LEARNING GREEK PHONETIC RULES USING DECISION-TREE BASED MODELS

Dimitrios P. Lyras, Kyriakos N. Sgarbas and Nikolaos D. Fakotakis
Wire Communications Lab, Electrical and Computer Engineering Department
University of Patras, Patras, GR-26500, Greece

Keywords: Grapheme to phoneme conversion, decision trees, machine transliteration, pronunciation, machine learning.

Abstract: This paper describes the application of decision-tree based induction techniques for automatic extraction of phonetic knowledge for the Greek language. We compare the ID3 algorithm and Quinlan's C4.5 model by applying them to two pronunciation databases. The extracted knowledge is then evaluated quantitatively. In the ten cross-fold validation experiments that are conducted, the decision tree models are shown to produce an accuracy higher than 99.96% when trained and tested on each dataset.

1 INTRODUCTION

The phonetic transcription of text is a crucial task for speech and natural language processing systems. In many cases, this task can be quite complicated as not all graphemes are always phonetically represented, and many graphemes may correspond to different phonemes depending on their context.

Concerning various proposed approaches to this problem, similarity-based and data-oriented techniques have yielded accuracy of 98.2% (Van den Bosch, 1993), while other methods such as neural networks (Rosenberg, 1987; Sejnowski and Rosenberg, 1987), direct lexicon access (Levinson et al., 1989; Xuedong et al., 1995), Hidden Markov Models (Rentzepopoulos and Kokkinakis, 1996), formal approaches (Chomsky and Halle, 1968; Johnson, 1972) and two-level rule-based approaches (Sgarbas et al., 1998) have also proved to be efficient. Linguistic knowledge based approaches to grapheme-to-phoneme conversion have been tried (Nunn and Van Heuven, 1993) yielding comparable results, whilst decision-tree learning approaches are also available (Dietterich, 1997). Memory-based approaches (Busser, G., 1999) are also considered to be remarkably efficient for the grapheme-to-phoneme conversion task.

In this paper we employ the ID3 divide-and-conquer decision tree algorithm and Quinlan's C4.5 decision tree learner model on the machine transliteration task.

2 ON DECISION TREES

Decision trees are important machine learning techniques that produce human-readable descriptions of trends in the underlying relationships of a dataset. As they are robust to noisy data and capable of learning disjunctive expressions, they are widely used in classification and prediction tasks. Two popular algorithms for building decision trees are ID3 and C4.5.

The ID3 algorithm uses *Entropy*, a very important measure from Information Theory that gives an indication on how uncertain we are about the data. The entropy of a target attribute is measured by:

$$Entropy(S) = \sum_{i=1}^{C} p_i \cdot log_2 p_i \qquad (1)$$

where p_i is the proportion of instances in the dataset that take the i-th value of the target attribute.

The *Information Gain* (2) calculates the reduction in Entropy (Gain in Information) that would result in splitting the data on an attribute A.

$$Gain(S, A) = Entropy(S) - \sum_{v \in A} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \qquad (2)$$

where v is a value of A and S_v is the subset of instances of S where A takes the value v.

P. Lyras D., N. Sgarbas K. and D. Fakotakis N. (2007).
LEARNING GREEK PHONETIC RULES USING DECISION-TREE BASED MODELS.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - AIDSS*, pages 424-427
DOI: 10.5220/0002381204240427

- So, by calculating the Information Gain for every attribute of the dataset, ID3 decides which attribute splits the data more accurately. The process repeats recursively through the subsets until the tree's leaf nodes are reached.
- The C4.5 is an extension of the basic ID3 algorithm designed by Quinlan (1993) addressing specific issues not dealt with ID3 (Winston P., 1992).

The C4.5 algorithm generates a classifier in the form of a decision tree. This method also provides the opportunity to convert the decision tree that is produced into a set of "L→R" rules, which are usually more readable by humans. The left-hand side (L) of the generated rules is a conjunction of attribute-based tests and the right-hand side (R) is the class. In order for the C4.5 algorithm to classify a new instance, it examines the generated rules until it finds the one whose conjunction of attribute-based tests (left side) satisfies the case.

The C4.5 algorithm can produce even more concise rules and decision trees by collapsing different values for a feature into subsets that have the form "A in $\{V_1,V_2,\ldots\}$".

Within the last 20 years many modifications have been made to the initial edition of C4.5 algorithm, improving its performance. In our experiments the 8th revision of the algorithm was used, setting the Gain Ratio as splitting criterion and a confidence level pruning of 0.25.

3 ALPHABET AND RULES

The Modern Greek alphabet consists of 24 letters, seven vowels (α, ε, η, ι, ο, υ, ω) and seventeen consonants (β, γ, δ, ζ, θ, κ, λ, μ, ν, ξ, π, ρ, σ, τ, φ, χ, ψ). Vowels may appear stressed (ά, έ, ή, ί, ό, ύ, ώ) and two of them (ι, υ) may have diaeresis or with or without stress (ΐ, ΰ, ϊ, ϋ). The consonant sigma (σ) is written as "ς" when it is the last letter of the word. Finally, a lot of these letters can be combined creating thus pairs of vowels (αι, αί, ει, εί, οι, οί, υι, υί, ού, ού) and pairs of consonants (ββ, γγ, γκ, δδ, κκ, λλ, μμ, μπ, νν, ντ, ππ, ρρ, σσ, ττ, τσ, τς, τζ) each one of which corresponding to a single phoneme.

In our experiments we avoided to use the SAMPA Greek phonetic alphabet, for the reasons explained in Sgarbas & Fakotakis (2005). Instead, our phonetic alphabet was based on Petrounias (1984), Babiniotis (1986) and Setatos (1974).

To efficiently represent the phonetic symbols to a computer compatible form, we chose a mapping of our phonetic units with one or more ASCII symbols, creating thus a CPA (Computer Phonetic Alphabet) which has many similarities to the one created by Sgarbas et al., (1998).

Table 1 shows a part of the correspondence between the International Phonetic Alphabet (IPA) (Robins, 1980) and the CPA used in our experiments. It also demonstrates an appropriate example for each case in graphemic and phonetic form.

Table 1: Phonetic correspondence between the IPA and the CPA used in our study.

#	IPA	CPA	EXAMPLE		
			Graphemic	Phonetic	Translation
1	a	a	**α**νήκει	**a**n'iKi	belongs
2	á	'a	**ά**νεμος	**'a**nemos	wind
3	e	e	**Ε**δώ	**e**ð'o	here
4	é	'e	**Έ**νας	**'e**nas	one
5	o	i	**η**μέρα	**i**m'era	day
6	ó	'i	**ί**σως	**'i**sos	maybe
7	i	o	**ο**σμή	**o**zm'i	smell (n)
8	í	'o	**ό**ταν	**'o**tan	when
9	u	u	**ου**ρά	**u**r'a	tail
10	ú	'u	**ού**τε	**'u**te	neither
11	p	p	**π**ηλός	**p**il'os	clay
12	t	t	**τ**ώρα	**t**'ora	now
13	k	k	**κ**αλός	**k**al'os	good
14	g	g	**γκ**ρεμός	**g**rem'os	pit
15	□	G	**γκ**έμι	**G**'emi	rein
16	f	f	**φ**ίλος	**F**'ilos	friend
17	θ	θ	**θ**έμα	**θ**'ema	subject
18	s	s	**σ**ώμα	**s**'oma	body
19	x	x	**χ**άρη	**x**'ari	grace
20	v	v	**β**λέπω	**v**l'epo	see
21	ð	ð	**δ**έμα	**ð**'ema	parcel
22	z	z	ζωή	**z**o'i	life
23	m	m	**μ**ισό	**m**is'o	half
24	ŋ	N	**ν**ιάτα	**N**'ata	youth

Then, two Greek datasets were created. Both datasets contained Greek words and phrases that complied with some of the fifty-two two-level phonological rules describing the bi-directional transformation for Modern Greek (Sgarbas et.al, 1995). The words and phrases contained into the first dataset complied with only seven of those fifty-two rules (rules 1-7), whilst those of the second dataset complied with nineteen of those rules (rules 1-10, 13-17, 29, 45-47).

Table 2: A part of the dataset that was used in our experiments.

CM2	CM1	CC	CP1	CP2	PM2	PM1	CP	PP1	PP2
*	*	**α**	β	γ	*	*	**a**	v	γ
*	α	**β**	γ	ó	*	a	**v**	γ	‘
α	β	**γ**	ó	*	a	v	**γ**	‘o	o
β	γ	**ó**	*	*	v	γ	**‘o**	*	*
*	*	**ε**	λ	ι	*	*	**e**	Λ	-
*	ε	**λ**	ι	ά	*	e	**Λ**	-	‘
ε	λ	**ι**	ά	*	e	Λ	**-**	‘a	a
λ	ι	**ά**	*	*	Λ	-	**‘a**	*	*

To prepare the data for the machine learning algorithms, we brought them together into a set of instances. Since each instance should contain all the necessary information for the grapheme whose phonetic unit was to be predicted, we selected as necessary attributes for our input pattern the grapheme at stake, the two graphemic symbols that precede it, the two that follow it, the two phonetic units that precede the phoneme that is about to be predicted and the two that follow it.

Table 2 shows a part of the dataset that was used in our experiments, containing the words: αβγό→[avγ’o] (egg) and ελιά→[eΛ’a] (olive). The abbreviation CC stands for the Current Character (grapheme); (CM2, CM1) represent its left context; (CP1, CP2) represent its right context; (CP) stands for the predicted phonemic unit; (PM2, PM1) represent its left context and (PP1, PP2) represent its right context. Whenever there did not exist any information for whichever of these features, an asterisk (*) was placed and in cases where a grapheme did not correspond to any phoneme, a dash (-) was placed.

4 QUANTITATIVE ANALYSIS

To obtain a more quantitative and qualitative picture of the experiments, we decided to split each of the two datasets that were used in our experiments into twenty four segments (one for every letter of the Modern Greek language), creating thus forty eight smaller datasets.

For the training and testing procedure, the WEKA implementation (Witten, 2005) of the aforementioned classification techniques, was used. The statistical technique chosen for our experiments was the 10-fold cross validation (i.e. each dataset was split into ten approximately equal partitions. Each time, one partition was used for testing and the remainder partitions were used for training. This procedure was repeated ten times so that all partitions were used for testing once). All the reported results were averaged over the ten folds.

Figure 1 graphically represents the experimental results. For the first group of datasets (Datasets I), ID3 achieved a performance varying from 82.1782% to 99.9606%, while C4.5 achieved a slightly better performance ranging from 85.1485% to 99.9494%. For the second group of datasets (Datasets II), the increase of the accuracy was even larger: ID3 scored from 94.3162% to 99.9685%, and C4.5 varied between 97.2746% and 99.9815%.

As the experimental results suggest, C4.5 demonstrated slightly higher accuracy than ID3 in the majority of the cases. Also, C4.5 was more than 50% faster in building the model than ID3, without any effects on the performance.

Another important observation is that the learning procedure seems independent of the rules with which the data comply. In particular, we may observe that the highest accuracy achieved for the first group of datasets (Datasets I) that complied with only the first seven phonological rules, was 99.9606%, whilst the best performance for the

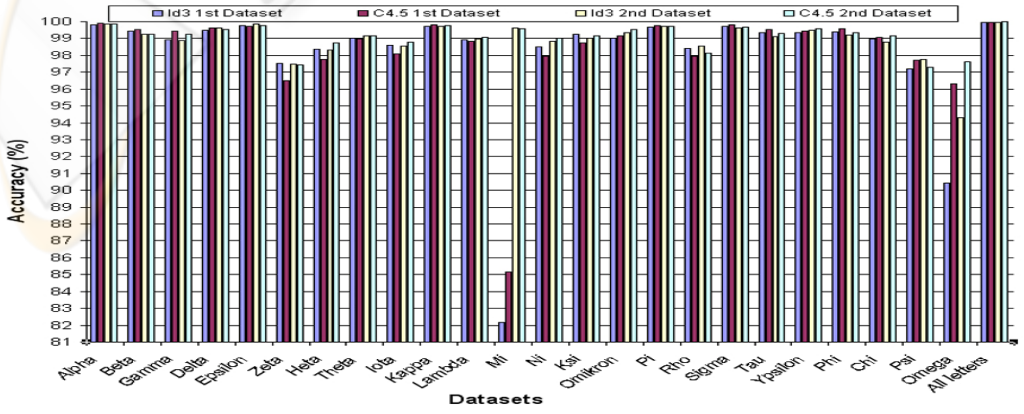

Figure 1: Graphical Representation of the performance of the evaluated classifiers for the first and the second group of Datasets (Datasets I and Datasets II).

second group of datasets (Datasets II) that complied with nineteen rules, was 99.9815%.

Finally, based on the experimental results, we deem that the accuracy of the learning procedure depends on the number of instances that are contained into the dataset. Specifically, the lowest accuracy for both groups of datasets was demonstrated by both algorithms when the dataset with the least number of instances ("Mi" for the first group of datasets and "Omega" for the second one) was applied, and the highest accuracy was achieved when the database with the most number of instances was applied (the dataset "All Letters" for both groups of datasets).

5 CONCLUSIONS

In this paper we presented a decision-tree based approach for learning Greek phonetic rules. A comparative evaluation of the ID3 divide-and-conquer decision tree algorithm and Quinlan's C4.5 learner model was performed, using two databases that contained respectively 31990 and 48238 Greek words and phrases.

The experimental results suggested that although both algorithms perform exceptionally well at the phonetic rule-learning task, the C4.5 classifier is a lot quicker. Furthermore, the phonetic rule-learning task was proven independent of the phonological rules according to which the database is constructed, but depends highly on the size of the dataset (i.e. the number of instances that are contained in it).

REFERENCES

Babiniotis, G., 1986. *Συνοπτική Ιστορία της Ελληνικής Γλώσσας*. Athens.

Busser, G., Daelemans, W., Van den Bosch, A., 1999. Machine Learning of word pronunciation: The case against abstraction. *In Proc. 6th European Conference on Speech Communication and Technology, Eurospeech 99,* Budapest, Hungary, pages 2123-2196.

Chomsky, N., and Halle, M., 1968. *The Sound Patterns of English*, Harper & Roe, New York.

Dietterich, T.G, 1997. Machine Learning research:Four current directions. *AI Magazine,* 18(4):97-136.

Johnson, C. D., 1972. *Formal Aspects of Phonological Description.* Mouton, Hauge.

Levinson, S.E., Liberman, M.Y., Ljolje, A., and Miller, L.G., 1989. Speaker Independent Phonetic Transcription of Fluent Speech for Large Vocabulary Speech Recognition. *ICASSP'89*, pp. 441-444.

Mitchell, T., "Decision Tree Learning", in T. Mitchell, *Machine Learning,* McGraw-Hill, 1997, pp. 52-78.

Nunn, A., van Heuven, V.J., 1993. *Morphon, lexicon-based text-to-phoneme conversion and phonological rules*. In V.J. Van Heuven and L.C.W. Pols, editors, *Analysis and synthesis of speech; strategic research towards high-quality text-to-speech generation.* Berlin, Mouton de Gruyter.

Petrounias, E., 1984. *Νεοελληνική Γραμματική και Συγκριτική Ανάλυση*. University Studio Press, Thessaloniki, Greece.

Quinlan, J. R., 1993. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann Publishers.

Rentzepopoulos, P., and Kokkinakis, G., 1996. Efficient Multilingual Phoneme-to-Grapheme Conversion Based on HMM. *Computational Linguistics,* 22:3.

Robins, R. H., 1980. *General Linguistics. An Introductory Survey.* 3rd Edition, Longman.

Rosenberg, C. R., 1987. Revealing the Structure of NETtalk's Internal Representations. In *Proc. of the 9th Annual Conf. Cognitive Science Society,* pp.537-554.

Sejnowski, T.J., Rosenberg, C.S., 1987. Parallel networks that learn to pronounce English text. *Complex Systems,* 1:145-168.

Setatos, M., 1974. *Φωνολογία της Κοινής Νεοελληνικής*, Papazisis, Athens.

Sgarbas, K., Fakotakis, N., and Kokkinakis, G., 1995. A PC-KIMMO-based Morphological Description of Modern Greek. *Lit. & Ling. Computing,* 10:189-201.

Sgarbas, K., Fakotakis, N., and Kokkinakis, G., 1998. A PC-KIMMO-based Bi-directional Graphemic/ Phonetic Converter for Modern Greek. *Literary and Linguistic Computing,* Oxford University Press, Vol.13,No.2, pp. 65-75.

Sgarbas, K., Fakotakis, N., 2005. A Revised Phonetic Alphabet for Modern Greek. In Proceedings of *SPECOM 2005, 10th International Conference on Speech and Computer,* 17-19 October 1005, Patras, Greece, pp.273-276.

Triantafyllidis, M., 1977. *Νεοελληνική Γραμματική*. ΟΕΔΒ, Athens.

Van den Bosch, A., Daelemans, W, 1993. Data-Oriented Methods for Grapheme-to-Phoneme Conversion. *Proceedings of EACL, Utrecht*, 45-53.

Winston, P., *Learning by Building Identification Trees*, in P. Winston, 1992. Artificial Intelligence, Addison-Wesley Publishing Company, pp.423-442.

Witten, I., Frank, E., 2005. *Data Mining: Practical Machine Learning tools and techniques,* 2nd Edition, Morgan Kaufmann, San Francisco.

Xuedong, H., Acero, A., Alleva, F., Hwang, M.-Y., Jiang, L., and Mahajan, M., 1995. Microsoft Windows Highly Intelligent Speech Recognizer: WHISPER. *ICASSP'95*, USA.