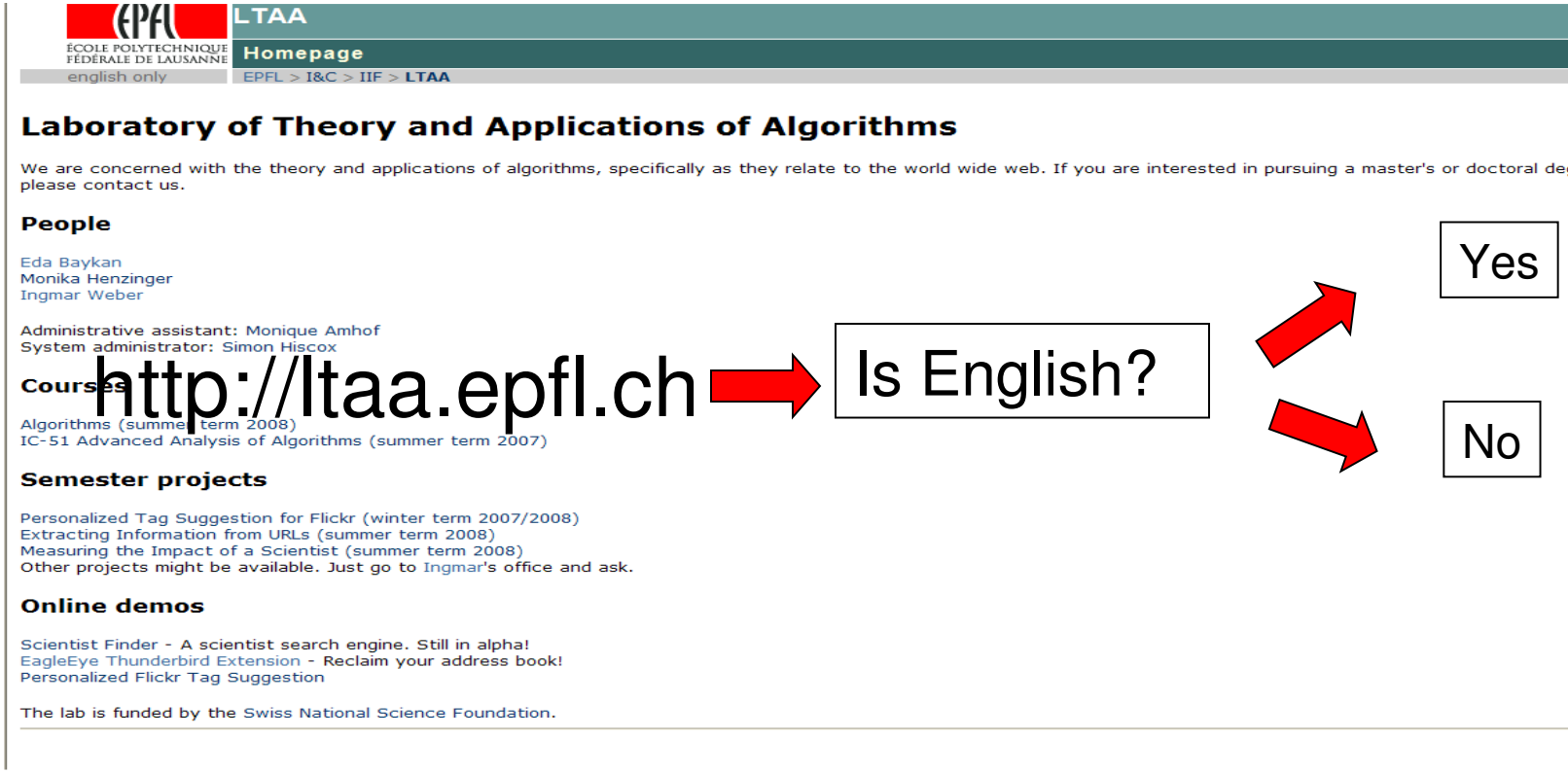
Web page language identification based on URLs

Eda Baykan
Monika Henzinger
Ingmar Weber
EPFL
Lab. of Theory And Applications of Algorithms
VLDB'08

Plan

- Korzyści
- Trudności
- Powiązane prace
- Wymagania systemowe
- Reprezentacja danych
- Algorytmy klasyfikacji
- Dane testowe / treningowe
- Miary oszacowania
- Wyniki

Identyfikacja języka strony WWW na podstawie jej adresu



Korzyści

- Crawlery – quota stron w danym języku
- Poprawa jakości identyfikacji języka na podstawie zawartości
- Wyszukiwarki specyficzne dla konkretnych języków - www.voila.fr, www.yandex.ru, www.fireball.de etc.
  - Marnowanie szerokości pasma

Korzyści c.d.

- Grupowanie / filtrowanie wyników wyszukiwania
- Ikonki językowe przy linkach do stron
- Personalizacja przeglądarek

Trudności

- Adresy stron mają mniej słów niż ich treść
- Adresy zawierają słowa zarówno po angielsku, jak i w ich rzeczywistym języku
  - http://www.bookings-belgium.com/region/be/brugseommeland.fr.html
- Adresy zawierają dużo sztucznych słów
  - google, yahoo, wikipedia
- Domeny .com i .org (np. www.wasserbett-test.com)
  - zawierają ~ 60-70% stron WWW
  - Posiadają strony w wielu językach
- Literówki

Powiązane prace

- Klasyfikowanie stron z hostów akademickich pod względem kategorii (przedmiot/wydział/projekt/student)
- Klasyfikowanie językowe stron na podstawie zawartości

System

- Obsługuje pięć europejskich języków
  - angielski, niemiecki, francuski, hiszpański, włoski
- Binarny klasyfikator dla każdego języka
  - Techniki uczenia maszynowego
    - Faza trenowania (real positive, real negative)
    - Faza testowania (real label, classification label)

Reprezentacja danych

- Mapowanie danych treningowych i testowych na wektory charakterystyczne numerycznych

Wektor charakterystyczny 1: Słowa

- Adres podzielony na znakach niealfanumerycznych
- Usuwane stringi długości < 2 oraz słowa specjalne: „www”, „index”, „html”, „htm”, „http” i „https”
- Przykład: http://www.internetwordstats.com/africa2.htm => internetwordstats, com, africa
- Liczniki wystąpień każdego tokena w adresach danego języka
- Plusy
  - Uczy się nazw domen
  - Zachowuje się lepiej przy dużej ilości danych treningowych
- Minusy
  - Zachowuje się kiepsko przy małej ilości danych treningowych
  - Nie dzieli „zlepionych” słów
    - np. wystąpienie słowa „cheapflights” nie pomoże przy „cheaphouses”

Wektor charakterystyczny 2:
Trigramy

- Spójny ciąg trzech liter
  - np. “*the*” => *angielski*
- Najpierw URL dzielony na tokeny, potem tokeny na trigramy
- Intuicja: powinno pomóc przy „sklejonych” słowach
- Dla każdego języka budowany ze zbioru treningowego rozkład trigramów
- Wybrane *k* najczęstszych trigramów lub wszystkie występujące >= *k* razy
- Strona należy do tego języka, do którego ma najbardziej podobny rozkład

Wektor charakterystyczny 2:

Trigramy

- Co to znaczy „najbardziej podobny”? Rank-order statistic / Relative Entropy
- Dobrze działa, gdy zbiór treningowy wzbogacony o treść stron WWW
- Plusy
  - Zachowuje się dobrze przy małej ilości danych treningowych
- Minusy
  - Nie uczy się nazw domen
  - Plącze się przy adresach wielojęzycznych

Wektor charakterystyczny 3:
Indywidualnie wybrane cechy

- Intuicja: to, co robi człowiek
  - ccTLD (Country Code Top Level Domain) / ccTLD+ - .org, .net ..) + binarna cecha – wystąpienie ccTLD w innych częściach adresu, np. http://fr.search.yahoo.com
  - Słowa ze słownika
    - OpenOffice Dictionary
    - słownik miast
    - słownik wyuczony na zbiorze treningowym - słowo dodane do języka X, jeśli:
      (1) pojawia się w > 0.01% adresów języka X
      (2) >80% adresów, w których słowo się pojawia, należy do języka X)
      oraz (3) ma długość > 2
      np. „arcor” dodany do słownika niemieckiego (provider)
  - Liczba łączników

Wektor charakterystyczny 3:
Indywidualnie wybrane cechy

- Łącznie 74 cechy
- Eliminacja cech zachłannym algorytmem wprzód – redukcja do 15 cech, spadek F-miary tylko o 0.03
- Plusy
  - Ograniczona ilość cech
  - Bardziej wydajne
  - Łatwiejsze w interpretacji
- Minusy
  - Wymaga dużej ilości danych treningowych

Algorytmy klasyfikacji

- Pięć osobnych binarnych klasyfikatorów (wyjątek: ccTLD)

Algorytm ccTLD

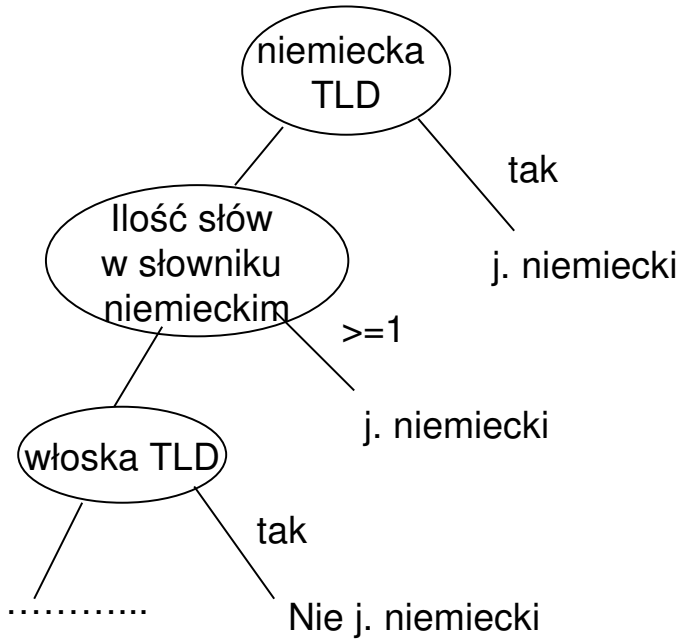
- Pobiera ccTLD z adresu i przypisuje do danego języka

ccTLD	Język
.us, .gov, .mil, .uk, .au, .ie, .nz	angielski
.de, .at	niemiecki
.fr, .tn, .dz, .mg	francuski
.es, .cl, .mx, .ar, .co, .pe, .ve	hiszpański
.it	włoski

- Nie wymaga danych treningowych

Drzewa decyzyjne

- Budowanie drzewa binarnego, węzły wewnętrzne – test pojedynczej cechy, liście - kategorie
  - Indywidualnie wybrane cechy



- Zachłanna zstępująca budowy drzewa (kolejna cecha do podziału wybierana tak, aby w węzłach potomnych coraz dokładniej można było określić kategorię - maksymalizacja przyrostu informacji)

Naiwny klasyfikator bayesowski

- Oparty na założeniu o wzajemnej niezależności cech (class conditional independence)
- Przykład X klasyfikujemy jako pochodzący z tej klasy Ci, dla której wartość P(Ci|X), i = 1, 2, ..., m jest największa
- Wydajnościowo porównywalny do drzew decyzyjnych oraz sieci neuronowych
- Duża dokładność i skalowalność nawet dla bardzo dużych danych

Relatywna entropia

- Odległość Kullbacka-Leiblera (nie jest metryką - nie jest symetryczna i nie spełnia nierówności trójkąta)
- Określa rozbieżność między dwoma rozkładami prawdopodobieństwa
- Przyjmuje wartości nieujemne, przy czym 0 wtedy i tylko wtedy, gdy porównywane rozkłady są identyczne
- Na zbiorze testowym algorytm uczy się rozkładu prawdopodobieństwa dla każdego języka
- Wektor charakterystyczny ze zbioru testowego jest konwertowany na rozkład prawdopodobieństwa i zostaje zaklasyfikowany do tego języka, dla którego relatywna entropia między jego wyuczonym średnim rozkładem a rozkładem tego wektora jest najmniejsza

Maksymalna entropia

- Gdy nic nie wiadomo, rozkład powinien być na tyle jednolity, na ile to możliwe (czyli mieć maksymalną entropię)
- Dane treningowe dostarczają więzów charakteryzujących wymagania dla rozkładu specyficzne dla kategorii
- Więzy – oczekiwane wartości cech
- Problem optymalizacji z więzami – iterative scaling

Dane

Data set	Language	Training size	Test size
Open Directory Project	English	145,000	4,910
	German	144,999	4,965
	French	144,996	4,961
	Spanish	144,974	4,878
	Italian	144,987	4,933
Search Engine Results	English	99,992	999
	German	99,572	992
	French	99,549	997
	Spanish	99,838	997
	Italian	99,786	997
Web Crawl	English	0	1,082
	German	0	81
	French	0	57
	Spanish	0	19
	Italian	0	21

Miary oszacowania

- Precyzja $\frac{|\{classified+\} \cap \{real+\}|}{|\{classified+\}|}$
- Recall - positive success ratio p(+|+).

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

- Negative success ratio p(–|–).
- F-miara

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

Ludzka wydajność

Test set	Language	P	$R = p(+\|+)$	$p(-\|-)$	F
Web Crawl	English	.73	.99	.63	.84
	German	.99	.70	.99	.82
	French	.99	.54	.99	.70
	Spanish	.99	.37	.99	.54
	Italian	.99	.76	.99	.86

Table 2: Aggregate numbers of the human performance on the web crawl test set. The language refers to the decision made by the humans: "Is the page written in language X or not?"

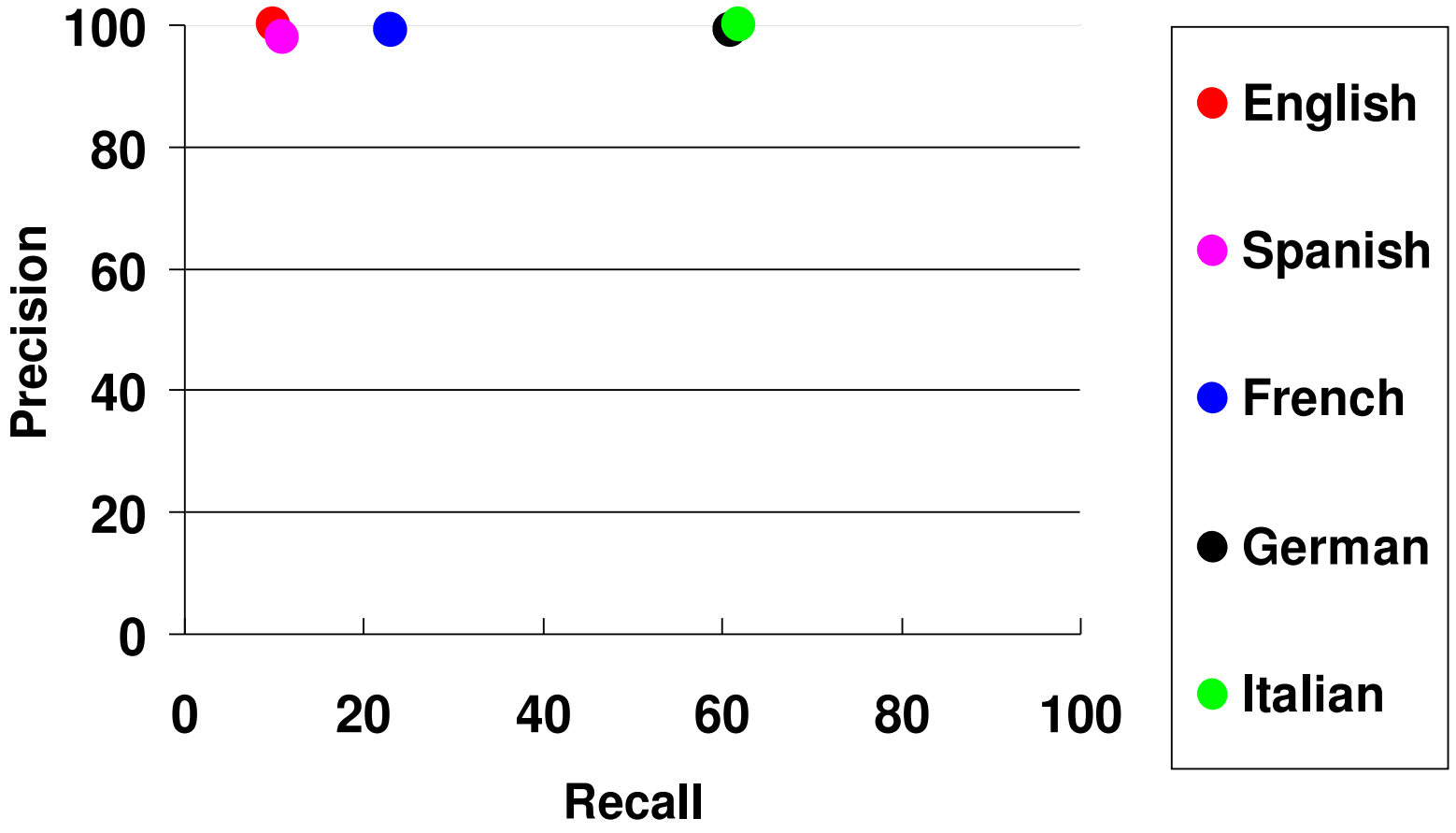
Macierz pomyłek

Test set lang.	Reported language by human evaluators				
	English	German	French	Spanish	Italian
En.	99%	0%	1%	0%	0%
Ge.	30%	70%	0%	0%	0%
Fr.	45%	0%	54%	1%	0%
Sp.	58%	0%	0%	37%	5%
It.	24%	0%	0%	0%	76%

Table 3: Confusion matrix for the human evaluation on the crawl test set, averaged over both evaluators. A single cell shows the percentage of URLs from language X (the row) for which language Y (the column) is reported. Note that for all languages the biggest confusion is with English, i.e., URLs "look" English, although the corresponding web page is not.

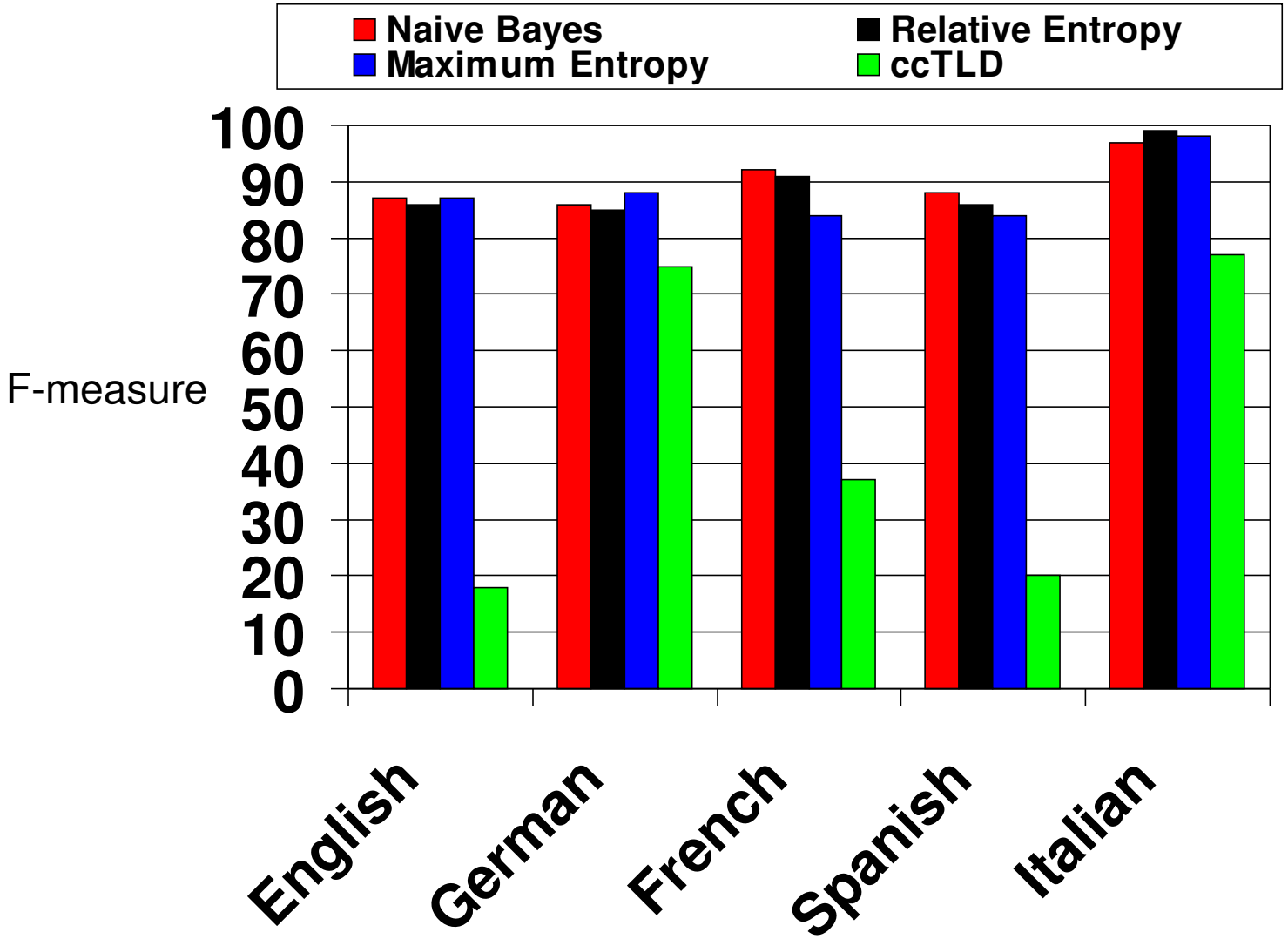
- Strony nieangielskie wykazują problem z miarą recall
- Przyczyna: angielski uważany za „język techniczny”
- `http://forum.mamboserver.com/archive/index.php/t-7062.html` – „typowa” strona niemiecka
- `http://www.priceminister.com/navigation/default/category/126541/l1/q` – „typowa” strona francuska

Wydajność ccTLD na zbiorze testowym web crawl

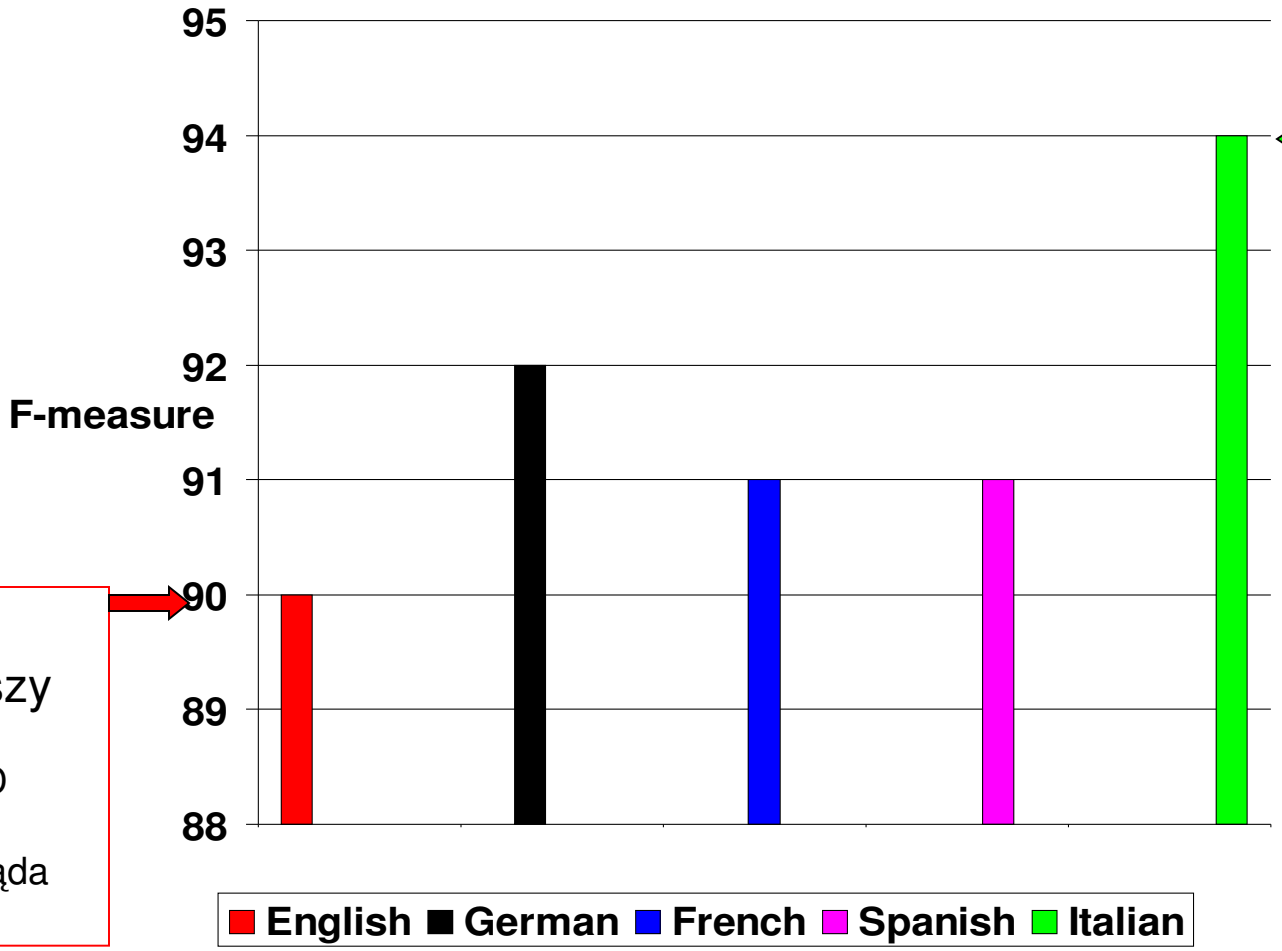


- Dlaczego recall niski? Domena .com zawiera dużo stron różnych języków

Wydajność wektorów słów na zbiorze testowym web crawl



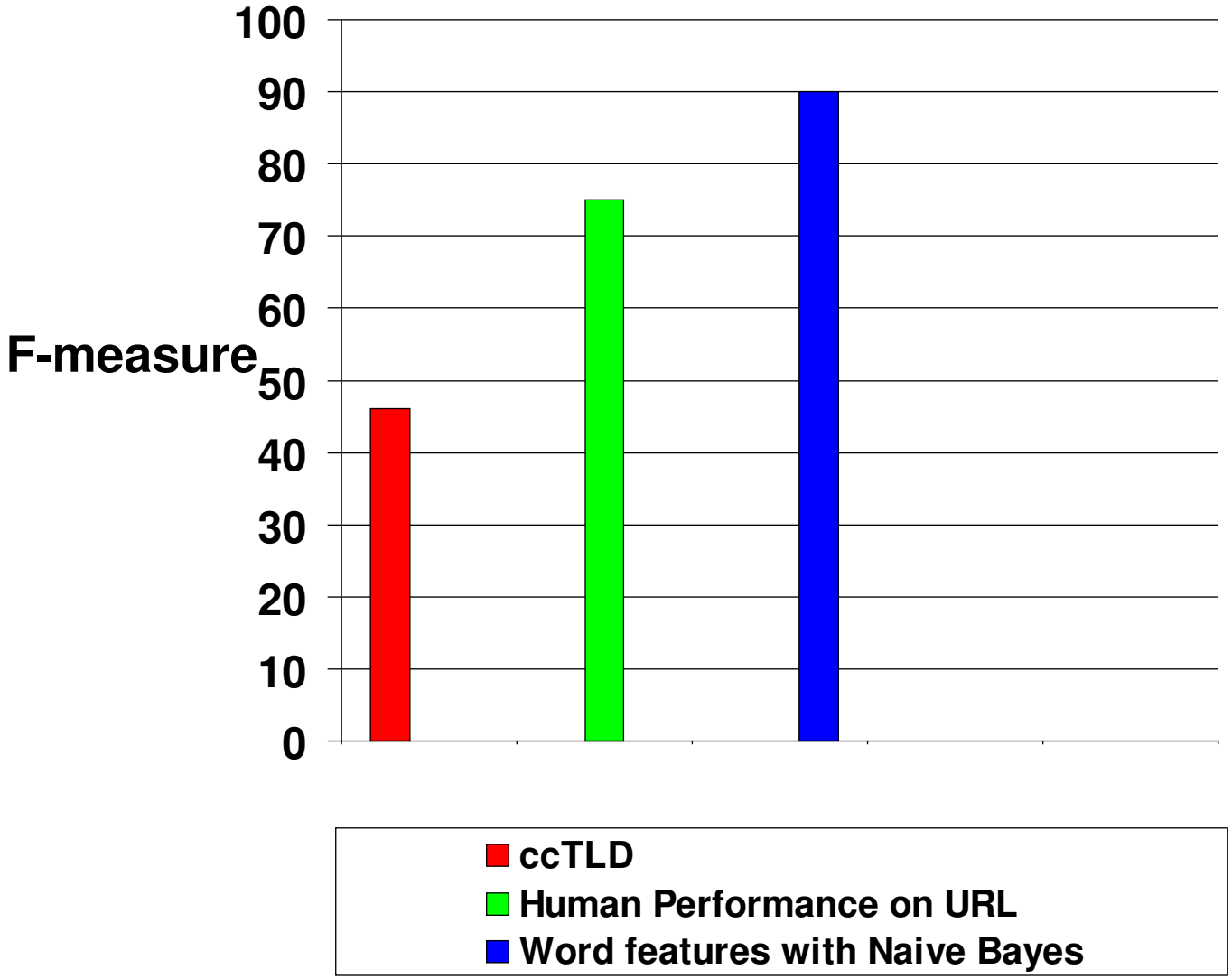
Wydajność wektorów słów uśredniona na zbiorach testowych z naiwnym klasyfikatorem bayesowskim



- Włoski najprostszy
  - Większość włoskich stron jest w domenie .it

- Angielski najtrudniejszy
  - Brak silnego sygnału ccTLD
  - Większość adresów wygląda na angielskie

Wydajność na zbiorze testowym web crawl uśredniona na językach

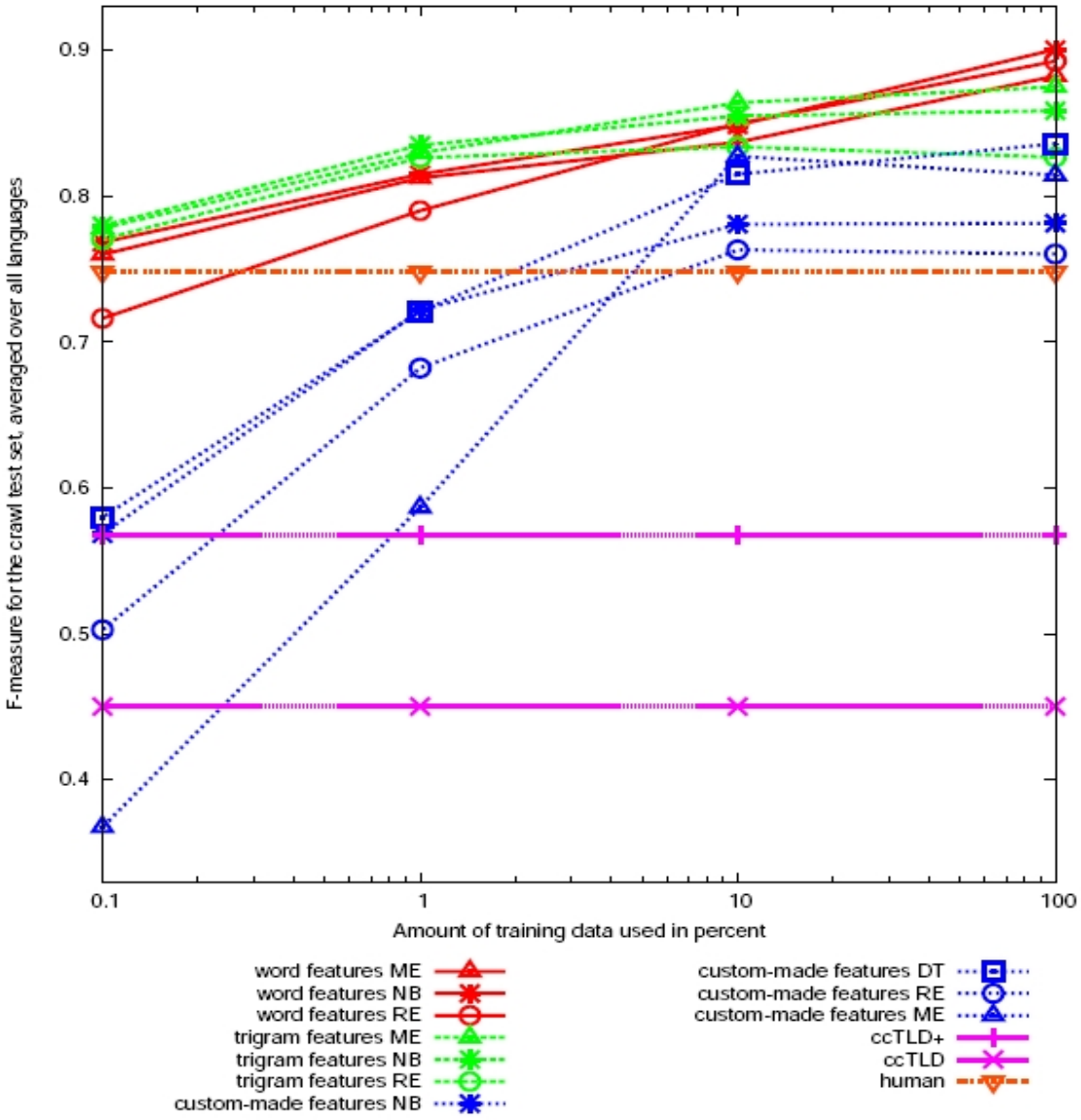


Wydajność najlepszych kombinacji algorytmów uśredniona na zbiorach testowvch

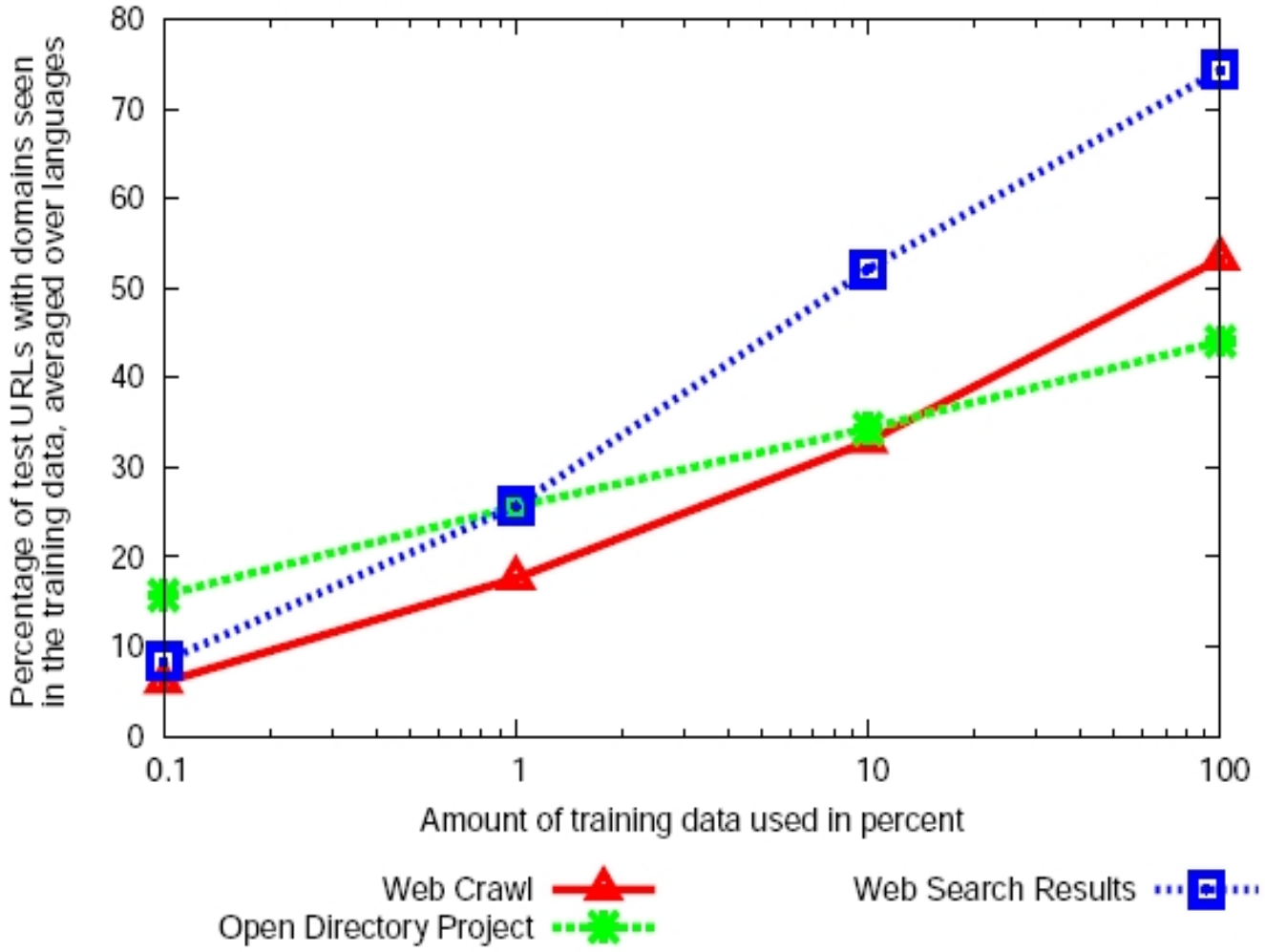
Classifier language	Test set			Average over test sets
	ODP	SE	Crawl	
English	.87	.95	.88	.91
German	.95	.97	.88	.93
French	.88	.94	.91	.91
Spanish	.89	.96	.93	.93
Italian	.90	.97	.97	.95
Average	.90	.96	.92	.93

Table 9: F-measure results when for each language the best combination of classifiers was chosen. These combinations were then used for all three test sets.

Zależność od ilości danvch



Spamiętywanie nazw domen



Trenowanie na zawartości

Alg.	English		German		French		Italian		Spanish	
	U	Co	U	Co	U	Co	U	Co	U	Co
NB	.87	.81	.94	.77	.86	.79	.86	.85	.87	.83
ME	.87	.81	.93	.70	.86	.79	.85	.81	.86	.83

Table 10: F-measure for Naive Bayes (NB) and maximum entropy (ME) on the ODP test set for URL-based (U) and content-based classifiers (Co) with word features trained only on the ODP training set.

Wnioski i praca na przyszłość

- Można zbudować wysokiej jakości klasyfikatory językowe wyłącznie z adresów stron WWW
- Najtrudniejsze: rozpoznawanie „angielsko-wyglądających adresów nie-angielskich stron”
- Ważniejszy zbiór cech niż algorytm klasyfikacji
- Wydajność na zbiorze cech zależy od ilości danych treningowych
  - dużo: słowa
  - mało: trigramy
- Może pomóc informacja o strukturze dowiązań
- Klasyfikacja tematyczna na podstawie adresu?...