AUTOMATIC ERROR CORRECTION IN TEXT

Victoria Bobicev, Anatol Popescu, Tatiana Zidraşco

Technical University of Moldova

Abstract: Letter sequence statistics in text is used to solve many text processing tasks. One of the most difficult tasks is automatic error correction in text. Errors appear in the text when it is either typed or scanned. Four types of errors can usually be found in a typed text: one letter missing, one letter extra, transposition of two letters and one letter wrong.

Absolutely different types of errors appear when a text is scanned. Generally speaking error types vary and depend on documents quality, font type and text recognition program.

To find the erroneous word a dictionary is usually used. Since many text elements are personal names, abbreviations, abridgements, firm names, they can not be found in the dictionary and do not need to be corrected. That is why a block determining these elements is necessary. We determined erroneous words according to their entropy on the letter trigram statistical model basis. We found that almost all words with the entropy higher than 4,5 are erroneous.

When the most frequent errors were analyzed the confusion table was created to determine the correct word. The word with minimal entropy is considered to be correct.

Key words: automatic error correction, letter statistics, entropy.

INTRODUCTION

The statistics of letter sequences in text is of considerable theoretical and practical interest. Letters appearance statistics is used in programs of text compression, cryptography, automatic language detection and diacritics restoration.

One of the most difficult and unsolved so far is the problem of error corrections in text. In **[Damerau, 1964]** it was found, that **80 %** of all typing errors in text are the results of one of the following four errors: one letter missing, one letter extra, transposition of two letters and one letter wrong. However, when correcting errors, a set of problems appear:

- detected error can be not an error at all, but, for example, an abbreviation, name of a person or foreign word;
- a error can be not detected at all, if, for example, instead of verb **is** preposition **in** has falsely appeared;

- a error can be of other type, for example, two words stuck together, that frequently happens, a replacement of two symbols by one (**clamper - damper**) or, on the contrary, one symbol by two (**automat-autoniat**).

STATISTICS OF OCR AND MISTYPING IN MY CORPORA

Corpus of texts used for the investigations in this work contains 885 documents that represent The Appeal Court decisions (http://moldova.wjin.net) further mentioned as **Hot;**

Frequencies of chains of two, three and more letters represent an interesting statistics that can not be presented here because of lack of space.

Corpus **Hot** represents apt illustration of scanning errors**. L**ooking up the words in the dictionary was the first step in the attempt of automatic error correction. Thus **7%** of words were defined as unknown that is not found in the dictionary. Most of them are geographical names, personal names etc., so we have to provide for a mechanism determining really erroneous words.

Scanning errors differ a lot from typing errors. For example, one does not see missing letters, if a letter is missing it is usually replaced by something else. It should be mentioned that many errors are often repeated, which helps us to create a confusion table. Still, some words contain several errors that complicates considerably the process of determining the correct word.

If a word is not found in the dictionary, it does not mean that the word is erroneous. To find erroneous words we used the entropy **[Shannon 1951].** The entropy is calculated by formula:

$$H(p,m) \approx -1/n \, (\log_2 P_m(i_1, i_2, \ldots i_n)) \qquad (1)$$

where $P_m(i_1, i_2, \ldots i_n)$ – fragment probability of the text having n elements $i_1, i_2, \ldots i_n$, calculated in the statistical model m. Entropy is a characteristics inverse to probability and measures mathematically how unexpected the given fragment is. In this case it shows the word improbability in text. Probability was calculated on the letter trigram basis. After a number of experiments made, it was found that correct word entropy is not higher than 4,5. The words with entropy higher than 4,5 are abbreviations, abridgements, personal names, foreign words and words with errors.

It is obvious that a subprogram is required in error correcting system in order to determine personal names, abbreviations and abridgements. Most of such systems **[Chinchor 1998]** are usually based on a list of rules, frames and patterns and also on the dictionary of personal names. The remaining undetected words are considered to be erroneous.

The next step is an attempt to correct the erroneous words. The confusion table proves to be very useful in this case. Typical errors for our documents are reflected in table 1.

Table 1. Typical errors for corpus documents.

Erroneous symbols	Correct symbols	Example
T	ţ	curtii - curţii
Ţ	t	hoţărîrea - hotărîrea
D	ci	dedziei - deciziei
Â	ă	stârii - stării
n'	ri	necăsăton'ţi - necăsătoriţi
I	l	parlamentuI - parlamentul
I	l	acţiuniie - acţiunile
!	l	codru! - codrul
K	ic	aplkarea - aplicarea

Having such table and calculating the word entropy, we determine the word with the minimal entropy as the correct one. The example of replacements and their entropy can be found in the table 2.

Table 2. Letter replacements and word entropy for error correction.

word	probability	entropy	
prindpiiie	2,00483E-17	4,92119	Initial word
principiiie	1,66900E-14	3,80527	
principliie	8,60778E-14	3,62321	
principilie	1,47272E-10	2,79702	
principiile	5,17026E-10	2,65765	Correct word
iiegaiă	1,37507E-20	7,68322	Initial word
ilegaiă	8,73934E-08	2,95754	
llegaiă	1,33632E-16	6,21139	
ilegală	1,81762E-07	2,84015	Correct word
propn'a	2,12362E-06	2,44611	Initial word
propria	1,17015E-05	2,17254	Correct word

It should be mentioned that words with lost diacritical marks almost always have low entropy and are considered to be correct by system (for example **catre, curtii**). On the contrary if an unnecessary diacritical mark is added the word entropy increases drastically.

CONCLUSION

One of the most difficult tasks of text processing is automatic error correction in text. Errors appear in the text when it is either typed or scanned. Four types of errors can usually be found in a typed text: one letter missing, one letter extra, transposition of two letters and one letter wrong. Absolutely different types of errors appear when a text is scanned. There are usually no missing letters, but rather a symbol instead of two letters. Also one symbol can be replaced by the other or one letter can be divided into two.

To find the erroneous word a dictionary is usually used. Since many text elements are personal names, abbreviations, abridgements, firm names, they can not be found in the dictionary and do not

need to be corrected. That is why a block determining these elements is necessary. As we had a small dictionary we determined erroneous words according to their entropy on the letter trigram statistical model basis. We found that almost all words with the entropy higher than 4,5 are erroneous.

When the most frequent errors were analyzed the confusion table was created to determine the correct word. The word with minimal entropy is considered to be correct. Unfortunately, sometimes even confusion table can not help in determining the correct word if there are several errors in the word.

It is necessary to emphasize that if system is trained on the texts of the same type with those requiring correction the result improves considerably. The best result is achieved when a part of texts is corrected manually and then the system is trained on these texts.

REFERENCES

[Марк 1913] А. А. Марков (старший). Исчисление вероятностей. Третье издание, пересмотренное и значительно дополненное. — СПб: Тип. Императ. АН, 1913. 382 с.

[Brown et al 1992] Brown P.F., Della Pietra S.A., Della Pietra V.J., Lai J.C., Mercer R.L. An estimate of an upper bound for the entropy of English. Computational Linguistics, 18(1):31-40.

[Chinchor 1998] N. Chinchor. 1998. Overview of MUC-7. In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference held in Fairfax, VA, April 29-May 1, 1998. www.muc.saic.com/muc_7_proceedings/overwiew.html

[Damerau 1964] Damerau F.J., 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3): 171-176.

[Shannon 1951] C. Shannon. Prediction and entropy of printed English. Bell Systems Technical Journal, 30:50–64, 1951.