Diacritic error detection and restoration via part-of-speech tags

Jerid Francom, Mans Hulden

Wake Forest University
323 Greene Hall, 27103 Winston-Salem, NC, USA
francojc@wfu.edu

University of Helsinki
P.O.Box 24 (Unioninkatu 40) FI-00014 Helsinki, Finland
mhulden@email.arizona.edu

Abstract

In this paper we address the problem of diacritic error detection and restoration—the task of identifying and correcting missing accents in text. In particular, we evaluate the performance of a simple part-of-speech tagger-based technique comparing it to other well-established methods for error detection/restoration: unigram frequency, decision lists and grapheme-based approaches. In languages such as Spanish, the current focus, diacritics play a key role in disambiguation and results show that a straightforward modification to an n-gram tagger can be used to achieve good performance in diacritic error identification without resorting to any specialized machinery. Our method should be applicable to any language where diacritics distribute comparably and perform similar roles of disambiguation.

Keywords: Diacritic restoration, Part-of-speech tagging, Romance languages, Spanish

1. Introduction

Lexical disambiguation is key to developing robust natural language processing applications in a variety of domains such as grammar and spell checking (Tufiş and Ceauşu, 2008), text-to-speech applications (Ungurean et al., 2008), among other direct applications. Effective diacritic restoration, usually a pre-processing task, is also essential to the accuracy and reliability of any subsequent text processing and ever more important as NLP investigations are applied to 'real-world' contexts in which normalized text cannot be assumed.

The primary areas of investigation on lexical disambiguation have focused on syntactic, or part-of-speech (**house/noun** vs. **house/verb**) and semantic, or word sense (**bug/insect** vs. **bug/small microphone**) ambiguities. Less attention has focused on ambiguities that arise from orthographic errors. An important component of the orthography of many of the world's languages, diacritic markings are often stripped or appear inconsistent due to technical errors—OCR errors, 8-bit conversion/stripping, ill-equipped keyboards, etc.—and human error—native speaker errors related to the level of formality and/or orthographic knowledge. For example, email communication in Spanish more often than not lack systematic diacritic markings, most probably for convenience. Also, of all lexical errors for high school and early college students in Spain orthographic accentuation is the most common (Paredes, 1999).

Whereas identified diacritic errors can be restored quite trivially in Spanish, a language in which all words have maximally one diacritic mark and because we rarely find more than two-way ambiguous words in the lexicon, reliable detection of diacritic errors is less straighforward. One the one hand, stripping diacritics from some words produce non-ambiguous non-words such as **según/*segun** (according to), **quizás/*quizas** (perhaps), etc. which can be resolved easily with access to a large lexicon and/or morphological analysis resources. On the other hand, stripping diacritics from other words produce semantic or syntactic ambiguity, such as pairs for the verb 'hablar' (to speak) **habló[3p/past]/hablo[1p/pres]**, **hablará[3p/fut]/hablara[3p/past/subj]**, noun pairs **secretaría/secretaria** (secretariat vs. secretary) or demonstrative/noun **estas/éstas** (these) or demonstrative/verb **estas/estás** (these/are[2p/pres]) pairs; real-world, context-dependent errors which are much more difficult to reliably identify.

In this paper we focus on a novel, potentially informative angle to the problem of diacritic error detection which leverages the observation that the disambiguation a modern HMM part-of-speech tagger performs can be extended to give a judgment on whether a particular word is correctly diacriticized. For our main target language, Spanish, the correction itself is a relatively straightforward task as almost all words have maximally one diacritic mark and because we rarely find more than two-way ambiguous words in the lexicon. Hence, in the following, we shall focus on the identification of incorrectly diacriticized words and assume the availability of auxiliary techniques to perform the correction.

2. Prior work

There are three key studies that are relevant to the current investigation spanning grapheme, word and tag-based approaches to diacritic error detection/restoration. First, Yarowsky (Yarowsky, 1994; Yarowsky, 1999) advocates a decision-list approach combining simple word form frequency, morphological regularity and collocational information to restore diacritics.[1] This decision-list implemen-

[1] The suffix-based approach described by Yarowsky (Yarowsky, 1994; Yarowsky, 1999) is not considered here as morphological form merely serves as a proxy for part-of-speech category, the primary variable of the current POS-tag approach.

tation is motivated in that each strategy in the decision list show complementary strengths and weaknesses. Results from Yarowsky's study, looking at Spanish in particular, appear to show high levels of accuracy for individual strategies and even higher levels when combining methods. Of note, the impressive results reported in these results may, in part, be enhanced due to the fact that the training and evaluation data are drawn from the same genre/register and most likely show homogeneity that cannot be expected in all training/evaluation tests and the strategies employed require fairly large training sets of orthographically correct(ed) text—a characteristic that potentially limits the application of these methods to diacritic restoration of text for less-resourced languages in which availability of correctly marked text is lacking.

A second approach to diacritic restoration explored in the literature, partly motivated by the resource-scarcity problem (Scannell, 2007), hypothesizes that the local graphemic distribution can provide sufficient information to cue effective diacritic restoration, even based on small training sets (Mihalcea, 2002). An implementation of a memory-based learning (MBL) algorithm, De Pauw et al. (De Pauw et al., 2007) reports mixed restoration accuracy rates for a variety of African, Asian and European languages, including the Romance Language French, using a grapheme-based learning approach. In the authors' assessment, memory-based restoration effectiveness hinges on LexDiff; a metric of orthographic ambiguity calculated by taking the ratio of the number of unique latinized forms over the total number correctly marked forms. However, it is not clear how robust the technique is as reported performance varies for languages matched for LexDiff for various training corpus sizes.

Most closely related to the current work, Simard and Deslauriers (Simard and Deslauriers, 2001) present a POS-tagger based approach for the restoration of French diacritics. In essence, the proposed solution is to try different variant 'candidate' diacritizations of an input sentence and use a Hidden Markov Model (HMM) tagger to produce a probability score for each 'candidate' restoration. The combination of diacritics with the highest likelihood, as judged by the POS tagger, is finally chosen as the correct restoration. As in the current approach POS-tag information is harnessed to provide cues for diacritic error detection. However this POS-candidate approach requires more machinery and is somewhat more involved than the current POS-tag approach, which all else being equal would be preferred given implementation considerations.

3. Current proposal

The core idea behind our POS based method is to train an n-gram tagger on a tagged corpus which is augmented with information about diacritic placement. More concretely, given a corpus C, we divide the corpus into two copies C_1 and C_2—one with diacritics stripped, and another with the correct diacritic placement intact. For each word/tag pair in the stripped corpus, we then augment the corresponding tag for each word with information about whether the word is correctly diacriticized (by adding a sequence *BAD* or *OK* to the original tag). Naturally, only some words in the stripped corpus will be marked *BAD*, while the rest will all carry the *OK*-tag. See Table 1 for an illustration of a hypothetical sentence 'According to them (he/she/you[formal]) (did) not buy anything there' from both corpora.[2] From this new corpus, which is twice the size of the original corpus, we train a Hidden Markov Model tagger in the usual way.

(1)		(2)	
Según	SPS00-OK	Segun	SPS00-BAD
ellos	PP3MP000-OK	ellos	PP3MP000-OK
no	RN-OK	no	RN-OK
compró	VMIS3S0-OK	compro	VMIS3S0-BAD
nada	RG-OK	nada	RG-OK
allí	RG-OK	alli	RG-BAD
.	Fp-OK	.	Fp-OK

Table 1: Example of generation of diacritic-stripped variant corpus for training, from original sentence (1). Tags are simply augmented with information about whether words are correctly diacriticized.

After the tagger is trained on the corpus, we in effect produce a POS tagger that not only is able to tag words where the diacritics are missing, but that also marks each word as having either correct or incorrect accent marks. That is, the output of the tagger would look like the two columns in Table 1.

The motivation behind such an approach is threefold. First, there is the obvious direct connection behind part-of-speech and diacritic-induced ambiguity: some words are ambiguous entirely along these lines (**completo** [N] (complete) / **completó** [V] (completed), **esta** [PRON] (this) / **está** [V] (is), etc). Secondly, as has been noted in the literature on decision lists, POS sequences themselves are very good indicators of a given correct accent placement, although the ambiguity may not distinguishable by local POS class alone: for example, the sequence **(1) PREP (2) que (3) ... -ara**, is a very strong clue to word **(3)** being in the subjunctive mood and thus diacriticized **-ara**, vs. the future **-ará** (**para que terminara / comprara / empezara/etc.**). Contrariwise, the sequence **(1) NOUN (2) que (3) ... -ara**, is usually indicative of the opposite choice: the future tense (**cosa que terminará / acabará / etc.**).[3] Thirdly, assuming the tagset is fine-grained enough, distinguishing between person and tense of verbs, for instance, we can evaluate common diacritic placement errors that are distinguishable along those lines: **hablo** [1p/pres] vs. **habló** [3p/past] (to speak), **toco**[1p/pres] vs. **tocó**[3p/past] (to touch), etc.

It should be noted that such discriminative power can be obtained by first part-of-speech tagging a text (keeping

[2] The widely-adopted EAGLES tagset for Spanish is used for part-of-speech annotation (see Freeling: http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html)

[3] We assume here that the relevant n-gram tagger that is trained has some suffix-based mechanism for guessing parts-of-speech for unknown words as is generally the case with better-performing taggers, such as TnT (Brants, 2000), or HunPos (Halácsy et al., 2007).

in mind that the tagger needs to be able to handle incorrectly diacriticized input), and then applying a separate decision list trained to identify such sequences as described above. However, by including the information about correct accent placement in the tags themselves, we can produce equally effective results without the decision list.

As the output of the system does not tell us—if there is an incorrectly accented word—how the correct word should read, further processing is necessary. By examining the augmented POS tag we only know whether a word is correctly or incorrectly diacriticized. The task of actually restoring the diacritics may require further methods that depend on the language and the types of ambiguity it contains. In the case of Spanish, one can restore most words correctly even without access to a dictionary. Since Spanish stress placement is predictable in the absence of diacritics, one can rule out a number of implausible corrections only using knowledge of Spanish stress rules and ambiguity types (such as **-ara/-ará** and **-o/ó** mentioned above). Having access to a simple dictionary or morphological analyzer would further resolve the restoration problem almost perfectly. This because almost all words in Spanish are maximally two-way-ambiguous in diacritization and only a few exceptions such as **esta/está/ésta** exist. However, more elaborate treatment would be required for other languages such as French where the diacritization behaves differently and is less predictable.

It is worth noting that while a standard trigram tagger would not be expected to yield accuracies over 97% in a plain POS tagging task, the diacritic restoration task is much simpler and we can expect to exceed POS tagging accuracies by a fair margin.

4. Evaluation

4.1. Alternative restoration methods

To evaluate the performance of the POS-based method, we compare its performance to a baseline frequency model, decision list model and grapheme-based algorithm, each documented in the previous literature and implemented as follows.

As a baseline model we have chosen the simple approach of simply choosing the most frequent orthographic word form if it is ambiguous. In the case that an input word is already diacriticized, we do not remove accent marks, working on the assumption that anything diacriticized is reliable. This baseline has been shown quite difficult to substantially improve upon for many languages, including Spanish (Yarowsky, 1994) and French (Simard and Deslauriers, 2001).

As a second alternative, we have implemented the decision list strategy described in Yarowsky (Yarowsky, 1994). The method involves collecting collocational information for each ambiguous word in a corpus, using pre-specified contexts, and then creating a decision list from the collocation counts. The decision lists are rigged so that the most reliable collocations are applied first. The types of collocations considered in the learning task are the following:

- Word to the left (-1w)
- Word to the right (+1w)
- The previous two words (-2w,-1w)
- The following two words (+1w,+2w)
- Any word in a ±20 word window (+-20w)

Once the collocation counts are collected from a corpus, the decision list is sorted by log-likelihood ratio so that more reliable rules are applied before less reliable ones. In the absence of any applicable rule, a 'default' rule will choose the most frequent accent marking for ambiguous words.

As an example of the types of rules output by the decision list method, one highly-ranked rule from the induced rule set reads as follows:

```
inicio (-1w,se) => inició   4.11087386417331
```

That is, **inicio** should be corrected as **inició**, if preceded by the word **se**.

For comparison, we have also evaluated the performance of a character-based restoration algorithm, available directly through the *Charlifter* program (Scannell, 2011) and based on the description in Wagacha et al. (Wagacha et al., 2006).

4.2. Training and evaluation data sets

In order to evaluate the performance of these various methods and contrast them with our current POS-tag approach we trained each on the same data set and then evaluated each model produced on three separate evaluation sets constituting varying types of diacritic error detection challenges. Our training data set was the POS-tagged portion of the Gigaword Corpus which consists of AFP newswire text (Graff, 2011), henceforth AFP. In total, it contains over 1.2 million word tokens (1,202,339). The AFP training set includes a somewhat simplified part-of-speech tag set that includes major grammatical category information—key for the POS-tag error detection algorithm.

The evaluation sets include a small subsection of the AFP data not included in the training set (23,078 tokens) and a subsection of a film dialogue corpus, currently under development, based on dialogue drawn from Spanish films (29,191 word tokens). The film data was selected from the larger film dialogue corpus given that in the original data acquired on the web, diacritic markings were only partially correct and reflect errors typically made by native speakers. The original film data set was hand-corrected by the authors in order to produce a gold-standard version for evaluation purposes. Finally, diacritic-stripped versions of both the AFP and film evaluation data sets were created and the original, partially-correct film data set was retained and included in the evaluation as a measure of the effectiveness of these restoration strategies in a real-world test case.

4.3. POS tagger method details

To apply the method proposed in this paper, we used the freely available HunPos tagger (Halácsy et al., 2007). First, the AFP corpus was augmented with the tags that mark stress placement correctness as described above, after which a trigram tagger was trained using the default

AFP Trained

Evaluation	Accuracy	Improvement
Baseline		
Film	91.97%	5.31%
Film (original)	95.27%	0.05%
AFP	98.79%	7.98%
Charlifter		
Film	92.03%	5.37%
Film (original)	92.03%	-3.19%
AFP	98.54%	9.75%
Decision List		
Film	93.26%	6.6%
Film (original)	**95.88%**	**0.66%**
AFP	98.92%	10.13%
POS		
Film	**93.27%**	**6.61%**
Film (original)	95.57%	0.35%
AFP	**99.05%**	**10.26%**

Table 2: Accuracy scores for all restoration methods and evaluation sets.

options.[4] This tagger was then used to provide information about whether a word in the test set was correctly or incorrectly stressed—i.e. the output of the tagger would simply be a sequence of tags, where each tag also indicates correctness as in Table 1 above.

5. Results

We applied the training models from each of the four detection/restoration methods (baseline, *Charlifter*, decision list and POS) to each of the three evaluation data sets (AFP, Film and Film (original)) and compared the restored versions to normalized versions reporting an overall accuracy as a percentage.

Base scores for the three evaluation sets differ with the original film data which already starts at over 95% accuracy—as expected, the stripped film set showing the most diacritic ambiguity at only 87% of all words being correctly stressed, and the AFP evaluation set closer to 89% accuracy from the outset as seen in Table 2. Applying a simple unigram frequency as a baseline, accuracy rates show relatively high improvement scores for the stripped data sets (AFP: 7.98%, film 5.31%). In fact, the improvement is quite dramatic for the AFP evaluation set which reaches 98.79% accuracy with the baseline strategy whereas the partially correct film data improves very minimally from its starting base score.

POS

Evaluation	Accuracy	Improvement
AFP		
Film	93.27%	6.61%
Film (original)	95.57%	0.35%
AFP	99.05%	10.26%
Ancora		
Film	94.53%	7.87%
Film (original)	96.89%	1.67%
AFP	97.4%	8.61%

Table 3: Accuracy and improvement scores for POS-restoration method trained on AFP and ANCORA data sets.

Comparing the other three methods to the baseline scores, we see that both decision list and POS approaches match or improve on baseline accuracy whereas *Charlifter* is at or below baseline accuracy overall. When evaluated on both stripped AFP and film data sets *Charlifter* fares somewhat better, in particular for the AFP evaluation set. Performance, however, on the original, partially-correct film data actually increases orthographic errors (-3.19%) suggesting that the language contained in the AFP training set fails to inform the grapheme-based algorithm precisely where residual human-based errors occur.

For the decision list and POS-tag methods overall accuracy scores are quite similar. Performance differences are negligible for stripped film at 93.26% and 93.27% respectively, whereas for the AFP data set the POS (99.05%) shows a slight advantage over the decision list (98.92%) and the opposite being true for the original film data, a difference of less than 1% (.31%).

5.1. POS-method sensitivity to POS-tag granularity

Findings from this set of experiments suggests that, minimally, the POS-tag approach achieves similar accuracy levels as the decision list approach on a large training data set with a relatively shallow (AFP) POS tag set. As an exploratory step to assess the extent to which part-of-speech information can be leveraged from training data sets with richer POS tag specifications, the same POS-tag implementation (POS) was trained using the ANCORA corpus (Taulé et al., 2008), a part-of-speech tagged corpus of Spanish drawn from newswire text consisting of roughly 100k word tokens and 'manually corrected', and evaluated on the same three sets as in the previous round of experiments. Notably, the annotation of the ANCORA corpus is much more fine-grained (including tense and person information missing from AFP) and thus one expects better results in the error detection/correction task. Naturally, this advantage will be somewhat neutralized by the small size of the training data (100k tokens vs. 1.2 million tokens in AFP).

Comparing accuracy and improvement results (Ta-

[4] The default options train a trigram tagger that also trains a separate suffix tree to help tag previously unseen words. This obviously has great bearing on the ability to generalize to errors in unseen words as many competing diacritizations are found on the suffixes of words, as in the **-ara/-ará** subjunctive vs. future example given above. The HunPos tagger, unlike standard trigram taggers which condition pos-word pairs only on the previous two tags, also conditions its tagging choice on the previous word.

ble 3) from the AFP and ANCORA trained POS-based restorer, we see that while AFP training/evaluation scores both show very high improvement (10.26%) and accuracy (99.05%), the POS-tagger restorer trained on ANCORA outperforms the AFP-trained POS-based restorer in accuracy and improvement for the film data sets by 1.29%. This latter result is informative in that while the lower performance on the AFP test set is to be expected because of a genre mismatch—naturally a restoration strategy trained on a separate part of the same corpus as it is evaluated on will yield better results—however, the better performance on the 'real-world' film corpus restoration task is directly attributable to the more fine-grained nature of the POS tags in the ANCORA corpus.

6. Discussion and further work

As the results show, leveraging a POS tagger to provide judgment on correctness of accentuation marks and diacritics is a strategy that appears to be competitive with other approaches to diacritic restoration. One naturally expects the best results whenever targeting a language that has high correlation between different diacritizations and the information provided by a POS tagger. In this scenario, Spanish (and potentially other Romance languages including French, Italian, and Portuguese) appear suited for this approach. The foremost advantage of the method is that it is very simple to apply, given access to a tagged corpus and a POS tagger. Also, the size of the language model induced from training data may be substantially smaller: our POS-tagger model occupied 10Mb while the decision list induced from the same data set occupied 257Mb. The adaptability of the method to other languages warrants further testing. Similarly, combinations of decision list and POS-tagger-based methods may be profitable, if the resources are available.

Additionally, many of the remaining errors in both the POS-tagger-based method and the decision list method are easily disambiguated by humans using strictly local contextual information. In other words, despite a fairly large training set, many 'obvious' generalizations remain unseen in the training data. This suggests that a hybrid method based on a set of hand-written rules may be profitably combined with the existing methods.

Also, the POS-based method presented here appears to generalize better across genres than a decision list -based method. This robustness is seen in the second experiment where training on a very small tagged corpus using a fine-grained tagset outperformed other methods when applied to the film dialogue restoration task. This flexibility is useful in that many of the scenarios where diacritic error detection is desirable involve a genre or sublanguage that is unpredictable and where little genre-specific training data is available.

7. References

Brants, T., 2000. TnT: a statistical part-of-speech tagger. In *Proceedings of the sixth conference on Applied natural language processing*. Association for Computational Linguistics.

De Pauw, Guy, Peter W. Wagacha, and Gilles-Maurice de Schryver, 2007. Automatic Diacritic Restoration for Resource-Scarce Languages. *Text, Speech and Dialogue*:170–179.

Graff, David, 2011. Spanish Gigaword Third Edition (LDC2011T12). *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA*.

Halácsy, P., A. Kornai, and C. Oravecz, 2007. HunPos: an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics.

Mihalcea, Rada F., 2002. Diacritics Restoration: Learning from Letters versus Learning from Words. *Computational Linguistics and Intelligent Text*.

Paredes, Florentino, 1999. La ortografía en las encuestas de disponibilidad léxica. *Reale*, 11:75–97.

Scannell, Kevin P, 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, incorporating Cleaneval*:5.

Scannell, Kevin P, 2011. Statistical unicodification of African languages. *Language resources and evaluation*.

Simard, Michel and Alexandre Deslauriers, 2001. Real-time automatic insertion of accents in French text. *Natural Language Engineering*, 7(02).

Taulé, M., M.A. Martí, and M. Recasens, 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*.

Tufiş, Dan and Alexandru Ceauşu, 2008. DIAC+: A professional diacritics recovering system. *Proceedings of the Sixth International Language Resources and Evaluation (LREC)*.

Ungurean, Cătălin, Dragoş Burileanu, Vladimir Popescu, Cristian Negrescu, and Aurelian Dervis, 2008. Automatic Diacritic Restoration For A TTS-based E-mail Reader Application. *Bulletin, Series C:*, 70:3–12.

Wagacha, Peter W., Guy De Pauw, and Pauline W Githinji, 2006. A Grapheme-Based Approach for Accent Restoration in Gikuyu. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*:1937–1940.

Yarowsky, David, 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*:88–95.

Yarowsky, David, 1999. Corpus-based Techniques For Restoring Accents In Spanish And French Text. *Natural language processing using very large*.