A word-based approach for diacritic restoration in Māori

John Cocks
Department of Computing and Mathematical Sciences
University of Waikato
Waikato, New Zealand
chaopay@hotmail.com

Te Taka Keegan
Department of Computing and Mathematical Sciences
University of Waikato
Waikato, New Zealand
tetaka1@gmail.com

Abstract

This paper describes a supervised algorithm for diacritic restoration based on naive Bayes classifiers that act at word-level. Classifications are based on a rich set of features, extracted automatically from training data in the form of diacritically marked text. The method requires no additional resources, which makes it language independent. The algorithm was evaluated on one language, namely Māori and an accuracy exceeding 99% was observed.

1 Introduction

The Māori language, along with other Polynesian languages, features a written diacritic mark above vowels, signifying a lengthened pronunciation of the vowel. Māori texts without diacritics are quite common in electronic media. The problem arises as most keyboards are designed for English and the process of inserting diacritics becomes laborious. In all but the most ambiguous cases, a native reader can still infer the writer's intended meaning. However, the absence of diacritics can still confuse or slow down a reader and it makes pronunciation and meaning difficult for learners of the language. For other languages using diacritics, such as German or French, this problem can typically be handled by a simple lexicon lookup procedure that translates words without diacritics into the properly marked format (Wagachar and Pauw, 2006). However, this is not the case for languages such as Māori where comprehensive lexicons are not publically available.

This paper proposes a machine learning approach to diacritic restoration that employs a naive Bayes classifier that acts at word-level. The proposed algorithm predicts the placement of diacritics on the basis of local word context. The algorithm is contrasted with a traditional grapheme-based algorithm, originally proposed by Scannell (2010), showing a significant increase in accuracy for diacritic restoration in Māori.

The remainder of the paper is organized as follows: In Section 2, previous work on diacritic restoration is discussed. Section 3 outlines the use of diacritics in Māori. Section 4 describes the dataset used in training and testing each model. Section 5 outlines the baseline models for diacritic restoration used in this paper. Section 6 discusses the Naive Bayes classifier. Section 7 and 8 describe the grapheme-based and word-based models, respectively. Section 9 discusses the results obtained from the baseline, grapheme-based and word-based models. Finally, future work is discussed in Section 10.

2 Previous Work

Until recently, the majority of research on diacritic restoration was directed at major languages such as German and French and less emphasis directed towards minority languages. These methods typically employ the use of large lexicons which

John Cocks and Te Taka Keegan. 2011. A word-based approach for diacritic restoration in Māori. In *Proceedings of Australasian Language Technology Association Workshop*, pages 126−130

are not publically available for resource scarce languages. In recent past, Pauw and Schryver (2009) presented a memory-based approach to diacritic restoration that act at the level of the morpheme for numerous African languages, reporting scores exceeding 90%. Scannell (2010) describes a similar approach, reporting a high degree of accuracy for numerous languages using training data in the form of a web-crawled corpus. Moreover, the diacritic restoration methods presented by Scannell (2010) report a score of 97.5% for Māori. This can be seen as an increase of 1% over the baseline method which chooses the most frequent pattern in the training set. In order to determine the feasibility of the approach proposed in this paper, the experiments outlined by Scannell (2010) are reproduced using a large, high quality corpus and the scores are contrasted with those obtained from the proposed word-level algorithms.

3 Diacritics in Māori

The Māori alphabet consists of 15 characters: 10 consonants and 5 vowels. Vowels in Māori can be pronounced both short and long, so in written form, long vowels carry a diacritical mark. In Māori texts where diacritics have been omitted, long vowels are predominately substituted for short vowels. Table 1 shows the complete set of vowels in Māori.

Short	a	e	i	o	u
Long	ā	ē	ī	ō	ū

Table 1: Short and long vowels in Māori

During substitution, genuine ambiguity arises when two or more distinct words have the same base word-form. To exemplify this ambiguity, consider the Māori word *wāhine* (women). The base word form after diacritics have been removed is *wahine* (woman – singular of wāhine).

4 Dataset

The diacritic restoration algorithms presented in this paper were trained and evaluated on a fully diacritically marked corpus containing approximately 4.2 million words. The corpus was compiled from a comprehensive collection of short stories, bible verses, dictionary definitions and conversational texts. Table 2 displays statistical data extracted from the corpus.

1. Words	4,281,708
2. Words with diacritics	859,083 (20.06%)
3. Words with 0 ambiguity	1,656,051 (38.68%)
4. Words with 1 ambiguity	2,346,874 (54.81%)
5. Words with 2 ambiguities	98,995 (2.31%)
6. Words with 3 or more ambiguities	179,788 (4.20%)

Table 2: Statistical corpus data

The second statistic shows on average, every fifth word in the corpus contains a diacritic. More interestingly, the third statistics shows approximately 39% of the words have no ambiguity and can be correctly restored with a simple lookup procedure; whereas an inflated 61% of the words are ambiguous, and cannot be correctly restored without classification.

5 Baseline Models

In order to determine the significance of the word-based algorithms, two baseline models are defined. The first baseline model assumes no diacritic markings exist. The second baseline model identifies candidate words for diacritic marking, and chooses the most frequent pattern observed in the training set. Candidate words are identified as sharing the same base word-form after diacritics have been removed. For example, the words *āna*, *ānā* and *anā* share the same base word-form *ana*. If two or more candidate words are observed equally, the model randomly chooses a candidate word.

6 Naive Bayes Classifier

In spite of their naive design, naive Bayes classifiers are widely used in various classification tasks in natural language processing. Naive Bayes classifiers are a set of probabilistic learning algorithms based on applying Bayes' theorem with the naive assumption of independence between features. Given a class variable *c* and a dependent feature vector *x1* through *xn*, Bayes' theorem states the following relation:

$$P(c|x1, ..., xn) \propto P(c)\prod_{i=1}^{n} P(xi|c) \quad (1)$$

P(c) is interpreted as the conditional probability of class *c* occurring, and *P(xi|c)* is interpreted as the conditional probability of attribute *xi* occurring given class *c*.

To find the most likely classification *cf*, given the attribute values *x1* through *xn*, equation (1) can be rewritten as:

$$cf = \arg\max\ P(c)\prod_{i=1}^{n} P(xi \mid c) \quad (2)$$

In practice, equation (2) often results in a floating point underflow as *n* increases. It is therefore better to perform the computation by adding logarithms of probabilities instead of multiplying probabilities as in (3).

$$cf = \arg\max \left[\log P(c) + \sum_{i=1}^{n} \log P(xi|c) \right] \quad (3)$$

7 Grapheme-Based Model

Scannell (2010) employs a naive Bayes classifier at the grapheme-level, reporting a high degree of accuracy for numerous languages. These classifiers are trained using various feature sets, each consisting of grapheme-based n-grams relative to the target grapheme. Each n-gram is represented by the vector (o, n), where o represents the offset of the n-gram from the target grapheme, and n represents the length of the n-gram. These feature sets are outlined below. Note that this paper proposes a new grapheme-level feature set: FSG5.

- FSG1: Features (-3, 1), (-2, 1), (-1, 1), (1, 1), (2, 1), (3, 1). That is the three monograms on either side of the target grapheme.
- FSG2: Features (-5, 1), (-4, 1), (-3, 1), (-2, 1), (-1, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1). That is the five monograms on either side of the target grapheme.
- FSG3: (-4, 3), (-3, 3), (-2, 3), (-1, 3), (0, 3), (1, 3), (2, 3). That is the two trigrams on either side of the target grapheme and the three trigrams containing the target grapheme.
- FSG4: (-3, 3), (-1, 3), (1, 3). That is the single trigram on either side of the target grapheme and the single trigram containing the target grapheme.
- FSG5: (-2, 5), (-3, 5), (-1, 5). That is the n-grams of length 5 centered on the target grapheme, and the two n-grams of length 5 starting at offsets -3 and -1.

8 Word-Based Model

This paper improves upon previously mentioned approaches to diacritic restoration by applying diacritic classification at the word-level as opposed to the grapheme-level. This approach extracts word-based n-grams relative to the target word. These features are outlined below:

- FSW1: Features (-1, 1). That is the monogram preceding the target word.
- FSW2: Features (-2, 2). That is the bigram preceding the target word.
- FSW3: Features (-3, 3). That is the trigram preceding the target word.
- FSW4: Features (1, 1). That is the monogram following the target word.
- FSW5: Features (1, 2). That is the bigram following the target word.
- FSW6: Features (1, 3). That is the trigram following the target word.
- FSW7: Features (-1, 1), (-2, 2). That is the monogram and bigram preceding the target word.
- FSW8: Features (1, 1), (1, 2). That is the monogram and bigram following the target word.
- FSW9: Features (-1, 1), (1, 1). That is the monogram on either side of the target word.
- FSW10: Features (-2, 2), (-1, 1), (1, 1), (1, 2). That is the monogram and bigram on either side of the target word.

- FSW11: (-1, 3), (-2, 2), (1, 2), (-1, 4), (-2, 4).

8.1 Naive Bayes Estimates

In order to apply a Naive Bayes classifier to the task of diacritic restoration, estimates for the parameters *P(c)* and *P(xi|c)* in equation (3) outlined above must be found. Assuming a diacritically marked text *T* is a sequence of words *wi* through *wn*, where *n* is the number of words in the text, T can be represented as:

$$T = w1, w2, ..., wn \quad (4)$$

Further, assume each word *wi* in *T* has an associated base word-form *bi*, where *bi* is the result of removing all diacritics from *wi*. Thus a text *T* has a base word-form sequence *Tb* associated with it, which can be written as follows:

$$Tb = b1, b2, ..., bn \quad (5)$$

Let *Wd* be the set of distinct words in *T* and let *Bd* be the set of distinct base word-forms in *Tb*. Further, let $f : B \rightarrow Ws$ be a function that maps a base word-form *bi* to a set of words *Ws*, where *Ws* is a subset of *Wd*, and each word in *Ws* has a corresponding base word-form equal to *bi*. The goal is to find, for each base word-form *bi* in *Tb*, the word *w* in *f(b)*, such that *w* maximizes the probability for all words in *f(b)*. Using Bayes theorem in (3), the prior probability for each word *w* in *f(b)* can be estimated by:

$$P(w) = \frac{Nw}{N} \quad (5)$$

Where *Nw* is the number of occurrences of word *w* in text *T*, and *N* is the total number of occurrences of each word in *f(b)* in text *T*. Further, the conditional probability for each word *w* in *f(b)* is estimated as:

$$P(w) = \frac{Nwi + 1}{Ni + n} \quad (6)$$

Where *Nwi* is the number of occurrences of word *w* with feature *i* in text *T*, and *Ni* is the total number of occurrences of each word *w* in *f(b)* with feature *i* in text *T*, and *n* is the number of words in *f(b)*. To avoid zero estimates, Laplace smoothing is employed.

9 Evaluation

To evaluate the accuracy of the algorithms, a 10-fold cross validation is used. For each experiment, the corpus is partitioned into ten subsets where one subset is used as test data while the remaining nine are used as training data. The experimental results shown in table 3 show that the word-based naive Bayes models significantly outperform the grapheme-based naive Bayes models. Evidently, the FSW11 feature set resulted in the highest accuracy of 99.01%. This can be seen as an increase of 1.9% over the second baseline method which chooses the most frequent pattern in the training data.

Feature Set	Accuracy (%) (proportion of words)
Baseline1	79.94
Baseline2	**97.11**
FSG1	79.94
FSG2	79.94
FSG3	84.45
FSG4	87.02
FSG5	**95.07**
FSW1	98.50
FSW2	98.33
FSW3	97.94
FSW4	98.28
FSW5	98.34
FSW6	98.01
FSW7	98.65
FSW8	98.54
FSW9	98.65
FSW10	98.85
FSW11	**99.01**

Table 3: Accuracy for the baseline, grapheme-based and word-based algorithms

A paired t-test was performed to determine if the increase in accuracy between Baseline2 and FSW11 feature set was significant. The mean increase in accuracy (M=1.8928, SD=0.0234, N=10) was significantly greater than zero, t(9)=255.68, two-tail p=1.08989E-18, providing evidence that FSW11 had a significant increase in accuracy over the Baseline2 feature set. A 95% C.I. about mean accuracy increase is (1.8761, 1.9096).

10 Conclusion and Future Work

This paper presented a method for diacritic restoration based on naive Bayes classifiers that act at grapheme and word level. The use of grapheme-based naive Bayes classifiers in the context of diacritic restoration has already been proposed earlier by Scannell (2010). The experiments presented in this paper extend upon the work by Scannell by proposing training naive Bayes classifiers at the word-level opposed to the grapheme-level. The results show that a word-based naive Bayes model can significantly outperform a grapheme-based naive Bayes model for diacritic restoration in Māori. This paper provides a case study for other Polynesian languages which are closely related to Māori. For future work, the algorithms outlined in this paper will be evaluated across several of these languages where appropriate training data exists in the form of diacritically marked text.

Acknowledgments

The research presented in this paper was made possible through the support of the University of Waikato and Ngā Pae o Te Māramatanga. A demonstration system for Māori diacritic restoration can be found at http://www.greenstone.org/macroniser.

References

Paul, G. and Schryver, M. 2009. African Language Technology: The Data-Driven Perspective.

Santic, N. and Snajder, J. 2009. Automatic Diacritics Restoration in Croatian Texts.

Scannell, K. 2010. Statistical Unicodification of African Languages. Department of Mathematics and Computer Science, Saint Louis University, St Louis, Missouri, USA.

Wagachar, P. and Pauw, G. 2006. A Grapheme-Based Approach for Accent Restoration in Gikuyu. School of Computing and Informatics, University of Nairobi, Nairobi, Kenya.

Yarosky, D. 1996. A Comparison of Corpus-Based Techniques for Restoring Accents in Spanish and French Text. Department of Computer Science, John Hopkins University, Baltimore, MD.