On Asymptotic Finite-State Error Repair*

M. Vilares[1], J. Otero[1], and J. Graña[2]

[1] Department of Computer Science, University of Vigo
Campus As Lagoas s/n, 32004 Ourense, Spain
{vilares,jop}@uvigo.es
[2] Department of Computer Science, University of A Coruña
Campus de Elviña s/n, 15071 A Coruña, Spain
grana@udc.es

Abstract. A major issue when defining the efficiency of a spelling corrector is how far we need to examine the input string to validate the repairs. We claim that regional techniques provide a performance and quality comparable to that attained by global criteria, with a significant saving in time and space.

1 Introduction

Although a lot of effort has gone into the problem of spelling correction over the years, it remains a research challenge. In particular, we are talking about a critical task in natural language processing applications for which efficiency, safety and maintenance are properties that cannot be neglected.

Most correctors assist users by offering a set of candidate repairs. So, any technique that reduces the number of candidates for correction will show an improvement in efficiency that should not have side effects on safety. Towards this aim, we focus on limiting the size of the repair region [2], in contrast to previous global proposals [1]. Our goal now is to evaluate our proposal, examining the error context to later validate repairs by tentatively recognizing ahead, avoiding cascaded errors and corroborating previous theoretical results.

2 Asymptotic behavior

We introduce some preliminary tests illustrating that our proposal provides a quality similar to that of global approaches with a significant reduction in cost, only equivalent to that provided by global approaches in the worst case. To do it, we choose to work with Spanish, a language with a highly complex conjugation paradigm, gender and number inflection. The lexicon has 514.781 words, recognized by a *finite automaton* (FA) containing 58.170 states connected by 153.599 transitions, from which we have selected a representative sample

* Research partially supported by the Spanish Government under projects TIC2000-0370-C02-01 and HP2002-0081, and the Autonomous Government of Galicia under projects PGIDIT03SIN30501PR and PGIDIT02SIN01E.

that follows the length distribution of the words in the lexicon. For each length-category, a random number of errors have been generated in random positions.

We compare our proposal with the Savary's global approach [1], to the best of our knowledge, the most efficient method of error-tolerant look-up in finite-state dictionaries. We consider the set of calculations associated to a transition in the FA, that we call *item*, as the unit to measure the computational effort. Finally, the *precision* will reflect when the correction attended by the user is provided.

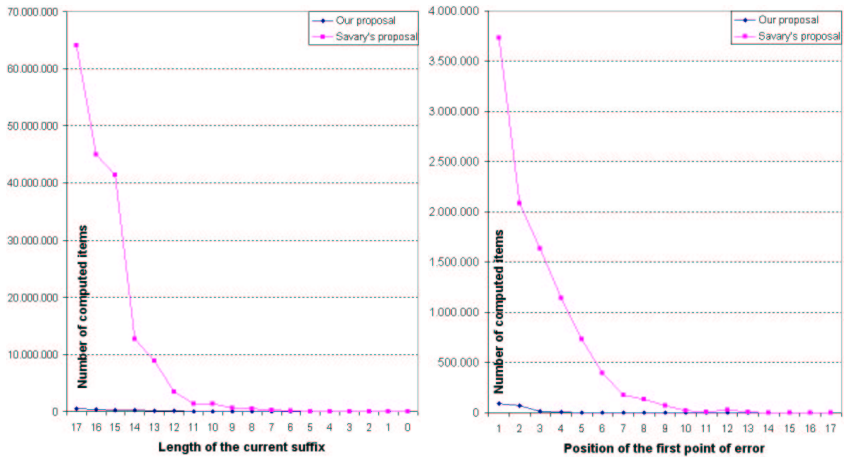

Fig. 1. Number of items generated in error mode.

Some preliminary results are compiled in Fig. 1. The graphic illustrates our contribution from two viewpoints. First, our proposal shows a linear-like behavior, in contrast to the Savary's approach that seems to be of exponential type, resulting in an essential property: the independence of the time of response on the initial conditions for the repair process. Second, the number of items is significantly reduced when we apply our regional criterion. These tests provided a precision of 77% (resp. 81%) for the regional (resp. global) approach. The integration of linguistic information should reduce this gap, less than 4%, or even eliminate it. In effect, our regional approach only takes now into account morphological information, which has an impact in the precision, while a global technique always provides all the repair alternatives without exclusion.

References

1. A. Savary. Typographical nearest-neighbor search in a finite-state lexicon and its application to spelling correction. *Lecture Notes in Computer Science*, 2494:251–260, 2001.
2. M. Vilares, J. Otero, and J. Graña. Regional finite-state error repair. In *Proc. of the Ninth Int. Conf. on Implementation and Application of Automata (CIAA'04)*, Kingston, Canada, 2004.