Language Translation for File Paths

Neil C. Rowe, Riqui Schwamm,

and Simson L. Garfinkel

Computer Science Department

U.S. Naval Postgraduate School
Monterey, California, USA

ncrowe@nps.edu

August 5, 2013

Why translate file paths?

- 3.6% of the language of paths in our Real Data Corpus (non-US) is not English or computer terms, ignoring punctuation and digits.
- Much of this non-English language is important for investigators as it represents user-created files.
- The language of the file name is often the language of the file.
- Translation of file names need not always be perfect since preliminary investigations only need to decide file relevance.
- Translation of everything first to English is the easiest.

Machine translation has a long history

- Automated translation dates from the 1960s.
- One approach is to store many translation cases.
  - This is done by Systran.
- Another approach is to learn statistical associations between words and phrases in matched corpora (copies of the same document in different languages).
  - This is done by Google Translate.
- 60% accuracy for automatic sentence translation is typical for European languages.
- However, file paths are shorter than sentences and translation of them should be more accurate.
- No one has focused specifically on file-path translation.
- A side problem is identifying the language of the text to translate, for which character-bigram statistics are traditionally used.

Obstacles to path translation

- Sending a whole path to a translator errs on some interpolated English words when English should be echoed.
  - Systran translated “Temporary Internet Files” on a Mexican drive as “Temporary Internet You Case Out”.
- 23.1% of our paths changed language at least twice. We must translate each directory segment separately.
  - Example : Documents and Settings/defaultuser/Mes documents/Ma musique/Desktop.ini.
- There often aren’t enough character bigrams in a path to adequately guess the language.
  - Tool LA-Strings thought "obj viewsspt viewssrc vs lk“ was most likely Latvian.
- Country of origin is not always a good predictor of the language – e.g. Chinese is all over the world.

Testbed: The Real Drive Corpus (RDC)

- Currently 3537 drives, with more arriving continuously, 92.1 gigabytes of metadata (in DFXML format)
- 1202 empty or unreadable, 1682 Windows operating systems, 47 MS-DOS, 19 Macintosh and Linux, 389 storage devices, 368 devices with cameras, 89 other mobile devices
- From 32 countries (only U.S. material is from our own group); largest number from India, with good representation of Israel, China, Singapore, Mexico, and Palestine
- Currently 94.5 million files, 48.9 million distinct paths; 20.7 million match NSRL hash values, 1.8 match hashsets.com hash values.

Our approach

1. Use SleuthKit/Fiwalk, convert to UTF-8.
2. Exclude paths without a word of at least three characters that is not known English or a computer code like “jpeg”.
3. Collect the words for each remaining directory over the corpus.
4. Infer the most likely language of these directories using five clues.
5. Infer the most likely language of each remaining directory segment of each path using three clues.
6. Translate segments using Systran, Google Translate, or word-for-word dictionary substitution.
7. Insert translated words into appropriate paths using analogous punctuation and case.
8. Put translated path into DFXML metadata with new tag <englishfilename>.

Example file path translations we produced

Applications/Microsoft Office X/Office/Assistenten-Vorlagen/Kataloge/Kapsel

was translated to:

Applications/Microsoft Office X/Office/Assistants-Were-present/Catalogs/Cap

top.com/تصميماتي/السلسلة المعلوماتية.jpg

was translated to:

top.com/My designs/The computer-based series.jpg

Note analogous punctuation and case to the originals.

Sources of dictionary/translation information

34 languages currently handled, 1.2 million words

- English wordlists (currently 403,000 words) from several online sources
- Wikipedia: Good for everyday words (but most one-letter and two-letter words excluded to handle code-like names like “ab8e6rs”)
- Google Translate output of the 32,015 English words occurring at least 10 times in the corpus (except when identical): Good for technical words
- Transliterations of 18 European languages
- Manual entry of common computer abbreviations
- Automated splitting of compound words: Both to recognize the language and get the translation

Coverage of major languages

Language	Wikitionary words	Translated common English words	Other dictionary sources	Systran translation	Google Translate
English (en)	-	-	269,205	-	-
Spanish (es)	74,978	24,101	2,080	X	X
French (fr)	83,654	20,096	491	X	X
German (de)	77,863	22,166	1,110	X	X
Dutch (nl)	56,881	22,200	50	X	X
Swedish (sv)	42,943	21,056	-	X	X
Finnish (fi)	93,867	24,716	6		X
Russian (ru)	91,394	30,665	-	X	X
Romanian (ro)	24,561	22,654	-		X
Greek (el)	33,081	26,805	127	X	X
Hebrew (he)	13,438	28,894	9,980		X
Arabic (ar)	19,479	30,218	-	X	X
Farsi (fa)	11,131	27,837	-		X
Urdu (ur)	-	20,209	-		X
Hindi (hi)	8,190	25,973	-	X	X
Thai (th)	22,213	26,346	-		X
Chinese (zh)	40,207	61,320	-	X	X
Korean (ko)	17,889	27,808	-	X	X
Japanese (ja)	39,532	30,337	-	X	X
Hausa (ha)	-	-	4,966		

Automatically finding compound words

- Automated analysis found 185,248 potential compound words to check, in the unknown words of the corpus by splitting them.
- To reduce false alarms, splits had to involve words of at least four characters (except for Chinese), where both were known words of the same language.
- English examples: arabportal, mainparts, cityhospital, seatdisplay.
- Recognizing foreign-language compounds permits automated inference of a translation.
- Examples: horadormir -> hour sleep, ventadirecta -> sale direct, producktregistierung -> manufacture registration, weichzeichnen -> flexible chart.

We must address transliteration

- Many users attempt to do their languages on a U.S. keyboard.
- This means they transliterate characters.
- The mapping is straightforward for European characters, but more complex otherwise.
- We create transliterated dictionaries to match with words in file paths, for the 18 most unproblematic languages.
- This is particularly helpful for Spanish and French.
- It didn’t work well for Arabic, which has many transliteration ambiguities.

Aggregation of directory words

- Directory: WINNT/Profiles/adrian/Menú Inicio/ Programas/ Accesorios/Multimedia on Mexican drives contained:
  - Control de volumen.lnk
  - Grabadora de sonidos.lnk
  - Reproductor de CD.lnk
  - Reproductor de medios.lnk
- Words extracted for this directory:

  control de volumen grabadora de sonidos reproductor de cd reproductor de medios
- All Ascii. But 11/12 words are in a Spanish dictionary, 2/12 are in an English dictionary, 3/12 are in an computer-term dictionary.
- So guess this directory is Spanish.
- Weight a language by inverse of log of size of its word list (following Zipf’s Law).

Character distributions (unigrams)

- We compute conditional probabilities of a language given its character based on the dictionaries. E.g: “a” with umlaut has probability 0.54 for Finnish, 0.30 for Swedish, 0.11 for German, 0.05 for other languages.
- Weight of a language: $\exp[(1/M)\sum_{i=1}^{M}\ln(\max(p_{i,L},c_{i,L})]$

  ranging over given words where p is the conditional probability and c is a lower bound for previously-unseen characters.
- We also assign characters to one of 20 categories by Unicode codepoint numeric range.
  - This enables us to assign never-seen characters to categories.
  - It also permits statistics on the categories for each language. This gives another way to identify the language.

Other methods to identify the language were tested

- LA-Strings: A character-bigram tool.
- Character type: 20 broad classes of characters.
- Country of origin: We used a standard table of language percentages by country.
- Keywords in the path: Certain words indicate language encodings, like standard abbreviations for languages.
- Inheritance from the languages of the directory above a given directory.

Combining the language clues

- Combining clues for a directory language L:

$$c_d w_{L,dictionary} + c_c w_{L,characters} + c_o w_{L,country} + c_k w_{L,keywords} + c_l w_{L,LA-Strings}$$

Justification: Clues may be missing, so situation is disjunctive.

- Combining clues for a path segment for L:

$$w_{L,dictionary} w_{L,characters} \sqrt{w_{L,directory}}$$

Justification: All three clues must be strong for a good candidate, so situation is conjunctive.

Testing clues in directory language identification

Factors	Raw accuracy	Adjusted accuracy
All	0.721	0.904
All without character types and inheritance	0.798	0.934
All without LA-Strings	0.694	0.904
All without dictionary lookup	0.662	0.836
All without character distributions	0.703	0.898
All without country	0.722	0.886
All without path keywords	0.793	0.897
Just dictionary lookup	0.649	0.857
Just character distributions	0.359	0.775

"Adjusted accuracy" combines transliterated with untransliterated, ignores confusion of English with untranslatable, and weights misidentification of English by 1/3. Conclusion: Character types and inheritance do not provide useful clues, LA-Strings maybe, others yes.

Confusion matrix for directory-segment language identification on random sample of 3518 directories

Overall adjusted accuracy was 93.7%. We got 93.5% on a different random sample of 29 million new drives, so there was little training bias. “t-” means transliterated. Rows denote true language.

	ar	de	en	es	fr	he	it	ja	ko	nl	ru	tr	zh	t-de	t-es	t-fr	t-he	t-hi	t-it	oth	un
ar	799		2																	8	3
de		14	3											36							10
en			301											2	4	1			1	3	273
es			4	107										2	370					11	85
fr			2		25											9					2
he						179															1
it			1				9								1						
ja								6													1
ko			1						17												1
nl			1							3											1
ru											2										
tr												17								8	6
zh			1										67								1
t-ar			2																	1	8
t-de														2							
t-es				1											80						1
t-fr																8					1
t-he																					1
t-hi			1											1						8	3
t-it			1												1				3		
oth	7		1									2								9	5
un			7												3	2	2			8	952

Translation methods tested

- Systran: Implemented using standalone code rather than as a service (though this involved difficult negotiations).
- Google Translate: Runs only as a service. We manually entered words and manually copied results. This approach must tediously address Google’s word limit.
- Word-for-word translation: Look up the words in our translation dictionary and append the results together in order.

Testing of path-segment translation

We used 200 examples each and judged results ourselves.
Conclusion: Google Translate is significantly better than the others.

Language / Measure	Spanish	French	Japanese
Word-for-word OK	.72	.74	.57
Systran OK	.65	.61	.75
Google Translate OK	.81	.80	.92
None OK	.07	.03	.04
Word-for-word best	.55	.65	.48
Systran best	.52	.55	.48
Google Translate best	.78	.75	.85

Example translations

Guessed language: words for translation	Word-for-word translation	Systran translation	Google Translate translation
Spanish: entren ser lider	come be head	they enter to be leader	come to be leader
Polish: magazyn kratownica	repository truss	warehouse grate	storage grid
Japanese: デスクトップの 表示	desktop display	Indication of desktop	Show Desktop
Arabic: مشكلة سقوط السارية	problem downfall applicable	Shaper of falling contagious	Problem of the fall of the applicable
Chinese: 陆行鸟饲 å x 手 x e x c	陆 行 鸟 饲 å x hand x e x c	Goes by land the bird to raise å x x e x c	The land line Torikai å x hand x e x c
French: premierbaiser pps	first kiss pps	premierbaiser pps	premierbaiser pps
French: tetes de vainqueurs pps	heads of winners pps	suck winners ps	heads of winners pps

File segments having a given number of words

Number of words	Percentage of translatable file segments having that number of words
1	20.5%
2	26.1%
3	17.9%
4	11.1%
5	7.4%
6	4.0%
7	2.4%
8	1.9%
>8	8.7%

So 46.6% of all translatable file segments are one-word or two-word, and their translations could be provided by dictionary lookup.
Also note: The fraction of translatable was not significantly different for user file paths.

Languages inferred per country

These came from 50 million files. Country is row, language is column.

	ar	de	en	es	fr	he	it	ja	ko	nl	pt	ru	sv	tr	zh	other	un
ae	28	105	40,560	69	5,879		20		16	5	22		24	38	353	88	144,218
bd		2	7,300	1	5		2		1,080		2		13	2	32	29	44,376
ca		3	7,389		6						1		25	4	70	6	38,134
cn		143	89,935	86	75		57	972	7	34	29	12	105	14	5,350	172	257,991
de		3,460	10,545	82	13	2	28		4	1	21		19	17	21	98	56,689
eg	655	19	3,687	17	18		5				2		13	2		61	28,001
gh		5	25,569	6	9	2	6		577		2		8	2	13	33	63,416
il	5	146	117,441	838	160	23,873	59		2	23	71		137	25	4,841	277	657,282
in	44	777	210,764	1,192	707	7	642	2		432	316	2	180	806	52	1,029	703,589
ma	1	5	1,226	7	258		8						7	1	1	7	16,578
mx		58	44,889	97,831	78	4	129		3	6	127		30	10	197	181	330,763
pk		1	5,730		3		14		90		2		1	1	6	17	31,091
ps	649	13	63,136	36	31	12	20			3	6		41	10		73	210,933
sg	0	89	82,766	42	14	9	10	13			9		10	6	59	40	204,875
tr	0	40	18,478	20	17	2	4			4	7		8	4,290		78	55,753
?	126,190	5,040	432,141	65,885	474	50	511	146	34	209	892	27	603	563	906	4,649	1,013,389

Conclusions

- Translation of file paths in harder that it seems.
- Success in translation requires handling each segment of a path separately.
- Success in language identification requires aggregating words from the same directory over a corpus.
- Surprisingly, character bigrams didn't help much. But dictionary lookup and unigrams did help.
- On translation quality, Google Translate was clearly the best. Systran performance was equalled by a simple word-for-word translation substitution.
- To translate to other languages, first translate to English and then to the target language.
- We will make freely available a Python package that does path translation using word-for-word substitution (input: DFXML metadata files).