



GOVERNMENT
USERS CONFERENCE

JUNE 8-9
2010
CHANTILLY, VA

Language Identification: The First Step in Processing Intelligence

Nobuo Otsuka

Senior Software Engineer

Basis Technology Corp.



Human Language Technology From Arabia to Afghanistan

What is Language Identification?

- High-speed, automatic detection of language and encoding of documents.
 - Documents in a single language may be encoded in different ways
 - E.g. Japanese – Shift-JIS, ISO-2022-JP encoding...etc.
- The very first step in processing any volume of foreign data.

Who Uses Language Identifiers?

- Search engine vendors Microsoft, Yahoo
- E-Discovery firms KPMG
- Web rating agencies comScore

Looks Like Russian, But...

Председник ruske vlade Владимир Путин позвао је, после разговора са својим пољским колегом Доналдом Туском, да се спорна питања из историје Другог светског рата оставе историчарима, а да се данас посвети пажња развоју односа између две земље.

“Vladimir Putin”

Cyrillic Script Languages

A group of languages share Cyrillic script.

- Russian
- Ukrainian
- Bulgarian
- **Serbian**
- Macedonian
- Uzbek

Looks Like Arabic, But...

ہندوستان کی خارجہ سیکرٹری نروپما راؤ نے کہا کہ پاکستانی وزیر اعظم کو دہشتگردی، دراندازی اور ممبئی پر حملوں کے ملزمان کے خلاف مقدمات کی 'سست رفتاری' کے بارے میں ہندوستان کے خدشات سے آگاہ کیا۔

Arabic Script Languages

A group of languages share the Arabic script.

- Arabic
- Persian
- Pushto
- Kurdish
- Urdu

RLI Language Identification Approach

- Rosette Language Identifier (RLI) uses a statistical approach that requires:
 - No dictionary
 - No character map
 - No hard code for grammar analysis

RLI Detection Coverage

- 55 Languages
 - European
 - Cyrillic
 - Middle Eastern
 - Indic
 - South East Asian
 - East Asian
- 35 encodings
 - ISO 8859-1, Cp1256, Shift-JIS, ...
 - UTF-8, UTF-16 for every language

RLI Statistical Approach

- Based on N-gram statistical model
- N-gram:
 - N “item” sequence extracted from document.
 - N=1: unigram
 - N=2: bigram
 - N=3: trigram
 - N \geq 4: N-gram
- N = 4 bytes for RLI

Extract N-grams

“their heir to the throne”

N-gram *frequency*

'the' 2

'heir' 2

'eir ' 2

'hron' 1

' thr' 1

' to ' 1

'thei' 1

'thro' 1

'the ' 1

'rone' 1

' he' 1

'one ' 1

Put N-grams Into Profiles

English/Cp1252

5000 N-grams

" the" 1.518770
"the " 1.271283
"and " 0.772795
"ing " 0.767108
" of " 0.763523
" to " 0.753107
" and" 0.725608
"tion" 0.651863
.
.
.

German/UTF-8

5000 N-grams

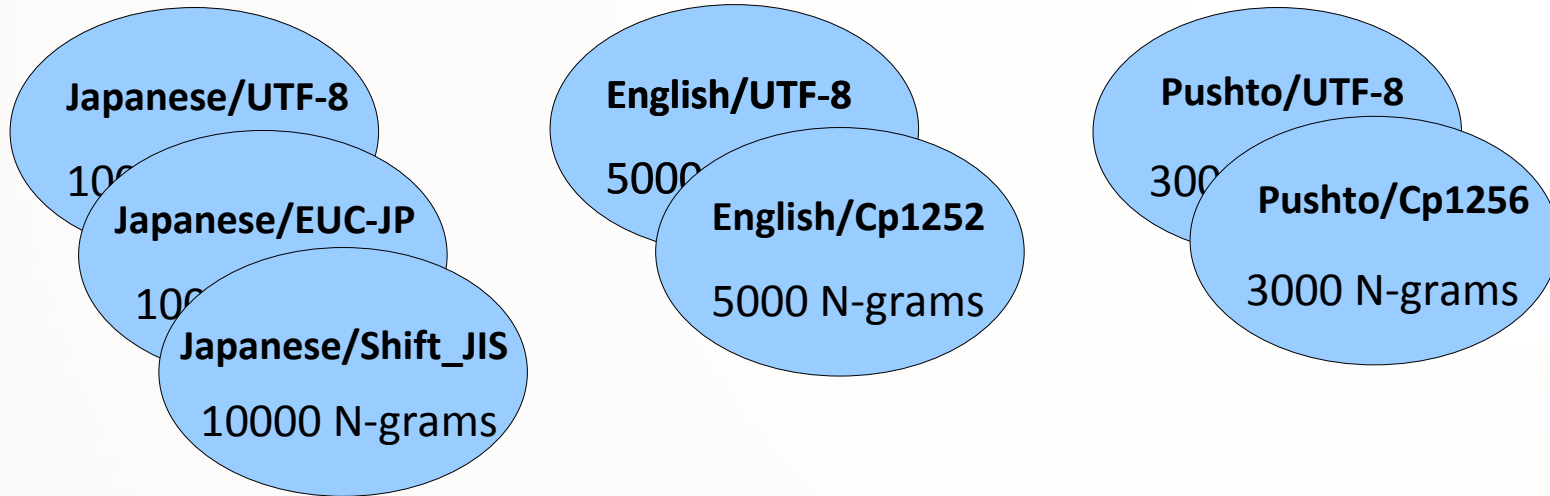
"ten " 0.69057
"und " 0.67099
"der " 0.64234
" und" 0.61716
"den " 0.54616
" der" 0.529223
"ung " 0.522831
.
.
.

French/Cp1252

5000 N-grams

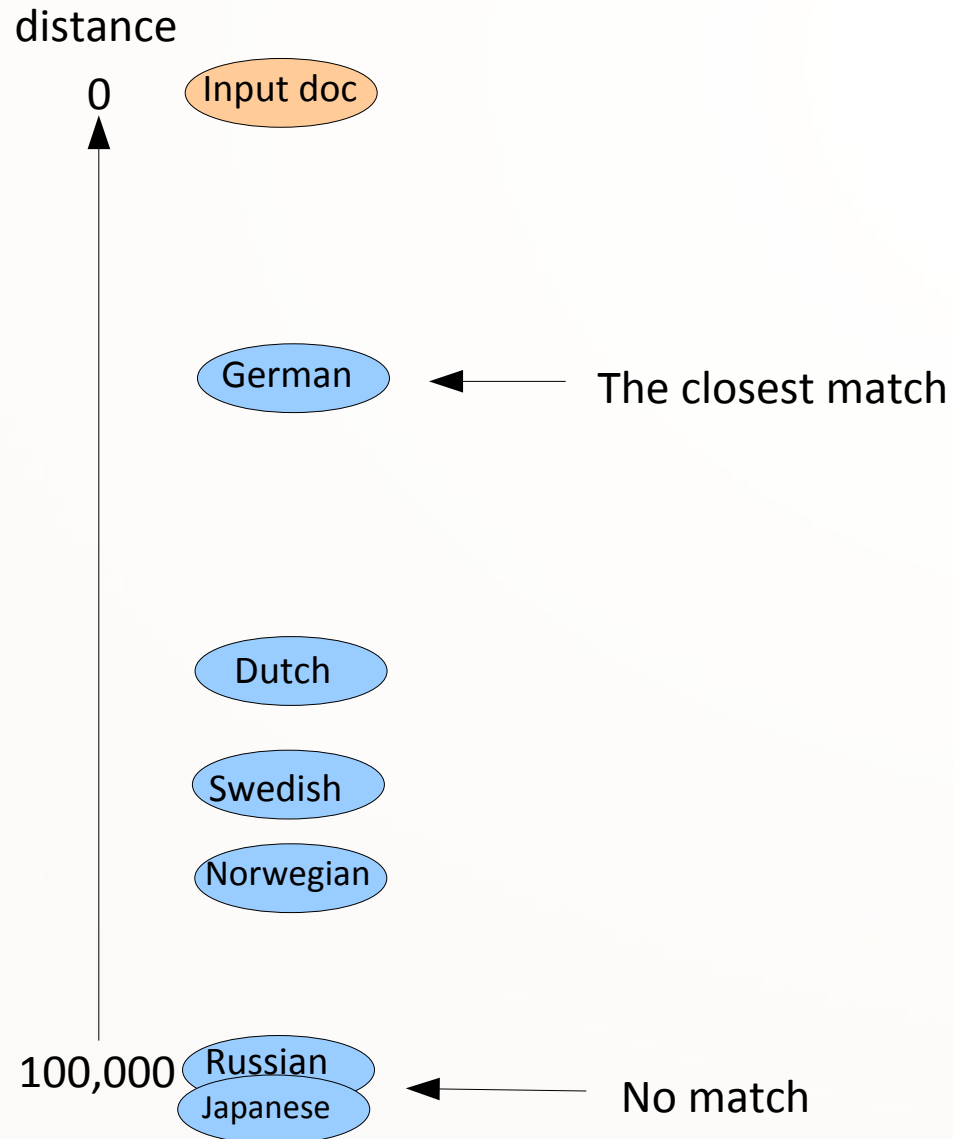
" de " 1.920552
"les " 0.777410
" et " 0.730199
" la " 0.713694
"des " 0.708883
"tion" 0.683449
" des" 0.674289
.
.
.

N-gram Profiles



- 155 profiles of language/encoding pair
- Trained on 3 million documents in the real world

N-gram Statistical Model



RLI System Size

- Binary size
 - 40k bytes per profile
 - 800K bytes core engine
- Runtime memory size
 - 230k bytes per profile
 - 15.3M bytes core engine
- Customize the number of profiles
 - Reduce memory foot print.
 - Faster

RLI Throughput

- 1500 docs/sec

based on document size: 5k bytes

5K = average web page doc size (without html tags)

Configuration:

CPU: AMD Athlon 64 dual 2.79 GHz,

512kBx2 L2 cache

Memory: 2.5GB RAM

Digits, Punctuations Don't Count

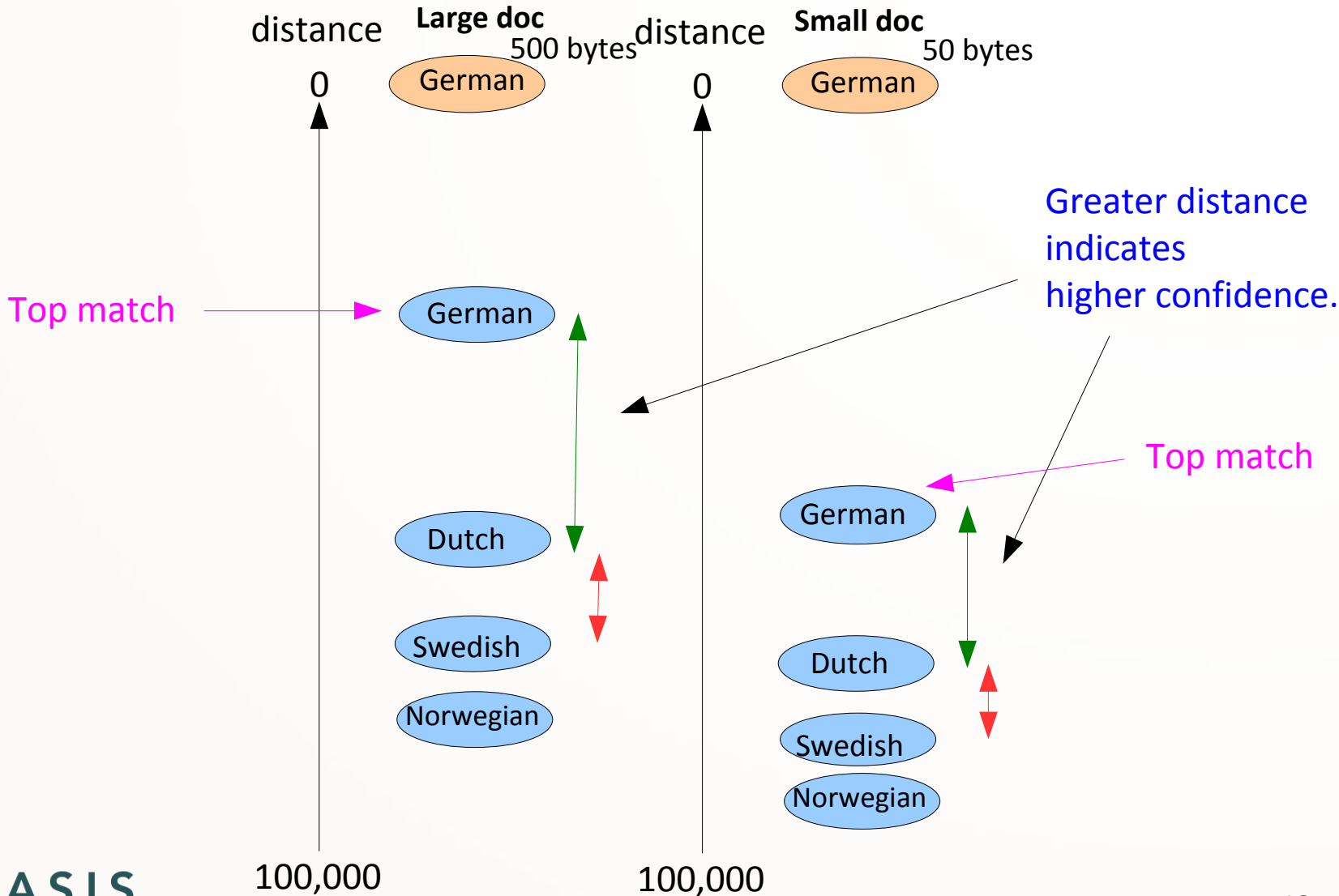
“DCR-TRV22K, DCR-TRV27, DCR-TRV30, DCR-TRV33K, DCR-TRV50
<http://www.haloscan.com/comments/nazlik/1241390718461393557/?src=hsr#323683>
<http://www.haloscan.com/comments/nazlik/1241390718461393557/?src=hsrs#32368>”

- No match result

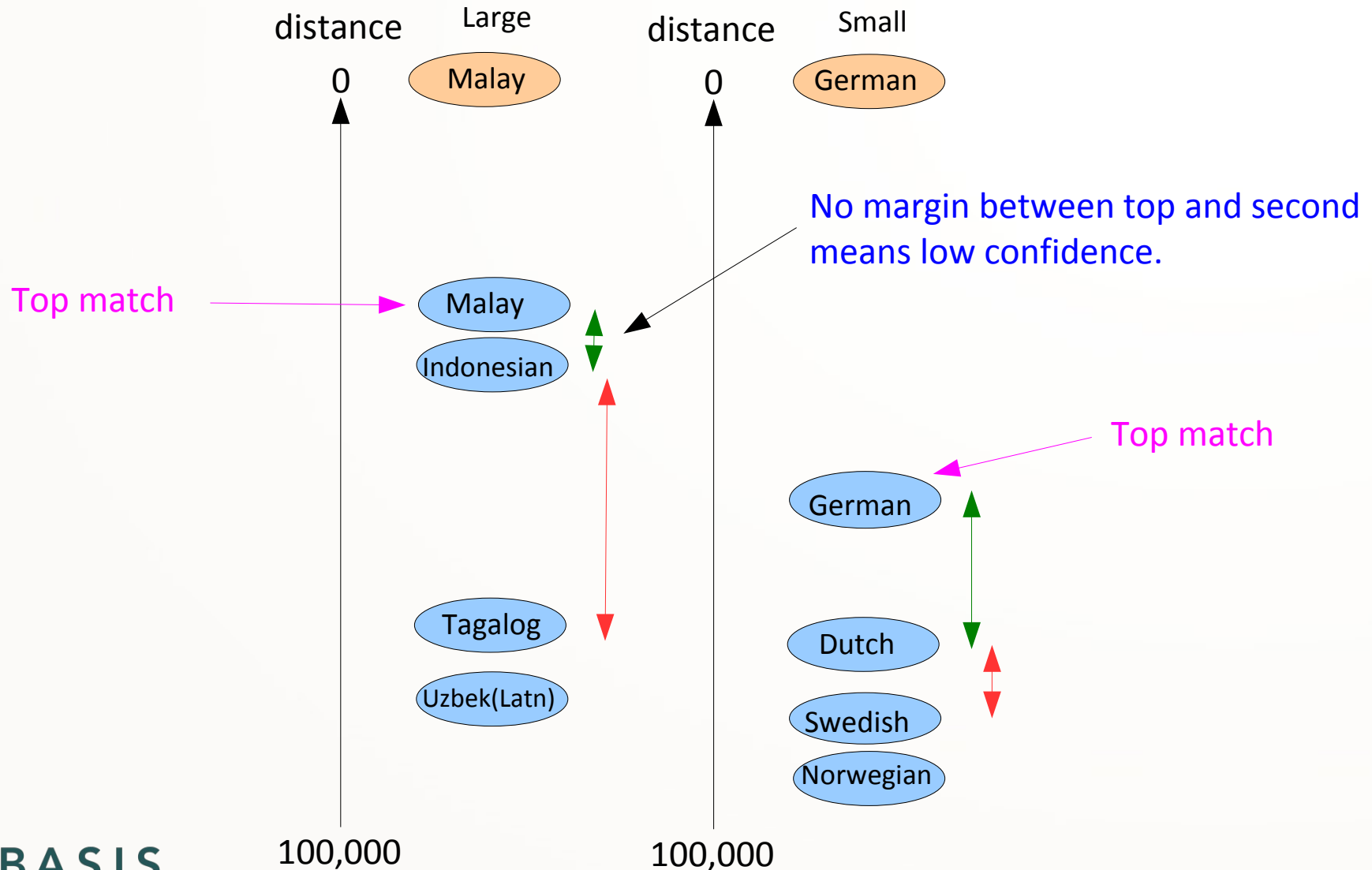
Names Don't Count

- “with Marisa” → English
- “con Marisa” → Italian
- “Marisa” → No match
- “胡錦濤主席が訪日する” → Japanese
 “Hu Jintao”

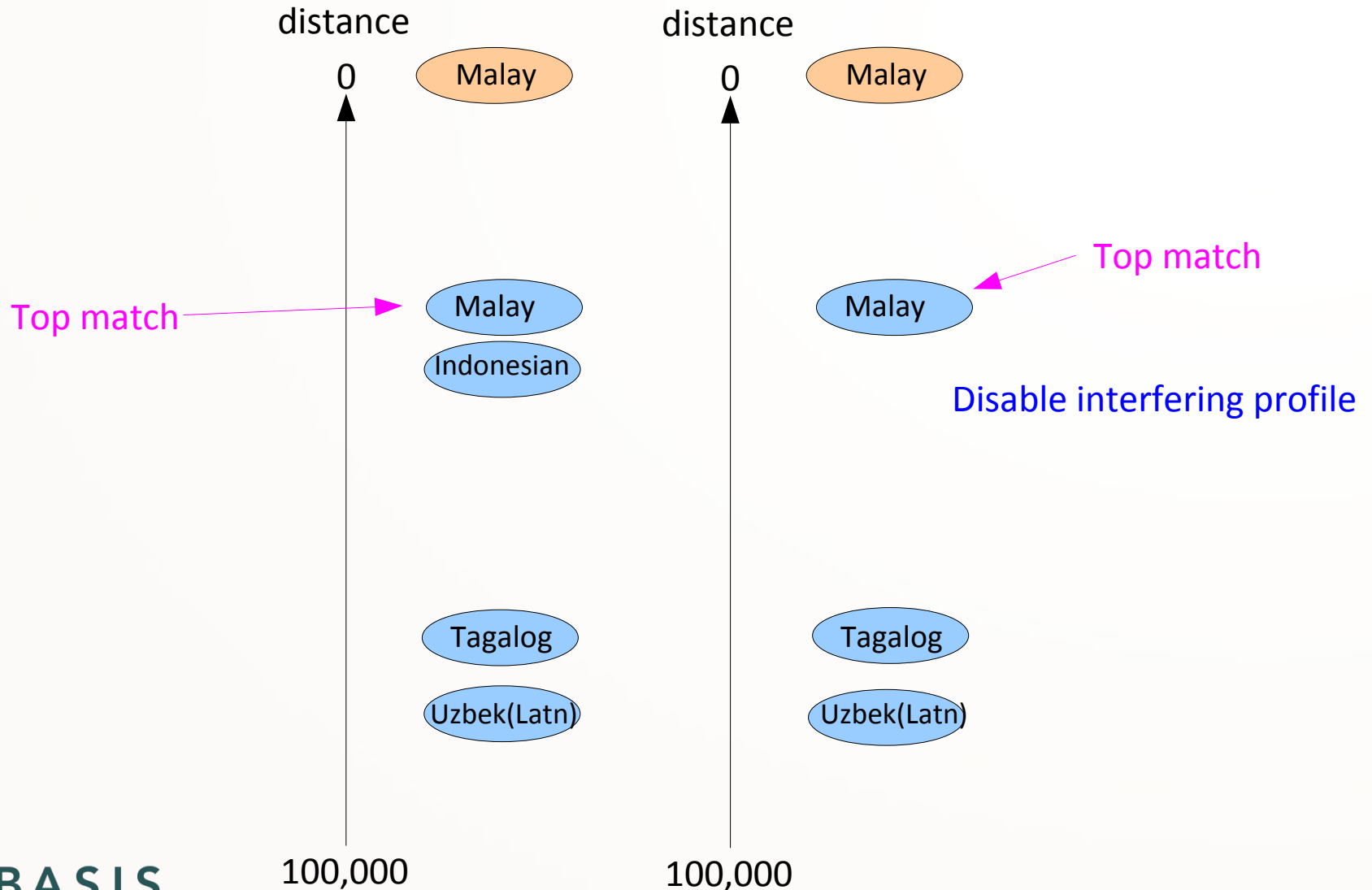
Input Size vs. Accuracy



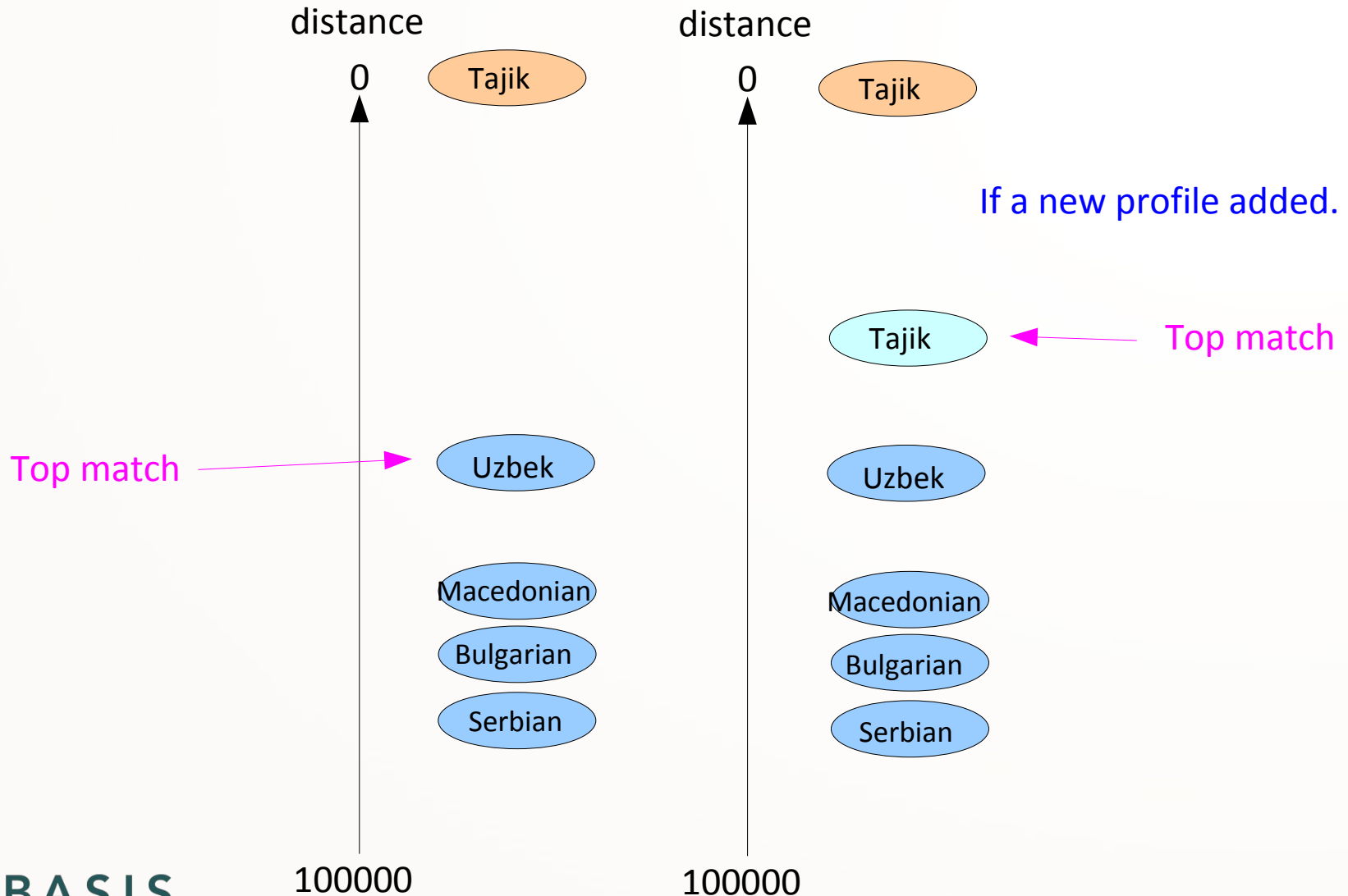
Confidence Level of Match Results



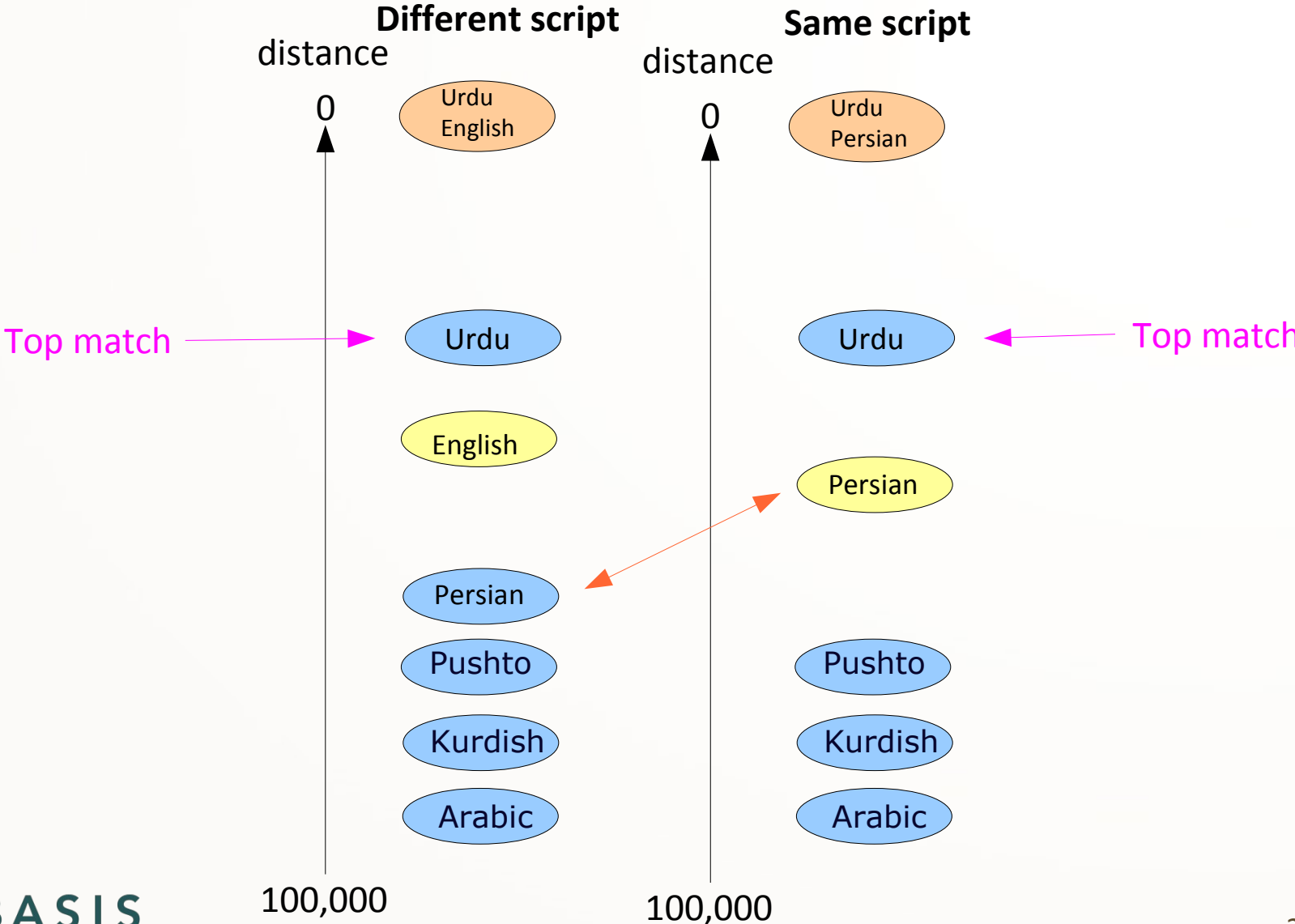
How To Improve Accuracy



What if Unsupported Languages are Input?



How To Detect a Multilingual Document



RLI Detection Accuracy

128 bytes input needed for 100% detection accuracy

Input size	Arabic	Persian	Greek	Spanish	French	Korean	Japanese
128 bytes	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
64 bytes	99.01%	97.68%	100.00%	97.81%	100.00%	100.00%	99.14%
32 bytes	94.50%	92.01%	100.00%	97.10%	97.69%	100.00%	98.60%
16 bytes	87.18%	78.39%	100.00%	85.14%	88.03%	100.00%	92.59%
8 bytes	79.92%	69.48%	99.93%	67.06%	78.33%	97.90%	78.39%

What's Next?

- Multilingual document accuracy improvement
- Sub-language detection
 - Arabic
 - Koranic
 - Chat
 - Dialects

RLI advantages

- Speed
- Wide language coverage
- Configurable language detection scope
- Confidence scores

Questions



For more information:
Visit www.basistech.com

Write to
info2010@basistech.com

Call 617-386-2090 or
800-697-2062

