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

[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 diac ically marked characters. Falling outside the scope of the standard Latin enc ing, these characters are often represented in digital language resources as t unmarked equivalents. This renders corpus compilation more difficult, as th languages typically do not have the benefit of large electronic dictionaries perform diacritic restoration. This paper describes experiments with a mach learning approach that is able to automatically restore diacritics on the basis local graphemic context. We apply the method to the African languages of Cilu Gĩkũyũ, Kĩkamba, Maa, Sesotho sa Leboa, Tshivenda and Yoruba and contr it with experiments on Czech, Dutch, French, German and Romanian, as wel Vietnamese and Chinese Pinyin.

1 Introduction

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

Typical diacritic restoration methods employ large lexicons to translate w out diacritics into the properly annotated format. This type of information however not digitally available for most resource-scarce languages, many make extensive use of diacritically marked characters. In this paper we descri ments with a machine learning approach that tries to predict the placement of on the basis of local graphemic context, thereby circumventing the need fo dictionary.

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

with those obtained on better resourced languages: Czech, Dutch, French Romanian and Vietnamese. To isolate its performance on predicting tonal dia also investigate the technique on Chinese Pinyin data.

We first look at previous work on diacritic restoration in Section 2, highli grapheme-based approach to diacritic restoration. Section 3 discusses the lang data sets used in the experiments. We then outline the experimental results in 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 a of retrieving diacritics of unmarked text and the related tasks of part-of-sp ging and word-sense disambiguation. Although complete diacritic restorati involves a large amount of syntactic and semantic disambiguation, this type c can typically not be done for resource-scarce languages. Moreover, these me heavily on lookup procedures in large lexicons, which are usually not availabl languages.

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

The general idea of the approach coined in [8,9,10] is that local graphem encodes enough information to solve this disambiguation problem. It project restoration as a standard classification problem, that can be solved by a mach ing 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 extrac graphemes of the words in the corpus. We illustrate this in Figure 1, using a from one of the target languages under investigation in this paper, i.e. the Gĩk

this case the diacritically marked character "ũ". The instances can then be us
a machine learning algorithm which can consequently classify new instances

Touted as language independent, the scalability of this technique to smal
and its applicability to non Indo-European data sets, has so far not extensi
investigated. Furthermore, the experimental results presented in [8,9] do not p
appropriate task-oriented evaluation of the approach. In this paper, we wish
these issues by adjusting the experimental setup of the technique and re-evalu
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 und
gation. While a detailed overview of the orthography of all these languages
beyond the scope of this paper, we will attempt to quantify the disambigua
lenges that our diacritic restoration method faces on the respective languages

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

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

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

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

(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 Maa dictionary[1] as our data set. We restricted the disambiguation problem characters (representing phonemes) and discarded tonal markings. The compl marked orthography includes more than 40 characters and can not be han a data set of this size.

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

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

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

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

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

large number of diacritically marked characters predict a complicated disam
problem, similar to Yoruba.

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

4 Experiments

4.1 Experimental Setup

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

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

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

4.2 Memory-Based Learning

The instances extracted from the training set are used to train a TiMBL class
an implementation of the machine learning technique of memory-based lear
scope of the experiments prevented a thorough exploration of parameter a
settings. The experimental results were obtained by using the standard settin
for an increased k-value of 3.

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

performance for these experiments, often significantly underperforming. Fur
previous experiments using trigram-based processing [10] showed a signific
racy increase for this task on the Gĩkũyũ data set. After rigid pre-process
lexicons, the trigram approach, typically providing more noise-robust outpu
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
different, more task-oriented evaluation. Whereas [8,9] provide accuracy sco
grapheme level, we opt to primarily evaluate the technique on the word lev
percentage of words in the test that have been predicted completely correctl
nevertheless also provides the average accuracy with which latinized graphe
been disambiguated.

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

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

Word	**Ci**	**Gĩ**	**Kĩ**	**Ma**	**Se**	**Ts**	**Yo**	**Cz**	**Ro**	**Fr**	**Du**	**Ge**	**V**
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
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

Grapheme	**Ci**	**Gĩ**	**Kĩ**	**Ma**	**Se**	**Ts**	**Yo**	**Cz**	**Ro**	**Fr**	**Du**	**Ge**	**V**
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
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

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

restoration problem is still far from solved for these languages. The traili
compared to the other African languages, are caused by the tonal markings
these languages. Tonal diacritics can simply not be solved on the level of the
Particularly the problem of floating tones needs to be resolved on the sente
The increase in accuracy reported on these languages is mainly due to the rest
diacritics that indicate phonemic alternatives.

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

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

4.4 Experimental Results: Plain Text

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

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

For languages with a larger corpus, like Sesotho sa Leboa, Czech and Rom
combined approach outperforms all other alternatives, but rather surprising

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	7
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	**8**
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**	8

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

As expected, the score for the lexicon lookup approach is quite low for the scarce languages of Cilubà, Gĩkũyũ, Kĩkamba, Tshivenda and Yoruba. Fo these, the grapheme-based approach also outperforms the combined approach nificant margin. This means that a typical training set for these resource-s guages does not yet contain enough lexical information to enable accura lookup approaches. This projects the grapheme-based approach as the more r critic restoration method for resource-scarce languages.

Also note that the word level accuracy scores on plain text are a lot hi those for unknown words. This is particularly true for the Chinese Pinyin da hypothesize that the artificially inflated scores are the effect of using smal specific corpora, with typically a restricted lexicon. This provides further sup claim that the diacritic restoration task is preferably to be evaluated on unkno 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 mach ing approach for diacritic restoration. We described a new experimental ap this task, that enables a more task-oriented evaluation of this particular disam problem. The difference in results between disambiguating unknown words a words provides some indication that previously reported results were overs also introduced the metric "lexical diffusion" that is able to predict the difficu diacritic restoration problem for a given language.

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

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

Acknowledgments and Demo

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

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

References

1. de Schryver, G.M.: Web for/as corpus: A perspective for the African languages. N
nal of African Studies 11/2, 266–282 (2002)
2. Wagacha, P., De Pauw, G., Getao, K.: Development of a corpus for Gĩkũyũ usi
learning techniques. In: Proceedings of LREC workshop - Networking the deve
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
In: Proceedings of the IEEE conference on Image Visualisation 2003, London, U
389. IEEE Computer Society Press, Los Alamitos (2003)
4. Bailey, D.: Creating a South African keyboard. In: Afrilex 2006, the user pe
lexicography, programme and abstracts, Pretoria, South Africa (SF)^2 Press, pp. 17
5. Yarowsky, D.: A comparison of corpus-based techniques for restoring accents in S
French text. In: Proceedings of the Second Annual Workshop on Very Large Corp
Japan, pp. 19–32 (1994)
6. Tufiş, D., Chiţu, A.: Automatic diacritics insertion in Romanian texts. In: Pro
the International Conference on Computational Lexicography, Pecs, Hungary, p
(1999)
7. Simard, M.: Automatic insertion of accents in French text. In: Proceedings o
Conference on Empirical Methods in Natural Language Processing, Granada, Sp
35 (1998)
8. Mihalcea, R.F.: Diacritics restoration: Learning from letters versus learning f
In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 339–348. Springer,
(2002)
9. Mihalcea, R.F., Nastase, V.: Letter level learning for language independent diacri
tion. 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 re
Gĩkũyũ. In: Proceedings of the Fifth International Conference on Language Res
Evaluation, Genoa, Italy, ELRA, pp. 1937–1940 (2006)
11. de Schryver, G.M.: Bantu Lexicography and the Concept of Simultaneous Fee
dissertation). Ghent University, Ghent, Belgium (1999)

13. Pala, K., Rychlý, P., Smrz, P.: DESAM - annotated corpus for Czech. In: Jeffery,
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
Linguistic Data Consortium, Philadelphia, PA (1993)
15. Ha, L.: A method for word segmentation in Vietnamese. In: Proceedings of the C
guistics 2003 Conference, pp. 282–287 (2003)
16. Daelemans, W., Zavrel, J., van den Bosch, A., van der Sloot, K.: TiMBL: Tilbur
Based Learner, version 5.1, Reference Guide. ILK Technical Report 04-02, Tilbu
sity (2004)