Digital Investigation 10 (2013) S78–S86

Contents lists available at SciVerse ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Language translation for file paths

Neil C. Rowe*, Riqui Schwamm, Simson L. Garfinkel

U.S. Naval Postgraduate School, Monterey, CA 93943, USA

Keywords:
Digital forensics
File paths
Machine translation
Dictionary
Character distribution
Unicode
Naive Bayes inference

A B S T R A C T

Forensic examiners are frequently confronted with content in languages that they do not understand, and they could benefit from machine translation into their native language. But automated translation of file paths is a difficult problem because of the minimal context for translation and the frequent mixing of multiple languages within a path. This work developed a prototype implementation of a file-path translator that first identifies the language for each directory segment of a path, and then translates to English those that are not already English nor artificial words. Brown's LA-Strings utility for language identification was tried, but its performance was found inadequate on short strings and it was supplemented with clues from dictionary lookup, Unicode character distributions for languages, country of origin, and language-related keywords. To provide better data for language inference, words used in each directory over a large corpus were aggregated for analysis. The resulting directory-language probabilities were combined with those for each path segment from dictionary lookup and character-type distributions to infer the segment's most likely language. Tests were done on a corpus of 50.1 million file paths looking for 35 different languages. Tests showed 90.4% accuracy on identifying languages of directories and 93.7% accuracy on identifying languages of directory/file segments of file paths, even after excluding 44.4% of the paths as obviously English or untranslatable. Two of seven proposed language clues were shown to impair directory-language identification. Experiments also compared three translation methods: the Systran translation tool, Google Translate, and word-for-word substitution using dictionaries. Google Translate usually performed the best, but all still made errors with European languages and a significant number of errors with Arabic and Chinese.

Published by Elsevier Ltd.

1. Introduction

Forensic examiners increasingly work with materials in unfamiliar human languages. Although some examiners have human linguists available to translate text, audio, and video into their native tongues, most do not. Human translation is expensive and not always timely (U.S. FBI, 2004). Automated translation of directories, file names, and other metadata could be a useful first step in an investigation. Computer users rely on named directories to organize their information, and they give their files descriptive names. Translation of them could enable recognizing similar activities or otherwise interesting behavior taking place in different linguistic parts of the world, and aid cross-language clustering of files.

This issue is important with our research data, the Real Drive Corpus. It currently contains more than 3000 drive images from 28 wide-ranging countries. It contains a wide range of languages, and not just the languages one would expect. For instance, drives from Israel contained significant amounts of Spanish and Chinese, while drives from United Arab Emirates contained significant amounts of French.

Machine translation of forensic file paths need not be perfect to be useful since many investigations focus on keyword lookup. In fact, merely identifying the languages

* Corresponding author.
E-mail address: ncrowe@nps.edu (N.C. Rowe).

1742-2876/$ – see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.diin.2013.06.009

of a file name without translating the words may help an investigation, as file contents are almost always in the same language as a file name, and file-name analysis could suggest what translators to call for the contents.

This work only addresses translating to English. This is the easiest target language since many operating-system file and directory names are in English. Nonetheless, there are many challenges with the remaining words.

1.1. Prior work

Prior work shows that drives can be characterized in many important ways by their metadata alone (Rowe and Garfinkel, 2012), but file paths in an unintelligible language can still impede investigation.

Machine translation has a long history (Wilks, 2009). The major approaches are case-based reasoning as in the Systran system, and statistical inference as in IBM's Candide system of the 1990s and its many descendants including Google Translate. Current systems are far from perfect; a figure of 50% accuracy on prose is often cited. But file paths use a limited language and translation success rates could be higher. Good success has been shown for instance for the constrained domain of news stories (Turchi et al., 2012).

Language identification is a key subproblem. Most prior work on it has focused on N-grams as in (Mishra et al., 2010) and LA-Strings (Brown, 2012), though word clues (Yang and Liang, 2010) and other mathematical techniques (Da Silva and Lopes, 2006) have been used. Language-detection products include Google's Compact Language Detector (McCandless, 2011), and Shuyo's Language Detection Library for Java (Shuyo, 2010).

Basis Technology has demonstrated the Odyssey Digital Forensics Search system (Basis Technology, 2013) which combines the company's multilingual named entity extraction technology with a search capability allowing a user to enter English words and search for their foreign-language equivalents, but it neither translates nor transliterates paths.

2. Making sense of mixed-language paths

Our approach is to obtain paths by first using SleuthKit to extract drive images, and then Fiwalk (now part of SleuthKit) to extract file metadata including file paths. SleuthKit's tsk_fs_dir_walk reads UCS-2 file names stored in the FAT32 and NTFS directory entries, and recodes them as UTF-8 sequences. FAT12, FAT16 and FAT32 file systems use OEM character sets and Code Pages to store short file names which SleuthKit uses to convert to UTF-8 Unicode. In a significant number of cases, SleuthKit did not produce either valid Unicode code points or the shortest possible encoding. We checked SleuthKit file names character-by-character and replaced invalid bytes with Python-style escape sequences. For example, the UCS-2 byte sequence "0xFF 0xFF" would be encoded as "\xFF\xFF", as U+FFFF is not a valid Unicode character.

Multiple tools for language handling are needed because the problem of translating file paths is difficult. Most translation tools are designed for large blocks of prose and take advantage of punctuation and syntactic rules. File paths use considerable but nontraditional punctuation, use little conventional syntax, and use frequent abbreviations and code words. Context is important in translation (Larson, 1984), so a big challenge is understanding the context of a word, but this is frequently unclear in paths. A path often contains different kinds of information in different places, and when multiple languages are used, they are rarely consistent through the path. 30.4% of the file paths in our corpus change language once in their sequence, and 23.1% change at least twice, ignoring untranslatable words. Consider these examples from our corpus:

- Documents and Settings/defaultuser/Mes documents/Ma musique/Desktop.ini
- Mis Documentos/SalvadorJP/Excel/GRUPOS.xls
- Documents and Settings/3742008/Configuración local/Datos de programa/Microsoft/Internet Explorer/.
- human/animation/weapon_pistol/major_pain/몄□윞□킙욶궲A□□□/pistol_pain_crawldeath.skc

2.1. Collecting word sequences for translation

The Systran and Google Translate software return input words unchanged if they cannot translate them. Thus our first attempt was to send Systran the entire path when languages are mixed, since most words in our corpus were English. This ran into trouble with words having different meanings in different languages. For instance, Systran translates "Temporary Internet Files" occurring in a Mexican file path into English as "Temporary Internet you case out" because "files" is the present subjunctive second person of the verb "filar" meaning "case out". Here the fact that these are all known English words should overrule an attempt to translate the phrase from Spanish. "Temporary Internet Files" appears frequently on our Mexican drives, and "files" is the most common English word in our corpus, so translating paths as a whole does not work well.

So it is important to handle each directory and file name in a path separately. We thus first segment directories, then segment at each punctuation mark or digit. The current Unicode specifications list 1216 punctuation marks (Unicode, 2013). The few words that contain punctuation (like "tutti-frutti") can be ignored without causing much trouble, except for apostrophe-s in English which is handled separately. The remaining characters can be translated and inserted back into the path to get a translated path.

2.2. Directory word aggregation

In our tests the LA-Strings language identifier demonstrated mediocre performance on short strings, apparently as a result of its inability to extract sufficient statistics. For instance, for the software terms "obj viewsspt viewssrc vs lk", LA-Strings thought the most likely language was Latvian, and for "cmap enutxt", Southern Dong. This problem could be reduced by aggregating the words of each directory over the corpus. In this it was important to keep separate word lists by country of origin of each drive, since