



What Language Is That?

Using the Rosette Language Identifier (RLI)

Nobuo Otsuka, Senior Software Engineer
June 14th, 2006

Basis Technology Corporation

P 617.386.2000
800.697.2062 (toll-free)
F 617.386.2020
W info@basistech.com
www.basistech.com



Overview of RLI

- Identify
 - 43 Languages
 - 33 native encodings
 - UTF-8 for every language
- Based on N-gram technology
- Plain text data detection
 - No HTML content-type/language information used



N-gram profiles

- 114 profiles of language/encoding pair
- Statistical quad-grams (3000-10000) per profile
- Trained from 3 million documents in real world



RLI System size

- DLL
 - 40K bytes per profile
- Runtime memory
 - 230K bytes per profile
 - 15.3M bytes core engine



Detection speed

Throughput:

2714K bytes / second

Input text: 2K bytes

Configuration:

CPU: Dual Intel Pentium 4 3GHz,

512KB L2 cache

Memory: 1GB RAM



Encoding detection

- N-gram match + Byte validation
- Demotion
 - Cp1252 → ISO-8859_1 → ASCII
- Promotion
 - Shift_JIS → Shift_JIS-2004(JISX0213)



Detection accuracy

- 128 bytes input needed for 100% detection accuracy

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



Short string detection accuracy

- Short string detection
 - RSS
 - HTML title
 - Mail Subject
- Accuracy test
 - detecting 800,000 HTML titles (Average 39 characters)
 - RLI 4.3.0 misidentified 22%
 - RLI 5.0 misidentified 9%



Detection Example 1 (small input)

Input text:

Commissione per l'accesso ai documenti amministrativi

Results:

Italian/US-ASCII Distance: **512** Match 37 of 42 N-grams Input Size: 48 valid chars ambiguous
English/US-ASCII Distance: **109** Match 33 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
Portuguese/US-ASCII Distance: **131** Match 26 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
Romanian/US-ASCII Distance: **131** Match 25 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
Spanish/US-ASCII Distance: **59** Match 23 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
French/US-ASCII Distance: 99355 Match 23 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
Dutch/US-ASCII Distance: 99392 Match 23 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
Finnish/US-ASCII Distance: 99674 Match 11 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
Danish/US-ASCII Distance: 99479 Match 19 of 42 N-grams Input Size: 48 valid chars invalid ambiguous
Norwegian/US-ASCII Distance: 99540 Match 17 of 42 N-grams Input Size: 48 valid chars invalid ambiguous

Number of input characters
excluding white spaces

Number of N-grams in input

Number of N-grams matched with profile

Vector distance between input and profile. 100000 means no match.



Detection Example 2 (large input)

Input text:

La prima voce delle competenze spettanti al parlamentare è l'indennità, quella che nel linguaggio comune è definita "stipendio". Seguono la diaria e i rimborsi: per le spese inerenti i supporti per lo svolgimento del mandato parlamentare, per le spese accessorie di viaggio e per i viaggi all'estero, per le spese telefoniche. Completano la scheda le voci sull'assegno di solidarietà (di fine mandato), sulle prestazioni previdenziali e sanitarie e sui trasporti. L'indennità, prevista dalla Costituzione all'art. 69, è determinata, in base alla legge n. 1261 del 31 ottobre 1965, in misura non superiore al trattamento complessivo massimo annuo lordo dei magistrati con funzioni di presidente di Sezione della Corte di Cassazione ed equiparate. Con delibera del 1993 il Consiglio di presidenza del Senato ha stabilito che tale misura fosse ridotta al 96% del predetto trattamento dei magistrati. Per effetto delle disposizioni contenute nella legge finanziaria 2006, l'importo lordo dell'indennità ha subito inoltre una riduzione pari al 10 per cento. L'indennità è corrisposta per 12 mensilità. L'importo mensile è pari ora a 5.419,46 euro (prima del "taglio" della finanziaria 2006 era pari a 5.941,91 euro), al netto della ritenuta fiscale (€ 3.555,63), nonché delle quote contributive per l'assegno vitalizio (€ 962,42), per l'assegno di solidarietà (€ 749,79) e per l'assistenza sanitaria (€ 503,59). Nel caso in cui il Senatore versi anche la quota aggiuntiva per la reversibilità dell'assegno vitalizio (2,15 per cento), l'importo netto dell'indennità scende a 5.178,86 euro.

Results:

Italian/windows-1252 Distance: 544	97154 Match 497 of 699 N-grams	Input Size: 1347 valid chars
Italian/UTF-8 Distance: 1514	97698 Match 494 of 699 N-grams	Input Size: 1347 valid chars ambiguous
Spanish/windows-1252 Distance: 57	99212 Match 294 of 699 N-grams	Input Size: 1347 valid chars invalid ambiguous
Catalan/windows-1252 Distance: 58	99269 Match 225 of 699 N-grams	Input Size: 1347 valid chars invalid ambiguous
Romanian/windows-1250 Distance: 28	99327 Match 283 of 699 N-grams	Input Size: 1347 valid chars invalid ambiguous
Spanish/UTF-8 Distance: 28	99369 Match 289 of 699 N-grams	Input Size: 1347 valid chars invalid ambiguous
Portuguese/windows-1252 Distance: 99399	Match 232 of 699 N-grams	Input Size: 1347 valid chars invalid ambiguous



Detection Example 3 (dialect - similar language)

Input text:

Penetapan harga dasar oleh Pemerintah India tersebut berguna bagi perhitungan bea masuk dan untuk mencegah hilangnya pendapatan akibat under invoicing yang dilakukan oleh importer.

Results:

Indonesian/US-ASCII Distance:	206	97782	Match 99 of 130 N-grams	Input Size: 157 valid chars ambiguous
Malay/US-ASCII Distance:	1542	97988	Match 98 of 130 N-grams	Input Size: 157 valid chars ambiguous
English/US-ASCII Distance:	72	99530	Match 47 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous
Tagalog/US-ASCII Distance:	35	99602	Match 29 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous
Romanian/US-ASCII Distance:	19	99637	Match 38 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous
Norwegian/US-ASCII Distance:		99656	Match 33 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous
Spanish/US-ASCII Distance:		99666	Match 35 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous
Finnish/US-ASCII Distance:		99760	Match 19 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous
French/US-ASCII Distance:		99791	Match 22 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous
Italian/US-ASCII Distance:		99692	Match 32 of 130 N-grams	Input Size: 157 valid chars invalid ambiguous



Detection Example 4 (Multilingual text)

Input text:

Posted by AyyA:: at :: 9:40 PM::

Walladah said...

بلغتي العرجاء تذكرت أغنية طفولتي من فيلم صوت الموسيقى

هل هناك أي رابط أم أن العتب على اللغة العرجاء؟

10:05 PM

AyyA said...

Walla ma7ad fahimni kithrich Princess, you got it right ;)

10:09 PM

Results:

Arabic/windows-1256 Distance: 95589 Match 35 of 133 N-grams Input Size: 197 valid chars
English/ISO-8859-1 Distance: 99523 Match 26 of 133 N-grams Input Size: 197 valid chars invalid ambiguous
English/UTF-8 Distance: 99523 Match 26 of 133 N-grams Input Size: 197 valid chars invalid ambiguous
ArabicTranslit/windows-1256 Distance: 99724 Match 17 of 133 N-grams Input Size: 197 valid chars invalid ambiguous
ArabicTranslit/UTF-8 Distance: 99773 Match 17 of 133 N-grams Input Size: 197 valid chars invalid ambiguous
ArabicTranslit/ISO-8859-1 Distance: 99773 Match 17 of 133 N-grams Input Size: 197 valid chars invalid ambiguous
Persian/windows-1256 Distance: 99819 Match 11 of 133 N-grams Input Size: 197 valid chars invalid ambiguous



What's new

- Pashto, Urdu, Kurdish, Somali
- Romanized Arabic, Persian
- Shift_JIS-2004 (JISX0213)
- Language scope options



Languages in the same script

- Arabic script
 - Arabic, Persian, Pashto, Urdu, Kurdish
- ASCII
 - English, Indonesian, Malay, Tagalog, Somali
- Chinese script
 - Chinese, Japanese, Korean



Setting language detection scope

- Improve
 - Performance
 - Runtime memory
 - Misidentification
 - Multilingual text detection

- XML option file

```
..
<language load="yes" name="Persian" />
<language load="yes" name="Pashto" />
<language load="yes" name="Kurdish" />
<language load="yes" name="Urdu" />
..
```

- Work best by turning off languages as a group
 - All Latin1 languages
 - CJK
 - Arabic scripted languages



Example (language scope option)

- Improve Indonesian accuracy by turning off Malay profile

Input size	Indonesian w/ Malay on	Indonesian w/ Malay off
8 bytes	40.23%	68.80%
16 bytes	52.20%	82.08%
32 bytes	65.72%	94.05%
64 bytes	69.15%	98.94%
128 bytes	76.04%	100.00%
256 bytes	77.08%	100.00%
512 bytes	87.50%	100.00%
1024 bytes	91.67%	100.00%
2048 bytes	100.00%	100.00%



On going improvements

- Improved accuracy
 - Short string
 - Indonesian vs. Malay
 - *enhanced profile uniqueness*
 - English misidentification
 - *profile cleansing*
- Performance
 - Maximizing CPU cache hits
 - *RLI 5.0. is 4 times faster than RLI 4.0*



What's coming next

- New encodings
 - GB18030 (GB2312 superset)
 - HKSCS (Big5 superset)
 - EUC-JISX0213
- New languages
 - Indian Languages:
 - Hindi*
 - Kannada*
 - Bengali*
 - Gujarat*
 - Tamil*



Questions / Demo



More information:
<http://www.basistech.com>
productinquiries@basistech.com