Automatic creation and evaluation of spell matrices for a specialised corpus

Ester E.E. Ytterbrink
Department of Computer Science
Lund University, Sweden
ester@ytterbrink.nu

Abstract

This paper describes the first step in developing a system for adopting a spelling program to a certain lingual environment. A adapted version of the noisy channel model is used. The corpus is a e-mail list for people with visual impairment and is mainly in Swedish. To evaluate the spelling correction a search engine on the internet is used. That, in turn, is evaluated. The result, 21% concurrency between the spell correction and the search engine is not very impressive. One reason could be that the corpus was rather small, 2.7 million words, and that no specialization of the word list was done.

1 Introduction

Kernighan (Kernighan et al., 1990) describes *correct*, a spell correction program that was trained with a corpus collected from the Associated Press (AP) newswire. In the modern society of internet communication with differentiated tools for accessing the internet there could be a need for specializing a spell correcting function. With a corpus from a special lingual context, defined either by the texts subject or by the means of editing the text, or by both, a spell program could be adopted. Kernighan (Kernighan et al., 1990) uses confusion matrices based on a Noisy Channel Model and that technology is used here as well. The difference is that, since the training of the spelling program is thought to be a part of the end user application, minimizing the need of human resources is of essence. The administrator of a internet community that wants to adopt the spell correction feature shall not need to find lingual experts to evaluate the results.

	intended word	typo
sub[x,y]	y	x
rev[x,y]	xy	yx
del[x,y]	xy	x
add[x,y]	x	xy

Table 1: Annotation in matrices

2 The Noisy Channel Model

As a person thinks of a word and then types it, it does not always result the word that was intended. A way to model this is the Noisy Channel Model. Most typos can be described by one of four transformations of the original word. They are:

deletion one letter too little was inserted

insertion one letter too much was inserted

reversion two letters were typed in reversed order

substitution one letter was substituted for another letter

These transformations can be applied to all letters, and together they form four matrices (Table 1). Some typos can only be modified to one other word, other can be transformed into several correct words. To find the most probable word the goal is to maximize $Pr(c)Pr(t|c)$. This is the probability that a word is the word c times the probability that the typo is t when the word c was the intended word.

matrix	relevant char frequency
sub	y
rev	xy
del	xy
add	x

Table 2: Relevant char frequency for matrices

Pr(c) is computed directly from the corpus as the frequency of the word c. The possible transformations are assembled into the confusion matrices. $Pr(t|c)$ is computed from the confusion matrices and a normalizing factor that is the corpus frequency of the intended char combination (Table 2). Example: The possibility that *solt* is really supposed to be *stolt* is the frequency of *stolt* in the corpus multiplied by del[s,t] and divided by the number of times that *t* comes after a *s* in the corpus. The possibility that *solt* is really supposed to be *kolt* is the frequency of *kolt* in the corpus multiplied by sub[s,k] and divided by the number of times that *k* occurs in the corpus. As a deletion or a addition is made in the beginning of a word *x* is annotated as @ and that is treated special in these matrices. Therefor a special statistics file is created for the initial letter of the words.

3 The technique

3.1 Cleaning the corpus

The mailing list archive contains the headers of each e-mail. There are also individually formated footers for the different users. That causes a lot of text that can not be used for the training set and must therefore be filtered out. As a e-mail is replied to the previous messages are sometimes kept. Therefore one text can appear multiple times. To avoid this a perl-script is adopted to find the actual content of the e-mails and sort out full sentences. Each sentence is only counted once, to avoid that copied text is accounted for twice or more. Words that only contain numbers are discarded from the training set.

3.2 Creating the underlaying statistics

Some statistics are also computed by the perl script.

- The frequency of each character. Special characters can be grouped. Answers the question: How many b are there in the corpus?
- The frequency of pairs of chars. Answers the question: How many times is a typed after c?
- The first letter in the words. Answers the question: How many words starts with a z?
- The frequency of the words. Answers the question: How many times does 'and' occur in the corpus?

3.3 The Matrices

To compute the matrices the words from the corpus must be sorted into the categories of those that are correctly spelled and those who are assumed contains spelling errors. To do this a word list is used. This project has used dsso-1.39 (dsso.se) which is a open source wordlist for Swedish. Then the four different transformations are applied to all positions in the rejected words. If a transformation results in a word the transformation is stored as a possible correction. When all the possible transformations are evaluated the rejected words are divided into three categories. The words with no found corrections are taken aside. The misspelled words where transformations can only result in one correct word are used to initialize the confusion matrices. The initialized confusion matrices are used to find the most probable transformation of the rejected words with more than one suggestion. As shown in equation (1) the transformation that maximizes $Pr(c)Pr(t|c)$ can be found by using the frequency ($freq(c)$) of the correct word, the value in the matrix that corresponds to the transformation that turns the typo into a correct word ($matrix[x][y]$) and the char frequency table that belongs to that transformation ($char[xy]$).

$$\begin{aligned} MAX\left(Pr(c)Pr(t|c)\right) &= \\ MAX\left(freq(c)\frac{matrix[x][y]}{char[xy]}\right) \end{aligned} \tag{1}$$

These are then added to a new version of the confusion matrices, together with the corrections with only one suggestion. The new version of the matrices are then used to iterate over all the words again and this is done until there is no changes between the iterations.

3.4 The evaluation

To make the evaluation low cost in human resources internet search services are used. Trigrams with a rejected word in the center are constructed from the test set. Only trigrams where the first and the third word existed in the word list was accepted. Suggestions for a correction are computed and ranked by $Pr(c)Pr(t|c)$. The correction with the highest probability is then compared with the suggestion from the internet based suggestion. The internet based evaluation uses a search service. It performs a search for the trigram with the rejected word, and its suggested corrections, and returns the word with the highest page count.

4 The Corpus

The corpus used is the archive from a mailing list for persons with a visual impairment. That effects the vocabulary used since the text, for example, contains special terms for accessibility tools and optical diseases. It also effects the spelling errors since many of the participants uses tools such as braille sticks and screen readers. The main language of the mailing list is Swedish, but Norwegian and English also occur. The corpus is divided into a training set and a test set.

4.1 Data for the training set

The number of words in the training set are *2,713,580*. Of these there are *98,375* unique words. The number of unique words from the corpus found in dsso are *43,396*. The rest are either words with no suggestion (*37,816*) or words that can be corrected (*17,164*) .

5 Results

5.1 Results from the evaluation

Out of 199 words, that could not be found in the corpus wordlist, the spell correction and the internet based evaluation agreed on 42 words. That is just over 21 % and not very good. One possible reason is that most of the words did not need correction. Among the words, where the grading methods did not agree, there were many who could be found in the dsso, but not in the corpus. There were also words in English and the other categories of correct words that are not found in the dsso that are listed in 5.2. Since the larger search services on the internet has limits for how many searches that are allowed in a certain amount of time, a smaller search service is used. For some of the trigrams there were no results at all. That made the method less accurate. To use this in a larger scale a cooperation with a company that provides a search service would be good.

5.2 Uncorrected words

The words that are correct, but not found in dsso, mostly falls under one of the following categories:

- words from another language
- proper nouns
- compound words
- special terminology
- abbreviations

When a word can not be corrected the most common reasons are these:

- words with need of more than one transformation.
  - Sounds that consists of more than one letter and is confused with each other: axel → aksel, information → informasjon
  - Multiple typing of the same character: eller → ellerrr
  - More than one case of possible double consonants, sometimes with extra typos: adresser → aaddresser, addreser
  - Numbers are not written as words. tresidig → 3sidig
- The word never occurs correctly spelled in the corpus. That is especially for words that exists in many forms.

6 Limitations

The corpus is small. There are other mailing lists that could have been added to make it larger, but that would increase the area of specialized words. It would also decrease the specialisation. The word list, dsso, is not complete. Many of the rejected words are correct.

7 Suggested improvements

7.1 The word list

A possibility to sort out the words in the corpus that might be correct in an intelligent way would make it easier to customize the word list. In Swedish and German compound words are common and can be created in many creative ways. If all the rejected words are divided into two parts, that then are tested, a list of possible compound words could be presented for a human administrator who could add the correct ones to a special word list. Words that are very common, but not found in the word list and where no corrections are found, should also be monitored manually.

7.2 Transformations

Phonemes that consists of more than one character could be treated as one unit in some cases. A number could be substituted for its lexical representation. Some common words that are hard to spell could be treated special.

References

Mark D. Kemighan, Kenneth W. Church, William A. Gale 1990. Alternation *International Conference On Computational Linguistics archive Proceedings of the 13th conference on Computational linguistics*, 2:205 - 210