Indirect Symbolic Correlation Approach to Unsegmented Text Recognition

G. Nagy[1], S. C. Seth[2], S. K. Mehta[3], Y. Lin[2]
[1]Rensselaer Polytechnic Institute, Troy, NY (nagy@ecse.rpi.edu)
[2]University of Nebraska-Lincoln (<seth, ylin>@cse.unl.edu)
[3]Indian Institute of Technology, Kanpur, India (skmehta@cse.iitk.ac.in)

Abstract

The new non-parametric approach to unsegmented text recognition builds two bipartite graphs that result from the feature-level and lexical comparisons of the same word against a reference string which need not include the query word. The lexical graph preserves the relative order of edges in the feature graph corresponding to correctly recognized features. This observation leads to a subgraph-matching formulation of the recognition problem. An initial implementation proves the robustness of the approach for up-to 20% noise introduced in the feature-level graph.

1. Introduction

During the last twenty years, most recognition engines for difficult to segment scripts have been built around Hidden Markov Models (HMMs). Parametric recognizers for unsegmented signals, like HMMs, are hard to train. In contrast, non-parametric classifiers, like Nearest-Neighbor, require only a labeled reference list. In this paper, we provide preliminary results in support of an entirely new method for non-parametric classification of unsegmented text. *Indirect symbolic correlation* is a general method for bringing lexical context into the recognition of unsegmented signals that represent words or phrases in printed form. It is applicable wherever segments of lexically labeled reference signals can be compared to unlabeled signals. The signal need only preserve the ordering of the alphabetic units within a word (and of words within a phrase).

The method is general in the sense that it does not depend on the signal representation or signal-matching algorithm except for the above constraint. *Indirect* means that the unknown signals can represent words for which no labeled signals are available. *Symbolic* means that the label of the unknown signal is determined by comparing the signal-level matches with lexical matches precomputed between the labels of the reference signals and a lexicon of admissible words. *Correlation* refers to both the signal-level and the lexical tally of ordered matches, which is usually accomplished by shifting one signal with respect to the other.

The nature of the underlying features is immaterial for the graph-level comparisons. Errors at the feature level can be compensated by extending the reference signal to increase the number of potential matches for each segment of the unknown signal.

Indirect Symbolic Correlation promises far-reaching benefits over HMM. It avoids parameter estimation with unstable Expectation Maximization. It can use the reference sample more efficiently than HMM, and immediately incorporate new samples into the recognition process.

2. Approach

In contrast with most pattern recognition methods, which compare an unknown signal with a labeled reference signal, we propose a graph-based *comparison of sequence comparisons*. To illustrate the concepts, consider an idealized example of cursive English script with only a six-letter subset of the alphabet.

Unknown words must belong to a lexicon of acceptable words. This list is available in some computer representation, such as UNICODE symbols. Here the *lexicon* consists of 9 words:

low, me, mole, mule, moll, mellow,
wool, loom, we.

Written samples of a subset of the lexicon are available. These samples are correctly labeled. Such a sample is usually called a training set, but here it is used only for comparisons, and called the *reference set*. There are five samples in the reference set. Every sample happens to be a different word. The reference sample does not need to be segmented at the word level; we show the inter-word blanks only for the sake of legibility.

The script font of the reference set is intended to suggest handwriting:

loom, mole, me, mule, moll.

The transcript of the reference set is also available:

```
loom, mole, me, mule, moll.
```

We wish to recognize an unknown word sample, such as

mellow.

This word does not appear in the reference set, but segments of the word, such as

m e me l ll o lo

will be similar to the corresponding segments of the reference set. In fact, some of the segments (here only individual letters) appear multiple times in the reference set. With a large reference set, there will be more matches that are multi-symbol. It is precisely the number of redundant comparisons that increase with the length of the reference set, which is the strength of this method. Even if many of the matches fail to be recognized, or there are false matches, adding samples to the reference list will reduce the error rate.

If one knew exactly where each match occurred then one could immediately identify each symbol, and hence the unknown word. However, the position of the matches cannot be known unless both the reference words and the unknown pattern are *segmented* at the character level. It is known from experience that accurate segmentation prior to recognition is a nearly impossible task in OCR.

2.1. Match Graphs

Figure 1 shows the two bipartite graphs that result from the feature-level and lexical comparisons of the same word against the reference string. The top graph (G') shows error-free matching of the feature string of the unknown word ("mellow") against the stored reference feature string. The bottom graph (G) shows the same comparison at the lexical level.

We next compare the signal-match graph G' to the lexical graph G. G' represents the correspondence of matching segments between the features of the unknown signal and a reference signal. G represents the lexical correspondences between the labels of the reference signal and a specific word of a potentially very large lexicon. New words are recognized because their constituent parts are matched by portions of the reference signal, and the order of the matches is unscrambled through the lexical comparison of the reference labels with the lexicon. An entire set of graphs similar to G are precomputed, one for each word in the lexicon. The unknown pattern is identified with the label of the lexical graph that best matches the signal graph G'.

The discretization at the feature level is arbitrary because different letters may have different widths. At the lexical level, each character has a count of 1. Although the numbering is different, ideally the sequence of matches is the same in the two comparisons of the same word. However, the lexical transcripts preserve only the relative order of the features, not their linear scale. We must therefore adopt a graph-theoretic approach for sequence comparisons instead of vector space operations.

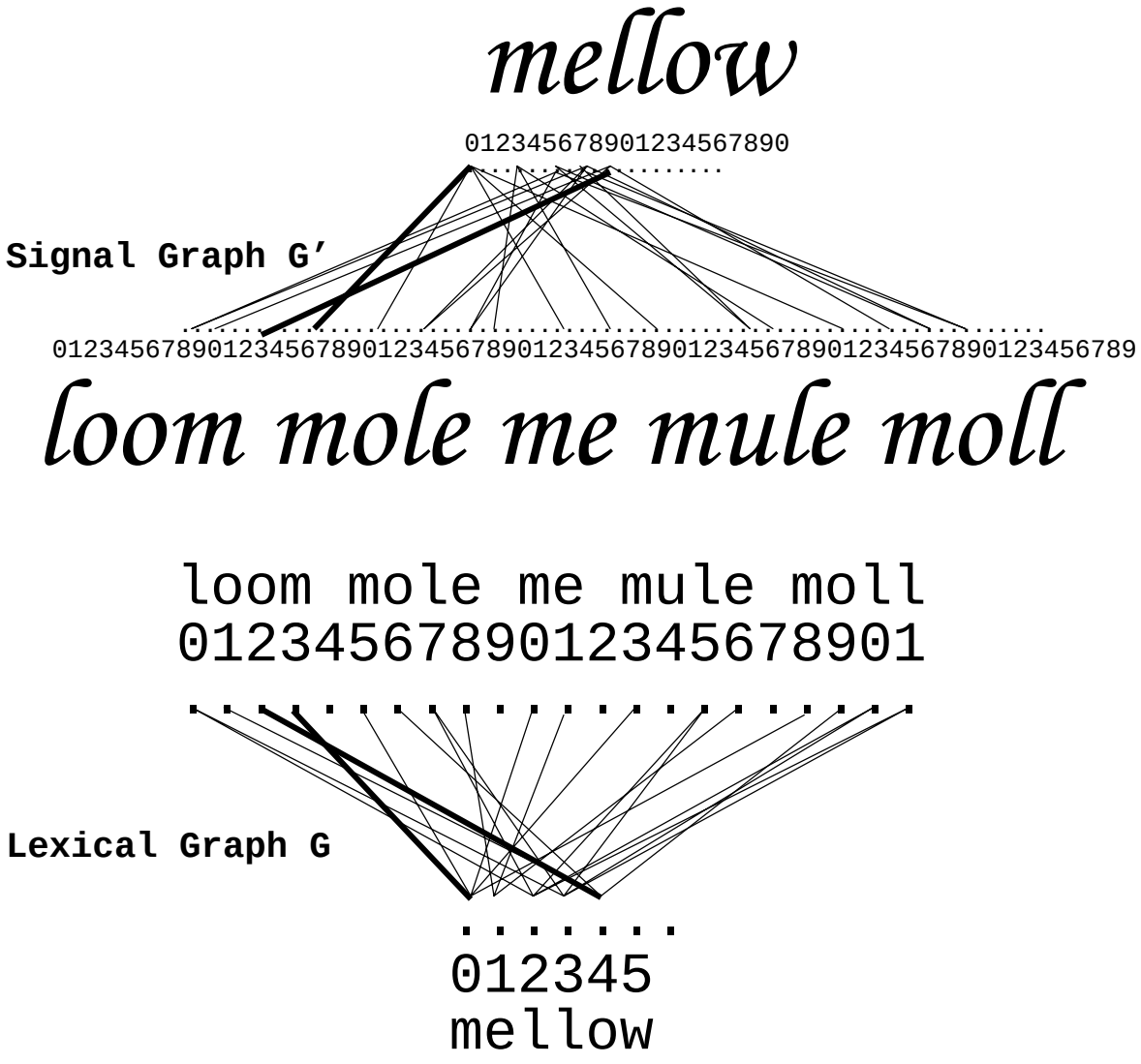
Recognizing the unknown sample requires finding a word in the lexicon that has an identical (or similar) sequence of matches as that of the unknown word with the reference feature string. In other words, when the match function is error-free, the string match set X_{mellow} is order-isomorphic to the signal match set X'_q, whereas the string match set of a different word, such as X_{wool}, is not. In Figure 1, the lexical comparison for "wool" is listed without showing its lexical graph.

The notion of *order isomorphism* can be interpreted as follows. The matches (2, 4) and (3, 0) in the string match set have the same order relation as the matches (6, 11) and (10, 0) in the signal match set. Thus, identifying (2, 4) with (6, 11) is *compatible with* identifying (3, 0) with (10, 0). (The corresponding graph edges are shown in bold in Fig. 1.) If every pair of pairs is compatible, then the two sequences are order isomorphic.

2.2. Order Isomorphism and Permutations

The order isomorphism problem can be restated as finding common patterns in two permutations. Graphically, a permutation of integers 1 to n can be represented as a bipartite graph, e.g. the permutation (2 4 1 3 6 7 5) of 1 to 7 will have the representation shown in Figure 2.

In general it is not necessary to use contiguous set of numbers to describe a permutation as long as we keep in mind the natural ordering of the numbers. So the same graph can also be represented by (5 11 3 8 16 20 14).



Lexicon L	{low, me, mole, mule, moll, mellow, wool, loom, we}
Unkown q(t)	mellow
Ref. Signal r(t)	low me mole mule moll
Ref. String r_s	low me mole mule moll
X′(q)	{(0,7), (0,9), (2,11), **(6,11), (10,0)**, (15,0), (19,11)…}
X_{mellow}	{(0,2), (0,3), (1,4), **(2,4), (3,0)**, (5,0), (6,4), (7,2), …}
X_{wool}	{(0,3), (1,1), (1,2), (2,1), (2,2), (6,1), (6, 2), (7,3), ...}
f(q)	mellow

Figure 1: example of feature-level and lexical comparisons.

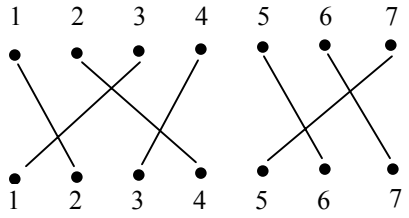


Figure 2: Permutation represented as a bipartite graph.

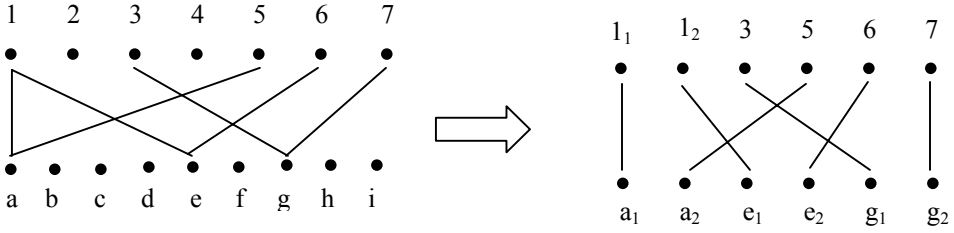


Figure 3: Conversion of a bipartite graph to permutation.

Both our signal and lexical graphs are bipartite but not always permutations because a node can have zero or multiple edges incident on it. However, we can convert them to permutations by splitting nodes with multiple edges and eliminating nodes with zero degree, as illustrated in the example in Figure 3.

The conversion process preserves the order isomorphism because the left-to-right order of edges encountered (at the top or bottom) does not change.

A sub-permutation is a subsequence of the permutation sequence. So (11 3 16 14) is a sub-permutation of (5 11 3 8 16 20 14). It is easy to observe that the graph associated with a sub-permutation is a sub-graph. Actually, there is a one-to-one correspondence between the sub-graphs and sub-permutations.

Now the order isomorphism problem can be restated as follows: Given two permutations, G1 and G2, not necessarily of the same size, find the largest common sub-graph of G1 and G2, where the size of a permutation graph is the number of its edges.

We will first restate the problem in the language of permutations:

Given two permutations $T = (T_0, T_1, \ldots, T_{n-1})$ and $P = (P_0, P_1, \ldots, P_{m-1})$ (corresponding to graphs G1 and G2 respectively). Find the *longest* sequence of pairs, $<T_{i1},P_{j1}>$, $<T_{i2},P_{j2}>$, …, $<T_{ik},P_{jk}>$, with the following properties:

(a) $0 \le i1 < i2 < \cdots < ik \le m-1,$
$0 \le j1 < j2 < \cdots < n-1,$ and

(b) Permutations $T' = (T_{i1},T_{i2},\ldots,T_{ik})$ and $P' = (P_{j1}, P_{j2},\ldots, P_{jk})$ are isomorphic.

where, T' (equivalently P') is the desired permutation. The pair of permutations will be said to represent the *best matching* permutation. In the sequel, without loss of generality, we will assume that $m \le n$.

Returning to our example in Figure 1, we obtain a compact permutation corresponding to X'_q in steps as follows:

$$(7\ 9\ 11_1\ 11_2\ 0_1\ 0_2\ 11_3) \Rightarrow (3\ 4\ 5\ 6\ 1\ 2\ 7)$$

Similarly, the permutations for X_{mellow} and X_{wool} will be (3 5 6 7 1 2 8 4) and (7 1 4 2 5 3 6 8). It can be verified that the best matching permutation between X'_q and X_{mellow} is of length 7 (all of X'_q) whereas the best

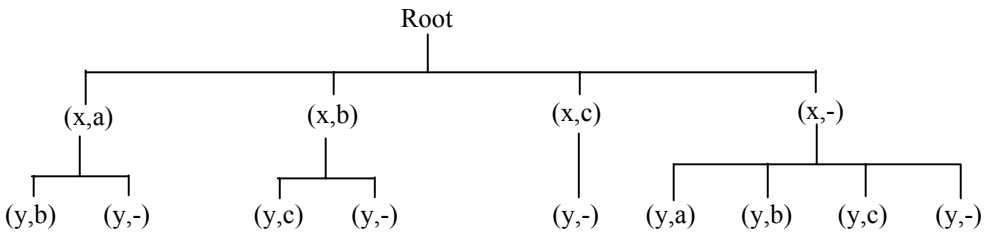


Figure 4: A host tree used for matching of permutations.

matching permutations between X'_q and Xwool is of length 5 (e.g. the sub-permutation (3 4 5 6 7) of X'_q).

3. Algorithm

Our algorithm systematically generates sequence of pairs that satisfy condition (a) by traversing a host tree (see Figure 4) in the depth-first order. Note that a node, such as (x, -), in the tree denotes a null matching. For every generated pair, condition (b) is checked for using the following observation about permutations:

Observation: Given two permutations A= $(a_1,a_2,...,a_k)$ and B = $(b_1,b_2,...,b_k)$, let t_q be the number of integers to the left of a_q which are greater that a_q, and s_q be the number of integers to the left of b_q which are greater than b_q. Then A and B are isomorphic iff $t_q = s_q$, for all q.

This observation also implies that if $(a_1,a_2,...,a_k)$ and $(b_1,b_2,...,b_k)$ are isomorphic then so are their corresponding sub-permutations. In particular, $(a_1,a_2,...,a_r)$ and $(b_1,b_2,...,b_r)$ are also isomorphic for any $r \leq k$. Now our approach would be to generate the host-tree from top to bottom (in depth-first order) and never generate a node that gives a sequence corresponding to non-isomorphic initial sub-permutations.

The example in Figure 5 shows the pruned host tree for permutations T = (6 3 2 5 1 7 4) and P = (1 5 3 4 2). In this tree we also did not generate nodes that would have resulted in a shorter than a valid sequence found earlier. Single integer in the parentheses is the length of the sequence associated with that leaf node.

The pruning of the above tree occurs at the nodes marked as *no gain.* The host tree is pruned at *no gain 1* because already three "-" pairs are formed and there is no possibility of getting a sequence of length greater than 2. In the *no gain 2* case, a sequence of length 4 has already occurred and since a parent of this node has a "-"node, no sequence of size 5 will occur in this sub-tree. In the *no gain 3* case, 1 has been matched with 5 and there are only 3 more integer after 5 in T so the sequences in this sub-tree can not be of length more than 4 and we already have seen a sequence of length 4. Thus, the resulting longest isomorphic sub-permutations in the example are T' = (2 5 7 4) and P' = (1 3 4 2).

The order-isomorphism problem can be shown to be NP-complete by reducing the pattern matching for permutations problem [1] to it. Our algorithm, however, is practical as long as the sizes of the two graphs are not too large.

4. Experimental Design

We report on preliminary experiments on a synthetic data set for the signal graphs designed to evaluate the proposed approach under the following assumptions:

(a) The signal graph represents the result of a bigram-matching process
(b) The matching process is relatively reliable -- although it may miss or misclassify some bigrams, it can identify a large fraction of them correctly.

In a related lexicon-based, indirect symbolic approach, El-Nasan et al. [2] used bigram occurrences as the basis for unsegmented text recognition. The approach proposed here imposes a stricter matching criterion than bigram occurrences: the bigrams should not only be common between the two words but also occur in the same order in the two graphs. This additional constraint can result in significantly shorter reference set.

Selection of Lexicon: We use the same 1000-word lexicon from the Brown Corpus [3] as used by El-Nasan and Nagy. The corpus contains 43,300 unique words in lower case letters, apostrophes, and a few quotation marks. The percentages of words with a

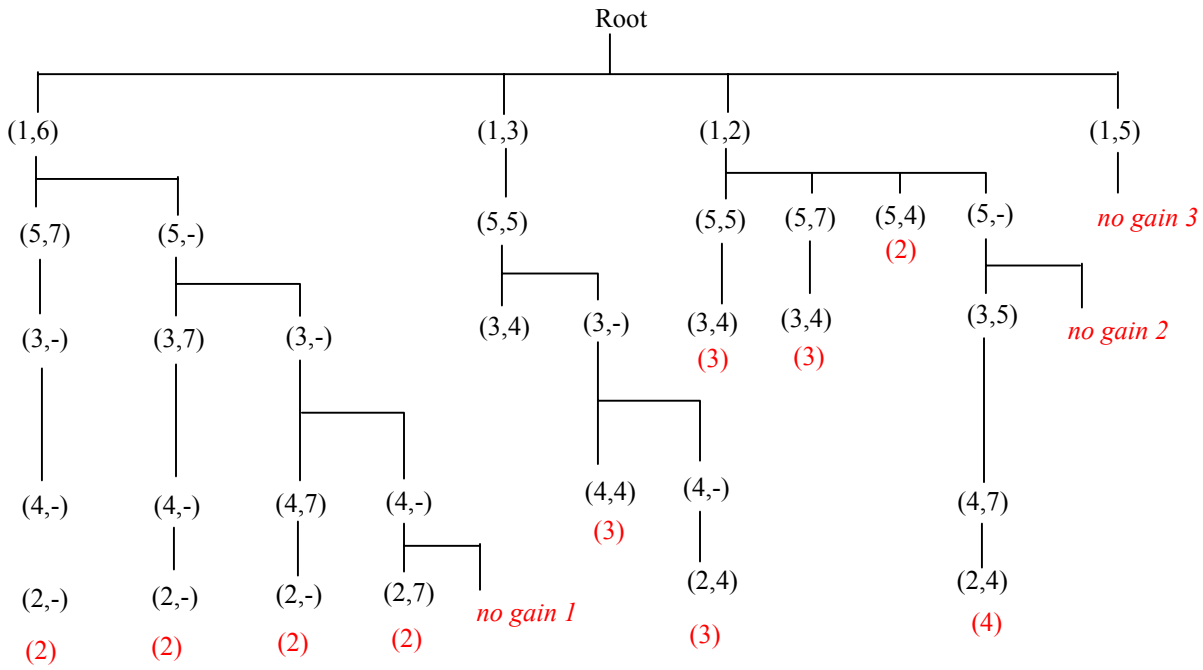


Figure 5: Pruned host tree for permutations T = (6 3 2 5 1 7 4) and P = (1 5 3 4 2).

unique set of letters, bigrams, and trigrams are 48.68%, 99.92%, and 99.99% respectively [2].

The Brown Corpus words are sorted in descending order with respect to their usage frequency. The lexicon words are the first 1000 words from the sorted list with a space character appended to the beginning and end of each word (for the purpose of doing word-level bigram analysis). As already stated, a query word is assumed to be in the lexicon.

Selection of Reference String: In our experiments, we construct reference strings by concatenating 1000 words of the lexicon in three different ways: original (sorted) order, reverse order, and random order. The first reference string starts with the most frequently used words, which are usually the *stop* words. The second reference string starts with less common words that are usually long.

The distribution of the lexical-graph size for the reference set of 1000 words is shown in Figure 6. The, average graph sizes is 305.

Noise Models: The purpose of modeling noise is to determine the effect of the frequency of feature-level errors on the overall word recognition rate. Feature level errors can be classified as follows:

False positive at position (i,j): This feature level error occurs when a bigram match is found between a query word feature string at position *i* and a reference word string at position *j*, whereas the transcript of the query word at the corresponding position *i* does not share this bigram with the transcript of the reference string at the corresponding position of *j*.

False negative at position (i,j): This feature level error occurs when a lexically shared bigram at position *i* of the query word and *j* of the reference string is not detected between the feature string of the query word and the feature string of the reference word at the corresponding position.

Further, we assume that the probabilities of false-positive and false-negative errors are position independent and denote them as $p(e|0)$ and $p(e|1)$ respectively.

We consider two noise models. Both involve one normalized parameter Q with a real value between 0 and 1 indicating the total amount of error. In the *symmetric* noise model, $p(e|0)$ and $p(e|1)$ are assumed to be equal and Q is the sum of these two probabilities.

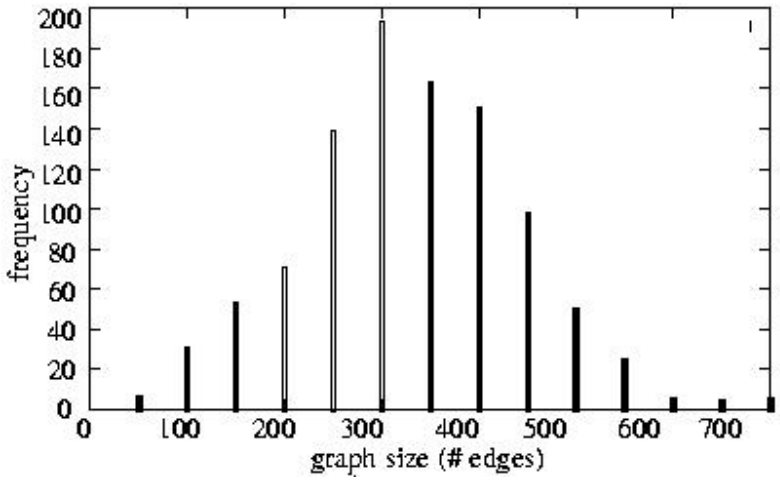


Figure 6: Distribution of the graph size of words in the lexicon

In the *weighted* noise model, the total probability Q is obtained as the weighted sum of the two probabilities:

$$Q = w_1 p(e|1) + (1 - w_1) P(e|0)$$

where the weights are defined as follows. We consider the complete bipartite graph in which the two sets of nodes correspond to character positions in the query word and the reference string. The lexical graph corresponding to the query word G_q is a subgraph of G_c. Let the size of a graph be the number of edges in the graph. Then the weight w_1 is the size of G_q, normalized with respect to the size of G_c. Typically this value is quite small because of the sparseness of G_q, therefore the second noise model heavily favors false positive errors.

Match Algorithm: The time complexity of our permutation-matching algorithm is highly data dependent. Generally, the matching time grows non-linearly with the sizes of the two graphs, as well as with their size differences. This observation suggests an iterative approach in which the reference string size grows progressively larger. Further, by choosing a prefix of the reference string initially and extending it to the right by a fixed amount in the subsequent iteration, we ensure that the two graphs used for matching in the previous step are subgraphs of the new graphs. In the results reported below, the initial reference string is chosen to be 10 words long and it is extended by another 10 words if it is necessary to take the next iterative step. At each step, only those lexicon words are matched that were not eliminated as mismatches during a previous step. Thus, the set of candidate matches can only diminish after each step.

During each iteration, the matching process starts by performing a simple check to eliminate matching two graphs of substantially different sizes. The acceptable size interval is determined differently for the each noise model, representing the 98% confidence value that the query graph is a noisy version of the lexical graph. Further pruning of the candidate matches occurs after the matching process. Every candidate has an associated matched subgraph associated with it. If the matched subgraph of a word is too small compared to the best match, that word is also eliminated from further matching. Again, the range of acceptable values is determined differently for each noise model according to another parameter representing the 98% confidence value. These algorithmic parameters can be adjusted according to the desired accuracy/speed tradeoff.

After an iteration, the matching lexicon words that survive define the current *ambiguity set*. Further iterations are tried only if the size of the ambiguity set is greater than one. We report the following outcomes of the algorithm based on the ambiguity set obtained at the end of the algorithm:

a) The ambiguity set has just one element and it corresponds to the correct match. We call it a *unique match*.
b) The ambiguity set has more than one element and it includes the correct word; we denote this situation as *rejection*.
c) Otherwise, the ambiguity set does not include the correct word; we consider this case as *misclassification*.

5. Experimental Results and Discussion

The match algorithm, as described in the last section, was implemented and run for the two noise models on a server with Sun Superscalar SPARC 9 750 MHz processors .

Figure 7 shows the coverage of uniquely matched words for the noise-free query graphs for three different ways of selecting the reference strings, as described above. The first two cases – *long words* and *stop words* – are identical to those considered by El-Nasan et al. [2] therefore this data can be compared directly with the result in Figure 2 of their paper. We note that while the general trend remains the same (the coverage for stop words rises slower than for long words) the rate of coverage in each case is faster. This demonstrates that the order isomorphism indeed results in requiring shorter reference strings for the same level of coverage.

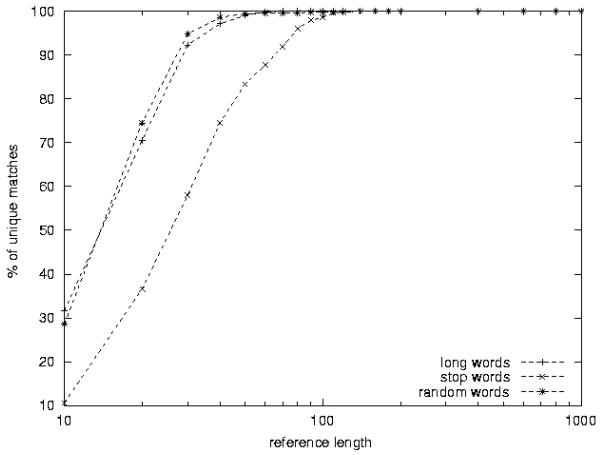


Figure 7: Coverage vs. reference-string length for noise-free data

A summary of the matching results according to the symmetric noise model appears in Table 1. These results were obtained for the reference string constructed by concatenating the lexicon words in a random order. In the table, Q is the noise parameter in the first column; the next three columns give the minimum, maximum, and median lengths of reference strings required in matching individual words; the last three columns indicate the matching performance in terms of correct recognition, misclassification, and rejection. We include the corresponding noise-free data (Q=0) for comparison. The maximum times required to match a word were 0.045 s, 215.0 s, and 76.2 s respectively for Q = 0, 0.1, and 0.2 respectively.
The results for the weighted noise model appear in Table 2.

Table 1: Results of matching for the symmetric noise model

Q	min ref length	max ref length	med ref length	unique	misclass	reject
0	10	140	20	100%	0%	0%
0.1	10	740	60	95.8%	1.5%	2.7%
0.2	10	620	70.5	87.1%	7.7%	5.2%

Table 2: Results of matching for the weighted noise model

Q	min ref length	max ref length	av ref length	unique	misclass	reject
0	10	140	20.6	100%	0%	0%
0.1	10	620	40	96.9%	1.9%	1.2%
0.2	10	680	40	83.8%	8.7%	7.5%

The correct recognition rate declines in both cases with the amount of added noise but is remarkably high for Q=0.1. We also note that with noise the required length of the reference string to match a word goes up significantly.

It is clear that for longer reference strings the error rate will be smaller, however, the exact nature of this dependence is yet to be explored. We conjecture that, even for very noisy data, the error rate can be arbitrarily reduced by increasing the size of the reference string.

References

[1] P. Bose, J. F. Buss, and A. Lubiw, "Pattern Matching for Permutations", Proc. Workshop on Algorithms and Data Structures, Lecture Notes in Computer Science 709, , Springer Verlag pp. 200-209, New York, 1993.

[2] A. El-Nasan, S. Veeramachaneni, and G. Nagy, "Word Discrimination Based on Bigram Co-occurences", 6th ICDAR, pp. 149-153, 2001.

[3] W. N. Francis and H. Kucera, "Brown Corpus Manual", available at: www.hit.uib.no/icame/brown/bcm.html, revised and amplified 1979.