A Simple Spanish Part of Speech Tagger for Detection and Correction of Accentuation Error

S. N. Galicia-Haro, I. A. Bolshakov, and A. F. Gelbukh

Laboratorio de Lenguaje Natural, C.I.C., I.P.N.
Av. Juan de Dios Bátiz. AP 75-476, CP 07738, Mexico City, Mexico
{igor, gelbukh, sofia}@pollux.cic.ipn.mx

Abstract. One of the most frequent kind of typographic errors specific to Spanish is connected with accentuation, namely, with omission of an obligatory stress mark or insertion of a superfluous one. If such an error transforms one word to another existing one, the latter cannot be detected by usual spell-checkers, since some context analysis is necessary. A simple procedure is proposed for this task. It relies on (1) some simple heuristics that determine linear context and (2) on a small list of pairs of words that differ only in accentuation mark. This idea is applied to numerous nouns or adjectives like *número* that pass to quasi-homonymous personal verb forms if they lose their stress marks.

1 Introduction

The usual spell-checkers consider each word out of its context. With such a strategy only those typographic and orthographic errors may be detected by the spell-checker, that change an existing word to a senseless letter string. As to the errors converting one existing word to another existing one, the words with them stay unnoticed and make the text in essence ungrammatical. For their reliable correction we propose some simple heuristic method to solve this problem.

In Spanish there is a peculiar source of typographic errors which are fully undetectable if the words are taken out of context. These errors are connected with Spanish accentuation rules. For example, the phrases *este artículo tiene ...* 'this article has ...', or *las páginas siguientes ...* 'the following pages ...' would be considered correct by a spell-checker even if the underlined words lost their stress marks: **este articulo tiene ...* *'this I join has ...', **las paginas siguientes ...* *'the you paginate following ...'. Indeed, words *articulo, numero*, and *paginas* are true Spanish verb forms, though fully unacceptable in the mentioned contexts.

Meanwhile, this type of errors is quite usual for foreigners. One of the authors, has made more than 60 accentuation errors in his first paper in Spanish [1], and only half of them were detected by the spell-checker of Word for Windows, version 6 [2]. Newer versions of Word for Windows use a powerful grammar checker to detect such errors, though this seems to be a task rather for a simple enough spell-checker, since for the user this is an ordinary typo.

V. Matoušek et al. (Eds.): TSD '99, LNAI 1692, pp. 219–222, 1999.

2 Linguistic Information

All error-prone Spanish words with stress marks may be divided to several non-intersecting groups. One group contains those nouns or adjectives that pass to personal verb form, if the stress mark is omitted. At the first stage of our study, the quasi-homonymous pairs "noun - verb" were gathered through the manual search in large Spanish dictionaries, such as of Academy of Madrid [3] and Anaya Group [4]. It gave us about 150 pairs then a special program was written for automatic extracting from large electronic dictionaries. Thus the full amount of pairs reached nearly 300.

The common features of the pairs are the following: 1) The pairs are Spanish words of middle statistical ranks, 2) An accentuated counterpart is a specific word form of a noun, an adjective or a vocable combining two homonyms, 3) A non-accentuated counterpart is a specific personal form of a verb in singular. So the components of each pair correlate as a word form to another word form. 4) Each accentuated form, independently of its part of speech is characterized with a specific combination of number and gender.

In our method we employ a kind of a part-of-speech tagger, that can set only one mark: "possible adjective or noun", leaving all the other words unmarked. Good part-of-speech taggers always take into account the linear context for disambiguation part-of-speech homonymy [5]. Fortunately, Spanish presents good conditions for detecting the noun or adjective context and there is grammatical agreement between the nouns, adjectives and even articles.

3 The Algorithm

The algorithm is based on a procedure described in the next section, that for a given noun or adjective $\widetilde{\omega}_0$ can determine whether the immediate 4-word context is suitable for it. We do not have such a procedure for verbs though. Then we use a technique used by some style checkers [6]: instead of checking the current situation in the text, a hypothesis is form about a possible error of the text, and check this hypothesis. The hypothesis is incompatible with the original user's text. If the hypothesis looks reasonable, a possible error is reported.

The algorithm scans the text. Each word is looked up in the two lists: the list of accentuated words, and the automatically compiled list of their non-accentuated counterparts. The characteristics of the found word, namely its gender and number, are retrieved from the list. Let the word under consideration be ω_0, and its immediate linear context be ω_{-1}, ω_1, ω_2, such that the word order is ω_{-1}, ω_0, ω_1, ω_2.

Then the work of the algorithm depends on the list in which the word was found: 1) If the word was found in the accentuated list, it is considered to be a noun or adjective and the suitability of the immediate context is checked; the variable $\widetilde{\omega}_0$ described there is set to the word ω_0 itself. If the context is *not* suitable for a noun or adjective, a possible error is reported. 2) If the word was found in the non-accentuated list, it is considered to be a verb. Since we cannot

check the context for a verb, a hypotheses is considered that the true intended word was the corresponding accentuated noun. The variable $\widetilde{\omega}_0$ is set to this accentuated counterpart, and the context is checked for this hypothetical word. If the context is suitable for it, the hypothesis is accepted and a possible error is reported.

4 Linear Context

For the algorithm, a procedure is necessary to check for the word 0 that is supposed to be a *noun* or *adjective* in the form of already known *gender* and number, whether a specific 4-word immediate linear context ω_{-1}, ω_1, ω_2, such that the word order is ω_{-1}, ω_0, ω_1, ω_2, is suitable for a noun or adjective with these gender and number. Here $\widetilde{\omega}_0$ is either the current word considered by the ω_2 main algorithm, or its accentuated counterpart.

Let Preps be the list of all simple (one-word) prepositions, Preps = {*a, de, con, por, sin, en, sobre, para, ...*}, and Dets be the list of quasi-determinatives that depends on the gender and number of ω_0 according to the following table:

	Singular	*Plural*
Masculine	un, el, este, ese, aquel, mi, tu, su, al, del, buen, mal, primer, gran	unos, los, estos, esos, aquellos, mis, tus, sus, buenos, malos, primeros, grandes
Femenine	una, la, esta, esa, aquella, mi, tu, su, buena, mala, primera, gran	unas, las, estas, esas, aquellas, mis, tus, sus, buenas, malas, primeras, grandes

Let us use the notation $u \tilde{} v$ for grammatical agreement of words u and v in gender and number, i.e., for the fact that the first word form has or could have the same gender and number as the second one. The word $\widetilde{\omega}_0$ is considered to be a noun or adjective properly used in the given context, and thus the word ω_0 is considered likely to be an error, if any of the following four conditions is satisfied:

1. $\omega_{-1} \in Dets \cup Preps$, or
2. $\omega_{-1} \tilde{} \widetilde{\omega}_0$ or
3. $\omega_1 \tilde{} \widetilde{\omega}_0$ or
4. $\omega_1 \in$ {más, menos}& $\omega_2 \tilde{} \widetilde{\omega}_0$

The tests should be carried out in the given order, that helps to cope with the combinations like el *número* y el *género* gramaticales, where the agreement is more difficult to check. Since the gender and number of $\widetilde{\omega}_0$ are known, to check the agreement in the conditions 2 to 4, it is enough to check the hypothesis that the corresponding word ω_{-1}, ω_1, ω_2 is compatible with the hypothesis about its number and gender.

5 Experimental Results

The algorithm was realized in a complete program, consisting of 27 subprograms in Pascal, including the scanner of the input text and the dialog with the user for interactive correction of the reported errors in the text.

The program was applied to several unprepared Spanish texts. Before processing, all the stress marks were removed from the text, and the text was corrected by Word for Windows, version 6. This guaranteed that the text contained only the errors that could not be detected out of context. To find such remaining errors, our program was applied. In the text [1] consisting of 9 pages, there were 35 such errors not detectable by a usual spell-checker. As much as 32 of them, or 91.4%, were detected by the program. Only one of the missed ones corresponded to quasi-homonyms from the list: **no es practica*, the correct form being *no es práctica*; two other were connected with the other words.

6 Conclusions

A simple method is proposed to detect and correct one very common type of errors in Spanish text, namely, absent or misplaced accentuation marks. Such errors arise under the following circumstances: 1) Typos or orthographic errors of native speakers, as well as "simplified" Spanish spelling, 2) Errors made by foreigners, and 3) Problems of technical nature, such as the absence of the special Spanish keyboard or problems of different encoding of the accentuated letters under different operating systems.

Because of these reasons, some large text corpora, especially the ones collected from Internet, have significantly large fragments with the accented marks totally or partially lost. Due to informal genre of such texts the use of full-featured grammar checkers is not effective for them; this makes the suggested simple approach attractive for processing of large corpora. Also, the ideas similar to the described method can be used for simple detection of other types of errors, especially those of agreement in number and gender.

References

1. Bolshakov, I.A.: El modelo morfólogico formal para sustantivos y adjetivos en el español. Computacion y Sistemas, 1 (1996).
2. Word for Windows95. User's Guide. MicrosoftCorp. (1995).
3. RAE Diccionario de la lengua Española. Real Academia Española, Edición en CD-ROM (1996).
4. Diccionario del Español contemporáneo. Grupo ANAYA, `http://www.anaya.es`.
5. Cutting, D., et al.: A Practical Part-of-Speech Tagger. In: Proceedings of the Third Conference on Applied Natural Language Processing. Trento, Italy (ACL) (1992).
6. Ashmanov, I.: Grammar and Style Corrector for Russian Texts (in Russian). In: Proc. Of International Workshop on Computational Linguistics and its Applications, Dialogue-95, Kazan, Russia (1995).