	2	1	1	23	4	0.43
yu	3	0	11	1	2	0	4	1	0	2	0	2	84	0.76
Precision	0.53	0.68	0.55	0.58	0.40	0.39	0.59	0.59	0.46	0.70	0.27	0.74	0.74	

Evaluation: CEJ corpus

- Chinese, English, and Japanese names (Li et al., 2007)
 - ~97k total names, average length 7.6 characters
- Demonstrates a higher baseline with dissimilar languages
- Linear SVM only (RBF and sigmoid were slow)



Application to machine transliteration



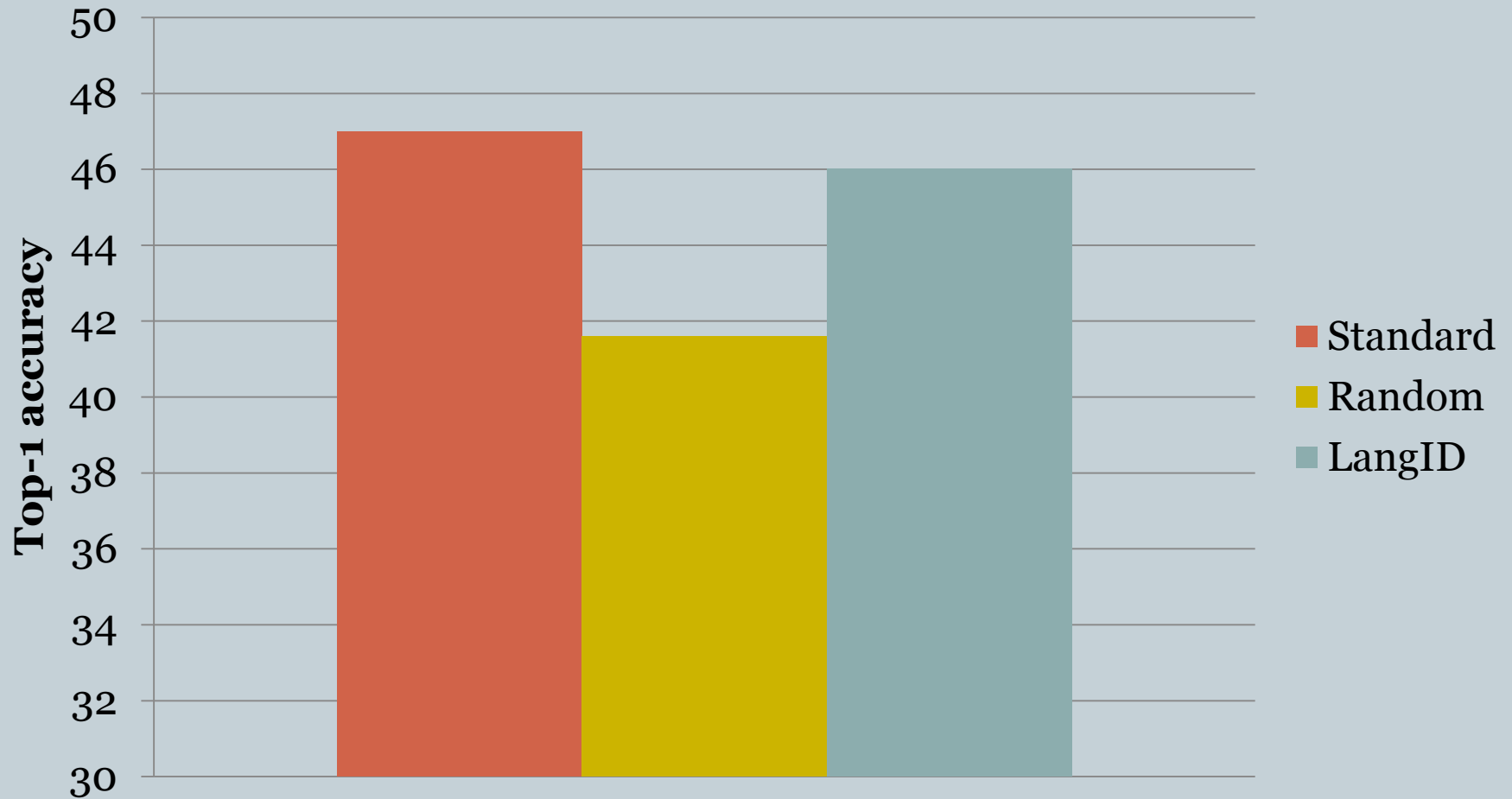
- Language origin knowledge may help machine transliteration systems pick appropriate rules
- To test, we manually annotated data
 - English-Hindi transliteration data set from the NEWS 2009 shared task (Li et al., 2009; MSRI, 2009)
 - 454 “Indian” names, 546 “non-Indian” names
 - Average length 7 characters
- SVM gives 84% language identification accuracy

Application to machine transliteration



- Basic idea: use language identification to split data into two language-specific sets
- Train two separate transliteration models (with less data per model), then combine
- We use DirecTL (Jiampojarn et al., 2009)
- Baseline comparison: random split
- Three tests:
 - DirecTL (Standard)
 - DirecTL with random split (Random)
 - DirecTL with language identification–informed split (LangID)

Application to machine transliteration



Conclusion



- **Language identification of names is difficult**
 - SVMs with n-grams as features work better than language models
- **No significant effect on machine transliteration**
 - But there does seem to be some useful information

Future work



- Web data
- Other ways of incorporating language information for machine transliteration
 - Direct use as a feature
 - Overlapping (non-disjoint) splits

Questions?

