Language Identification in the Wild

Timothy Baldwin

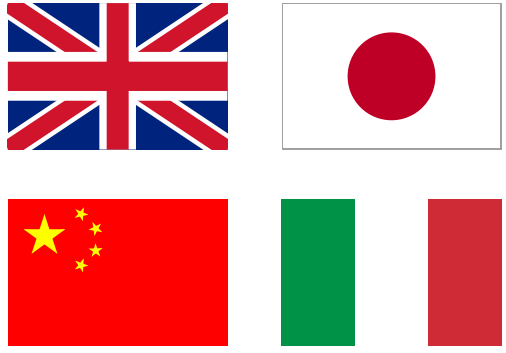
THE UNIVERSITY OF MELBOURNE

Talk Outline

What is Language Identification ("LangID")?

Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural-language processing is a very attractive method of human-computer interaction.

自然言語処理は、人間が日常的に使っている自然言語をコンピュータに処理させる一連の技術であり、人工知能と言語学の一分野である。計算言語学も同じ意味であるが、前者は工学的な視点からの言語処理をさすのに対して、後者は言語学的視点を重視する手法をさす事が多い。データベース内の情報を自然言語に変換したり、自然言語の文章をより形式的な（コンピュータが理解しやすい）表現に変換するといった処理が含まれる。



自然語言處理是人工智慧和語言學領域的分支學科。在這此領域中探討如何處理及運用自然語言；自然語言認知則是指讓電腦「懂」人類的語言。自然語言生成系統把計算機數據轉化為自然語言。自然語言理解系統把自然語言轉化為計算機程序更易于處理的形式。

L'Elaborazione del linguaggio naturale, detta anche NLP (dall'inglese Natural Language Processing), è il processo di estrazione di informazioni semantiche da espressioni del linguaggio umano o naturale, scritte o parlate, tramite l'elaborazione di un calcolatore elettronico.

How Hard can it be?

- What is the language of the following document:

  Så sitter du åter på handlar'ns trapp och gråter så övergivet.

How Hard can it be?

- What is the language of the following document:

  Så sitter du åter på handlar'ns trapp och gråter så övergivet.

  Swedish

Getting Wilder ...

- What is the language of the following document:

  Revolution is à la mode at the moment in the country, where the joie de vivre of the citizens was once again plunged into chaos after a third coup d'état in as many years. Although the leading general is by no means an enfant terrible per se, the fledgling economy still stands to be jettisoned down la poubelle

Getting Wilder ...

- What is the language of the following document:

> *Revolution is à la mode at the moment in the country, where the joie de vivre of the citizens was once again plunged into chaos after a third coup d'état in as many years. Although the leading general is by no means an enfant terrible per se, the fledgling economy still stands to be jettisoned down la poubelle*

English

... And Wilder Still ...

- What is the language of the following document:

  Jawaranya ngu yidanyi ngaba ngu yardi yaniya cool drink ninaka nanga alangi-nka.

... And Wilder Still ...

- What is the language of the following document:

  Jawaranya ngu yidanyi ngaba ngu yardi yaniya cool drink ninaka nanga alangi-nka.

  Wambaya

... And the Actual Wildness of the Problem

- What is the language of the following document:

```
11100000101110111001000011110000010111
01110010100011100000101110111001001l0
```

Talk Outline

Preliminary LangID Setting

- To get going, let's assume that:
  1. documents are all from a homogeneous source = **closed-world domain assumption**
  2. all documents are monolingual = **single label assumption**
  3. documents are of a certain length = **length assumption**
  4. we know every language = **closed-world class assumption**

Preliminary LangID Setting

- To get going, let's assume that:
  1. documents are all from a homogeneous source = **closed-world domain assumption**
  2. all documents are monolingual = **single label assumption**
  3. documents are of a certain length = **length assumption**
  4. we know every language = **closed-world class assumption**
- Here and throughout, we'll model the task as a supervised classification problem

Preliminary LangID Setting

- To get going, let's assume that:
  1. documents are all from a homogeneous source = **closed-world domain assumption**
  2. all documents are monolingual = **single label assumption**
  3. documents are of a certain length = **length assumption**
  4. we know every language = **closed-world class assumption**
- Here and throughout, we'll model the task as a supervised classification problem
- To get us going, let's play around with simple multinomial naive Bayes over byte *n*-grams

Machine Learning in LangID

- Supervised classification task
  - Model of labeled training data used to label "unseen" data
- Similar to text classification ("TC")
  - vector space models, naive Bayes models, logistic regression
- Text representation:
  - TC: bag of words
  - LangID: bag of byte-sequences

Source(s): Hughes et al. [2006]

Text Represenation for Language Identification

Input: `language␣identification`

1-grams: `l, a, n, g, u, a ...`

2-grams: `la, an, ng, gu, ua, ag ...`

3-grams: `lan, ang, ngu, gua, uag, age ...`

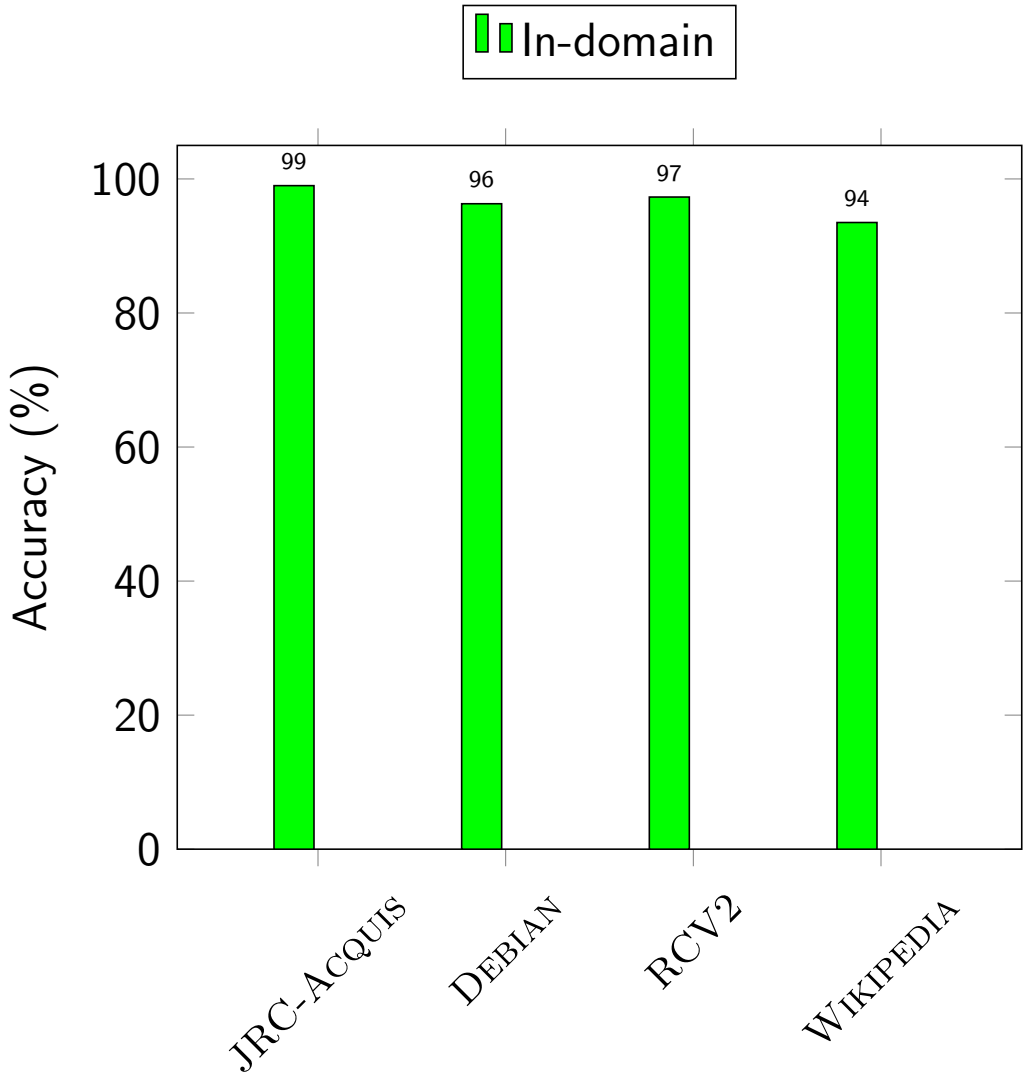
4-grams: `lang, angu, ngua, guag, uage, age␣ ...`

Datasets for Preliminary Experiments

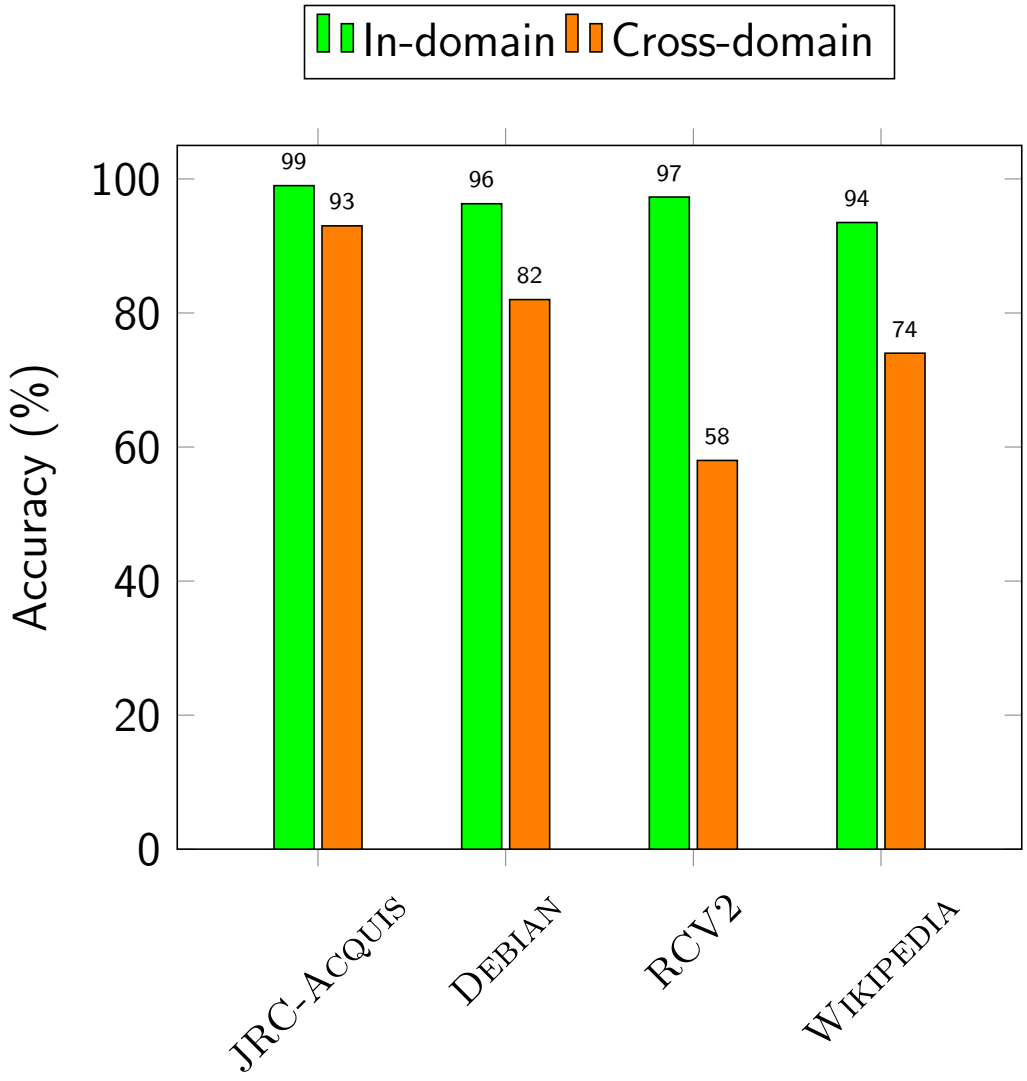
Dataset	Docs	Langs	Doc Length (bytes)
JRC-Acquis	20000	22	18478.5±60836.8
Debian	21735	89	12329.8±30902.7
RCV2	20000	13	3382.7±1671.8
Wikipedia	20000	68	7531.3±16522.2

Source(s): Steinberger et al. [2006], Lui and Baldwin [2011]

In-domain Results



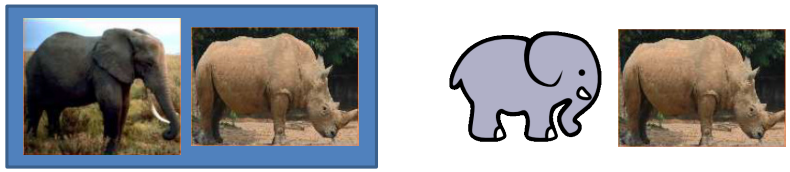
In-domain vs. Cross-domain Results



What's Happening? Transfer Learning

Supervised Learning

Inductive Transfer Learning



Transductive Transfer Learning

Finding

- In-domain, supervised language identification (over longish, monolingual documents) is relatively trivial for even basic learning algorithms, but overfitting tends to be chronic

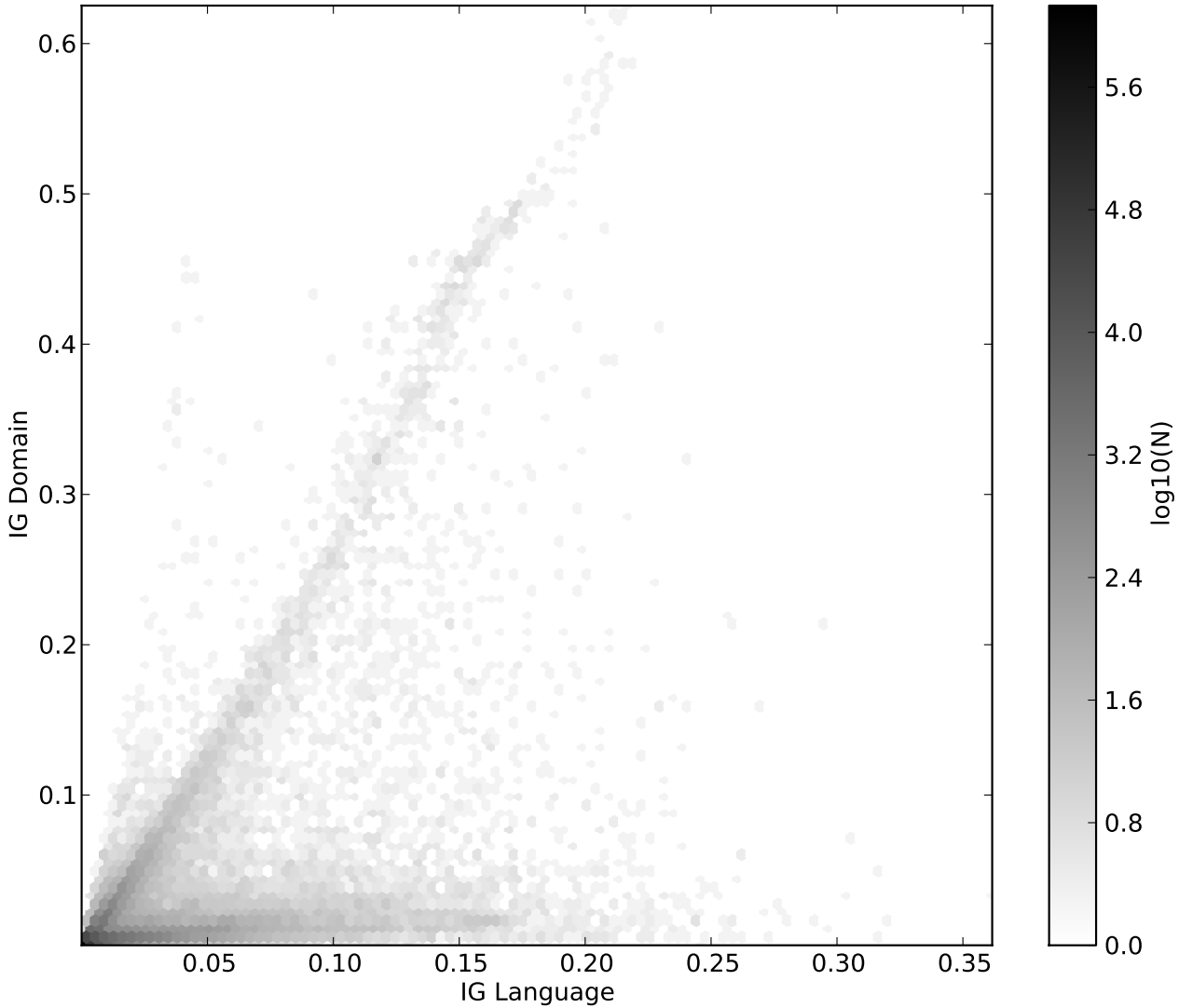
Source(s): Lui and Baldwin [2011, 2012]

Talk Outline

Cross-domain Feature Selection

- Goal: identify *n*-grams with:
  - high language association
  - low domain association
- Measure relationship using information gain (= difference in entropy after partitioning the data in a given way)

Language- vs. Domain-based IG



What does this Mean?

- There are two distinct groups of features: (1) $\mathcal{IG}$ for language is strongly correlated with that for domain; and (2) $\mathcal{IG}$ for language is largely independent of that for domain

the second of these is what we are interested in

- Automatically detecting language- (and not domain-) associated features:

$$\mathcal{LD}^{all}(t) = \mathcal{IG}^{all}_{lang}(t) - \mathcal{IG}_{domain}(t)$$

- Also consider a variant which is debiased to training data volume per language:

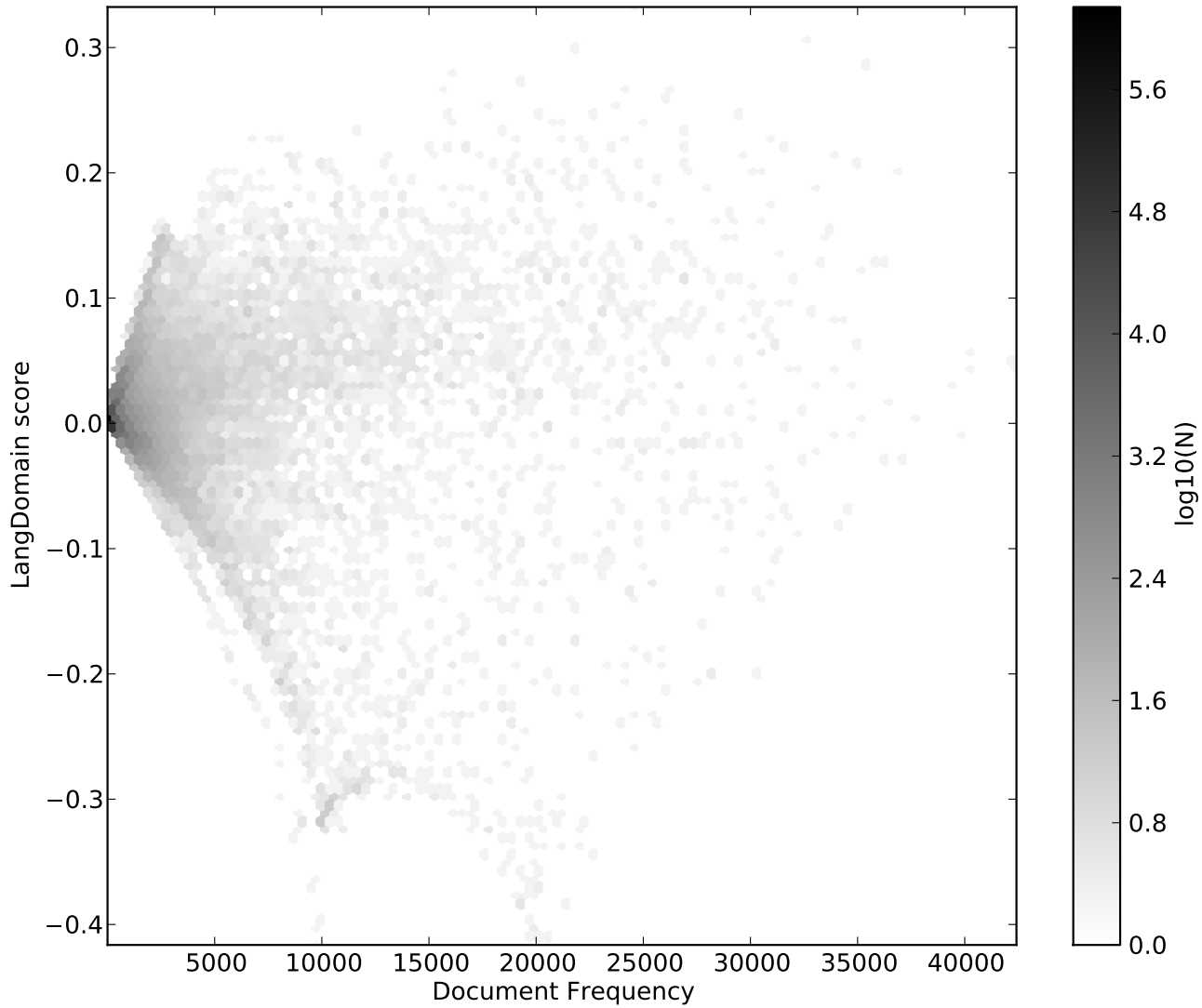
$$\mathcal{LD}^{bin}(t|l) = \mathcal{IG}^{bin}_{language}(t|l) - \mathcal{IG}_{domain}(t)$$

Source(s): Lui and Baldwin [2011]

Computational Bottleneck

- Calculating the $\mathcal{IG}$ for low-order n-grams is fine, but it quickly becomes intractable for larger values of n
- Ideally, we want a method which scales to (very) large numbers of features/high n-gram orders, with little computational overhead

DF vs. $\mathcal{LD}$



Observation

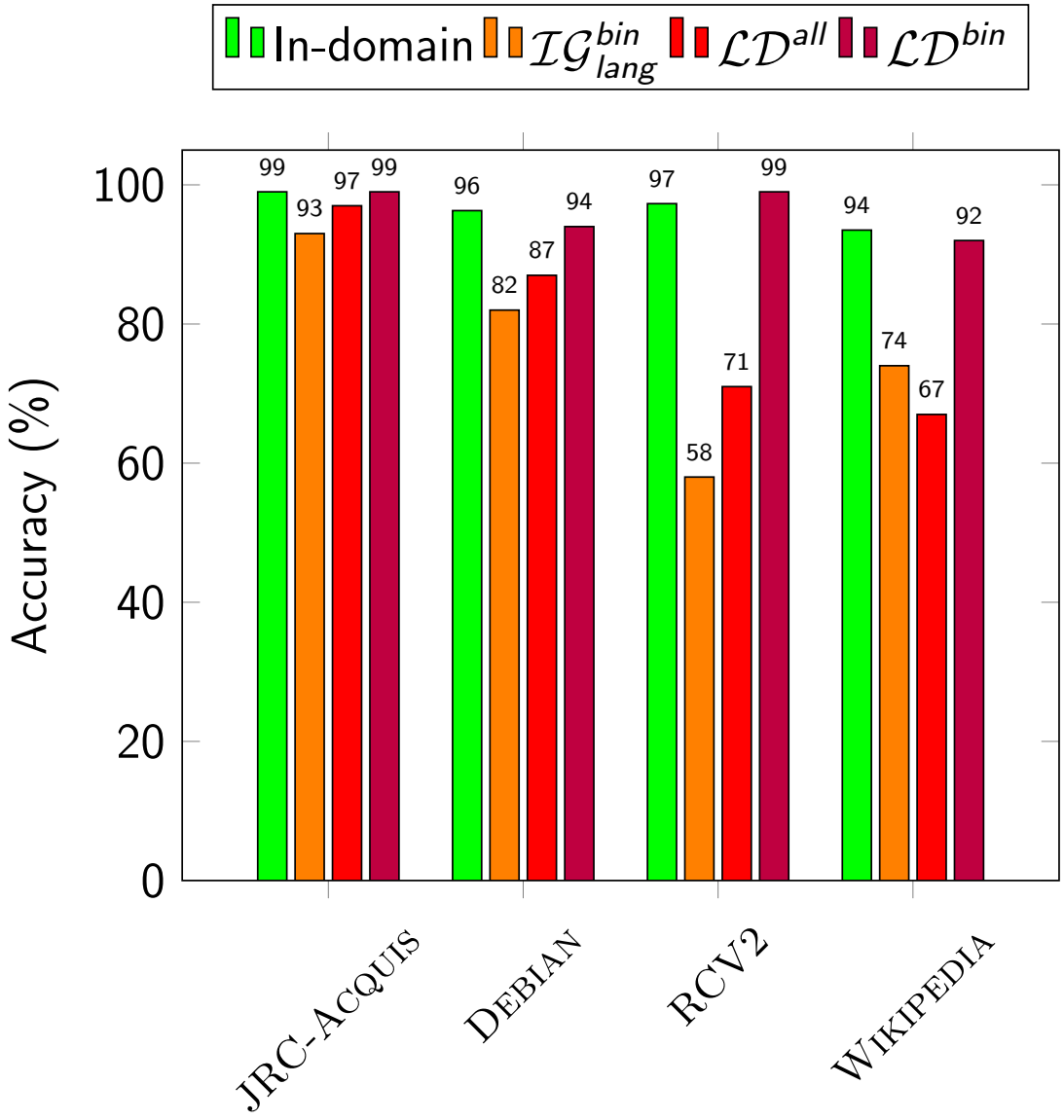
- Low DF is a good predictor of low $\mathcal{LD}$ (but not vice versa)
- As $\mathcal{DF}$ is much cheaper to compute, we first identify the 15000 features with highest $\mathcal{DF}$ for a given n-gram order, and assign a $\mathcal{LD}$ score of 0 to all features outside this set
- The final feature representation is a combination of the top-N features for each a predefined set of n-grams

Other Details

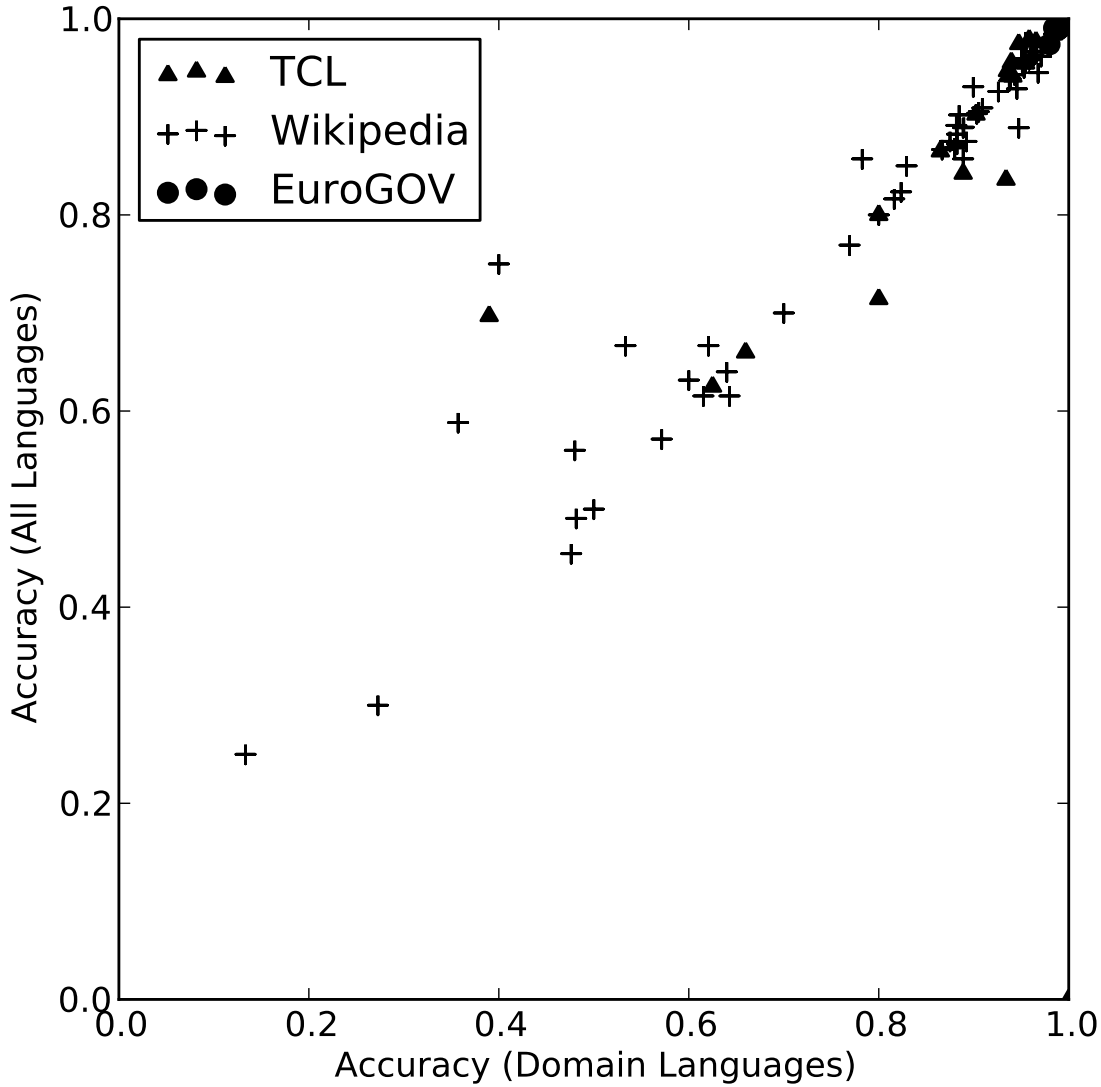
- Byte-based tokenisation
- Multinomial naive Bayes learner
- 1–9-grams (although little gained in going beyond 4-grams)
- Four training/test datasets, plus three test-only datasets (N-EuroGOV, N-TCL and Wikipedia)

Source(s): Baldwin and Lui [2010], Lui and Baldwin [2011, 2012]

In-domain vs. Cross-domain Results



Per-language Accuracy as $|L|$ Increases



Findings

- Assuming we have a reasonable number of different domains, $\mathcal{LD}^{bin}$ is an effective approach to feature selection, and when paired with DF ranking, can be scaled to very, very large feature sets, and large numbers of languages
- More training languages tends not to hurt performance, and in some instances actually helps
- Reference implementation:

  `https://github.com/saffsd/langid.py`

Talk Outline

Time to Get Wilder: Twitter LangID I

- All looks good to here, but it is well documented that langid performs worse over short documents [Baldwin and Lui, 2010]
- One source of short documents ($\leq$140 code points) of particular interest is Twitter, which is famously multilingual [Semiocast, 2010, Hong et al., 2011, Bergsma et al., 2012, Baldwin et al., 2013]

Time to Get Wilder: Twitter LangID II

- Challenges in LangID on Twitter:
  - short message length
  - informal register
  - lexical variation
  - linguistic diversity
  - limited labeled corpora
- Open question:

  how accurate are existing pre-trained ("off-the-shelf") language identification systems when applied to Twitter messages?

Source(s): Lui and Baldwin [2014]

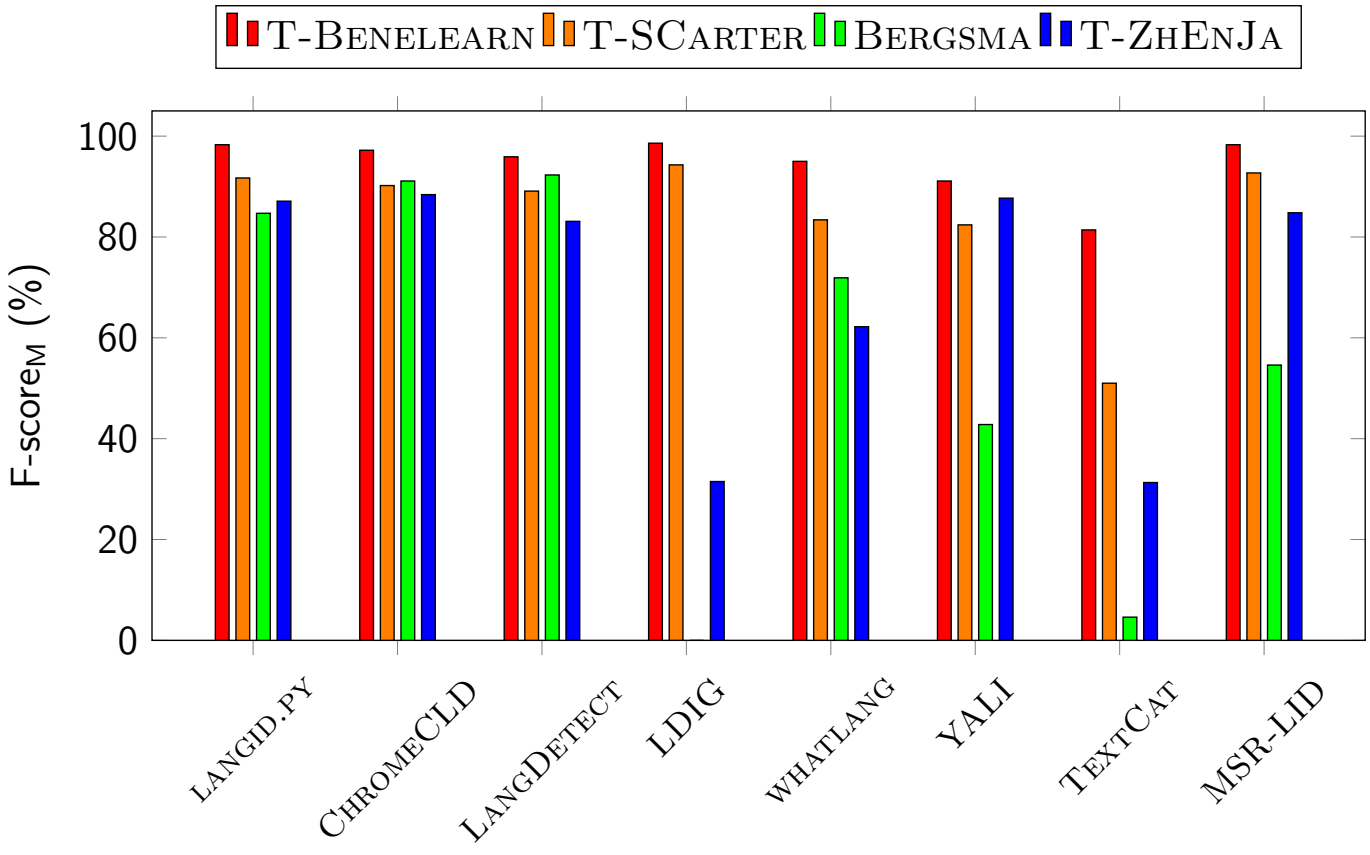
Off-the-Shelf Language Identifiers

langid.py	[Lui and Baldwin, 2012]
ChromeCLD	[McCandless, 2010]
LangDetect	[Nakatani, 2010]
LDIG	[Nakatani, 2012]
whatlang	[Brown, 2013]
YALI	[Majliš, 2012]
TextCat	[Scheelen, 2003]
MSR-LID	[Goldszmidt et al., 2013]

Manually-labelled Datasets

Dataset	# Docs	Languages
T-BENELEARN [Tromp and Pechenizkiy, 2011]	9066	(×6) de es en fr it nl
T-SCARTER [Carter et al., 2013]	5000	(×5) de es en fr nl
BERGSMA [Bergsma et al., 2012]	13190	(×9) ar bg fa hi mr ne ru uk ur
T-ZHENJA [Lui and Baldwin, 2014]	3016	(×3) en ja zh

Evaluating Off-the-Shelf LangID Systems on Twitter

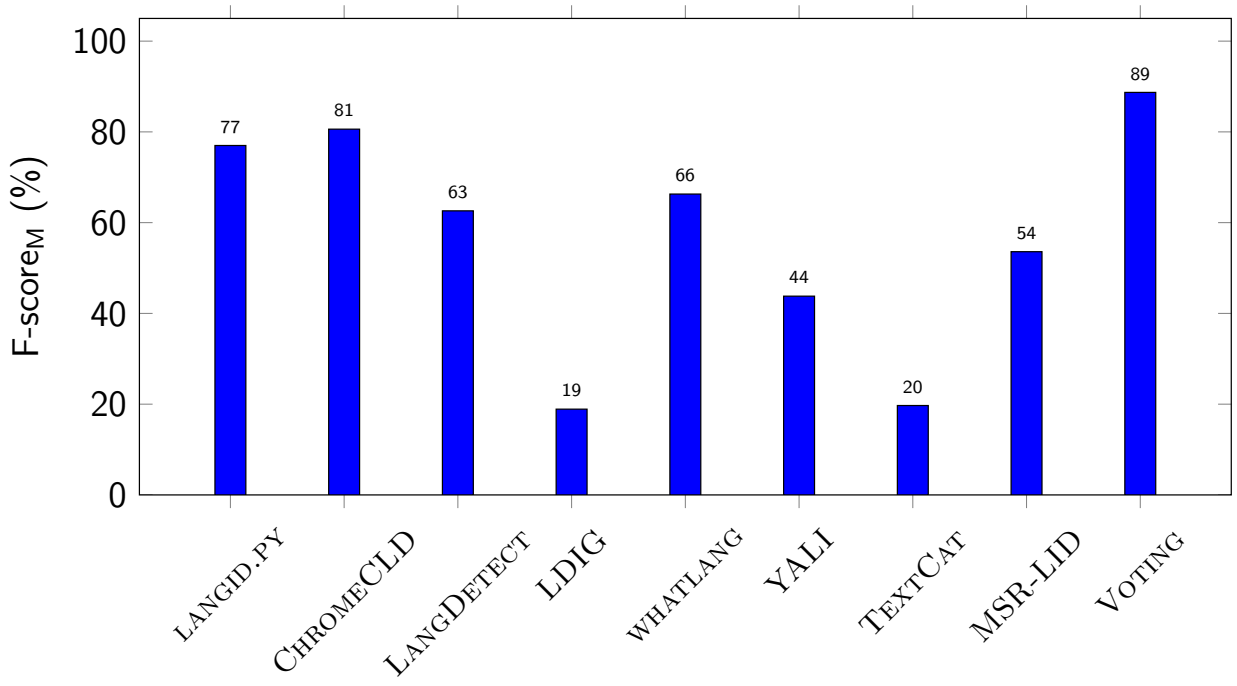


Which LangID System to Use?

- The best off-the-shelf systems for Twitter are LANGID.PY, CHROMECLD and LANGDETECT
- Based on experimentation over system combination based on simple majority vote, 3-system vote between LANGID.PY, CHROMECLD and LANGDETECT is a good choice

TwitUser Dataset

- Larger-scale evaluation over 65 languages, and a total of 26011 msgs from 2914 users:



Simple Improvements

- Cleaning (Tromp and Pechenizkiy [2011]):
  - remove links, usernames, hashtags, smilies
  - very easy to implement
  - no difference on Twitter-specific systems (MSR-LID, LDIG)
  - small improvement on other systems ($< 2\%$)

Simple Improvements

- Bootstrapping (Goldszmidt et al. [2013]):
  - requires re-training classifier
  - tested with LangDetect, TextCat, langid.py
  - not consistently better than off-the-shelf

Simple Improvements

- LangID priors (Carter et al. [2013]):
  - did not test
  - requires processing a large background collection
  - Bontcheva et al. [2013] report positive results
  - user identity priors will be artificially effective on TwitUser

Findings

- Twitter LangID far from a solved task
- No single off-the-shelf LangID system is perfect, but targeted system combination generates a reasonable system
- Twitter-specific tweaks have some impact
- Twitter API language predictions are not perfect

Source(s): Lui and Baldwin [2014]

Talk Outline

What about Multilingual Documents? I

What about Multilingual Documents? II

- Reasons for documents being multilingual:
  - code-switching
  - translations
  - boilerplate/interface language
  - multiple users
- Why detect multilingual documents?
  - pre-filtering for monolingual NLP
  - mining bilingual texts for MT
  - low-density languages on the web
- Task setting in this section:

  estimate the relative language proportions for a given mono/multi-lingual document

Source(s): Lui et al. [2014]

Generative Mixture Models

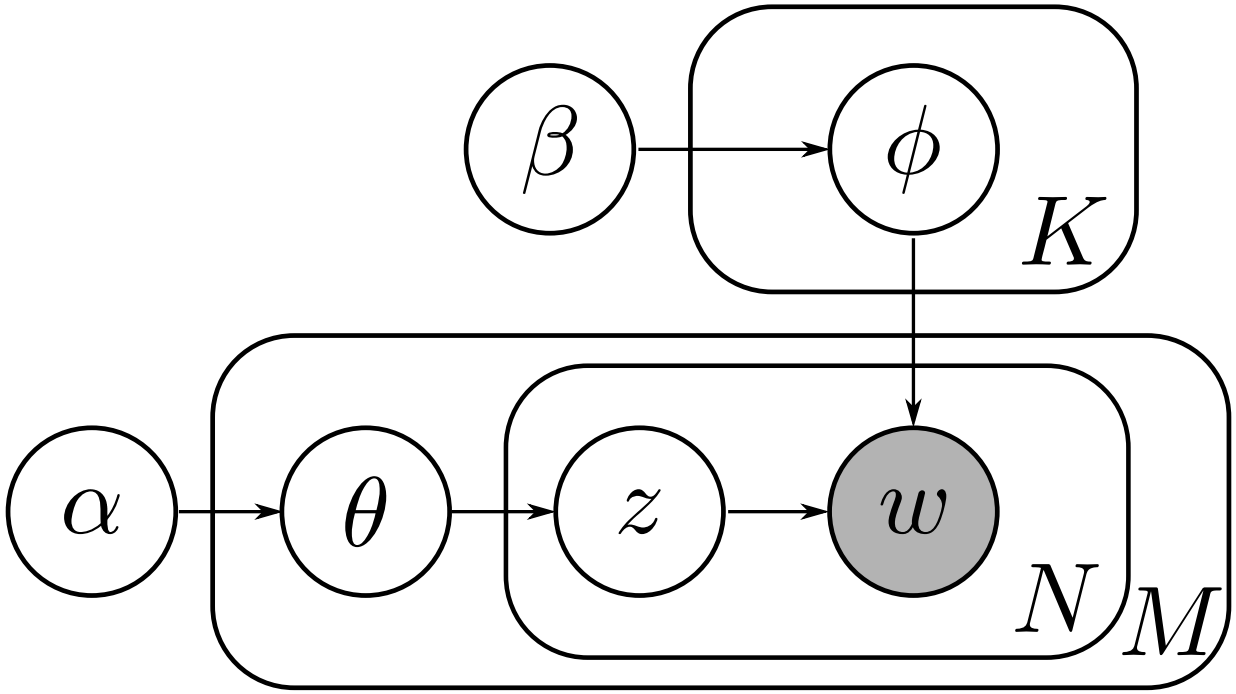
$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

$P(w_i)$: Probability of the i^{th} token

$P(w_i|z_i = j)$: Probability of w_i given label j

$P(z_i = j)$: Probability of the i^{th} label being j

Latent Dirichlet Allocation (LDA)



Source(s): Blei et al. [2003]

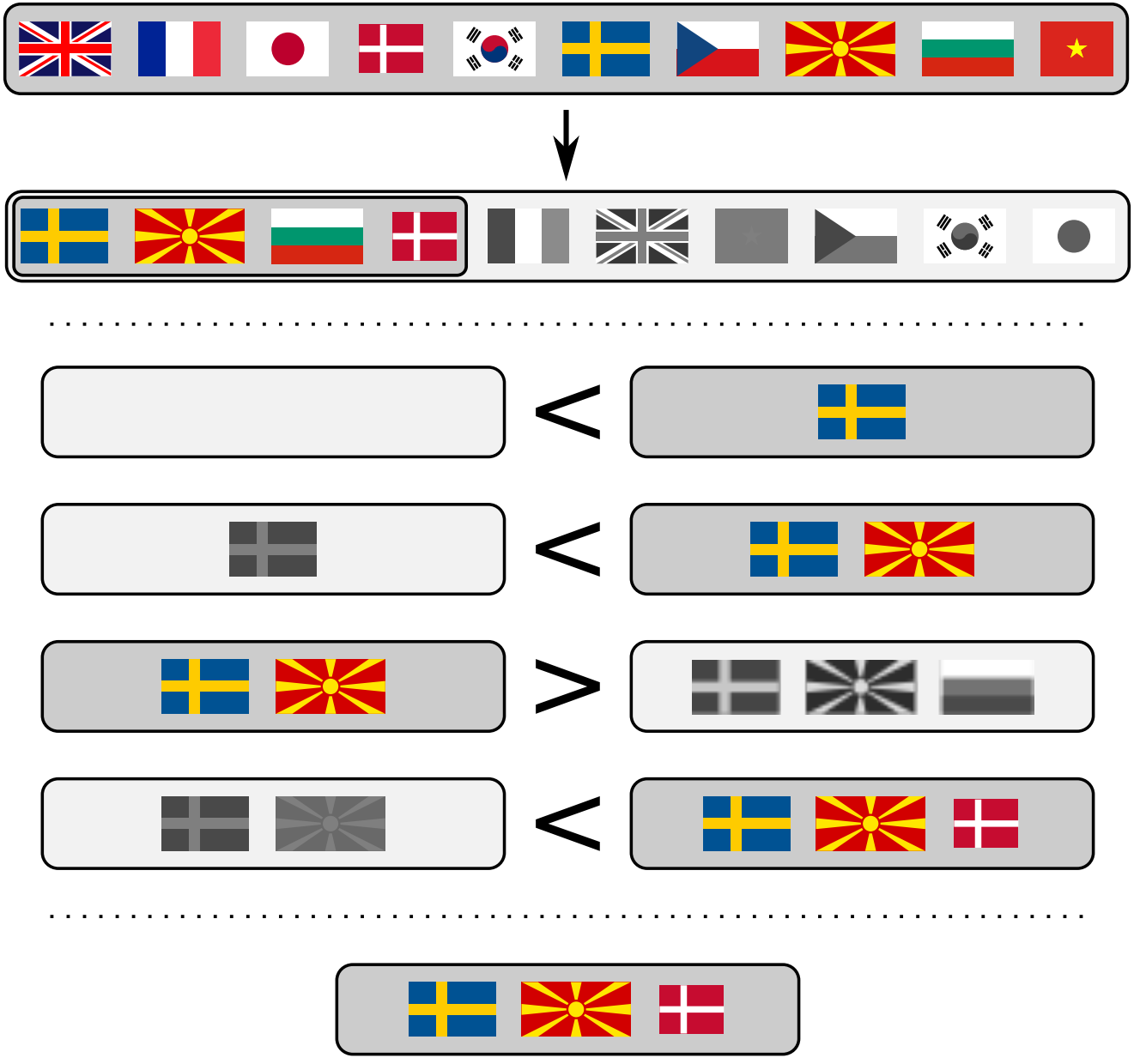
Gibbs Sampling: LDA vs. Multilingual LangID

		LDA	**Multilingual LangID**
labels	z	topics	language
tokens	w	words	language-indicative byte sequences
token–label distribution	ϕ	infer using Gibbs sampler	estimate using training data
label–document distribution	θ	infer using Gibbs sampler	

How many Languages?

- Use Gibbs sampling to infer the most likely mixture of languages over a fixed set:
  - L is the set of languages in the training data
  - find the most likely subset of languages $\lambda \subset L$

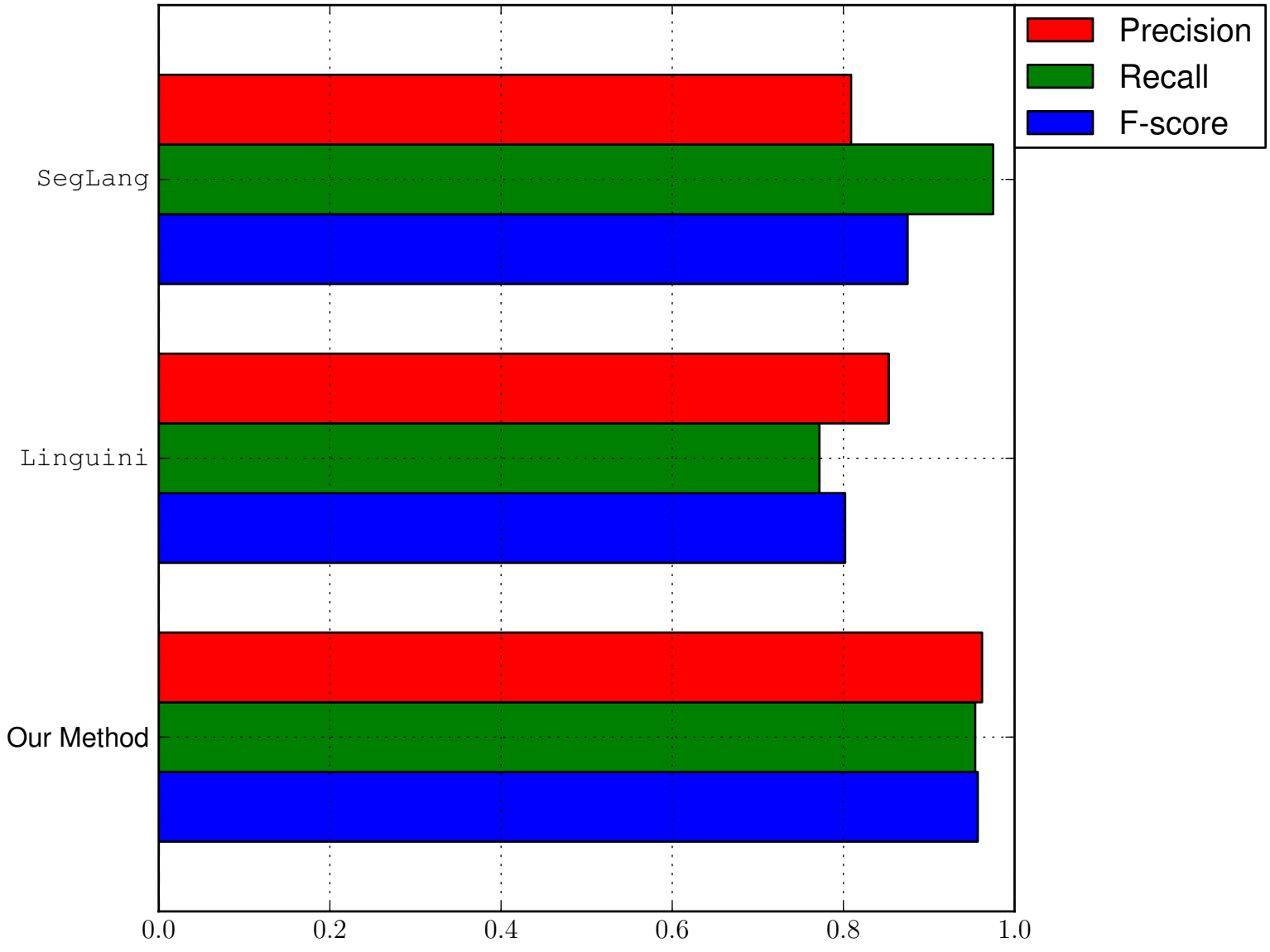
Source(s): Lui et al. [2014]



WikipediaMulti dataset

- synthetic dataset
- mixture of monolingual and multilingual documents
- K languages per document ($1 \leq K \leq 5$)
- samples lines from Wikipedia
- train: 5000 monolingual documents
- dev: 1000 documents for each K
- test: 200 documents for each K

WikipediaMulti results



Findings

- Extension of LANGID.PY to include prediction of the language mix in a document (and a Gibbs sampler to predict the language proportions), which outperforms benchmarks on synthetic and real-world data
- Reference implementation:

  `https://github.com/saffsd/polyglot`

Source(s): Lui et al. [2014]

Talk Outline

Conclusions

- LangID can be as easy or as hard as you want it to be, but for data in the wild, it is hard, esp. when:
  - large number of languages
  - shorter documents
- Still the question of what to do with open-world label set (i.e. unknown languages)
- Inter-language bias covered in this work, but interesting recent work on intra-language bias [Jurgens et al., 2017]

Acknowledgements

- The following collaborators contributed to this work: Steven Bird, Baden Hughes, Jey Han Lau, Marco Lui, Andrew MacKinlay and Jeremy Nicholson
- This work was supported by the Australian Research Council and NICTA

References I

Timothy Baldwin and Marco Lui. Language identification: The long and the short of the matter. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 229–237, Los Angeles, USA, 2010.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan, 2013.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. Language identification for creating language-specific Twitter collections. In *Proceedings the Second Workshop on Language in Social Media (LSM2012)*, pages 65–74, Montréal, Canada, 2012.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. TwitIE: An open-source information extraction pipeline for microblog text. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2013)*, Hissar, Buglaria, 2013.

References II

Ralf Brown. Selecting and weighting n-grams to identify 1100 languages. In *Proceedings of the 16th international conference on text, speech and dialogue (TSD 2013)*, Plzeň, Czech Republic, 2013.

Simon Carter, Manos Tsagkias, and Wouter Weerkamp. Microblog language identification: overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215, 2013.

Moises Goldszmidt, Marc Najork, and Stelios Paparizos. Boot-strapping language identifiers for short colloquial postings. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013)*, Prague, Czech Republic, 2013.

Lichan Hong, Gregorio Convertino, and Ed H. Chi. Language matters in Twitter: A large scale study. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM 2011)*, Barcelona, Spain, 2011.

Baden Hughes, Timothy Baldwin, Steven Bird, Jeremy Nicholson, and Andrew MacKinlay. Reconsidering language identification for written language resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 485–488, Genoa, Italy, 2006.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Volume 2: Short Papers*, pages 51–57, 2017.

References III

Marco Lui and Timothy Baldwin. Cross-domain feature selection for language identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 553–561, Chiang Mai, Thailand, 2011.

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea, 2012.

Marco Lui and Timothy Baldwin. Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 17–25, Gothenburg, Sweden, 2014. URL `http://www.aclweb.org/anthology/W14-1303`.

Marco Lui, Jey Han Lau, and Timothy Baldwin. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2(Feb):27–40, 2014.

Martin Majliš. Yet another language identifier. In *Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 46–54, Avignon, France, 2012.

Michael McCandless. Accuracy and performance of Google's compact language detector. blog post. available at `http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-googles.html`, 2010.

References IV

Shuyo Nakatani. Language detection library (slides). http://www.slideshare.net/shuyo/language-detection-library-for-java, 2010. Retrieved on 21/06/2013.

Shuyo Nakatani. Short text language detection with infinity-gram. blog post. available at http://shuyo.wordpress.com/2012/05/17/short-text-language-detection-with-infinity-gram/, 2012.

Frank Scheelen. *libtextcat*, 2003. Software available at http://software.wise-guys.nl/libtextcat/.

Semiocast. Half of messages on Twitter are not in English — Japanese is the second most used language. Technical report, Semiocast, 2010.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, and Dániel Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, Geona, Italy, 2006.

Erik Tromp and Mykola Pechenizkiy. Graph-based *n*-gram language identification on short texts. In *Proceedings of Benelearn 2011*, pages 27–35, The Hague, Netherlands, 2011.