Detection of Foreign Words and Names in Written Text

By

Bashir U Ahmed

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing

at

School of Computer Science and Information Systems
Pace University

May 2005

UMI Number: 3172339

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Dissertation Signature (Approval) Page

We hereby certify that this dissertation, submitted by Bashir U. Ahmed, satisfies the dissertation requirements for the degree of Doctor of Professional Studies and has been approved.

Dr. Charles Tappert Chairperson of Dissertation Committee	10 May 2005
Dr. Fred Grossman Dissertation Committee Member	10 May 2005
Dr. Sung-Hyuk Cha Dissertation Committee Member	10 May 2005

School of Computer Science and Information Systems
Pace University 2005

Abstract

Detection of Foreign Words and Names in Written Text

By

Bashir U Ahmed

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Professional Studies
in Computing
at
School of Computer Science and Information Systems
Pace University

May 2005

Tremendous research effort has gone into the field of natural language processing and understanding during the last half of the 20th century, yet it remains as one of the most challenging problems. Nevertheless, many companies have commercialized language processing applications such as Text-to-Speech (TTS) systems, domain-specific Automatic Speech Recognition (ASR) systems, Machine Translation (MT) systems, and limited-domain, speaker-dependent Speech-to-Speech (STS) prototype systems. While the quality of all natural language processing applications has improved steadily, they remain far from being perfect.

Due to globalization and the widespread use of the Internet, occurrences of foreign entities – foreign words, names, locations, and events – in written text are becoming commonplace. This further complicates the automatic processing of natural language text. Identification of such foreign entities is important for several reasons. For example, TTS systems will usually fail to pronounce such entities properly in their native language as they try to sound them using the lexicon of the base language. These foreign entities also cause problems for MT programs where translation is initially done word by word.

This dissertation simplifies the Naïve Bayesian, N-gram based text-classification algorithm by taking its logarithm, naming it the Cumulative Frequency Addition (CFA) algorithm, and applies it to three text processing tasks: language identification, name-to-nationality identification, and detection of foreign words and names. In the language identification task CFA yields 100% accuracy on string sizes greater than 150 characters. In the name-to-nationality task, it yields 86% accuracy on a 14 country database and 96% on a 7 country database within the top three choices. Identifying the nationality, or at least the language group, of a person from his/her name, can be important not only for proper pronunciation but also for purposes of forensics studies and national security. Finally, in the task of detecting foreign words we obtain 66.9% accuracy. This is the first study to apply natural language processing techniques to the latter two tasks.

Acknowledgments

I am a dreamer and I dreamt really big for my thesis topic wanting to develop a Universal Language Translator. I must thank my advisors for bringing me back to reality. First, I would like to thank Dr. Charles Tappert, my thesis committee chairman, who has diligently invited scholars and industry experts in the latest technological fields, which has further sparked my interest in natural language processing. I also thank Dr. Fred Grossman for the reality check he provided, as his constant and justifiable rejections of my initially unachievable topics helped shape my work into something that could be accomplished. I must also thank Dr. Sung-Hyuk Cha for helping me shape up the thesis idea and also providing the initial set of training data.

As it took great advice from several great advisors to finish the intellectual work, it took similar amount of support and encouragement at the home front to get through the wee hours at night. I am thankful to my seven years old daughter, Serina, for being so understanding while working on my thesis chapters and always reminding me that I was the best Dad in the "whole wide world". My three years old daughter, Samina, provided ultimate companionship 12 past midnight rolling over my back while I was working on my chapters lying down on the bed. I surely will miss those moments and cherish them the rest of my life. The arrival of my son, Aynis, during the third year of my doctoral research could have been a show breaker, but it did not and it became a joy because of one person – my wife. I could not thank my wife, Maria, enough for being so understanding and always wanting to see me finish the thesis. Maria's utmost confidence, "You can do this" and unselfish support helped me making my dream a reality.

My dream of going for a doctoral program began a long time ago – my grandfather seeded the dream when he promised my mother that he would finish his Ph. D. with my elder brother. My grandfather passed away without fulfilling his dream. I dared to take this challenge because I knew I would have all the needed support. I am thankful to God for giving me so much in life – my wife's determination to see me finish the program, and my brother (my toughest competition inspired by sibling rivalry), Nasir, whose unwavering mental support boosted my confidence. I am ever so grateful to my mother, Mazeda Akter, for raising me so well and believing in me so much. I am thankful to my uncle, Mohammad Z. Haque, for sponsoring us to come to America and for being so supportive in every step of the way. I would like to thank my mother & father-in-law for believing and giving me the confidence I needed to finish. My gratitude goes to all of my brothers & sisters and their husbands and wives for always giving encouragement and helping me get through this program. I could not have finished this work if I did not have the friendship and mental support from my fellow classmates, especially Melanie Johnson and Barbara Edington for being such a good listener, Steven Golikov and Borming Chiang for providing technical support when needed.

Finally, I thank Pace University for developing this DPS program and making my dream a reality.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1. Overview

Due to the vast variability in characters, pronunciations, accent and letter-to-sound rules, natural language processing (NLP) task is viewed as incredibly difficult [5]. There are 82 distinct languages listed as the world's major spoken languages, each with its own set of linguistic rules and each spoken by more than 10 million primary or alternative speakers [63]. In recent years, much progress has been made in NLP research largely due to the rapid growth in computing power. Research effort in NLP in both academic and commercial settings has increased greatly due to globalization and the need to communicate internationally. This resulted in many commercial speech recognition, language identification, language translation, and text-to-speech systems. This trend has created awareness of language processing (Speech recognition, language translation and text-to-speech) as a viable commercial tool.

Globalization is manifested in e-mails, corporate documents, and newspaper articles composed in mixed languages. Investigation of newspapers articles by Pfister et.al. [50] and our own, found numerous inclusion of English text in written German, Tagalog and Swedish newspapers. Our own investigation of internal e-mails from a German company

with sister location in the United States, found numerous German inclusions in English e-mails. This is an unavoidable reality because when a native German speaker composes an e-mail in English, some German words get included without the author's conscious knowledge. All our findings confirm that the phenomenon of mixed-linguality occurs more frequently than one would normally realize. This places a great difficulty on the text-to-speech, Machine Translation and Information Retrieval systems.

Natural sounding text-to-speech systems require accurate phonological, prosodic, morphological and syntactic knowledge, which are language specific [70]. World languages differ widely from each other on these properties. Thus, to be able to read a mixed lingual text, a text-to-speech system must know the identity and context of each word in the text before pronouncing them accurately. Language identification, which can be done using either voice input or text input, is the pre-requisite of any such system. In this study, a language identification system was implemented and applied to solve the foreign word detection problem.

Corporate America is constantly looking to increase its bottom line by finding ways to increase productivity, and many corporations are trying to leverage Automatic Speech Recognition and text-to-speech technology for customer service. Another area where text-to-speech is being used is accessibility applications – many government organizations, such as IRS in the USA and Australian Government, started to use ASR/text-to-speech to train employees who are visually impaired or have typing disabilities. Many e-mail and voice mail providers, such as Verizon and SBC

communications, provide voice enabled e-mail options where users can check their e-mail via telephone or check their voice mail from their computer. Both the voice mail and e-mail contents are read by text-to-speech software. However, as text-to-speech reading is highly language dependent, a monolingual text-to-speech system fails to provide natural sounding reading of user's voice mail or e-mail that are embedded with foreign words and names. In most cases, embedded foreign words are read in garbled manner [51]. An automatic detection and tagging of foreign words in written text can play an important role in the quality of a text-to-speech system in which it acts as the communicator to the end user. For example, when a text-to-speech system is turned on in English mode and it encounters a German word, if the system can detect that the next word is German, it can then automatically switch to German lexica and pronounce the word naturally as a normal human reader would do. This could increase the user acceptance of text-to-speech systems and make commercial application more appealing. The goal of this research is to provide a mechanism to tag the foreign words, so that a text-to-speech system will have the option to switch language when appropriate. In addition to text-to-speech, Machine Translation and Information Retrieval applications will also benefit from this research.

1.1.1. Implications of Automatic Processing of Natural Language Text

Automatic processing of natural language text has huge implications in commercial applications. For example, a document service provider may encounter numerous electronic requests in varieties of languages and sorting those requests manually would be extremely inefficient or impossible. Using automatic language identification, the service provider can build different request queues sorted by languages and then forward these

queues to appropriate sub-contractors such as a sister company in another country where people understand the language. This arrangement would allow a document service provider to become a truly global company. Automatic processing of natural language text is even more important when the input needs to go through some sort of transformation along the process such as text to speech and vice-versa. In order to implement a Voice Enabled Universal Language Translator, which does not exist yet, given any text or voice sample, the translator should be able to speak it in any other human language of the world. One approach for such a system could be the translation of the source language text to the target language text, and then read the output using text-to-speech. If the source text is in pure form, this process may work very well. However, if the source text is embedded with words from a third language, then the translation will be incorrect if those foreign words are not recognized and addressed properly. In this thesis, an algorithm to tag foreign words and names in native text was described. This was done first by identifying the base language from unknown text using CFA, and then applying a new algorithm to detect any foreign words that do not belong to the base language. Findings from this research would enable easier implementation of true Polyglot text-to-speech and Machine Translation systems where any free form source text can be read or translated accurately without the pre-existing knowledge of the source text.

1.1.2. Difference Among Languages

While all the Latin character based languages share a great common part – most of the 26 letters of the English alphabets are found in all western European languages- they vary to some extent with additional characters such as the accented characters and also in text

patterns. However, the differences become very pronounced when one talks about the spoken form of these languages. In spoken form, we can guess the language by the phonemes. A phoneme is the term defined by linguists to classify speech into a number of abstract categories for grouping together subsets of speech sound. For example, American English has about 40 phonemes. Even though no two speech sounds, or phones, are identical, all of the phones classified into one phoneme category are similar enough so that they convey the same meaning [39]. But there is much overlap of the phoneme sets, and there can be differences in the way the same phoneme is interpreted in two different languages. Each language has its own letter-to-sound rules [68]. In English, letters /l/ and /r/ are two different phonemes. Then the frequency of occurrence of phones and the phonotactic rules in languages can also differ significantly. For example, phoneme clusters /sr/ and /sp/ are quite common in Tamil and German, but not in English. Prosodic features, rhythm, and intonation, also vary among languages [15]. English is known to be stress-timed, and French is known to be syllable-timed. All these unique features are the subject of research for automatic identification of languages by machine.

1.2. The Problem

The problem this thesis addresses falls into the area of natural language processing and understanding. The title of the thesis “Detection of foreign words and names in written text” summarizes the problem accurately. Existing Text-to-speech and machine translation program performs well when processing monolingual text (text written in only one language). However, these systems perform poorly when the input text contains

words and names from foreign languages that cannot be found in the base dictionary of the main text. Therefore, this thesis investigates the problem of foreign words detection in written text.

1.2.1. Precise Problem Definition

Input: a text string – a sequence of words from one or more languages.

Output: the input text string with the language of each word identified.

Assumptions: There is a clearly dominant or main language of the input text string, and we call it the base language. Although this does not cover all possibilities, this is the usual case for this problem. Words not belonging to the base language are considered "Foreign" and identified.

Example: Given a test string containing 9 English words, 4 German words, and 3 French words. According to our definition, English would be considered the dominant language while German and French would be considered foreign languages. Given the input described above, the target output we tried to achieve is a tagged string as follows

Input:

"Sentence artificially constructed for der Zweck der Demonstration, la traduction pas be accurate, please pay attention"

Output:

<Eng> Sentence artificially constructed for <Ger> der Zweck der Demonstration </Ger> <Fre> la traduction pas. </Fre> be accurate, please pay attention </Eng>

The idea is to tag the overall sentence as English but containing German and French words as foreign words. This means there will be a beginning <Eng> tag and an ending </Eng> tag plus the beginning and ending tags for other languages inside the body of the output. Any word or sentence fragment inside the body of the output that is not explicitly tagged would be assumed to be English. This would allow automatic natural language processing application like Text-to-speech or machine translation programs to start processing in English mode and then switch to the German or French only when needed. With this approach, any common words – words that can be found in multiple languages – would default to the base language.

The problem described above is a sub-problem of a more general one - detection of language shift in written text where no notion of native or foreign is considered. Under this general scenario a text can contain any number of words from any number of languages and the challenge is to tag each word or sentence fragments containing contiguous set of words that belong to each language. A simplified version of this general problem, the one that is most common, was implemented by identifying the base language of the majority of the text, and then identifying the foreign words relative to that language.

In addition to the foreign words, an extension of this work involves the identification of the nationality or language group of proper names. This is relevant because written text often contains proper names, which require special processing. Knowing the nationality or language of a proper name can be helpful for TTS and machine translation programs.

1.2.2. Precise Problem Definition of the Extension Work

Input: A proper name such as person, organization, event, etc., in written form.

Output: The proper name identified as belonging to a country and subsequently associating the country to a language.

Assumption: The name has been tagged as a name, and then CFA was used to identify the nationality or at least the language group of the name. Note that the focus here was not on the name delimitation task, but on the identification of the nationality of the names. In the absence of a name recognition algorithm, however all sequential words that start with capital letters were assumed to be names, except for the initial words of each sentence if not immediately followed by words with capital letters.

Example: Given a sentence such as "My advisor is Sung-Hyuk Cha, and he is very technical," according to the problem definition, Sung-Hyuk Cha will be identified as a name and will be tagged as Korean. The word "My" would not be considered as a name.

1.3. State-of-the-Art of Text-To-Speech

1.3.1. Text-To-Speech: What For?

Text-to-speech is the creation of audible speech from computer readable text. The qualities of some commercial Mono/Multi Lingual Text-to-speech systems improved a great deal and they are being used in different critical applications. Red Planet, the Warner Bros. movie released on 11/09/2000, used AT&T's Text-to-speech as the voices of the space ship (female) and the astronauts' space suits (male). The movie "I am Sam"

starring Sean Penn and Michelle Pfeiffer, used AT&T's "Crystal" as the voice of Pfeiffer's voice-activated cell phone. In addition to its use in the movie industry, Text-to-speech systems are being investigated in the following areas [24]: Telecommunication Services, Language Education, Aid to handicapped persons, Talking books and toys, Vocal Monitoring, Multimedia, man-machine communication, and Fundamental and applied research.

1.3.2. How Current Text-To-Speech Read Text?

Synthesized speech is both a triumph of technology and the fruition of a very old dream. The first "acoustic-mechanical speech machine" was introduced in 1791 by the Viennese researcher Wolfgang von Kempelen. The machine simulated the major consonant and vowel sounds with an array of vibrating reeds, like a musical instrument. But not until the advent of electronics did machines truly begin to mimic human voices [8]. In the 1950s, researchers labored to model the acoustics of the human vocal tract and the resonant frequencies, or *formants*, it generates. This approach eventually led to workable but robotic results—certainly nothing a public-relations person would call customer ready. Stephen Hawking's voice synthesizer is the most famous example. Such a voice is not good enough for normal usage.

In the 1970s, researchers at what was then Bell Labs turned to a "concatenative" approach: Instead of trying to generate a human voice from scratch, they would start with an existing voice—several hours' worth of standard English sentences spoken by a clear-voiced person—and design a computer program to splice and re-splice it to say whatever words they wanted said.

The computer program first parsed the prerecorded sentences into consonant and vowel sounds, called phonemes—perhaps 50 or 60 in the early iterations. Then the phonemes were reassembled to form new words. The recorded word *cat,* for instance, could be deconstructed into the phonemes *k, ae,* and *t,* which could then be rearranged to form *tack.* It worked, and it was a definite improvement over robot-speak, but it wasn't any closer to sound of Peter Jennings or whatever Journalist you like. Fifty-odd phonemes simply couldn't capture the subtle intonations of spoken language.

In the mid-1990s, armed with a new generation of supercomputers, AT&T researchers began amassing a vast digital "voice warehouse" of phonemes. Instead of one t sound for the computer program to choose from, there might be 10,000. By having so many sounds, it offers a little more spontaneity. By parsing phonemes into "half-phones" it is possible to offer subtler possibilities for recombination. Voice synthesis now entails properly labeling the half-phones—10,000 versions of the "t_1" sound, 10,000 versions of the "t_2" sound, and so on—then creating a computer algorithm to smoothly string them into words and sentences. By assembling a simple word like *cat* from its half-phones—("k_1, k_2, a_1, a_2, t_1, t_2")—involves billions of combinatorial decisions and presents a massive computer-processing problem [8].

Allistar Conkie of AT&T is generally credited with devising a workable solution, now known as unit-selection synthesis [8]. His solution was to assign "costs" to the innumerable choices and combinations of half-phones. Charting the "least expensive"

path through the chorus of half-phones became simply a math problem for the computer to work out. Most costs crop up where two half-phones meet and attempt to join. The computer can measure the pitch, loudness, and duration (in milliseconds) of each one and compare them. If the total energies of each are vastly different, linking them would produce a disagreeable click or pop, so the link is rated as "expensive," and the computer avoids it. Some linkages are far less likely to occur than others, as in real spoken English, certain "k_2" sounds are almost never followed by certain "a_1" sounds. Those links could be deemed costly, too, and the computer could avoid them altogether. The word cat could theoretically call upon 10,000 ways of linking the "k_2" and "a_1" sounds. In practice, though, fewer than 100—a manageable number of choices for the computer to handle—can pass as reasonable facsimiles of human sounds.

There are many other niggling problems to deal with in Text-to-speech, such as how to teach the speaking computer to distinguish between written words like *bow* (as in "bow and arrow") and *bow* (as in the bow of a ship), or to recognize that minus signs aren't the same as hyphens [8]. This problem is complicated even further by the embedding of foreign words and names that require language specific letter-to-sounds rules for proper pronunciation.

Despite all the difficulties, speech synthesis and Text-to-speech technology has improved a lot. AT&T's Crystal, IBM's Via Voice, Dragon Naturally Speaking are some of the well known Text-to-speech software available in the commercial arena. Voices in Text-to-speech systems do not yet sound entirely natural. In short phrases ("I'd like to buy a ticket to Stockholm"), they can pass for a human, albeit an officious one. But longer