NLP for Social Media: Language Identification

Pawan Goyal

CSE, IITKGP

July 29, 2016

Language Mix: Corpora

Social media sources

- TWITTER: micro-blog posts from Twitter
- COMMENTS: comments from YouTube
- BLOGS: blog posts from Spinn3r dataset
- FORUMS: forum posts from popular forums
- WIKIPEDIA: documents from English Wikipedia

Language Mix: Corpora

Social media sources

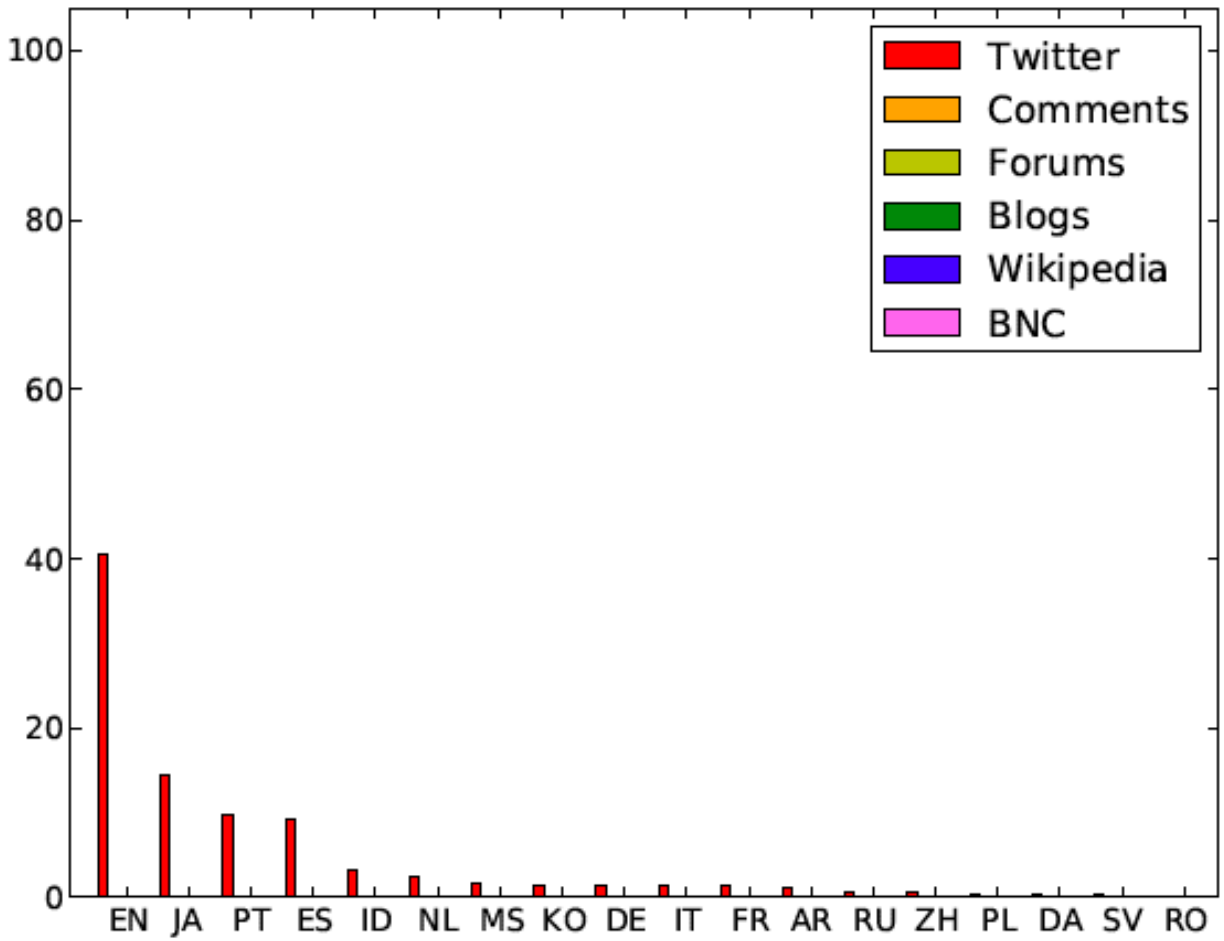
- TWITTER: micro-blog posts from Twitter
- COMMENTS: comments from YouTube
- BLOGS: blog posts from Spinn3r dataset
- FORUMS: forum posts from popular forums
- WIKIPEDIA: documents from English Wikipedia

A random sample of 4K sentences was taken from all these sources.

Reference: Baldwin et al. *How Noisy Social Media Text, How Diffrnt Social Media Sources?* IJCNLP 2013.

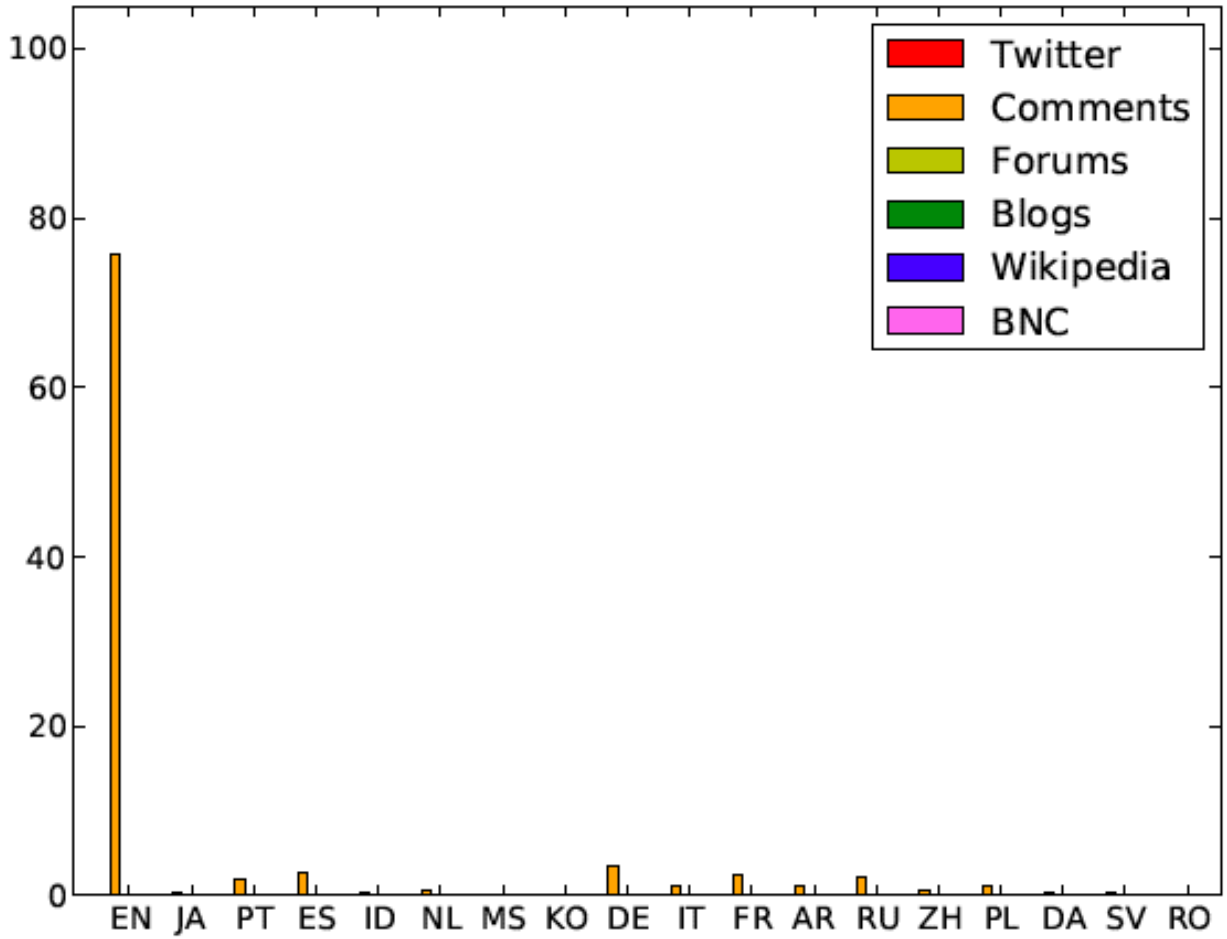
Language Mix

Twitter:

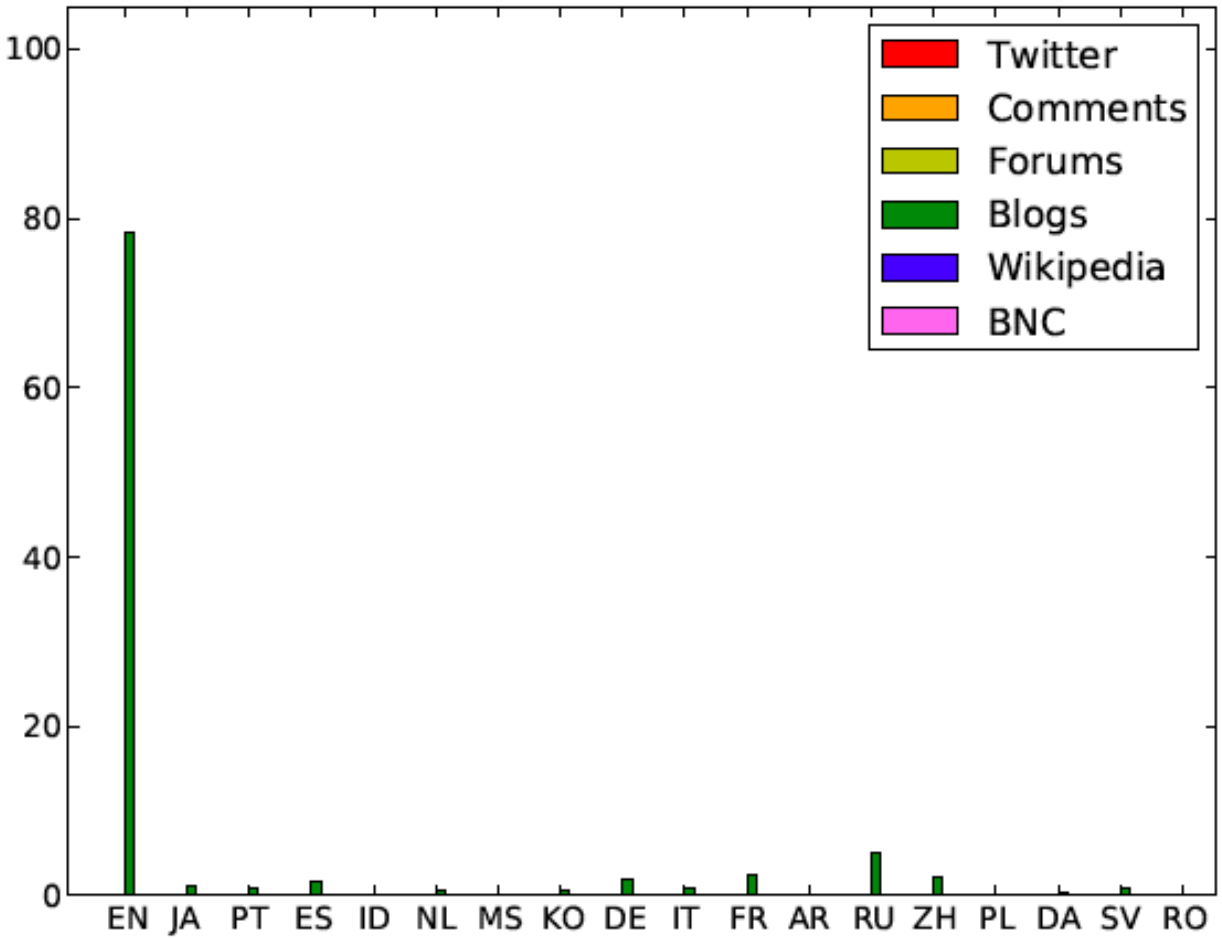


Language Mix

COMMENTS:

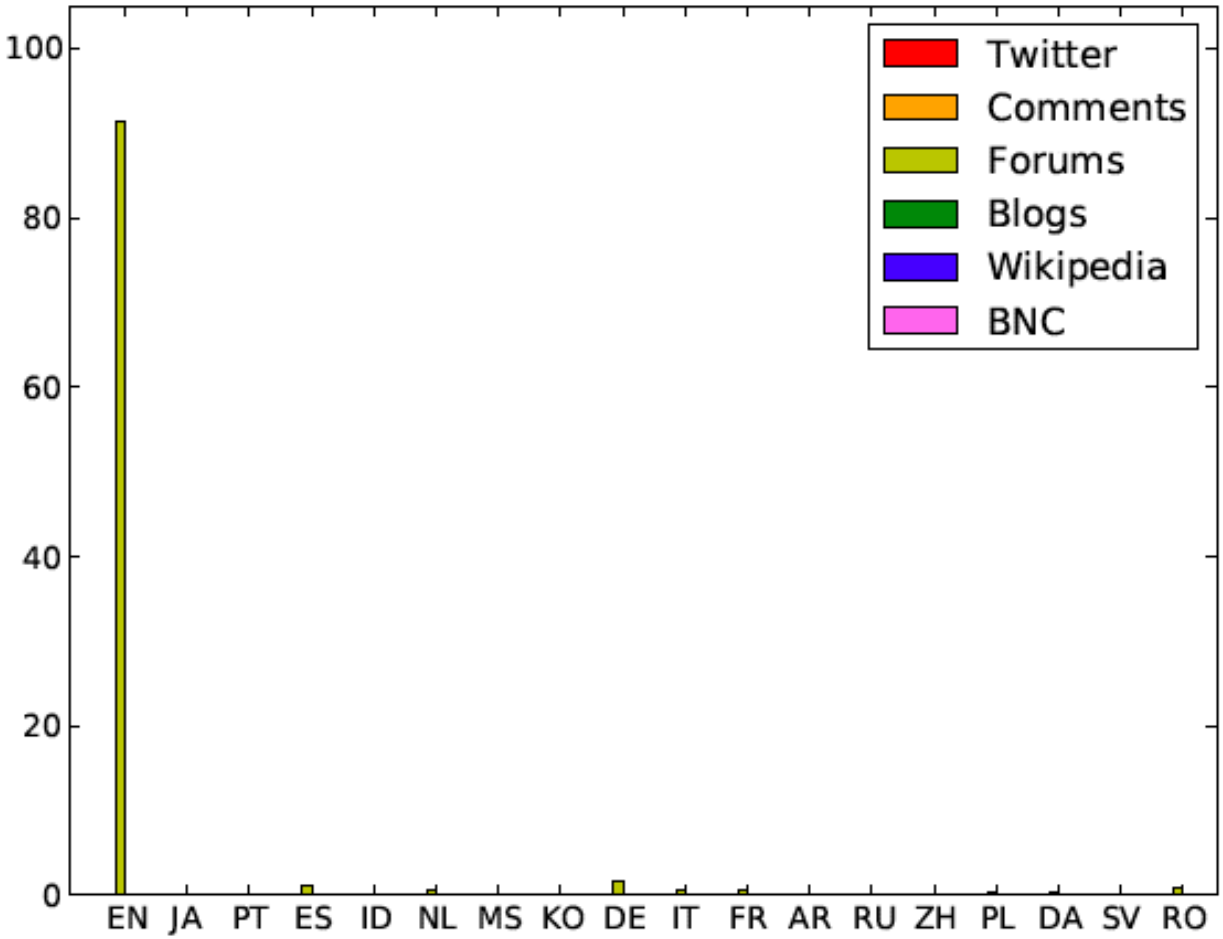


Blogs:


100
80
60
40
20
0
EN JA PT ES ID NL MS KO DE IT FR AR RU ZH PL DA SV RO
Twitter
Comments
Forums
Blogs
Wikipedia
BNC

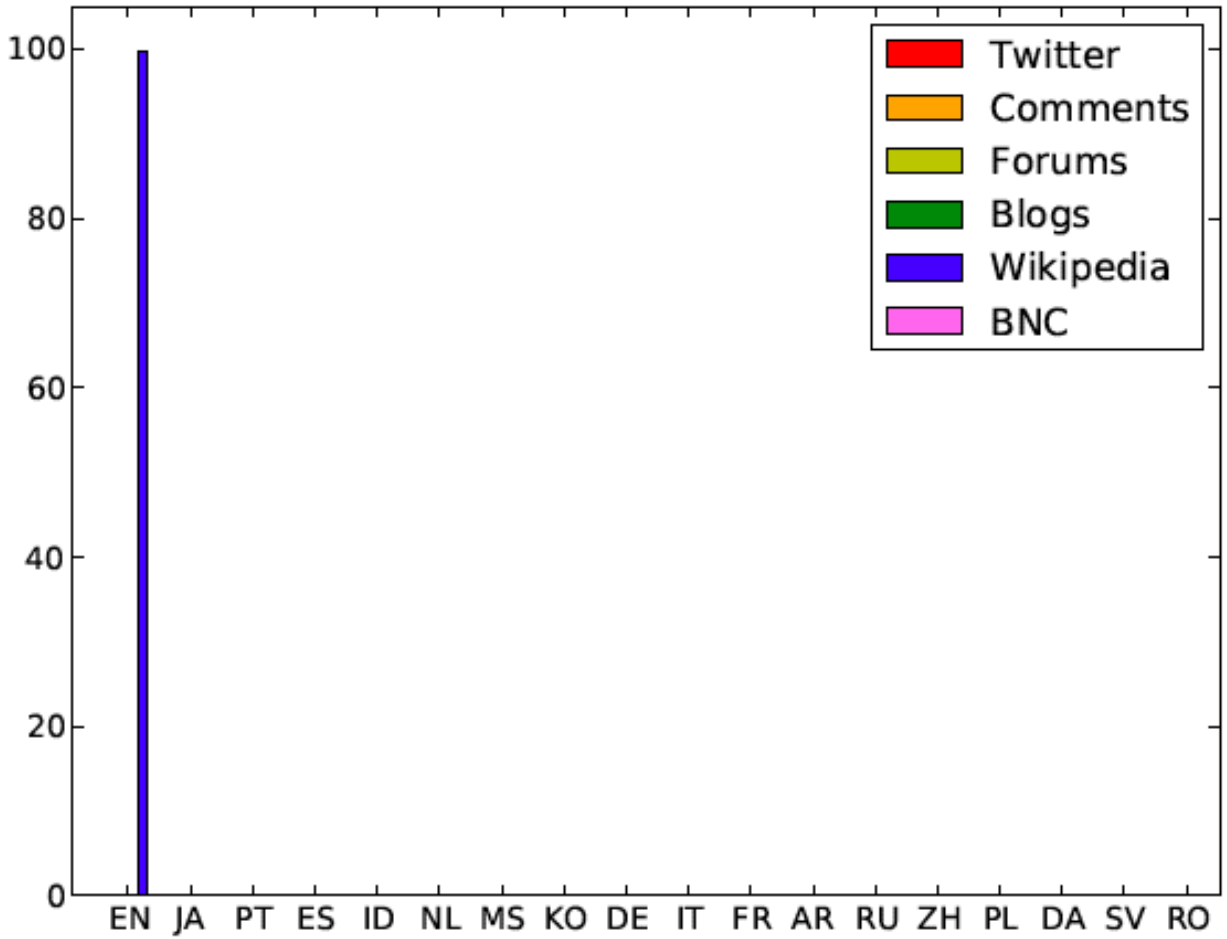
Language Mix

FORUMS:



Language Mix

WIKIPEDIA:



Language Mix: Overall Findings

- Twitter most multilingual ($> 50\%$ non-EN), followed by Comments, blogs and forums
- All 97 languages modeled by *langid.py* found in Twitter and Comments

Lexical Analysis: Average Word and Sentence length

Corpus	Word length	Sentence length
TWITTER-1	3.8±2.4	9.2±6.4
TWITTER-2	3.8±2.4	9.0±6.3
COMMENTS	3.9±3.2	10.5±10.1
FORUMS	3.8±2.3	14.2±12.7
BLOGS	4.1±2.8	18.5±24.8
WIKIPEDIA	4.5±2.8	21.9±16.2

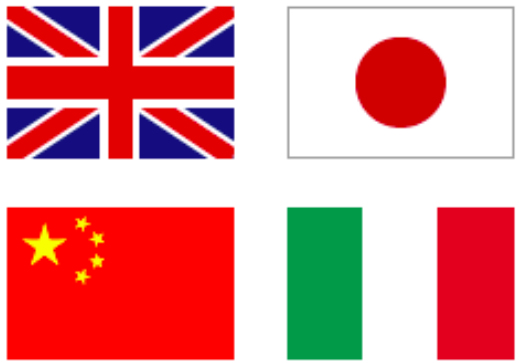
Lexical Analysis: Out-of-vocabulary words

Corpus	Word length	Sentence length	%OOV
TWITTER-1	3.8±2.4	9.2±6.4	**24.6**
TWITTER-2	3.8±2.4	9.0±6.3	**24.0**
COMMENTS	3.9±3.2	10.5±10.1	**19.8**
FORUMS	3.8±2.3	14.2±12.7	**18.1**
BLOGS	4.1±2.8	18.5±24.8	**20.6**
WIKIPEDIA	4.5±2.8	21.9±16.2	**19.0**

Language Identification

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。



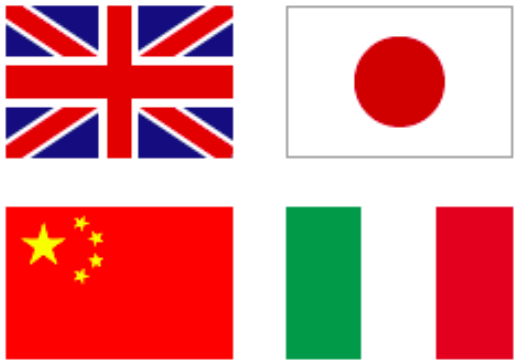
自然語言處理是人工智慧和語言學領域的分支學科。在這此領域中探討如何處理及運用自然語言；自然語言認知則是指讓電腦「懂」人類的語言。自然語言生成系統把計算機數據轉化為自然語言。自然語言理解系統把自然語言轉化為計算機程序更易于處理的形式。

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.

Language Identification

RT @ThotsOnTees: Its not rocket science.....Man was designed to fail.So to those that av their trust in Man,goodluck...mine is on GOD!

実行はいつなされるんですか？ RT “@a_X_o: 制服ブレイのメールの返事きてさ「いやらしすぎる」だけ帰ってきたんだけどもうなんか嫌だ”



@Luii_S2_KiSeop 哈哈哈 四次元这名号很match他xDDDDD 姐不是official kissme么？？要怎样才能成为官方km??

#Campiglio stellata, freddo, neve dura, ma sufficiente. Rossi, Alo e Massa ok, Hayden spalla ancora immobile e dolorante.Domani #StudioSport

LangID for Twitter

Why?

Twitter is highly multilingual

LangID for Twitter

Why?

Twitter is highly multilingual

What are the challenges?

- *Short message length:* individual documents are generally short.
- *Lexical variation:* There is a lot of fluidity in how a given word is spelled.
- *Linguistic diversity:* A rich mix of languages can be found, with no "closed-world" guarantee.
- *Limited labelled corpora:* language-labelled corpora of social media data are few.

Granularity-level

Granularity-level

Huge traffic restrictions for PM's visit to #blast site mean deserted roads in #Hyderabad. "Itna sanaata kyon hai bhai?"

Granularity-level

Huge traffic restrictions for PM's visit to #blast site mean deserted roads in #Hyderabad. "Itna sanaata kyon hai bhai?"

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

Granularity-level

Huge traffic restrictions for PM's visit to #blast site mean deserted roads in #Hyderabad. "Itna sanaata kyon hai bhai?"

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

document level language identification

We start with the case where the whole tweet belongs to one language only.

LI: Approaches

Unicode Block

Idea: Different languages use different scripts.

Unicode Block

Idea: Different languages use different scripts.

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 char
U+0300..U+036F	Combining Diacritical Marks	112	112	Inherited
U+0370..U+03FF	Greek and Coptic	144	135	Coptic (14 characters), Greek (117 characters), Common (4 chara
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 (236 characters), Common (7 characters), Inherited (12 characters)
U+0700..U+074F	Syriac	80	77	Syriac

LI: Approaches

Unicode Block

Idea: Different languages use different scripts.

Unicode Block

Idea: Different languages use different scripts.

U+0900..U+097F	Devanagari	128	128	Devanagari (124 characters), Common (2 characters), Inherited (2 characters)
U+0980..U+09FF	Bengali	128	93	Bengali
U+0A00..U+0A7F	Gurmukhi	128	79	Gurmukhi
U+0A80..U+0AFF	Gujarati	128	85	Gujarati
U+0B00..U+0B7F	Oriya	128	90	Oriya
U+0B80..U+0BFF	Tamil	128	72	Tamil
U+0C00..U+0C7F	Telugu	128	96	Telugu
U+0C80..U+0CFF	Kannada	128	88	Kannada
U+0D00..U+0D7F	Malayalam	128	114	Malayalam
U+0D80..U+0DFF	Sinhala	128	90	Sinhala

LI: Approaches

Unicode Block

Idea: Different languages use different scripts.

e.g., English, French, German, Spanish use *Basic Latin*, while Russian, Bulgarian, Serbian use *Cyrillic*.

LI: Approaches

Unicode Block

Idea: Different languages use different scripts.

e.g., English, French, German, Spanish use *Basic Latin*, while Russian, Bulgarian, Serbian use *Cyrillic*.

Issues

Still, many languages use the same block.

LI: Approaches

Unicode Block

Idea: Different languages use different scripts.

e.g., English, French, German, Spanish use *Basic Latin*, while Russian, Bulgarian, Serbian use *Cyrillic*.

Issues

Still, many languages use the same block.

Issues

How many of you use devanagari for messaging in Hindi?

LI: Approaches

Dictionary based

- Compute the intersection with each of the language lexicon.
- Declare the highest matching lexicon as the winner.

LI: Approaches

Dictionary based

- Compute the intersection with each of the language lexicon.
- Declare the highest matching lexicon as the winner.
- May also work with a subset, such as stop words.

LI: Approaches

Dictionary based

- Compute the intersection with each of the language lexicon.
- Declare the highest matching lexicon as the winner.
- May also work with a subset, such as stop words.

Issues

Coverage, short text, cognates

LI: Supervised Approaches

Input

- A document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$
- A training set of m hand-labeled documents $(d_1, c_1), \ldots, (d_m, c_m)$

LI: Supervised Approaches

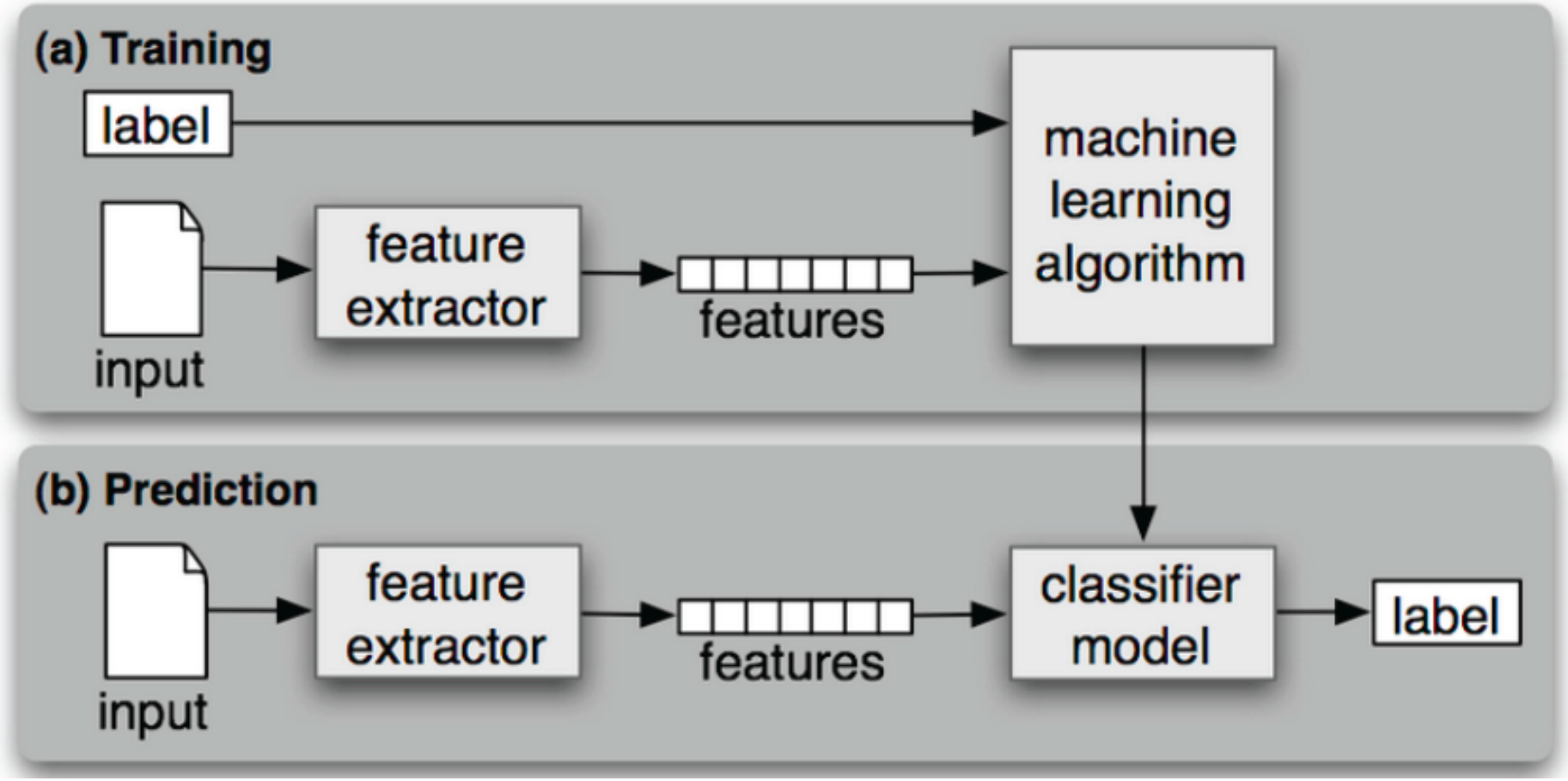
Input

- A document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$
- A training set of m hand-labeled documents $(d_1, c_1), \ldots, (d_m, c_m)$

Output

A learned classifier $\gamma : d \rightarrow c$

Supervised Machine Learning



Bayes' rule for documents and classes

For a document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Bayes' rule for documents and classes

For a document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naïve Bayes Classifier

$$c_{MAP} = \arg\max_{c \in C} P(c|d)$$

$$= \arg\max_{c \in C} P(d|c)P(c)$$

$$= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c)$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n|c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Conditional Independence

Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c_j.

$$P(x_1, x_2, \ldots, x_n|c) = P(x_1|c) \cdot P(x_2|c) \ldots P(x_n|c)$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n|c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Conditional Independence

Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c_j.

$$P(x_1, x_2, \ldots, x_n|c) = P(x_1|c) \cdot P(x_2|c) \ldots P(x_n|c)$$

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{x \in X} P(x|c)$$

Learning the model parameters

Maximum Likelihood Estimate

$$\hat{P}(c_j) = \frac{doc - count(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Learning the model parameters

Maximum Likelihood Estimate

$$\hat{P}(c_j) = \frac{doc - count(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Problem with MLE

Suppose in the training data, we haven't seen one of the words (say *pure*) in a given language.

$$\hat{P}(pure|Hindi) = 0$$

Learning the model parameters

Maximum Likelihood Estimate

$$\hat{P}(c_j) = \frac{doc - count(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

Problem with MLE

Suppose in the training data, we haven't seen one of the words (say *pure*) in a given language.

$$\hat{P}(pure|Hindi) = 0$$

$$c_{NB} = \arg\max_c \hat{P}(c) \prod_{x \in X} \hat{P}(x_i|c)$$

Laplace (add-1) smoothing

$$\hat{P}(w_i|c) = \frac{count(w_i,c)+1}{\sum_{w \in V}(count(w,c)+1)}$$

$$= \frac{count(w_i,c)+1}{(\sum_{w \in V}(count(w,c)) + |V|}$$

Homework

Get $\geq 15k$ tweets from Twitter Streaming API and check:

- Are all tweets LangID tagged (what %)?
- How many different language tags?

Then run langid.py and check:

- how many different language tagged?
- what % langid.py and Twitter's API agree/disagree?
- what kind of tweets/languages do they disagree?

Now take some of the posts and comments from a public facebook page and see if langid.py detects the language correctly.