

# Web page language identification based on URLs

Eda Baykan

EPFL

Laboratory of Theory & Applications of Algorithms

Yahoo Barcelona - 16.02.2009

## Web page language identification



LTAA

ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

Homepage

english only

EPFL > I&C > IIF > LTAA

### Laboratory of Theory and Applications of Algorithms

We are concerned with the theory and applications of algorithms, specifically as they relate to the world wide web. If you are interested in pursuing a master's or doctoral degree in one of these areas, please contact us.

#### People

Eda Baykan  
Monika Henzinger  
Ingmar Weber

Administrative assistance: Monika Henzinger  
System administration: Ingmar Weber

#### Courses

Algorithms (summer term 2008)  
IC-51 Advanced Analysis of Algorithms (summer term 2007)

#### Semester projects

Personalized Tag Suggestion for Flickr (winter term 2007/2008)  
Extracting Information from URLs (summer term 2008)  
Measuring the Impact of a Scientist (summer term 2008)  
Other projects might be available. Just go to Ingmar's office and ask.

#### Online demos

Scientist Finder - A scientist search engine. Still in alpha!  
EagleEye Thunderbird Extension - Reclaim your address book!  
Personalized Flickr Tag Suggestion

The lab is funded by the Swiss National Science Foundation.

<http://ltaa.epfl.ch>

Is English?

Yes

No

## Why useful?

- Download quota of search engines
- Improving quality of language identification from content
- Language specific search engines
  - [voila.fr](http://voila.fr) → French search engine
  - Bandwidth waste

3

## Why useful?

- Grouping results of a web search engine
- Language icons over the links of a web page

4

## Why difficult?

- URLs have fewer words than content
- URLs have words in English and in their true language
  - <http://www.bookings-belgium.com/region/be/brugseommeland.fr.html>
- URLs have huge vocabulary with artificial words
  - google, yahoo, wikipedia
- .com & .org
  - contain ~ 60-70% of the web pages
  - contain web pages from many languages

5

## Outline

- System & Techniques
- Data
- Evaluation measures
- Performance

6

# System

- Worked on 5 European languages
  - English, German, Italian, French, Spanish
- Built binary classifiers for each language
  - Machine learning techniques
    - Training (real positive, real negative)
    - Testing (real label, classification label)
- Feature vectors
- Classification algorithms

7

# Feature vector 1: **Words**

- Breaks URLs at non-alphabetical characters
  - Bag of words
- All words from training URLs are labeled to ML system
- Advantages
  - Learns domain names
  - Performs best with lots of training data
- Disadvantages
  - Performs poorly with small amount of training data
  - Does not break composite words
    - e.g cheapflights does not help with cheaphouses

8

## Feature vector 2: Trigrams

- Consecutive sequence of three letters
  - e.g. can learn that “*the*” is English
- Intuition: Should help for composite words
- Works well for content based training
- Advantages
  - Performs well on small amount of training data
- Disadvantages
  - Can not learn domain names
  - Gets confused with mixed language URLs

9

## Feature vector 3: Custom Made

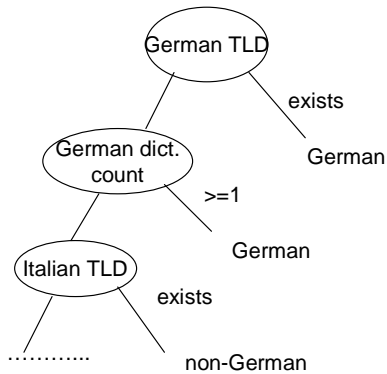
- Intuition: This is what a human would do
  - TLDs (ccTLDs, .org, .net ..)
  - words in a dictionary (trained dic., OpenOffice dic., cities dic.)
  - # of hyphens
- Advantages
  - Limited number of features
  - More efficient
  - Easier to interpret
- Disadvantages
  - Requires large amount of training data

10

# Classification algorithms

- Decision Tree

- Custom-made features



- Naive Bayes, Relative Entropy, Maximum Entropy

- All feature vectors

- SVM, KNN

11

# ccTLD algorithm

- Get the ccTLD of a URL and assign a language according to ccTLD

ccTLD	Language
.us, .gov, .uk, .au, ...	English
.de, .at	German
.fr, .tn, .dz, .mg	French
.es, .cl, .mx, .ar, ...	Spanish
.it	Italian

- Does not require any training data

12

# Data

- Train
  - Search engine results (**SER**)
    - 100k URL for each language
  - Open Directory Project (**ODP**)
    - 125k URL for each language
- Test
  - SER
    - 1k URL for each language
  - ODP
    - 5k URL for each language
  - Web Crawl test set
    - 1k random URL from 100 million web crawl
    - Human evaluation by content → Real label
    - Hardest (not in train set)
- Removed web pages written in many languages from train & test

13

# Evaluation measures

■ Recall  $\frac{|\{\textit{classified +}\} \cap \{\textit{real +}\}|}{|\{\textit{real +}\}|}$

■ Precision

$$\frac{|\{\textit{classified +}\} \cap \{\textit{real +}\}|}{|\{\textit{classified +}\}|}$$

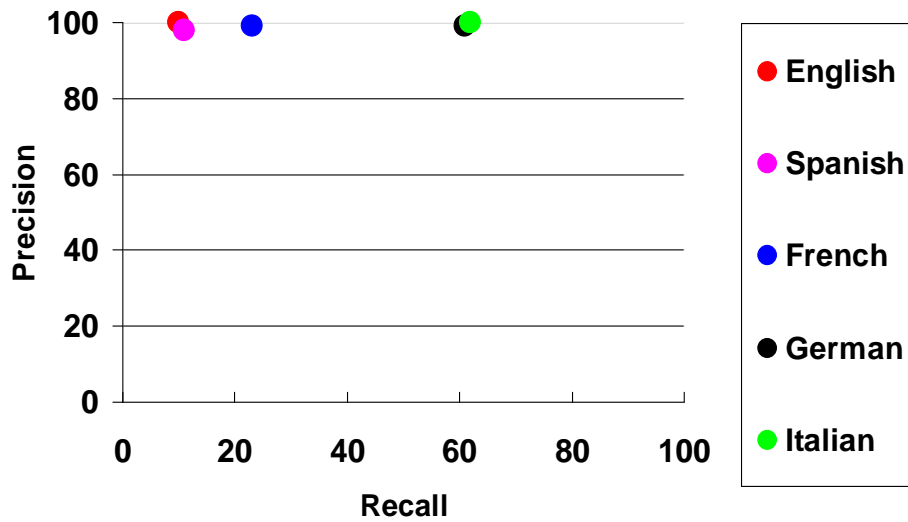
■ F-measure

$$\frac{2 \times \textit{recall} \times \textit{precision}}{\textit{recall} + \textit{precision}}$$

- Always lower than normal avg.

14

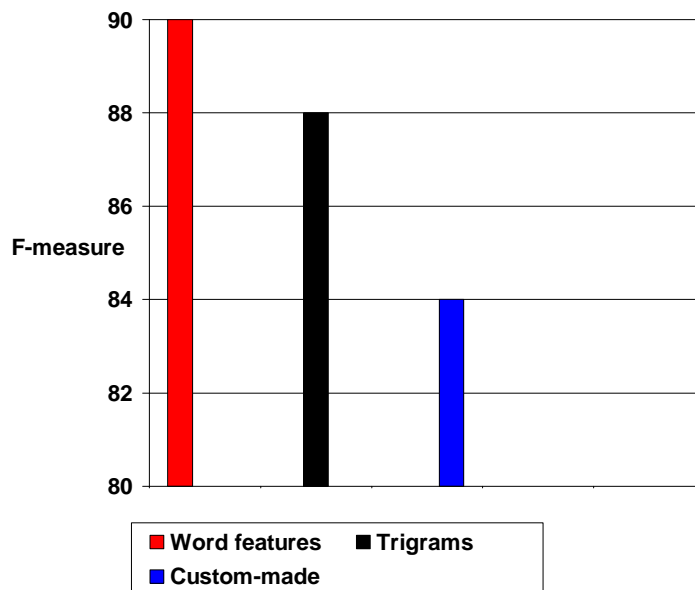
## ccTLD performance on web crawl test



- Why recall low? [.com](#) has many languages

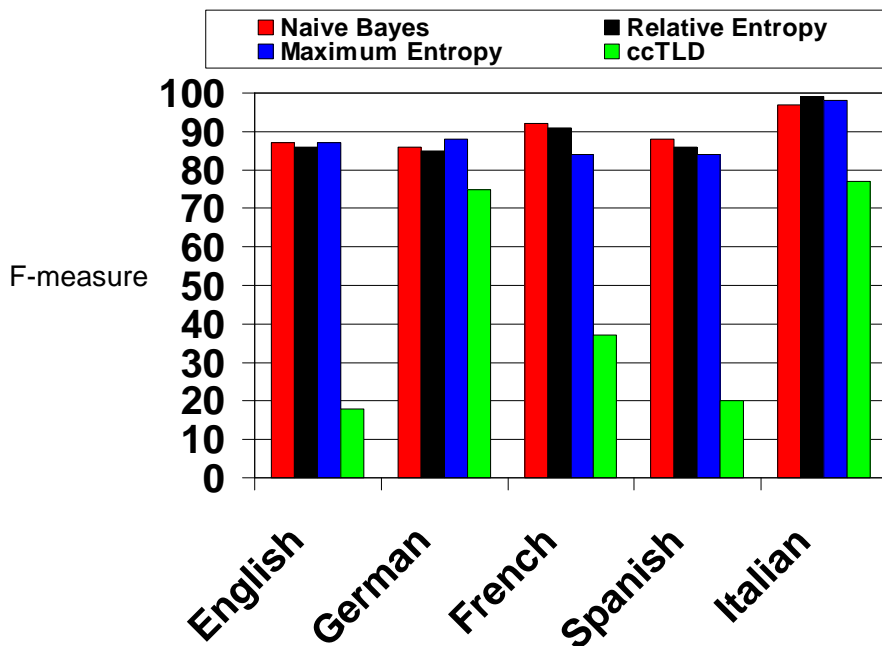
15

## Performance of features on web crawl test



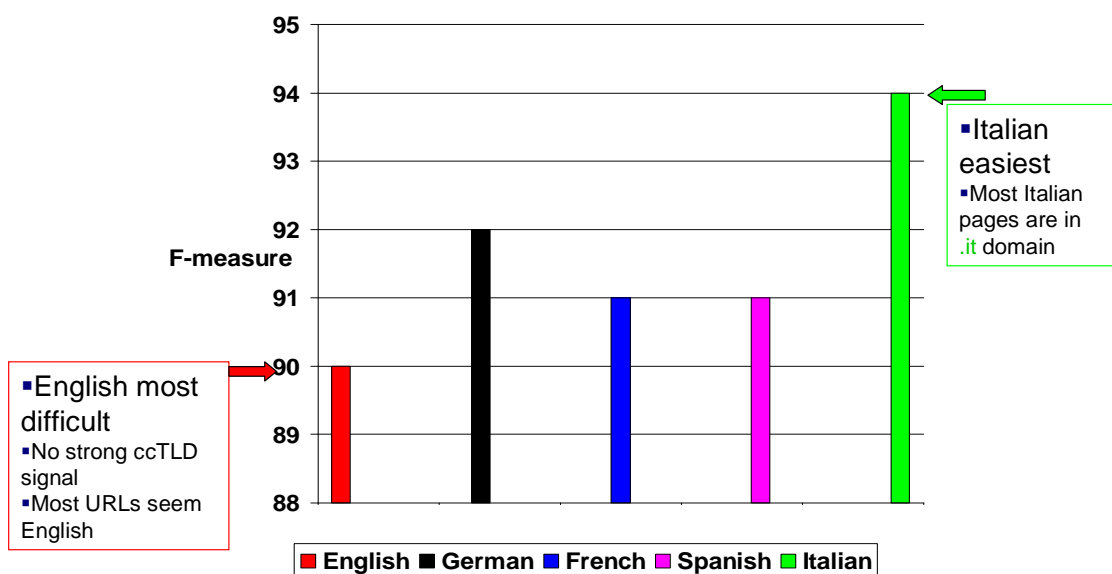
16

## Word features performance on web crawl test



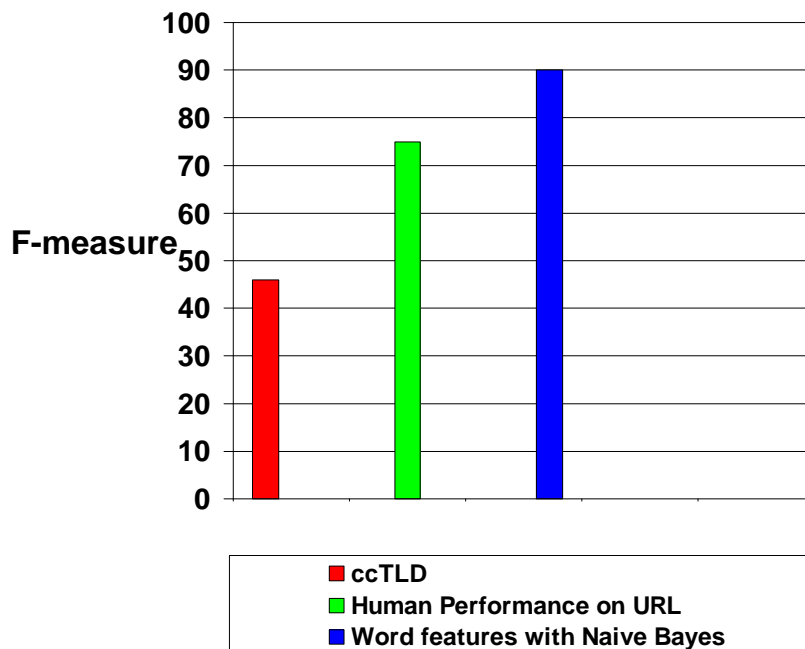
17

## Word features performance avged over test sets with Naive Bayes



18

## Performance on web crawl test avged over languages



19

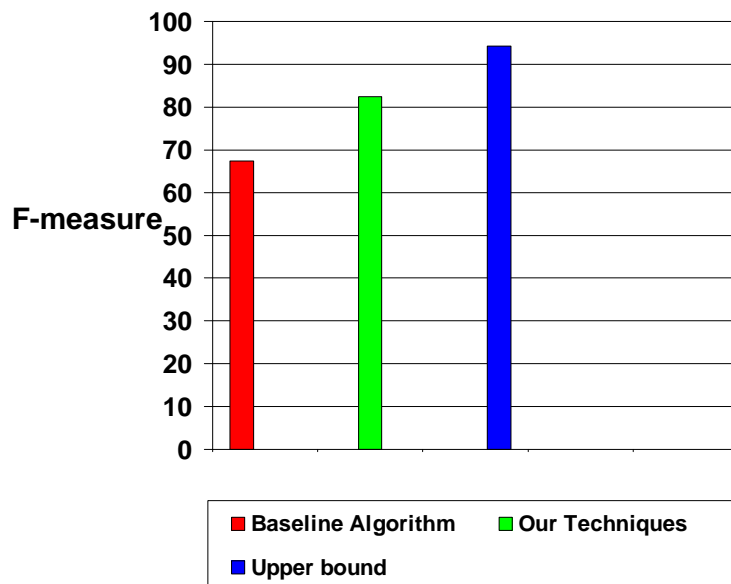
## Conclusions & Future Work

- High quality language identifiers for web pages can be built from URLs
- Feature set more important than classification algorithms
- Performance of feature set depends on amount of training data
  - Lots of training data: words
  - Small amount of training data: trigrams
- Biggest challenge confusion with English
- Inlink information can help
- Web page classification by topic from URL

20

## Topic Classification only from URL

Performance on ODP avged over topics



21

## Questions & Answers

- Thanks

22