*Clive Souter, Gavin Churcher, Judith Hayes, John Hughes & Stephen Johnson**

Natural Language Identification using Corpus-Based Models

Abstract

This paper describes three approaches to the task of automatically identifying the language a text is written in. We conducted experiments to compare the success of each approach in identifying languages from a set of texts in Dutch/Friesian, English, French, Gaelic (Irish), German, Italian, Portuguese, Serbo-Croat and Spanish.

The three techniques we chose to investigate are:

i) *Unique character string identification.*

This involved finding (empirically or using linguistic 'competence') short strings of characters which are unique to each language.

ii) *Frequent word recognition.*

Another method we explored was to extract frequency ordered wordlists, and choose, say, the top 100 words for each language. Unseen text would then be analysed word by word, looking up each candidate in the list for each language, and adding to a running total or likelihood for each. At any time, or at the end of the text, we can return the most likely language.

iii) *Bigraph/trigraph based recognition.*

All possible two- and three-letter combinations are extracted from the training texts, along with their frequencies in each language. Unseen text is then analysed by similarly splitting up the text into ordered bi/trigraphs, and a running total probability for each language maintained. As in method ii), we can return the most likely language at any stage.

Each method was implemented (using the POP11 language), by training the model on roughly 100 kilobytes of text and tested on text samples which had been set aside at the outset. We varied the length of the text samples, to see how performance was affected.

The results showed the unique character string identification to be very poor, since the test samples weren't long enough to contain the unique strings (many of which are quite rare). The bigraph recognition was 88% successful, being surpassed by the 'most-

* *Clive Souter, Gavin Churcher, Judith Hayes, John Hughes & Stephen Johnson*
School of Computer Studies
University of Leeds
Leeds LS2 9JT (UK)

common-word' approach, which correctly identified the language in 91% of the test samples. However, the most successful approach was trigraph recognition, with 94%. Test length results showed an erratic improvement in success rate as the test sample got longer, with some instances of decreased success with increasing length of sample. For the bigraph model, optimal success (100%) was reached on text samples of 200 characters or more, whereas for the trigraph model, 100% success could be gained on samples of more than 175 characters.

We then went on to monitor the effect of convergence on the bi/trigraph models, to see if a fully converged model performed better than one in which only a part of the possible graphs had been learned. Perhaps surprisingly, the two models differed in this respect: The bigraph model performed best having learned only 75% of the possible bigraphs in the languages. Whereas, for a trigraph model, only 25-50% of the graphs need to be learned to achieve optimal recognition. It appears that for both models, learning the very rare graphs only serves to decrease performance (by making the models for each language more similar). Some applications of these very simple language models are in language classification, enhanced performance for speech and handwriting recognisers, and perhaps in spelling checking. We explored the correlation between the frequencies of both bi- and tri-graphs for each language, to see whether the spelling conventions of different languages reflect the accepted historical links between languages of different families. Results can be displayed graphically using dendrograms, and generally back up the family trees of proto-Indo-European that historical linguists have proposed. English and French appear to be less tied to their supposed Germanic and Latin roots than is traditionally accepted.

1. Introduction

Before any processing of a language such as lexical look-up or parsing can begin, it is first necessary to know which language we are dealing with. In many cases, this is obvious because there is a human expert present to identify the text. However, we may wish to identify the origin of a text and not have access to such an expert. This paper describes an experiment in the development and use of bigraph and trigraph models for automatically recognising written natural languages. The models are extracted from corpora of different languages, and then employed to identify new texts probabilistically. The models could be used in other applications, such as optical character or handwriting recognition, and spelling checking. In this study, nine languages are used; Friesian, English, French, Gaelic, German, Italian, Portuguese, Serbo-Croat, and Spanish. Machine-readable samples of each of these languages were obtained from the Oxford Text Archive at Oxford University. Approximately one tenth of the data was reserved for testing

purposes, and the remainder used for training the language models. The original training sample sizes in characters are given in Table 1 below:

Table 1. Preliminary text sizes (in bytes)

It was not known at the start of the experiment whether these samples are of sufficient size to extract an adequate probabilistic model. Indeed, part of the experiment is to determine (for each language) when the learning of new bigraphs and trigraphs converges, so an adequate sample size can be determined. We selected files for training the language models which were each approximately 100 kb (with the obvious exception of Friesian). It so happened that these samples were large enough for a bigraph model to converge for all but the Friesian data. Further text needed to be collected to train the trigraph model.

It is also intended when testing the language recogniser to ascertain after how many characters the identification process converges, and no further text need be read.

The training samples had to be automatically and manually edited to 'comment out' non-alphabetic characters and introductory description which was not part of the text. We decided to strip out any notation for accents from the texts, as such notation was inconsistent, even within one language. This was rather unfortunate, as accents are features which do distinguish languages.

2. Methods for identifying languages

Language identification can be achieved using several approaches:

2.1. Unique character strings

The simplest identification technique might be to find a string of characters in the Latin alphabet which are unique to a particular language. For instance, we might suppose that word initial LL is unique to Welsh, or that CZY is unique to Polish. The task is to find strings of characters

which are unique to each of the languages we wish to identify. This could be done by soliciting the opinion of expert linguists, who may be able to suggest candidates to be tested against the real data, or by searching through text samples iteratively to find a string of characters which only occurred in one language. Set out in Table 2 are some pairs of characters (bigraphs) which were found empirically to be unique to one language in the original text samples used.

Table 2. Bigraphs unique to the languages shown

Clearly, there are problems with just using two-letter combinations, since native English speakers know that *nx* occurs in the words *anxious* and *anxiety.* These words must simply have not been found in the English training material. There were no two-letter combinations unique to Portuguese found in the training data, so it would be impossible to identify Portuguese using this method. Furthermore, there is no guarantee that a string which is unique to one of the nine languages used here would remain so were other languages to be included. Longer sequences of characters are needed to uniquely identify a language, but the problem with such longer sequences is that they tend to occur more rarely. Consequently, several hundreds or even thousands of sentences may have to be read before the language can be identified uniquely using this technique.

2.2. Frequent word recognition

An alternative method is to extract a frequency ordered list of the words in each language from the training material. Then, the most frequent

words in each language can be used as a test list against which the words in a new, unknown sentence can be matched. Some of the words in the unknown text will hopefully be found in the test list. For instance, the word *the* would be found to belong to at least English and French, but would be more frequent in English. Having read in the whole of the unknown text, a profile of possible languages can be constructed, and probability values for each language calculated. This technique should provide a solution even with small quantities of the language to be identified, such as just one sentence. However, its accuracy is strongly dependent on the size of the test list, and may still require several sentences of unknown material before reaching a reliable solution.

2.3. Bigraph/Trigraph based recognition

A compromise between these first two methods is to extract all the possible two- and three-letter combinations from the training material, along with frequencies of these combinations for each language. We will use the terms bigraph and trigraph to refer to such combinations, to avoid confusion with the terms bigram and trigram, which are used for word (and often grammatical tag) combinations. Part of the table of two-letter combinations is shown below (Table 3), with the frequencies for each language represented as a percentage of the total number of bigraphs read in the training sample of that language. Note that the blank space character has been used as a 'letter' in this experiment:

Table 3. Sample bigraphs and their percentage frequencies

A similar table for trigraphs has been extracted from the training material. These tables can then be used to identify the language of an unknown sentence of input. We anticipated that these will prove to be a more accurate method than those mentioned in sections 1 and 2. In the unknown sentence, every two (or three) letter combination will provide a probability profile for each of the nine languages, and after only a few letters, a potential solution can be obtained. The accuracy of the solution will increase as more unknown words are read in. Using a trigraph rather than a bigraph model should also improve the accuracy of the language identifier. The success of each of the identification methods described here depends directly on the 'representativeness' of the training language samples, or at least on how similar the genre of the test material is to that of the training material. All the original training material came from written media or literary texts.

3. Designing the recognisers

The process of recognition using unique strings is straightforward, so it will not be described.

For the 'most-common-words' approach, the recogniser read input a word at a time, looked up the word in the table of frequent words, and if the word was found to belong to any of the languages, simply increased a counter for that language by one. At any point during the recognition process, it would be possible to return the current most likely language, but the recogniser proceeded to the end of the test sample before delivering its result.

For the bigraph and trigraph-based recognisers, quite a naive statistical approach was adopted. After each graph was read in, the table of percentages for each language (which had been extracted from the training data) was consulted, and the percentages simply added to a running total for each language. The running total is itself converted to a percentage of the grand total for all languages when returning a result. Again, at any point during the recognition process, it would be possible to return the current most likely language, but the recogniser proceeded to the end of the test sample before delivering its result. We are aware that this statistical approach is very basic, but are interested to see how well a simple model can perform.

4. Recognition Results

Each of the four methods of identification described above have been implemented using POP11. The test data consisted of 4 files each containing 45 samples (five from each language). Two files contained short text samples and two longer samples (on average 10 and 70 UNIX 'words' respectively). Similarly, two files contained unseen data and two contained data from which the models had been trained (seen data). The results of the tests for each method of recognition over all languages are summarised in Table 4 (percentage success rates).

Table 4. Recognition results for preliminary data

As we suspected, the unique strings method proved unsuccessful, only achieving 24% success overall, because in many cases the test material did not contain any of the unique letter sequences. The bigraph method and the most- common-word method were both quite successful on the test data, and each strangely performed better on unseen data than that on which it had been trained. We can only surmise that the unseen test data by chance happened to be more representative of the training data as a whole than the test samples chosen from the training data. The trigraph model performed best, achieving the highest results in each category. In fact, for the long unseen test file, the trigraph model was 100% successful in identifying the 45 language samples. Some languages were more easily recognised than others, as can be seen in the Tables 5 and 6, showing bigraph and trigraph recognition only, for which the recogniser was set to give the rank order of the target languages, most likely first. For instance, the bigraph result for Friesian shows that 19 times out of 20 Friesian was correctly recognised, but on one occasion, it came second behind some other language.

Table 5. Bigraph results per language

Table 6. Trigraph results per language

The recognition of Portuguese using the bigraph model was very poor. We have already noted that there are no bigraphs which are unique to Portuguese in the training data. In most cases the recogniser offered Spanish instead as the most likely language. However, the recognition of Portuguese improves markedly using the trigraph model, at the expense of the Spanish. In general, the effect of using a trigraph recogniser is to improve accuracy overall, with the exception of Spanish, and a very slight downturn in the recognition of French and Serbo-Croat.

5. Studying the effect of convergence in the training models

It was clear that not all the bigraphs had been learned using training files of about 100 kilobytes, since we could identify bigraphs which we knew existed in a particular language but had not been captured. So we set about monitoring the convergence of both the bigraph and the trigraph model. We arbitrarily defined convergence as follows: we monitor in training how many graphs are learned for every 1,000 graphs read. For a chosen number of 1,000 graph samples of text, we calculate a moving average of the total new graphs learnt. The chosen number is usually set at 3 or 4. If the moving average is zero for 3 or 4 samples of 1,000 graphs, we declare the model to have converged.

For example, if we observe the following totals for a series of 1,000 graph samples:

100 80 85 60 20 10 0 0

Then if we set the moving average window to be only 2, the resulting moving average totals would be:

90 82.5 72.5 40 15 5 0

Using moving averages rather than raw totals has the effect of smoothing out peaks in the data. We can vary the moving average level to more than 3 if we wish to allow a very long tail in the convergence, or set it to less than 3 if we wish to have a smaller model (for instance in the case where the text genre contains many acronyms, which would cause the model to converge slowly). Note that this definition does not preclude an incomplete graph model, since a sudden increase in new graphs could occur, for example when beginning training on text from a new genre.

We began monitoring convergence with the English text, to give us an estimate of how much text would be needed for the bigraph and the trigraph model. Using the definition of convergence given above, the English trigraph model converged after reading 346,000 trigraph tokens. As a consequence, it was clearly necessary for us to supplement our training data for seven of the nine languages. We had no shortage of English or French material, but could not be sure whether a particular model would be the same size for each language, or whether the model would converge at different rates for different languages. We obtained further on-line texts for each language (except Friesian) primarily by

collecting ‘ftp’-able material from archives and from the USENET bulletin boards for each language. Material collected from bulletin boards had to be edited to remove headers and verify that it contained only the target language. It was not possible to expand our collection of Friesian, so we began a new collection of modern Dutch. (At one point we included Dutch and Friesian in the Dutch training data, but discovered that in testing, all Friesian examples were being classified as English, so we decided to separate them out from the modern Dutch, and ignore them for the rest of the experiment).

Table 7 shows the total text available for the remaining experiments, once we had collected supplementary material:

Table 7. Total amount of training text available (in characters/ words)

Tables 8 and 9 give details of training convergence of the bigraph and trigraph models for the nine languages.

Table 8. Convergence of bigraph models for each language

Table 9. Convergence of trigraph models for each language

The bigraph models converge to varying totals of bigraph types for each language, ranging from only 269 (Italian) to 513 (Dutch). A language model which used all the possible bigraph combinations would contain 729 bigraph types: (26 + 1) squared, including the blank space as a letter. The largest possible trigraph model is 19,683 graphs. Figures are also given to show the quartile ranges, and how many bigraph tokens had to be read before each model converged. The number of bigraphs read cannot be converted directly into characters in the source file, because of the fact that accent markers and other non-alphabetic text characters are ignored by the training (and the testing) programs.

Note that the number of bigraphs needing to be read for convergence varies from 20000 (Spanish) to 94000 (English). It is not the case that a larger graph model will necessarily require a greater number of graph tokens to be read before reaching convergence. For example, the Dutch bigraph model converged at a total of 513 bigraph types, after reading 60000 tokens. Whereas smaller models such as English (509 types from 94000 tokens) and French (386 types from 61000 tokens) required more material to be read before convergence. Figures 1 and 2 below show the learning of bi/trigraph types (on the y axis) for English according to the number of tokens read (on the x axis multiplied by a factor of 1,000).

Figure 1. Bigraph learning for English

Figure 2. Trigraph learning for English

6. How does convergence affect language recognition?

Apart from studying convergence in the training models, we also wanted to observe the effect of a converging model on recognition success. Would a fully converged model result in improved recognition of unknown texts? A further variable to be taken into account is the length of the test material. The test material for the preliminary experiments was divided into short (approximately 60 character) texts and longer (approximately 420 character) texts. Preliminary results showed that the longer texts were more successfully identified.

We constructed some new test material from the additional training material, (and in the case of Dutch, exclusively from the modern Dutch) and divided the material into samples which were 50, 75, 100, 125, 150, 175 or 200 characters long. Each sample contained 36 texts, four from each language. Both the bigraph and trigraph recognition programs were run on this new test material, with results shown in Tables 10 and 11 below:

We can conclude from these figures that a fully or 75% converged bigraph model is slightly more accurate than one which has not yet reached this stage of convergence. However, the trigraph model appears to be more successful for language recognition when it has reached between 25-50% of its fully converged size. Learning the rarer trigraphs only adds noise to the model, and reduces its ability to distinguish between languages. The trigraph model is more successful than the bigraph model at each stage of convergence.

The effect of longer test samples improving recognition success is fairly clear in the bigraph model, being shown most strongly in the 75% and fully converged models.

However, no clear pattern emerges from the trigraph recognition. The overall success rate for longer test samples is lowered by the poor performance of the fully and 75% converged models. The picture is clearer when considering only the 25% converged model, which does tend to improve performance on longer test samples. The bigraph and trigraph models appear therefore to perform differently with respect to length of test input and convergence. Of the approaches discussed here, the ideal model to use for language recognition would seem to be a partly (25-50%) converged trigraph model, which begins to achieve 100% accuracy on texts of at least 175 characters (approximately 30 UNIX words) in length.

Table 10. Bigraph results for different length test samples

Table 11. Trigraph results for different length test samples

7. Investigating the effects of convergence in identification

Using the definition of convergence for identifying texts (as given in section 5) a number of tests were carried out to observe the behaviour of the identification process using texts of arbitrary length. For each language, ten samples were chosen pseudo-randomly from texts available. Choosing the best performing bigraph and trigraph models, the number of characters required before the system converged on the correct solution was noted. The bigraph model used was 75% of the converged model and the trigraph model used was 25% of the converged model.

The purpose of the tests is to establish limits to the identification process. Hence, lower bounds on the number of characters read before convergence should be declared can be determined.

7.1. Test results

Tables 12 and 13 show the maximum, minimum and average number of characters read for each language, by when convergence in recognition had been reached.

Table 12. Number of characters read before correct convergence for bigraph model

Table 13. Number of characters read before correct convergence for trigraph model

The tables show mixed results with a trend of fewer characters required for the trigraph model. There were two notable exceptions to this however; Italian and Serbo-croat required the same or fewer characters for the bigraph recogniser to succeed.

Calculating the average for the bigraph and trigraph results yields approximately 20 and 15 characters respectively. The lower bounds on the number of characters read before convergence is tested for can be set to these values. This will tend to ensure that the recogniser reaches the likeliest answer. For example, if convergence is tested for after every 5 characters then the first point after which a trigraph model could converge would be after 30 characters read. Since the purpose of this method is to reduce the amount of time the process takes to come to a correct solution for a large text, then thirty characters (7-8 UNIX words) would considerably reduce the identification time.

8. Using the bi/trigraph models in other applications

Apart from their obvious uses in language identification, the bi/trigraph models may be utilised in other application areas, including language classification, optical character recognition, handwriting recognition and spelling checkers. We will consider here only language classification. We can observe how closely related the writing systems of the different languages are by obtaining correlation coefficients between each language. We entered the two models (part of one of these was shown above in Table 3) into the Minitab statistical package and generated correlation values for the bigraph and trigraph models extracted from the preliminary training data. Tables 14 and 15 show correlation coefficients for the preliminary training data, but similar tables have been extracted for the fully converged models derived from the extended training data.

Table 14. Bigraph frequency correlations (at 75% of convergence).

Table 15. Trigraph frequency correlations (at 25% of converged model)

The correlation data can be represented more graphically using a clustering algorithm (here we used Ward's method) and displaying the clustering in the form of a dendrogram. Figures 3 and 4 show the changes in the clustering as the bigraph and trigraph models pass though 25%, 50% and 75% to full convergence.

Figure 3. Dendrograms for bigraph models

Figure 4. Dendrograms for trigraph models

The nearer the node joining two branches is to the right side of the diagram, the stronger the relationship between the letter combinations in those two languages. For the bigraph model, the clustering becomes stable at 75% of the size of the converged model, with Spanish and Portuguese being most closely paired. Interestingly, French and English form a group of their own related to Gaelic, and French is some distance from the other Romance languages. In the trigraph model, we still see

the strong relationship between Spanish and Portuguese, but this time they are linked to Italian and French, as one might hope from a historical linguistic viewpoint. German and Dutch are again strongly paired, with Serbo-Croat standing largely on its own at each stage.

The clusterings do tend to adhere to Indo-European family tree which have been proposed by historical linguists (e.g. Yule 1985; 168), but the looseness of the relationship between English and the other Germanic languages, and between French and the other Romance languages comes as a surprise. They also agree largely with the findings of Batagelj et al (1992) who used Ward's algorithm to cluster many more languages on the basis of only 16 chosen words!

9. Conclusions

We have explored four approaches to language identification, and found that a trigraph model is the most successful for recognising the languages included in the study. A fully converged bigraph model is more successful than one which has yet to converge, but it still just outstripped by a simple '100 most-common-words' approach. However, a far from converged trigraph model was the most successful of all. Using a trigraph model which had reached only 25-50% of its converged size on texts of at least 175 characters resulted in faultless recognition, even with a relatively simple statistical model for combining the graph frequencies. The effect of increasing the length of the test material did tend to improve recognition success, but selecting a model at the right level of convergence was the most important factor in achieving a high recognition rate.

Bigraph and trigraph models can be used to classify languages along the lines of a historical linguistic family tree for Indo-European languages, and generally support the links expressed in such trees, with some exceptions in the classification of French and English.

References

Batagelj, V., T. Pisanski and D. Kerzic (1992): Automatic clustering of languages. In: *Computational Linguistics* Vol. 18, No. 3.

Yule, G. (1985): *The study of language*. Cambridge University Press.