Automatic Diacritic Restoration for Resource-Scarce Languages

Guy De Pauw[1], Peter W. Wagacha[2], and Gilles-Maurice de Schryver[3,4]

[1] CNTS - Language Technology Group, University of Antwerp, Belgium
guy.depauw@ua.ac.be
[2] School of Computing and Informatics, University of Nairobi, Kenya
waiganjo@uonbi.ac.ke
[3] African Languages and Cultures, Ghent University, Belgium
[4] Xhosa Department, University of the Western Cape, South Africa
gillesmaurice.deschryver@ugent.be

Abstract. The orthography of many resource-scarce languages includes diacritically marked characters. Falling outside the scope of the standard Latin encoding, these characters are often represented in digital language resources as their unmarked equivalents. This renders corpus compilation more difficult, as these languages typically do not have the benefit of large electronic dictionaries to perform diacritic restoration. This paper describes experiments with a machine learning approach that is able to automatically restore diacritics on the basis of local graphemic context. We apply the method to the African languages of Cilubà, Gĩkũyũ, Kĩkamba, Maa, Sesotho sa Leboa, Tshivenda and Yoruba and contrast it with experiments on Czech, Dutch, French, German and Romanian, as well as Vietnamese and Chinese Pinyin.

1 Introduction

Language corpus compilation for resource-scarce languages is often done by web crawling the (limited) available content on the Internet [1] or by scanning and "OCRing" hard copy resources [2]. This poses a problem for languages that have diacritically marked characters in their orthography. Despite an increasing awareness of encoding issues, OCR research on orthographically rich languages [3], and the development of specialized computer keyboards [4], many of the digital and digitized language resources use the standard Latin alphabet, with accented characters represented by their unmarked equivalents. While language users can perform real-time disambiguation of unmarked text while reading, a lot of phonological, morphological and lexical information is lost this way, that could be useful in the context of language technology.

Typical diacritic restoration methods employ large lexicons to translate words without diacritics into the properly annotated format. This type of information source is however not digitally available for most resource-scarce languages, many of which make extensive use of diacritically marked characters. In this paper we describe experiments with a machine learning approach that tries to predict the placement of diacritics on the basis of local graphemic context, thereby circumventing the need for a digital dictionary.

V. Matousek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 170–179, 2007.

The focus in this paper will be on seven African languages: Cilubà (Congo, Central Africa), Gĩkũyũ, Kĩkamba and Maa (Kenya, Eastern Africa), Sesotho sa Leboa and Tshivenda (South Africa) and Yoruba (Nigeria, Western Africa). We contrast the results with those obtained on better resourced languages: Czech, Dutch, French, German, Romanian and Vietnamese. To isolate its performance on predicting tonal diacritics, we also investigate the technique on Chinese Pinyin data.

We first look at previous work on diacritic restoration in Section 2, highlighting the grapheme-based approach to diacritic restoration. Section 3 discusses the languages and data sets used in the experiments. We then outline the experimental results in Section 4 and conclude with some pointers to future work in Section 5.

2 Grapheme-Based Diacritic Restoration

Most of the automatic diacritic restoration methods [5,6,7] tackle both the actual task of retrieving diacritics of unmarked text and the related tasks of part-of-speech tagging and word-sense disambiguation. Although complete diacritic restoration ideally involves a large amount of syntactic and semantic disambiguation, this type of analysis can typically not be done for resource-scarce languages. Moreover, these methods rely heavily on lookup procedures in large lexicons, which are usually not available for such languages.

Mihalcea (2002) describes an alternative diacritics restoration method that uses a machine learning technique operating on the level of the grapheme [8,9]. By backing off the problem from the word level to the grapheme level, it opens up the possibility of diacritic restoration for languages that have no electronic word lists available. Applied to Romanian, Czech, Hungarian and Polish, the technique achieves very high accuracy scores of up to 99% on the grapheme level [9]. Similar work on Gĩkũyũ [10] has likewise yielded encouraging results.

The general idea of the approach coined in [8,9,10] is that local graphemic context encodes enough information to solve this disambiguation problem. It projects diacritic restoration as a standard classification problem, that can be solved by a machine learning algorithm.

Left	Left	Left	Left	Left	Focus	Right	Right	Right	Right	Right	Class
-	-	-	-	-	**m**	b	u	r	i	-	m
-	-	-	-	m	**b**	u	r	i	-	-	b
-	-	-	m	b	**u**	r	i	-	-	-	ũ
-	-	m	b	u	**r**	i	-	-	-	-	r
-	m	b	u	r	**i**	-	-	-	-	-	i

Fig. 1. Training Instances for the Gĩkũyũ word “mbũri” (goat)

To this end, training instances in the form of fixed feature vectors are extracted for the graphemes of the words in the corpus. We illustrate this in Figure 1, using an example from one of the target languages under investigation in this paper, i.e. the Gĩkũyũ word

"mbũri" (goat). Using a sliding window, the instance describes eleven features for each grapheme: an ambiguous focus letter, e.g. the Latin character "u", the left context of the focus grapheme and its right context. These features are associated with a class, in this case the diacritically marked character "ũ". The instances can then be used to train a machine learning algorithm which can consequently classify new instances.

Touted as language independent, the scalability of this technique to small data sets and its applicability to non Indo-European data sets, has so far not extensively been investigated. Furthermore, the experimental results presented in [8,9] do not provide an appropriate task-oriented evaluation of the approach. In this paper, we wish to address these issues by adjusting the experimental setup of the technique and re-evaluating it on a more varied array of languages and data sets.

3 The Data Sets

In this section we will outline the available data sets for the languages under investigation. While a detailed overview of the orthography of all these languages would fall beyond the scope of this paper, we will attempt to quantify the disambiguation challenges that our diacritic restoration method faces on the respective languages.

Table 1 provides some quantitative information for the data sets. For Dutch, German and Maa we used the readily available word lists. For each of the other languages, we extracted a word list of unique word forms (column **Types**) from a language corpus, consistently discarding English word forms often found in web crawled corpora. Table 1 further describes the number of non-Latin characters (column **n**) found in the word list and the percentage of words with at least one diacritic (column **T(d)**).

The most informative quantification of the diacritic disambiguation problem is the "lexical diffusion" metric (**LexDif**). To arrive at this value, we first convert all types to latinized word forms, whereby sometimes multiple types converge to the same Latin form. The **LexDif** value is then calculated by dividing the number of types by the number of latinized word forms. It thus expresses the average number of orthographic alternatives per Latin form. Since our grapheme-based technique can only predict one single possible alternative for a given latinized word form, this column describes the degree of resolvability of our approach: the higher the lexical diffusion value, the more inherently unsolvable the diacritic restoration problem.

Cilubà. The manually compiled corpus [11] for this Congolese Bantu language includes almost twenty non-Latin characters. Tonal marking in the orthography causes high values for the **T(d)** and **LexDif** metrics, indicating a significant disambiguation challenge.

Gĩkũyũ and Kĩkamba. These closely related Kenyan Bantu languages have manually compiled corpora available to them [2]. Both have two frequently used diacritically marked characters. The languages are tonal, but tone is not marked in the orthography. Previous diacritic restoration work on Gĩkũyũ [10] showed the grapheme-based approach to be effective for this language, despite the extensive use of diacritics in the orthography.

Table 1. Information on data sets used in the experiments: number of **tokens** and **types** in the corpus; number of diacritically marked characters (**n**); percentage of types with one or more diacritics (**T(d)**); average number of possible orthographic instantiations of the same Latin form (**LexDif**)

Language	Tokens	Types	n	T(d)	LexDif
Cilubà	144.7k	20.0k	17	71.8	1.17
Gĩkũyũ	14.8k	9.1k	2	64.9	1.03
Kĩkamba	38.3k	9.7k	2	65.7	1.07
Maa	22.2k	22.2k	11	46.9	1.05
Sesotho sa Leboa	6.9M	157.8k	1	23.3	1.04
Tshivenda	249.0k	9.6k	5	18.2	1.03
Yoruba	65.6k	4.2k	21	61.3	1.26
Czech	123.9k	105.8k	15	66.3	1.05
Romanian	3.3M	146.9k	5	39.9	1.05
French	23.2M	258.6k	19	21.0	1.04
Dutch	301.9k	301.9k	18	1.5	1.00
German	365.6k	365.6k	4	23.9	1.03
Vietnamese	2.6M	50.9k	26	61.3	1.21
Chinese Pinyin	73.5k	12.0k	25	97.1	1.12

Maa. For this Kenyan Nilotic language, spoken by the Maasai, we used the online Maa dictionary[1] as our data set. We restricted the disambiguation problem to eleven characters (representing phonemes) and discarded tonal markings. The complete tonally marked orthography includes more than 40 characters and can not be handled with a data set of this size.

Sesotho sa Leboa. As one of the eleven official languages of South Africa, this Bantu language has a considerable corpus [12]. With only one diacritically marked character and no tonal markings, the **LexDif** column nevertheless indicates a surprisingly hard disambiguation problem.

Tshivenda. As one of the smaller official Bantu languages of South Africa, a more modest corpus was manually assembled for the purposes of this paper. The orthography contains quite a few non-Latin characters, but has no tonal marking.

Yoruba. The **LexDif** value for this Nigerian Defoid language indicates a similar challenge as for Cilubà, also counting a considerable number of special characters and tonal markings. The corpus material was compiled from sources supplied by Paa Kwesi Imbeah (kasahorow.org) and Kevin Scannell (web crawler "An Crúbádán").

Indo-European languages. For the experiments on Czech we used a word list extracted from the DESAM corpus [13]. The Romanian data set is the same used for the experiments in [9]. The word list for French was extracted from a corpus of French newspaper text (Le Monde). For Dutch and German, we used the readily available lexical databases of CELEX [14].

[1] http://darkwing.uoregon.edu/~dlpayne/Maa%20Lexicon/lexicon/main.htm

Vietnamese. The data set for this Mon-Khmer language was compiled by Le An Ha [15]. The orthography employed in this corpus makes heavy use of diacritics, marking both phonemic variants and tonal characteristics. The high **LexDif** value and the large number of diacritically marked characters predict a complicated disambiguation problem, similar to Yoruba.

Chinese Pinyin. This data set[2] contains a latinized version of the Mandarin Chinese orthography. The diacritics only mark tone, no phonemic variations. Experiments on this data set will allow us to isolate the performance of the technique on predicting tonal diacritics.

4 Experiments

4.1 Experimental Setup

Given that the grapheme-based diacritic restoration approach can principally predict only one single alternative, it simulates a (unigram) lexicon lookup approach. In a practical context, one would therefore be expected to combine the lexicon lookup approach for known words and use the grapheme-based approach for out-of-vocabulary words. This consequently means it should be evaluated primarily on the basis of its performance on unknown words.

In the experiments described in [8,9], instances for graphemes are extracted from a corpus of plain text. The individual instances are then divided into a training set and test set. Making this division on the grapheme level, rather than the word level, means that there will be a significant amount of instances in the test set that have an exact match in the training set. While the experimental results reported in [8,9] are solid, we believe that this methodology does not constitute an appropriate evaluation of the diacritic restoration problem, since the performance on unknown words cannot be established in this manner.

We therefore opt for a significantly different experimental setup, that will allow for a more task-oriented evaluation. Rather than first processing the corpus and dividing the individual instances into a training and test set, we randomly divide the lexicon of unique word forms into ten parts. For each experiment during the 10-fold cross validation, we extract instances from nine partitions, used to train the machine learning algorithm, and evaluate it on the instances extracted from the test set, consisting of unknown words (Section 4.3). In a final experiment (Section 4.4) we also measure performance on plain text data.

4.2 Memory-Based Learning

The instances extracted from the training set are used to train a TiMBL classifier [16], an implementation of the machine learning technique of memory-based learning. The scope of the experiments prevented a thorough exploration of parameter and feature settings. The experimental results were obtained by using the standard settings, except for an increased k-value of 3.

[2] Compiled from http://www.inference.phy.cam.ac.uk/dasher

Interestingly, while other machine learning algorithms like maximum entropy learning and support vector machines are typically able to outperform memory-based learning on many NLP tasks, these algorithms were not able to improve on TiMBL's performance for these experiments, often significantly underperforming. Furthermore, previous experiments using trigram-based processing [10] showed a significant accuracy increase for this task on the Gĩkũyũ data set. After rigid pre-processing of the lexicons, the trigram approach, typically providing more noise-robust output, was no longer observed to yield significant increases in accuracy.

4.3 Experimental Results: Unknown Words

Following up on the new experimental setup described in Section 4.1, we also provide a different, more task-oriented evaluation. Whereas [8,9] provide accuracy scores on the grapheme level, we opt to primarily evaluate the technique on the word level, i.e. the percentage of words in the test that have been predicted completely correctly. Table 2 nevertheless also provides the average accuracy with which latinized graphemes have been disambiguated.

The baseline model identifies candidate graphemes for diacritic marking and chooses the most frequent solution observed in the training set. For French and Dutch for instance these invariably equal to the unmarked characters. This trivial baseline already achieves a very high accuracy for Dutch and Tshivenda (Table 2) because of the limited use of diacritics in these languages. While the disambiguation problem in Sesotho sa Leboa seems limited with only one diacritically marked character, the baseline results confirm the difficulty of the problem.

Table 2. Word level and grapheme level accuracy scores on unknown words (**Ci**: Cilubà, **Gĩ**: Gĩkũyũ, **Kĩ**: Kĩkamba, **Ma**: Maa, **Se**: Sesotho sa Leboa, **Ts**: Tshivenda, **Yo**: Yoruba, **Cz**: Czech, **Ro**: Romanian, **Fr**: French, **Du**: Dutch, **Ge**: German, **Vi**: Vietnamese, **Ch**: Chinese Pinyin)

Word	**Ci**	**Gĩ**	**Kĩ**	**Ma**	**Se**	**Ts**	**Yo**	**Cz**	**Ro**	**Fr**	**Du**	**Ge**	**Vi**	**Ch**
Baseline	28.2	48.7	58.4	53.1	76.2	81.8	35.4	33.7	60.6	75.2	98.5	78.3	29.4	6.7
MBL	36.6	74.9	73.5	58.6	90.1	89.3	40.6	74.4	83.2	88.2	99.6	92.7	63.1	31.5

Grapheme	**Ci**	**Gĩ**	**Kĩ**	**Ma**	**Se**	**Ts**	**Yo**	**Cz**	**Ro**	**Fr**	**Du**	**Ge**	**Vi**	**Ch**
Baseline	69.8	58.9	66.7	76.8	50.6	87.2	54.0	83.2	92.5	93.8	99.7	83.1	65.8	40.4
MBL	77.4	83.1	80.4	85.4	80.9	92.9	68.2	95.2	97.3	97.2	99.9	94.3	82.7	69.0

The grapheme-based memory-based learning approach (**MBL** in Table 2) is able to improve both word level and grapheme level accuracy scores for all data sets, with a particularly encouraging increase in accuracy for Gĩkũyũ, Kĩkamba, Sesotho sa Leboa, Czech, Romanian and Vietnamese. Note how for Czech and Romanian a modest increase of accuracy on the grapheme level has a major impact on the accuracy on the word level. Interestingly, the grapheme accuracy scores for Czech and Romanian are well below those reported in [8,9]. Since we use the same machine learning algorithm and same data, we hypothesize that the difference is due to evaluating the task on

unseen words, rather than evaluating it on graphemes, extracted from a combination of known and unknown words.

While the results for Cilubà and Yoruba have improved significantly, the diacritic restoration problem is still far from solved for these languages. The trailing results compared to the other African languages, are caused by the tonal markings present in these languages. Tonal diacritics can simply not be solved on the level of the grapheme. Particularly the problem of floating tones needs to be resolved on the sentence level. The increase in accuracy reported on these languages is mainly due to the restoration of diacritics that indicate phonemic alternatives.

This hypothesis is further corroborated by the results on Chinese Pinyin. Diacritics in this data set solely mark tone. While there is a significant increase using the machine learning approach, the results are still severely lacking. Note that the **LexDif** metric (Table 1) was able to predict the trailing results for Cilubà, Yoruba and Chinese Pinyin.

A special case is the language pair Gĩkũyũ and Kĩkamba. Closely related with a very similar orthography, we conducted some combination experiments. In the first experiment, we isolated a Kĩkamba test set and added the Gĩkũyũ data set to the Kĩkamba training set. Word-level accuracy decreased 5.4% compared to a plain Kĩkamba training set (67.1% vs 72.5%). A reverse experiment with a Gĩkũyũ test set yielded a decrease of 6.1% (67.4% vs 73.5%). In a second set of experiments, we solely used Gĩkũyũ training data to classify the Kĩkamba test set and vice versa. Word-level accuracy on the Gĩkũyũ test set was 55.8%, and 52.3% on the Kĩkamba test set. Since these results indicate the orthography of the languages is to some extent similar, re-using the data may bootstrap a basic diacritic restoration method for other closely related languages such as Kĩembu or Kĩmerũ.

4.4 Experimental Results: Plain Text

For the languages for which we had a plain text corpus available (all except Maa), we conducted some experiments measuring the effectiveness of our technique on a text containing both known and unknown words. Table 3 displays the results for these experiments. The baseline model for this experiment implements the lexicon lookup method (**LLU**). In this approach, the training set lexicon is used to translate the unmarked words in the test set into the associated diacritically marked words using a unigram model. Particularly for languages with a large training lexicon, this is the baseline to beat. The second method is the grapheme-based memory-based learning approach (**MBL**). The third method combines the two, using lexicon lookup for known words, and MBL for unknown words (**LLU+MBL**).

The results show that for Dutch and German, the lexicon lookup model scores quite well. For the former, this is almost a solved problem. Not surprisingly, the smaller lexicon for French yields a more modest score for the plain text test set. Using the MBL method, there is only a small decrease for French, Dutch and German compared to the lexicon lookup approach. These results are encouraging, since they give an indication of the relative accuracy of the grapheme-based approach, compared to the standard lexicon lookup approach.

For languages with a larger corpus, like Sesotho sa Leboa, Czech and Romanian, the combined approach outperforms all other alternatives, but rather surprisingly, despite

Table 3. Word level accuracy scores on plain text

Word	Ci	Gĩ	Kĩ	Se	Ts	Yo	Cz	Ro	Fr	Du	Ge	Vi	Ch
LLU	77.0	77.3	79.4	97.6	97.7	67.8	61.8	94.0	89.1	99.9	96.2	74.5	78.5
MBL	**85.3**	**92.4**	**91.6**	99.2	**99.4**	**76.8**	89.2	96.5	88.3	99.8	95.3	73.5	**83.9**
LLU+MBL	79.6	91.5	90.4	**99.4**	99.2	68.5	**90.1**	**96.6**	**89.3**	**99.9**	**96.8**	**75.5**	80.3

the considerable size of the training lexicon, MBL still significantly outperforms the lexicon lookup method.

As expected, the score for the lexicon lookup approach is quite low for the resource-scarce languages of Cilubà, Gĩkũyũ, Kĩkamba, Tshivenda and Yoruba. For each of these, the grapheme-based approach also outperforms the combined approach by a significant margin. This means that a typical training set for these resource-scarce languages does not yet contain enough lexical information to enable accurate lexicon lookup approaches. This projects the grapheme-based approach as the more robust diacritic restoration method for resource-scarce languages.

Also note that the word level accuracy scores on plain text are a lot higher than those for unknown words. This is particularly true for the Chinese Pinyin data set. We hypothesize that the artificially inflated scores are the effect of using small domain-specific corpora, with typically a restricted lexicon. This provides further support to the claim that the diacritic restoration task is preferably to be evaluated on unknown words, to truly measure its effectiveness in a practical context.

5 Conclusion and Future Work

In this paper we have presented experiments with a grapheme-based machine learning approach for diacritic restoration. We described a new experimental approach to this task, that enables a more task-oriented evaluation of this particular disambiguation problem. The difference in results between disambiguating unknown words and known words provides some indication that previously reported results were overstated. We also introduced the metric "lexical diffusion" that is able to predict the difficulty of the diacritic restoration problem for a given language.

Focusing on resource-scarce African languages, we showed that the machine learning approach is indeed to a great extent language independent. But while the method is able to predict diacritics for phonemic variants of the same Latin character with a high degree of accuracy, there are considerable issues when dealing with languages that mark tonality in the orthography. Future research will extend the technique to predict multiple variants of the same latinized word form, combined with contextual sentence models to trigger the correct tonal pattern of a word.

Since for most African languages there is an almost one-to-one mapping between phoneme and grapheme, an effective diacritic restoration method for African languages is almost tantamount to grapheme-to-phoneme conversion. Particularly given the more than encouraging results on processing plain text, the machine learning approach presented in this paper warrants further investigation on a larger array of African languages.

In the meantime, we are confident that the proposed diacritic restoration method can significantly speed up corpus development for the resource-scarce languages under investigation in this paper, as it provides an effective tool to process and enhance unmarked digital language resources.

Acknowledgments and Demo

The research presented in this paper was made possible through the support of the VLIR-IUC-UON programme. The first author is funded as a Postdoctoral Fellow of the Research Foundation - Flanders (FWO). We would like to thank Martine Coene, Paa Kwesi Imbeah (kasahorow.org), Rada Mihalcea, Kevin Scannell, Le An Ha, Mercy Nevhulaudzi, M.J. Mafela, Pauline Githinji and Ruth Wambua for their co-operation.

Demonstration systems of the diacritic restoration method presented in this paper, are available at http://aflat.org.

References

1. de Schryver, G.M.: Web for/as corpus: A perspective for the African languages. Nordic Journal of African Studies 11/2, 266–282 (2002)
2. Wagacha, P., De Pauw, G., Getao, K.: Development of a corpus for Gĩkũyũ using machine learning techniques. In: Proceedings of LREC workshop - Networking the development of language resources for African languages, Genoa, Italy, ELRA, pp. 27–30 (2006)
3. Hussain, F., Cowell, J.: Amharic character recognition using a fast signature based algorithm. In: Proceedings of the IEEE conference on Image Visualisation 2003, London, UK, pp. 384–389. IEEE Computer Society Press, Los Alamitos (2003)
4. Bailey, D.: Creating a South African keyboard. In: Afrilex 2006, the user perspective in lexicography, programme and abstracts, Pretoria, South Africa $(SF)^2$ Press, pp. 17–18 (2006)
5. Yarowsky, D.: A comparison of corpus-based techniques for restoring accents in Spanish and French text. In: Proceedings of the Second Annual Workshop on Very Large Corpora, Kyoto, Japan, pp. 19–32 (1994)
6. Tufiş, D., Chiţu, A.: Automatic diacritics insertion in Romanian texts. In: Proceedings of the International Conference on Computational Lexicography, Pecs, Hungary, pp. 185–194 (1999)
7. Simard, M.: Automatic insertion of accents in French text. In: Proceedings of the Third Conference on Empirical Methods in Natural Language Processing, Granada, Spain, pp. 27–35 (1998)
8. Mihalcea, R.F.: Diacritics restoration: Learning from letters versus learning from words. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 339–348. Springer, Heidelberg (2002)
9. Mihalcea, R.F., Nastase, V.: Letter level learning for language independent diacritics restoration. In: Proceedings of CoNLL-2002, Taipei, Taiwan, pp. 105–111 (2002)
10. Wagacha, P., De Pauw, G., Githinji, P.: A grapheme-based approach for accent restoration in Gĩkũyũ. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy, ELRA, pp. 1937–1940 (2006)
11. de Schryver, G.M.: Bantu Lexicography and the Concept of Simultaneous Feedback (MA dissertation). Ghent University, Ghent, Belgium (1999)

12. de Schryver, G.M.: Corpus-based statements of meaning versus descriptions of actual language use in dictionaries. Culture, Language and Identity (CLIDE) Seminar, University of the Western Cape, Bellville, South Africa (2007)
13. Pala, K., Rychlý, P., Smrž, P.: DESAM - annotated corpus for Czech. In: Jeffery, K.G. (ed.) SOFSEM 1997. LNCS, vol. 1338, pp. 523–530. Springer, Heidelberg (1997)
14. Baayen, R.H., Piepenbrock, R., van Rijn, H.: The CELEX lexical data base on CD-ROM. Linguistic Data Consortium, Philadelphia, PA (1993)
15. Ha, L.: A method for word segmentation in Vietnamese. In: Proceedings of the Corpus Linguistics 2003 Conference, pp. 282–287 (2003)
16. Daelemans, W., Zavrel, J., van den Bosch, A., van der Sloot, K.: TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide. ILK Technical Report 04-02, Tilburg University (2004)