A Morphological Analysis Based Method for Spelling Correction

Aduriz I., Agirre E., Alegria I., Arregi X., Arriola J.M, Artola X., Díaz de Ilarraza A., Ezeiza N., Maritxalar M., Sarasola K., Urkia M.(*)
Informatika Fakultatea, Basque Country University. P.K. 649. 20080 DONOSTIA (Basque Country)
(*) U.Z.E.I. Aldapeta, 20. 20009 DONOSTIA (Basque Country)

1 Introduction

Xuxen is a spelling checker/corrector for Basque which is going to be comercialized next year. The checker recognizes a word-form if a correct morphological breakdown is allowed. The morphological analysis is based on two-level morphology.

The correction method distinguishes between orthographic errors and typographical errors.

- Typographical errors (or misstypings) are uncognitive errors which do not follow linguistic criteria.
- Orthographic errors are cognitive errors which occur when the writer does not know or has forgotten the correct spelling for a word. They are more persistent because of their cognitive nature, they leave worse impression and, finally, its treatment is an interesting application for language standardization purposes.

2 Correction Method in Xuxen

The main problems found in designing the checking/correction strategy were:

- Due to the high level of inflection of Basque, it is impossible to store every word-form in a dictionary; therefore, the mainstream checking/correction methods were not suitable.
- Because of the recent standardization and widespread dialectal use of Basque, orthographic errors are more likely and therefore their treatment becomes critical.
- The word-forms which are generated without linguistic knowledge must be fed into the spelling checker to check whether they are correct or not.

In order to face these issues the strategy used is basically the following (see also Figure 1).

Handling orthographic errors

The treatment of orthographic errors is based on the parallel use of a two-level subsystem designed to detect misspellings previously typified. This subsystem has two main components:

- Additional two-level rules describing the most likely changes that are produced in the orthographic errors. Twenty five new rules have been defined to cover the most common orthographic errors. For instance, the rule `h:0 => V:V_V:V` describes that between vowels the h of the lexical level may dissapear in the surface. In this way `bear`, typical misspelling of `behar` (to need), will be detected and corrected.
- Additional morphemes linked to the corresponding correct ones. They describe particular errors, mainly dialectal forms. Thus, using the new entry `tikan`, dialectal form of the ablative singular, the system is able to detect and correct word-forms as `etxetikan`, `kaletikan`,... (variants of `etxetik` (from the house), `kaletik` (from the street), ...)

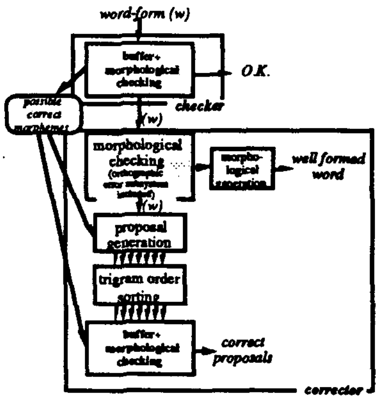
word-form (w)
buffer+ morphological checking
O.K.
checker
possible correct morphemes
(w)
morphological checking (orthographic error subsystem included)
morpho-logical generation
well formed word
(w)
proposal generation
trigram order sorting
buffer+ morphological checking
correct proposals
corrector

Figure 1 - Correcting strategy in Xuxen

When a word-form is not accepted by the checker the orthographic error subsystem is added and the system retries the morphological checking. If the incorrect form can be recognized now (1) the correct lexical level form is directly obtained and, (2) as the two-level system is bidirectional, the corrected surface form will be generated from the lexical form.

For example, the complete correction process of the word-form `beartzetikan` (from the need), would be the following:

```
  beartzetikan
      ↓ (1)
behar tze tikan(tik)
      ↓ (2)
  behartzetik
```

Handling typographical errors

The treatment of typographical errors is quite conventional and performs the following steps:

- Generating proposals to typographical errors using Damerau's classification.
- Trigram analysis. Proposals with trigrams below a certain probability treshold are discarded, while the rest are classified in order of trigramic probability.
- Spelling checking of proposals.

To speed up this treatment the following techniques have been used:

- If during the original morphological checking of the misspelled word a correct morpheme has been found, the criteria of Damerau are applied only to the unrecognized part. Moreover, on entering the proposals into the checker, the analysis starts from the state it was at the end of the last recognized morpheme.
- The number of proposals is also limited by filtering the words containing very low frequency trigrams.