There has been no comparable systematic work on multilingual text documents, although there has been some work based on an Optical Character Recognition (OCR) system, such as by Tan *et al.* (1999). One attempt at multilingual identification was by Prager (1999). His Linguini system uses a vector space based monolingual identifier to also find out the component languages of a document and the relative proportions of each. Artemenko *et al.* (2006) tried a method for identifying the languages in a document and have reported an accuracy of 97% for this task. But neither of them identified languages of segments.

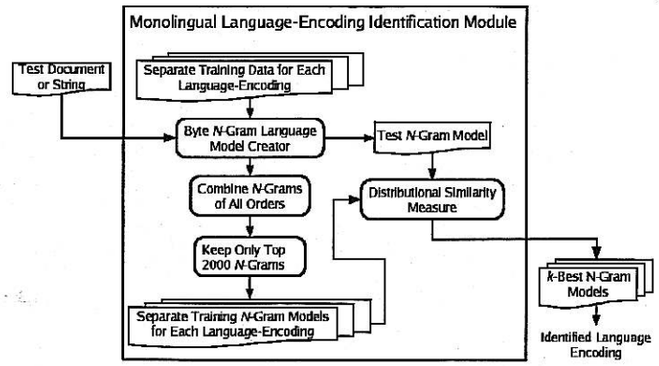

Figure 1. Monolingual Language Identification

5. Monolingual Identification

We use a monolingual identifier as a black box for multilingual identification. The method used by us for monolingual identification is based on Singh's work (Singh 2006), using symmetric cross entropy as the similarity measure. Such a monolingual identifier effectively calculates a distributional similarity score between two n-gram models. The system is trained by preparing byte based n-gram models from the training data. Then n-grams of all orders are combined and sorted by rank. Only the top N n-grams (where $N = 2000$) are retained because they are the characteristic n-grams for a language (Cavnar and Trenkle 1994).

For the given test data or string, we prepare a similar n-gram model and combine the n-grams of all orders. However, unlike for training models, we keep all the n-grams. This is because the test strings will be usually small: in our case as small as a word. This n-gram test model is then compared with all the training (or reference) n-gram models and similarity scores are calculated using symmetric cross entropy:

$$(1) \quad sim(p,q) = \sum_{x=y} (p(x) * log\ q(y) + q(y) * log\ p(x))$$

where p and q are the two distributions, or in the present case, n-gram models; x and y are the variables (n-grams) in the two distributions or n-gram models, respectively.

Now we can select the most likely language-encoding pair(s) based on this n-gram model similarity score. We kept the order of byte n-grams as 6 (instead of the usual 4 or 5) in our experiments because our method for multilingual identification depends on identifying even the small words correctly.

The process for monolingual identification has been shown in figure 1.

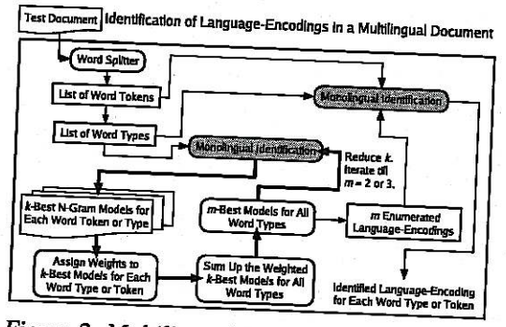

Figure 2. Multilingual Language Identification

6. Multilingual Language-Encoding Identification

There are two aspects of generalized multilingual language-encoding identification, one of which makes the problem much harder than monolingual identification. This is the fact that it is hardly possible to get representative training data for multilingual documents due to the very nature of the problem. However, the second aspect makes the problem solvable even without multilingual training data. This is the *limited ambiguity assumption* mentioned earlier: there are likely to be only two or three languages in a document in most cases.

Language-encoding identification, monolingual or multilingual, can be seen as a classification task and it can be argued that we should use some sophisticated pattern classification technique like maximum entropy for solving this task. However, the previous work and our own experience suggests that a simpler n-gram model similarity based method is more suitable for this purpose. Also, techniques like maximum entropy would require annotated multilingual training and testing data which is not easy to prepare. The advantage is that only a small amount of training data (2500-10000 words) per language-encoding is enough and we do not need any specially selected features. The data need not even be very clean. A small amount of test data (5-15 words) is enough for accurate identification even if fairly high level of diversity (60 varied language encodings pairs) is assumed.