The use of error tags in ARTFL's *Encyclopédie*: Does good error identification lead to good error correction?

Derrick Higgins
Department of Linguistics
University of Chicago

Abstract

Many corpora which are prime candidates for automatic error correction, such as the output of OCR software, and electronic texts incorporating markup tags, include information on which portions of the text are most likely to contain errors.

This paper describes how the error markup tag <?> is being incorporated in the spell-checking of an electronic version of Diderot's *Encyclopédie*, and evaluates whether the presence of this tag has significantly aided in *correcting* the errors which it marks. Although the usefulness of error tagging may vary from project to project, even as the precise way in which the tagging is done varies, error tagging does *not necessarily* confer any benefit in attempting to correct a given word. It may, of course, nevertheless be useful in marking errors to be fixed manually at a later stage of processing the text.

1 The *Encyclopédie*

1.1 Project Overview

The goal of this project is ultimately to detect and correct all errors in the electronic version of the 18th century French encyclopedia of Diderot and d'Alembert, a corpus of ca. 18 million words. This text is currently under development by the Project for American and French Research on the Treasury of the French Language (ARTFL); a project overview and limited sample of searchable text from the *Encyclopédie* are available at:

`http://humanities.uchicago.edu/ARTFL/projects/encyc/`.

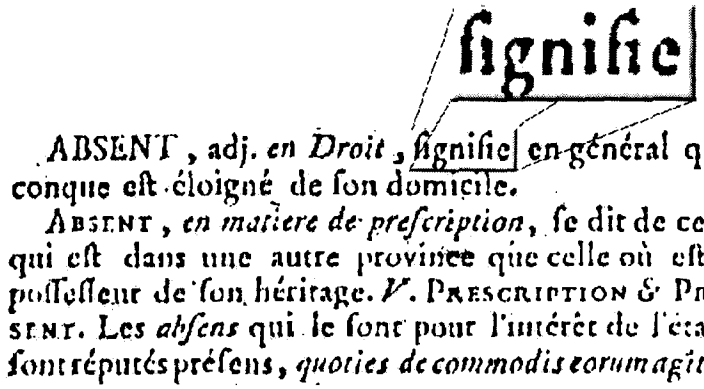
Andreev et al. (1999) also provides a thorough summary of the goals and status of the project.

The electronic text was largely transcribed from the original, although parts of it were produced by optical character recognition on scanned images. Unfortunately, whether a section of text was transcribed or produced by OCR was not recorded at the time of data capture, so that the error correction strategy cannot be made sensitive to this parameter. Judging from a small hand-checked section of the text, the error rate is fairly low; about one word in 40 contains an error. It should also be added that the version of the text with which I am working has already been subjected to some corrective measures after the initial data capture stage. For example, common and easily identifiable mistakes such as the word *enfant* showing up as *ensant* were simply globally repaired throughout the text. (The original edition of the *Encyclopédie* made use of the sharp 's', which was often confused with an 'f' during data entry—cf. Figure 1.)

At present, my focus is on non-word error detection and correction, although use of word n-grams seems to be a fairly straightforward extension to allow for the kind of context-sensitivity in error correction which has been the focus of much recent work (cf. Golding and Roth (1999), Mays et al. (1991), Yarowsky (1995)).

The approach I am pursuing is an application of Bayesian statistics. We treat the process by which the electronic text was produced as a noisy channel, and take as our goal the maximization of the probability of each input word, given a string which is the

Figure 1: Example text from the *Encyclopédie*. Note the similarity between the 'f' and the problematic sharp 's' in *signifie*


ſignifie

ABSENT, adj. *en Droit*, ſignifie en général quiconque eſt éloigné de ſon domicile.

ABSENT, *en matiere de preſcription*, ſe dit de celui qui eſt dans une autre province que celle où eſt le poſſeſſeur de ſon héritage. *V.* PRESCRIPTION & PRÉSENT. Les *abſens* qui le ſont pour l'intérêt de l'état, ſont réputés préſens, *quoties de commodis eorum agitur.*

output of the noisy channel. If we represent the correct form by w_c, and the observed form by w_o, our goal can be described as finding the w_c which maximizes $p(w_c|w_o)$, the probability that w_c is the intended form, given that w_o is the observed form.

By Bayes' rule, this can be reduced to the problem of maximizing $\frac{p(w_o|w_c)p(w_c)}{p(w_o)}$. Of course, the probability of the observed string will be constant across all candidate corrections, so the same w_c will also maximize $p(w_o|w_c)p(w_c)$.

The term $p(w_c)$ (the *prior* probability) can be estimated by doing frequency counts on a corpus. In this case, I am using an interpolated model of Good-Turing smoothed word and letter probabilities as the prior.

The term $p(w_o|w_c)$ is called the *error model*. Intuitively, it quantifies the probability of certain kinds of errors resulting from the noisy channel. It is implemented as a confusion matrix, which associates a probability with each input/output character pair, representing the probability of the input character being replaced by the output character. These probabilities can be estimated from a large corpus tagged for errors, but since I do not have access to such a source for this project, I needed to train the matrix as described in Kernighan et al. (1990).

Cf. Jurafsky and Martin (1999) for an introduction to spelling correction using confusion matrices, and Kukich (1992) for a survey of different strategies in spelling correction.

1.2 Treatment of <?>

A number of different SGML-style tags are currently used in the encoding of the *Encyclopédie*, ranging from form-related tags such as <i> (for italic text), to semantic markup tags such as <article>, to the error tag <?>, the treatment of which is the focus of this article. The data entry specification for the project prescribes the use of <?> in all cases in which the keyboard operator has any doubt as to the identity of a printed character, and also when symbols appear in the text which cannot be represented in the Latin-1 codepage (except for Greek text, which is handled by other means). Other instances of the <?> tag were produced as indications of mistakes in OCR output.

Some examples of the use of the error tag from the actual corpus include the following:

<?>	for a Hebrew ℵ
<?>dans	for *dans*
J'<?>i	for *J'ai*
ab<?>ci<?><?>es	for *abscisses*
d'autre<?>aladies	for *d'autres maladies*

The first is a case where <?> was used to mark an untypeable character. In the sec-

ond case, it was somehow inserted superfluously (most likely by OCR). In the third row, <?> stands in for a single missing character, and in the fourth it does the same, but three times in a single word. Finally, in the last row the error tag indicates the omission of multiple characters, and even a word boundary.

In fact, as Table 1 shows, words which include the error tag generally have error types which are more difficult to correct than average. Our confusion matrix-based approach is best at handling substitutions (e.g., *onfin* → *enfin*), deletions (*apellent* → *appellent*), and insertions (*asselain* → *asselin*), and cannot correct words with multiple errors at all.[1] "Unrecoverable" errors are those in which no "correction" is possible, for example, when non-ASCII symbols occur in the original.

The fact that the error tag is used to code such a wide variety of irregularities in the corpus makes it difficult to incorporate into our general error correction strategy. Since <?> so often occurred as a replacement for a single, missing character, however, I treated it as a character in the language model, but one with an extremely low probability, so that any suggested correction would have to get rid of it in order to appreciably increase the probability of the word.

In short, <?> is included in the confusion matrix as a character which may occur as the result of interference from the noisy channel, but is highly unlikely to appear independently in the language. This approach ignores the many cases of multiple errors indicated by the error tag, but these probably pose too difficult a problem for this stage of the project anyhow. The funding available for the project does not currently allow us to pursue the possibility of computer-aided error correction; rather, the program must correct as many errors as it can without human intervention. Toward this end, we are willing to sacrifice the ability to cope with more esoteric error types in order to improve the reliability of the system on other error types.

The actual performance of the spelling correction algorithm on words which include the error tag, while comparable to the performance on other words, is perhaps not as high as we might initially have hoped, given that they were already tagged as errors. Of the corrections suggested for words without <?>, 47% were accurate, while of the corrections suggested for words with <?>, 29% were accurate.[2] Actually, if we include cases in which the program correctly identified an error as "unrecoverable", and opted to make no change, the percentage for <?> suggestions rises to 71%.

It may seem that these numbers in fact undermine the thesis that the error tagging in the *Encyclopédie* was not useful in error correction. I.e., if the correction algorithm exhibits the correct behavior on 47% of untagged errors, and on 71% of tagged errors, it seems that the tags are helping out somewhat. Actually, this is not the case. First, we should not give the same weight to correct behavior on unrecoverable errors (which means giving up on correction) and correct behavior on other errors (which means actually finding the correct form). Second, the tagged errors are often simply 'worse' than untagged errors, so that even if the OCR or keyboard operator had made a guess at the correct form, they would have easily been identifiable as errors, and even errors of a certain type. For example, I maintain that the form `ab<?>ci<?><?>es` would have been no more difficult to correct had it occurred instead as `abfciffes`.

2 Conclusion

In sum, the errors which are marked with the <?> tag in the electronic version of the

[1] Actually, it does have a mechanism for dealing with cases such as `ab<?>ci<?><?>es`, in which the error tag occurs multiple times, but stands for a single letter in each case.

[2] I admit that these numbers may seem low, but bear in mind that the percentage reflects the accuracy of the *first guess* made by the system, since its operation is required to be entirely automatic. Furthermore, the correction task is made more difficult by the fact that the corpus is an encyclopedia, which contains more infrequent words and proper names than most corpora.

	Substitution	Deletion	Insertion	Word-breaking	Multiple	Unrecoverable
Contains <?>	37.4%	0%	2.2%	0%	16.5%	44%
Does not contain <?>	58.5%	11.6%	6.8%	12.9%	10.2%	0%

Table 1: Breakdown of error types, according to whether the word contains <?>

Encyclopédie encompass so many distinct error types, and errors of such difficulty, that it is hard to come up with corrections for many of them without human intervention. For this reason, experience with the *Encyclopédie* project suggests that error tagging is not necessarily a great aid in performing automatic error correction.

There is certainly a great deal of room for further investigation into the use of metadata in spelling correction in general, however. While the error tag is a somewhat unique member of the tagset, in that it typically flags a subpart of a word, rather than a string of words, this should not be taken to mean that it is the only tag which could be employed in spelling correction. If nothing else, "wider-scope" markup tags can be helpful in determining when certain parts of the corpus should not be seen as representative of the language model, or should be seen as representative of a distinct language model. (For example, the italic tag <i>. often marks Latin text in the *Encyclopédie.*)

Ultimately, I believe that what is needed in order for text tagging to be useful in error correction is a recognition that the tagset will influence the correction process. Tags which are applied in such a way as to delimit sections of text which are relevant to correction (such as names, equations, and foreign language text), will be of greater use than tags which represent a mixture of such classes. Error tagging in particular should be most useful if it does not conflate quite distinct things that may be "wrong" with a text, such as illegibility of the original, unrenderable symbols, and OCR inaccuracies. Such considerations are certainly relevant in the evaluation of emerging text encoding standards, such as the specification of the Text Encoding Initiative.

References

Leonid Andreev, Jack Iverson, and Mark Olsen. 1999. Re-engineering a war-machine: *ARTFL's Encyclopédie. Literary and Linguistic Computing*, 14(1):11–28.

Denis Diderot and Jean Le Rond d'Alembert, editors. 1976 [1751–1765]. *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers.* Research Publications, New Haven, Conn. Microfilm.

Andrew R. Golding and Dan Roth. 1999. A winnow-based approach to context-sensitive spelling correction. *Machine Learning*, 34(1):107–130.

Daniel Jurafsky and James Martin. 1999. *Speech and Language Processing: An Introduction to Speech Recognition, Natural Language Processing and Computational Linguistics.* Prentice Hall.

M. D. Kernighan, K. W. Church, and W. A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, volume 2, pages 205–211.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

Eric Mays, Fred J. Damerau, and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing & Management*, 27(5):517–522.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd An-*

nual Meeting of the Association for Computational Linguistics, volume 33, pages 189–196.