A Random Forests Text Transliteration System for Greek Digraphia

Alexandros Panteli and Manolis Maragoudakis
University of the Aegean, Samos, Greece
{icsd06136,mmarag}@aegean.gr

Abstract. Greeklish to Greek transcription does undeniably seem to be a challenging task since it cannot be accomplished by directly mapping each Greek character to a corresponding symbol of the Latin alphabet. The ambiguity in the human way of Greeklish writing, since Greeklish users do not follow a standardized way of transliteration makes the process of transcribing Greeklish back to Greek alphabet challenging. Even though a plethora of deterministic approaches for the task at hand exists, this paper presents a non-deterministic, vocabulary-free approach, which produces comparable and even better results, supports argot and other linguistic peculiarities, based on an ensemble classification methodology of Data Mining, namely Random Forests. Using data from real users from a conglomeration of resources such as Blogs, forums, email lists, etc., as well as artificial data from a robust stochastic Greek to Greeklish transcriber, the proposed approach depicts satisfactory outcomes in the range of 91.5%-98.5%, which is comparable to an alternative commercial approach.

Keywords: Greek Language, Transliteration, Data Mining, Random Forests, Non-Deterministic

1 Introduction

Greeklish is a term which originates from the words *Greek* and *English*, signifying a writing style in which Greek words are written using the Latin alphabet. Other synonyms for Greeklish are *Latinoellinika* or *ASCII Greek*. This phenomenon is not only appearing within the Greek domain, it is linguistically identified as *Digraphia* [1]. Digraphia is either synchronic, in the sense that two writing systems coexist for the same language, or diachronic, meaning that the writing system has changed over time and has finally been replaced by a new one. Examples of digraphia are common in a variety of languages that do not adopt the Latin alphabet or Latin script (e.g. Greek, Serbian, Colloquial Arabic, Chinese, Japanese etc.). Serbian is probably the most noticeable modern instance of synchronic digraphia, in which Serbian texts is found to be written concurrently in the Cyrillic script and in an adapted Latin-based one. *Singlish*, a word similar to Greeklish refers to an English-based creole used in Singapore which employs transliteration practices, as well as vocabulary modifications and additions from the English language. As a final point, we should mention the case of the Romanian Language, where there has been a full adoption of Latin-based writing style instead of the original Cyrillic one. The same principle is also appearing in the Turkish and other Central Asian countries of the former Soviet Union.

There is a significant amount of research papers that deal with the application and the acceptability of Greeklish as a writing style. The most representative amongst them was introduced by [2], in which a sequence of issues is studies such as the degree of penetration of Greeklish in textual resources, the acceptance rate of them, etc. More specifically, some descriptive statistical results mention that 60% of users have been reported to use Greeklish in over 75% of the contexts they submit. In addition, 82% of the users accept Greeklish as an electronic communication tool while 53% consider this style as non-appealing, 24% concern it as a violation or even vandalism of the Greek Language and 46% have reported to face difficulties in the reading of such texts. As regards to the latter, other research works study the reading time for the comprehension of words and sentences written in the Greek and Greeklish texts. The results indicate that the response time is lower when the text is written in Greek (657ms mean value) than when it is written using characters of the Latin alphabet (886ms mean value) [3].

Although an official prototype has been proposed (ELOT 743:1982 [4]) and already approved and used by the British council, the majority of users follow empirical styles of Greeklish styles, mainly categorized into four distinct groups:

- *phonetic transcription*. Each letter or combination of letters is mapped into an expression with similar acoustic form.
- *optical transliteration*. Each letter or combination of letter is mapped into an expression that optically resembles the former. For example the Greek letter θ is usually mapped into *8*, due to its optical similarity.
- *keyboard-mapping conversion*. in this group, many letters are mapped to Latin ones according to the QWERTY layout of the Greek keyboard. For example, θ is mapped to its corresponding key in the Greek/English keyboard which is *u*.

Additionally, Greeklish writing suffers from the presence of “iotacism”, a phenomenon which is characterized by the use of the Latin character “i” for the transliteration of the Greek symbol sets “ι”, “η”, “υ”, “ει”, and “οι” since they are all pronounced as “I” according to the SAMPA phonetic alphabet.

Based on the aforementioned issues, the present work is a Data Mining approach towards an efficient Greeklish-to-Greek transliteration tool, based on a state-of-the-art ensemble classification algorithm of Random Forests. This is, according to our knowledge, the first attempt to this domain in a non-deterministic manner. The use of Greeklish is now considered of an issue of high controversy and it is banned from numerous web sites and forums (e.g. Greek Translation Forum, Athens Wireless Metropolitan Network Forum, etc.). Therefore, a robust and affective transcriber is considered of high importance in order for users not to be excluded from web discussions and other social networking activities.

2 Previous works in Greek Digraphia

A lot of work has been done in the field of Greeklish-to-Greek conversion. The most representative approaches including E-Chaos [5], Greeklish Out [6], Greek to Greeklish by Innoetics [7], All Greek to me![8] and deGreeklish [9]. The first two approaches are not using a vocabulary and they are mainly based on manual rules, refined by the user and adjustable to include more in the future. The second implementation does not mention its scientific parameters as it is a commercial application, however, the company mention 98% using language models as the core mechanism. The third and fourth approaches are based on a more sophisticated methodology, namely Finite State Automata (FSA), which make the mapping of each letter more straightforward. The latter system is implemented as a web service in PHP and C++ and addresses a novel search strategy in the directed acyclic graph. Note that all of the above approaches use deterministic implementation, either using hand-coded rules or FSA, or other user-defined methods.

3 Random Forests

Nowadays, numerous attempts in constructing ensemble of classifiers towards increasing the performance of the task at hand have been introduced [10]. A plethora of them has portrayed promising results as regards to classification approaches. Examples of such techniques are Adaboost, Bagging and Random Forests. Random Forests are a combination of tree classifiers such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. A Random Forest multi-way classifier Θ(x) consists of a number of trees, with each tree grown using some form of randomization, where x is an input instance [11]. The leaf nodes of each tree are labeled by estimates of the posterior distribution over the data class labels. Each internal node contains a test that best splits the space of data to be classified. A new, unseen instance is classified by sending it down every tree and aggregating the reached leaf distributions.

Each tree is grown as follows:

- If the number of cases in the training set is N, sample N cases at random but with replacement, from the original data. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
- Each tree is grown to the largest extent possible. Therefore, no pruning is applied.

4 Experimental Design

The training data (a set of Greeklish characters with the corresponding Greek characters) were created using Stochastic Greek2Greeklish Transcriber. For each character a separate instance is created. For the small scale experiments a data set of ~12000 instances was used, these instances were created from some random articles from in.gr The large scale experiments use “OpenThesaurus – Green synonyms thesaurus OpenOffice.org edition” (under GNU general public license), which consists of ~84000 Greek words (with duplicates), this thesaurus combined with the 12000 instance dataset yields a dataset of over 0.7million instances. Using multiples of words produces a model with higher accuracy. This happens because each instance of a word would have a slightly different Greeklish conversion thus creating a better prediction model. The training data is converted to vectors suitable for data mining using the n-grams method. For each pair of Greek, Greeklish characters (e.g. (g,G)) an k-dimensional vector is created using n preceding characters of the Greek character (g), the Greek character g and m proceeding characters (n+m = k-1). The corresponding supervisory signal is of course the character G.

Using a large k impacts not only the accuracy of the classifier (as explained later) but the training and classification time. For example if a tree based classification algorithm is used the dimensionality of the training data affects the size of the produced tree and the time it takes to build it. Fortunately there are some clues that guided us to choose an optimal vector length (discussed later).

As regards to the creation of instances, before the input data is converted to instances all characters are turned to lowercase, this does not affect in any way the process since the capitalization rules are known. Apart from the change in case, all diaeresis are removed since they are rare and could impact the accuracy more (of all other classes) than not using them. All punctuation marks and whitespace are ignored and will be preserved unaltered. All words are independent from each other in the sense that no instance has characters from more than one word, the value for characters beyond the current word are filled in with the character ‘*’ which represents whitespace. The reasoning behind this is the same with the removal of diaeresis, since words with double word stresses are rare and context dependent and generally word stressing is independent for each word.

A training instance consists of a number of features which represent the next n characters in the word, the previous k characters and the current character. The corresponding class is the Greek character in the same position as the current Greeklish character. Special care has been taken for Greek character pairs (e.g. diphthongs) that are equivalent phonetically to one Latin character.

5 Experimental Evaluations

Using the dataset as mentioned above, a series of experiments was conducted using WEKA and RapidMiner benchmarks. For reasons of thorough evaluation, we have compared Random Forests (in practice, both implementations, either Random Input or Random Combination Forests proven to behave similarly, with little variation amongst them) against K-Nearest neighbor (IB1 and IB3 respectively), Decision Trees (J48), Naive Bayes (NB) and Bayesian Networks (BN) classification algorithms. The obtained the results are tabulated in the following table:

Table 1. The cell values represent correct prediction percentage using 10 fold cross validation.

	Algorithm					
Window size	J48	RF	IB1	IB3	BN	NB
[+2-2]	87.59	90.25	85.94	84.28	85.03	84.80
[+2-4]	89.99	92.70	87.12	84.28	86.66	86.13
[+3-3]	89.90	93.34	88.02	85.11	87.02	86.27
[+4-2]	87.90	91.20	84.99	82.72	85.51	84.80
[+2-6]	90.13	92.63	82.76	80.31	86.31	85.85
[+3-5]	89.96	93.21	85.25	82.49	86.60	85.97
[+4-4]	90.08	97.4	86.73	83.25	86.75	85.92
[+5-3]	90.09	98.43	85.79	82.97	87.27	86.22
[+6-2]	88.18	91.31	81.62	79.71	85.29	84.55
[+3-6]	90.10	93.14	93.25	81.06	86.42	85.83
[+5-4]	90.15	93.36	85.25	82.31	86.63	85.95
[+4-5]	90.01	93.34	85.06	81.90	86.61	85.85
[+6-3]	90.08	93.31	83.72	81.46	86.84	86.12
[+5-5]	90.08	93.24	83.98	81.49	86.54	85.72

The results shown in the above table are visualized in the following figure (Fig.1).

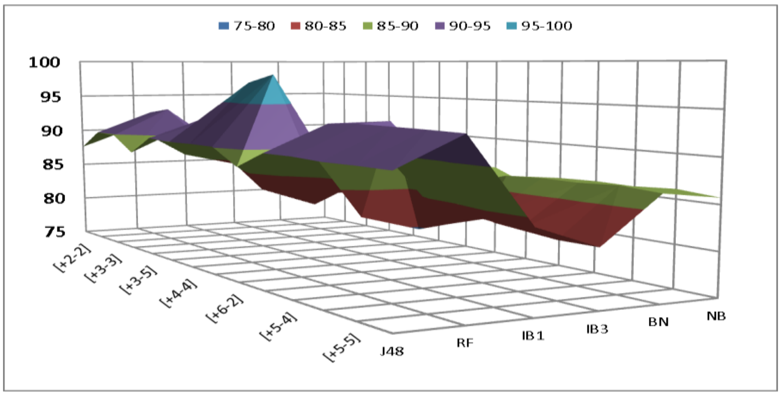

Fig. 1. Results on the set of benchmark algorithms, in terms of prediction accuracy.

The dataset used for the creation of the classifier, as mentioned, consists of about 84000 Greek words (including duplicates). This dataset was obtained by using a window size of [+4-4] to extract the n-grams. Character classes with 1/10th the number of instances are only 1% more accurate. As shown by a study done by Hatzigeorgiu et al. [12] the average length of a Greek word is between 6-7 characters, so using a much greater than this number of grams should yield worst results since the data would be sparse (remember that if the length of the word is smaller than the number of features whitespace is added). Our results confirm this by having an accuracy peak at 9 features (considering 3 previous characters and 5 next, plus the current character ([+5-3]). A second observation is that accuracy decreases if the number of characters considered is highly asymmetrical. This is to be expected since instances at the beginning or the end of the word (depending on if more next or previous characters are considered) will have a lot of whitespace, thus resulting in sparse data. The accuracy measurements obtained from the dataset persist when analyzing using the large dataset. Random Forest showed an overall accuracy of over 98%. This percentage of course refers to a single character being classified correctly.

Conclusions

This work dealt with the importance issue of implementing a Greeklish to Greek transliteration tool, which differs from existing approaches in two ways. The former lies to the fact that no vocabulary is used, therefore the proposed approach is robust to slang and other linguistic idioms, while the latter lies to the fact that it is a non-deterministic, Data Mining approach, which could encompass a variety of user's writing styles and be independent of manually defined, empirical rules. Evaluations against numerous other Data Mining classification approaches have supported our claim that Random Forests (using both of their existing utilizations) are well suited for the task at hand and behave competitive or better that existing deterministic, commercial implementations

References

1. Dale, Ian R.H.. "Digraphia". International Journal of the Sociology of Language 26: 5–13 (1980)
2. Androutsopoulos, J.: Latin-Greek spelling in e-mail messages: Usage and attitudes. In: Studies in Greek linguistics, pp.75-86 (2000) (*in Greek*)
3. Tseliga, T., Marinis, T.: On-line processing of Roman-alphabeted Greek: the influence of morphology in the spelling preferences of Greeklish. In: 6th International Conference in Greek Linguistics, Rethymno, Crete, 18-21 September (2003)
4. ELOT, Greek Organisation of Standardization (1982)
5. e-Chaos: freeware Greeklish converter, http://www.paraschis.gr/files.php
6. Greek to Greeklish by Innoetics, http://services.innoetics.com/greeklish/
7. Chalamandaris, A., Protopapas, A., Tsiakoulis, P., Raptis, S.: All Greek to me! An automatic Greeklish to Greek transliteration system. In: Proceedings of the 5th Intl. Conference in Language Resources and Evaluation, pp. 1226–1229 (2006)
8. DeGreeklish, http://tools.wcl.ece.upatras.gr/degreeklish
9. Greeklish Out!, http://greeklishout.gr/main/
10. Leo Breiman. Random forests.Machine Learning Journal, 45:532 (2001)
11. Igor Kononenko. Estimating attributes: analysis and extensions of Relief. In Luc De Raedt and Francesco Bergadano, editors, Machine Learning: ECML-94, pp. 171-182. Springer Verlag, Berlin (1994)
12. Hatzigeorgiu, N., Mikros, G., Carayannis, G.: Word length, word frequencies and Zipf's law in the Greek language. Journal of Quantitative Linguistics, vol. 8, pp. 175-185 (2001).