Statistical Augmentation of a Database for Use in Optical Character Recognition Software Evaluation

Ann E. M. Brodeen, Frederick S. Brundick
U.S. Army Research Laboratory

Malcolm S. Taylor
OAO Corporation

Abstract

In this paper, we consider a statistical approach to augment a limited database of groundtruth documents for use in evaluating optical character recognition (OCR) software. We require groundtruth documents to assign a performance measure to the OCR component of the Forward Area Language Converter (FALCon) system. A modified moving-blocks bootstrap procedure is used to construct surrogate documents for this purpose which prove to serve effectively, and in some regards, indistinguishably, from groundtruth. The proposed method is validated through a rigorous statistical procedure.

Introduction

The Forward Area Language Converter (FALCon) is a portable, field-operated, translation system designed to assist in intelligence collection. It enables an operator with no foreign language training to convert a foreign language document into an approximate English translation for an assessment of military relevance. The principal components of FALCon are an optical scanner, an optical character recognition (OCR) module, and a machine translation (MT) module. In order to assign a performance measure to the FALCon system, measures of effectiveness of the components must be developed and then aggregated into an overall measure. The focus of this paper is limited to evaluation of the OCR module.

A current procedure for determining a quantitative measure of the efficacy of an OCR product is as follows: A selection of carefully prepared source-language documents, called groundtruth, is stored in the computer; hardcopy of the same document set is then scanned into bitmap images; the OCR software partitions a gross bitmap image into homogeneous zones that are processed according to content. For zones that are identified as text, specialized scoring software then compares the OCR output against the corresponding groundtruth to produce accuracy statistics, usually including percentage agreement for both words and characters, and a confusion matrix.[*]

A central database of groundtruth documents, accepted as a baseline, would enable the evaluation of OCR products to proceed from a common benchmark.

---

[*] A confusion matrix displays the number of character insertions, substitutions, and deletions required to reconcile the groundtruth and OCR output files.

Unfortunately, such a database does not exist, making the comparison of OCR software more difficult and any conclusions drawn more tentative. Fundamental questions regarding sample size requirements, and suitable document composition for such a database, remain to be addressed.

Collection of a corpus that is sufficient for evaluation of an OCR product is likely to remain, even in the best of circumstances, a burdensome task. Access to a sufficient number of source-language documents, representative of the document classes of interest, may not be feasible; and, even if obtained, the expensive and time-consuming process of preparing groundtruth remains. To address this problem, we are proposing a statistical approach to corpus generation based on a small set of source-language documents. Coincident with the statistical inquiry, substantial work involving language transliteration must be accomplished.

Time Series Model

Consider the passage of Serbian text shown in Figure 1. Every character—letters, punctuation marks, interword spaces—is represented numerically in the computer. The set of character and numeric equivalents (the mapping) is called a codeset. For a specific language, the codeset representation may not be unique. Russian, for example, has four commonly used 8-bit encodings and some Asian languages even more [1]. A representation of the Serbian text in Figure 1 for a particular codeset assignment is shown in Figure 2.

In Figure 2, the first 80 letters (emboldened in Figure 1) of the Serbian text are portrayed. The vertical dashed lines mark the location of interword spaces, which have been removed, along with most punctuation, to facilitate our methodology. The x-axis indexes the order of occurrence of the characters in the text, and the corresponding codeset values (numeric equivalents suppressed for presentation purposes) are plotted along the y-axis. If the characters are processed sequentially, then we can assign to each character an associated time epoch, and Figure 2 can be considered as a time series representation of the first 80 letters. The scale of measurement for the y-axis is nominal; an alternative codeset, if appropriate, would lead to a different graphic representation with no attendant loss or gain of information.

In attempting to generate a corpus, we would like a core of authentic documents to serve as a basis from which to generate additional pseudodocuments. An analogous situation, arising in the analysis of time series data collected as part of a clinical study, has been described and addressed using the bootstrap [2, 3].

Bootstrap Application

In this section, we present an abridged description of the bootstrap procedure, modified for application to the textual model. Notice the time series has an inherent structure: the time series represents a block of text—it is not a random sequence. Moreover, the words themselves are subject to lexical constraints; hence, the patterns they assume in the codeset representation have meaning. These word patterns are, however, interrupted with great frequency; the interword spaces play the role of interventions in time series modeling. As a consequence, the time series has local structure contributed by the word patterns but little in the way of global structure due to the high frequency of interventions.

"Нецес остати пусто Невесиње равно, но цес бити оно сто си вазда било: расадник Српства и колевка лава!" "Србија мора постати велико радилисте и родилисте!" Ово су само два изватка из скорасњих говоранција Вука Драсковица. Анахроницна, срцепарајуца реторика, примерена политицарима осамнаестог века, представља данас најбољи пример лази-говора или језика-маске. А онај ко на себе стави маску лази-говора, пре или касније, изабраце и лаз као основни политицки принцип. Нико није изрекао толико лази, подвала и лазних доказа о Косову као г. Драсковиц и његова телевизија. Раније смо ту анахроницну реторску маску примали као некакав његов особењацки избор, као сто примамо нецији цудацки стил у одевању. Требало је за његову реторику реци оно сто је одувек и била - да је обицан киц.
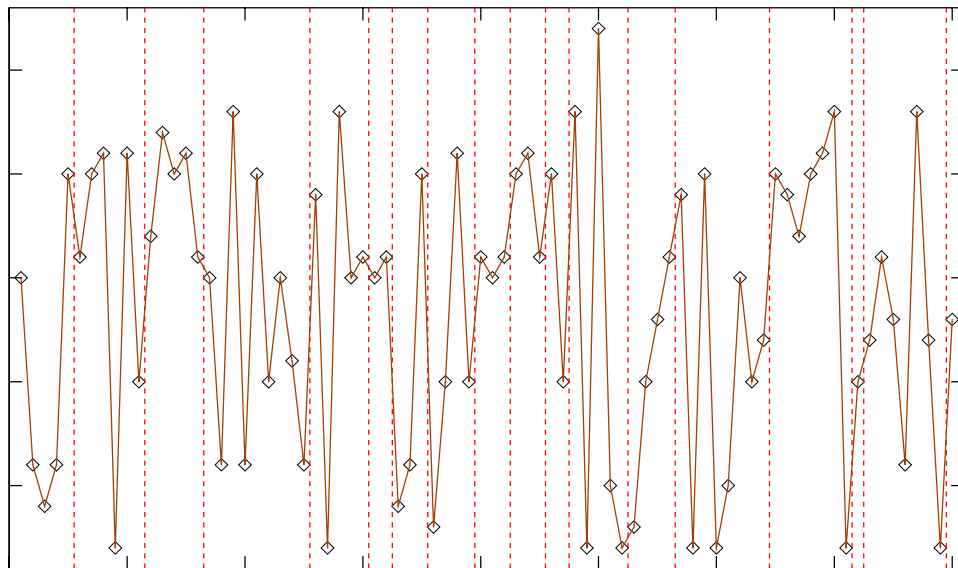
Figure 1. Serbian Text

Figure 2. Time Series Representation

Denoting the time series as a sequence of ordered pairs $(x_1, y_1), (x_2, y_2), \ldots,$ (x_n, y_n), we begin the bootstrap procedure by choosing a random location within the time series, say (x_r, y_r). Starting with (x_r, y_r), we copy the subsequence $(x_r, y_r), (x_{r+1}, y_{r+1}), \ldots, (x_{r'}, y_{r'})$ into an array. The length of the subsequence, $r' - r + 1$, is determined by sampling from the distribution of word-lengths found in the authentic document. A second random location, (x_s, y_s), is then determined, and a second subsequence, $(x_s, y_s), (x_{s+1}, y_{s+1}), \ldots, (x_{s'}, y_{s'})$, is copied and appended to the subsequence already in the array. Figure 3 illustrates a situation in which three subsequences have been chosen, two of them overlapping.[†] The overlap does not create a problem since the sampling is done with replacement. This process continues until terminated by a stopping rule. At that point, a bootstrapped time series, the first 80 values of which are shown in Figure 4, has been produced. The shaded regions appearing in Figure 3 are aligned in Figure 4 in order of their occurrence. Inverting the codeset mapping, subject to inherent lexical modeling constraints, yields the bootstrap document shown in Figure 5.
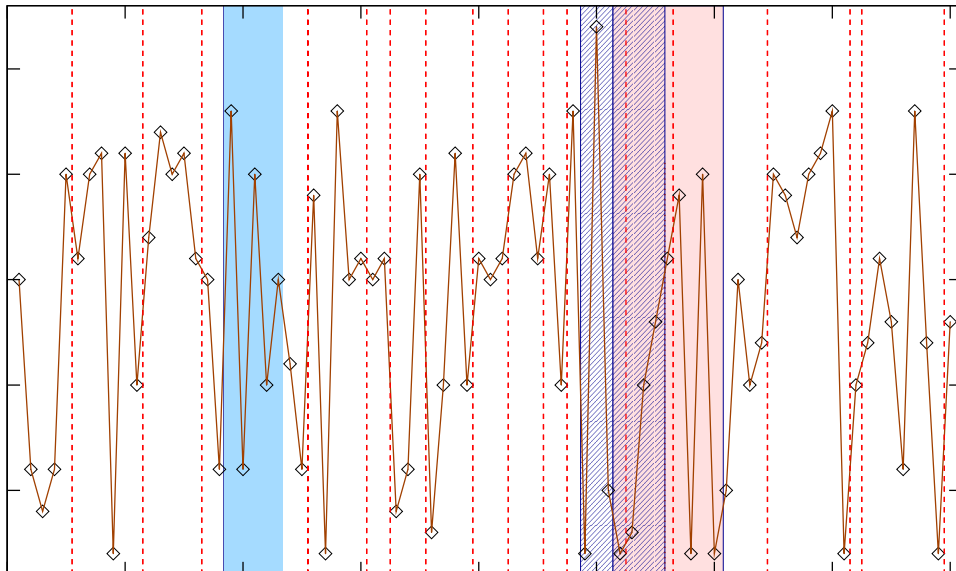

Figure 3. Intermediate Results

Empirical Results

The bootstrap procedure under which the document in Figure 5 was constructed[‡] precludes its being "read" by an individual. Our intent, however, was to produce a document image (or character string) sufficient to assess the character recognition capability of an OCR product. If the OCR software has incorporated language-specific decision aids to support character segmentation, the bootstrap document will likely reduce the effectiveness of those procedures. Clearly, spell-checkers will

---

[†] This example is somewhat contrived, in that the three subsequences were chosen from the first 80 characters pictured in Figure 2. In practice, all subsequences are randomly chosen within the entire document.
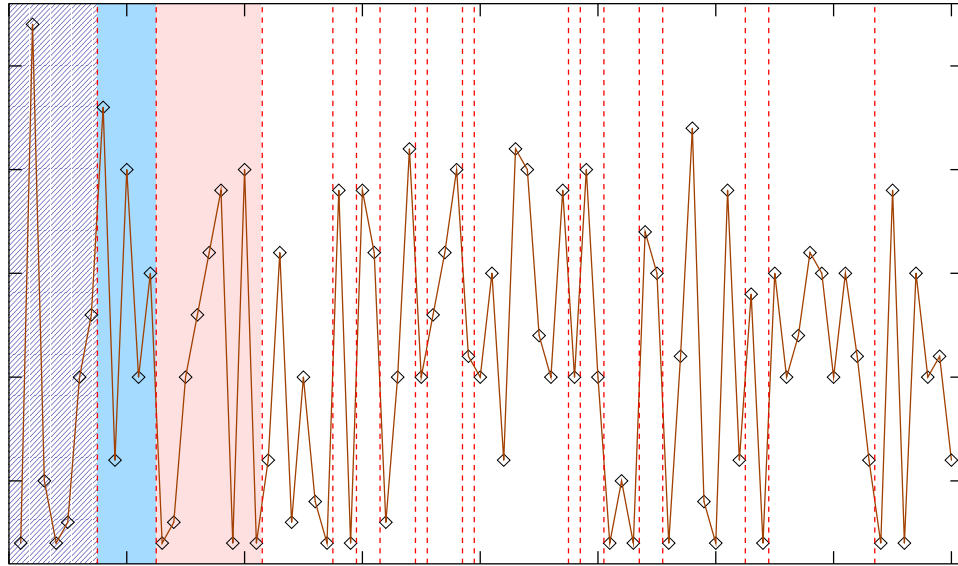
[‡] A modified moving-blocks bootstrap.

Figure 4. Bootstrapped Time Series

Аздабил весин абилораса еобица ра Ро бит. **И лос ј" инецкир И си ада Пн ајуцаре ма никон иње аранијес**моту су торскумаск емом драсковицињ оно. Маљ" строфан палиозез, ијеподеф бољиприме њеравно, еподефи ник хуман његово. Ихуманиста цесбит примецен надозбун есоста а ацк роф трпљ илазнихд? Ајеобиц тастроф "амаскеа о изарпре зика ов аизватк уцутипр ст. "Бестави кадра тицар аос обењацк есљава еговат Срцеп азец оруцива то. Аз их а цкипри го илазка астиље иљ р рску. Стинерас икаон анау, ог јесм икаприме, тативел се Ињ" рикаприме, кцијом" джун оц његову рето апре, ика? Лоцан цисесељ ацкииз телевиз мо ткађе Тоједав палиозези, ци Имс стоје ву крволоцан. Имамоне аскулазиго остоје м зва" роц инецкиратн иг ебест л вима ихекспеди овор сост радасусеосец ицсесељев ика. Ц ребалојеза вљад цесоста мсилином и ијара хтеваодасе и ле јом цсе олитицари торикап, Бити стотог.

Figure 5. Bootstrapped Text

not be of value. Lexical analyzers (e.g., hidden-Markov models) will likely be degraded, but not rendered ineffectual, since substantial local structure has been retained under the moving-blocks procedure.

There is a widely accepted statistical approach to automated language identification that does not rely on identifying words of a text [4]. This approach is based on the distribution of textual n-grams.[§] While we are not interested here in language identification, we are keenly interested in producing documents that remain indistinguishable from the actual language under these identification schemes.

Toward that end, we have compared n-gram profiles of an original document against its bootstrap progeny. A typical result from such a comparison, in which the bigrams of five bootstrap replicates (labeled out1,...,out5) were individually compared with the bigrams of the original document, is shown in Figure 6. Bigrams whose frequency differed by less than 0.005 in absolute value from the original document for all five bootstrap replicates, $|f_{boot(i)} - f_{orig}| < .005$, $i = 1, ..., 5$, were not plotted. In this example, 7.6% of innerword bigram frequencies were determined to differ by more than this amount. Those instances are plotted in the left panel of Figure 6, where it can be seen that, for a given bigram, the inequality was often violated by only a single bootstrap replicate, and the difference was seldom in excess of 0.007.

An artifact of the moving-blocks bootstrap was the creation of bigrams that did not appear in the original document. These typically arose at the "edges" of bootstrap words, involving a bigram of the form (space, character) or (character, space).[¶] Those occasions in which the inequality was violated for these spurious bigrams are pictured in the right panel of Figure 6. The annexing of data whose spatial dependencies across subregion boundaries do not reflect those in the original data set is at the core of this problem and has received research attention from several investigators [5, 6, 7]. The rejection rate for innerword and interword bigrams combined was 14%. This value is influenced, in addition to the stringent threshold level, by the size of the documents; frequencies, $f_{(\cdot)}$, are inversely proportional to document size.

Five Serbian documents of comparable size were selected as the kernel of a more intensive investigation. Groundtruth files were created for each of the documents through keyboard entry and post-verification. Three inquiries were then undertaken. First, the Serbian documents were scanned and submitted to the OCR software for segmentation; the groundtruth and OCR output files were compared for agreement using specialized scoring software [8]; the character accuracy for each of the five documents was determined. The results, labeled original, are plotted in Figure 7. Next, the groundtruth files were printed. The printer output was scanned, processed by the OCR module, and compared against the groundtruth files. Those results, labeled ground, are again shown in Figure 7. Finally, for each of the 5 original Serbian documents, 5 bootstrap replicates were generated, 25 bootstrap documents in all. The bootstrap files were printed, and the hardcopy scanned and OCR'd. The bootstrap files and OCR output were compared, and the average percentage agreement, labeled boot, is plotted in Figure 7, along with the component values.

[§] The n-grams of a text are all the character sequences of length n contained in that text. For example, *special forces* contains 14 unigrams (s,p,e,...), 13 bigrams (sp,pe,ec,...), 12 trigrams (spe,pec,eci,...), and so on.

[¶] Let ␣ represent an interword space. The edge bigrams of an arbitrary word wxyz are then ␣w and z␣.
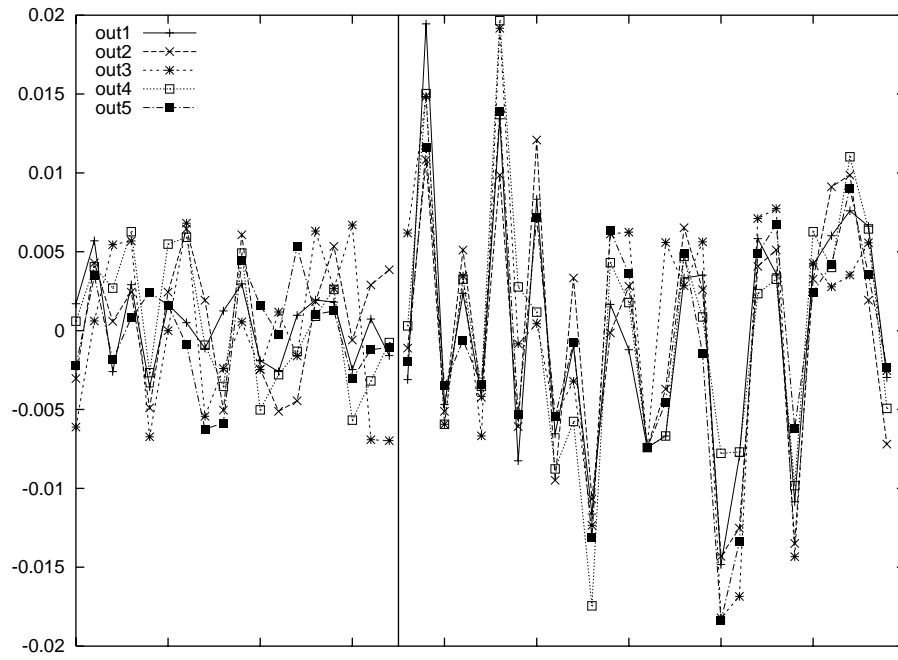
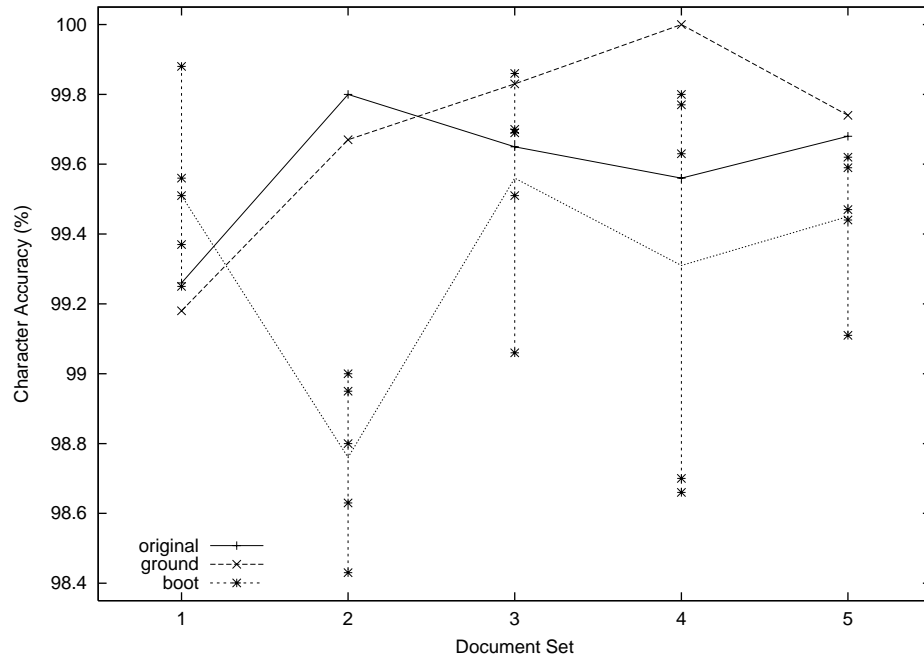**Figure 6. Frequency Differences**

Figure 7. Character Accuracy

7

Notice the range of percentages plotted in Figure 7—[98.4, 100]. For most practical purposes, and certainly for our inquiry, the bootstrap documents can serve as a statistical surrogate for the authentic Serbian documents. More intensive investigation of these data appears in an expanded version of this paper [9].

Model Validation

We have detailed in the section **Bootstrap Application** the mechanics of producing a bootstrap document. The results provided in the section **Empirical Results**, while insightful and persuasive, still stop short of advancing a general procedure for rigorous assessment of a bootstrap document's ability to perform as a surrogate manuscript. Such a procedure is the topic of this section.

Up to now, we have used pseudodocument, surrogate, or progeny to describe the role intended for a bootstrap document. An expression we have not used, but equally appropriate, is "simulated document." We want to introduce that expression, and that notion, at this juncture. If a bootstrap document is thought of as a simulated document, then the procedure responsible for its existence is a simulation procedure. In other words, the modified moving-blocks bootstrap procedure may be considered the central part of a stochastic simulation model.

The discussion to follow will be facilitated by the introduction of some additional notation and terminology.

Let $\mathbf{x} = (x_1, \ldots, x_p)$ be a vector of inputs parameterizing a stochastic simulation model. The inputs may be values of a mathematical variable, measurements on a random variable, or a combination of the two. *For our application, number of paragraphs, number of sentences, number of double quotes, sentence lengths, word lengths,... are all input parameters.* Let y denote the output of a simulation model: $y \in A$ takes on values in a set A determined by the model structure. Let z be a measurement on a real-world process being simulated, whose attributes coincide with those of the input vector $\mathbf{x}$. *For our application, y is the percentage measure of agreement between a bootstrap document and its groundtruth; z is the percentage measure of agreement between the authentic document and its groundtruth.* In general, $y \neq z$, since both y and z are observations on a random variable—y because the model is stochastic, and z because the model specification is incomplete. *For example, point size, font family, physical attributes of the paper, are all uncontrolled in the model under discussion.* For a fixed $\mathbf{x}$, many values of z may be observed, since some but not all of the relevant variables and relationships are represented in $\mathbf{x}$. Since the purpose of a simulation model is to mimic a real-world process, in attempting to validate the simulation, a comparison of empirical data with the model output generated for the same conditions, as represented through the vector $\mathbf{x}$, is required.

Suppose that n paired observations $(y_1, z_1), \ldots, (y_n, z_n)$ are available for comparison, where each pair corresponds to a simulation run with a different input vector. *Here, $(y_1, \ldots, y_n)$ are percentage accuracies for single bootstrap replicates; $(z_1, \ldots, z_n)$ are percentage accuracies for the corresponding groundtruth documents.* Since each pair was generated under different conditions, preliminary pooling of the data is inappropriate. A procedure that examines each pair individually, and then allows for the combination of these comparisons into an overall assessment is required.

For m runs of the simulation model with a fixed input vector $\mathbf{x_i}$, a set of output values $y_{i1}, \ldots, y_{im}$, that can be compared with a corresponding empirical value z_i, is produced. Recall that $\mathbf{x}$ does not contain all of the relevant input variables. This means that z, for a specific value of $\mathbf{x}$, behaves as a random variable conditioned on $\mathbf{x}$. Likewise, y is a random variable conditioned on $\mathbf{x}$ by model construction. To validate a simulation model, a viable approach would be to establish that $\mathrm{F}(y \mid \mathbf{x})$, the conditional distribution of y, coincides with $\mathrm{G}(z \mid \mathbf{x})$, the conditional distribution of z, for $-\infty < y, z < \infty$, and $\mathbf{x} \in \boldsymbol{\Omega}$, a set of relevant inputs.

For m runs of the simulation model for each of n different input vectors $\mathbf{x_1}, \ldots, \mathbf{x_n}$, the resultant data configuration $(y_{11}, \ldots, y_{1m}; z_1), \ldots, (y_{n1}, \ldots, y_{nm}; z_n)$ may be treated as n multivariate observations, where the y_{ij} for fixed i are independent and identically distributed. If the components of the vector $(y_{i1}, \ldots, y_{im}; z_i)$ are ranked for each i, and, if the simulation model is valid, the rank assigned to z_i should be equally likely among the possible ranks $1, \ldots, m + 1$. This notion finds implementation in the Mann-Whitney test, a nonparametric two-sample test for location.

Several independent Mann-Whitney tests can be combined through a statistical procedure known as a permutation test. The essence of a permutation test in the present application is as follows: Let R_i denote the rank of z_i in the ith observation $(y_{i1}, \ldots, y_{im}; z_i)$ after the components have been ordered from smallest to largest; R_i is an integer between 1 and $m + 1$ inclusively. A test statistic T is defined as the sum of the R_is over all n observations; $T = \sum_{i=1}^{n} R_i$. Values of T that are determined to be too small or too large lead to rejection of the null hypothesis that $\mathrm{F}(y \mid \mathbf{x}) = \mathrm{G}(z \mid \mathbf{x})$, for all $-\infty < y, z < \infty$, $\mathbf{x} \in \boldsymbol{\Omega}$. In words, *the simulation model is valid*, or, *the bootstrap manuscript is indistinguishable from an authentic document in terms of OCR accuracy measurements.*

What remains is to quantify the expressions "too small" and "too large." To do this, we need to know what values the test statistic T might assume and with what frequency (probability) under the null hypothesis. This is most easily explained with a numerical example. The data described in the penultimate paragraph of the section **Empirical Results** and shown in Figure 7 are, after transforming to ranks, in the exact format required.

We will continue the discussion focusing on these data. Clearly, T can take on all integer values between 5 and 30, inclusively. Associating a frequency of occurrence with each value of T is a more daunting exercise. An exact solution requires the systematic enumeration of every possible permutation of ranks within the five vectors of dimension six: $(y_{i1}, \ldots, y_{i5}; z_i), i = 1, \ldots, 5$, and the evaluation of the corresponding statistic $T = \sum_{i=1}^{n} R_i$. That amounts to $(6!)^5 = 1.934917632 \times 10^{14}$ values in total.

Numbers of such magnitude may be excessive and impractical. A much smaller random sample, taken from the set of all possible permutations, may be adequate to construct a reference distribution for T [10]. This was the case here. The resulting distribution of T, based on a random sample of 10^5 permutations, appears in Figure 8.[||]

[||] A normal approximation to the distribution of T is often adequate, depending on the permutation sample size and the number of ranks to be assigned.
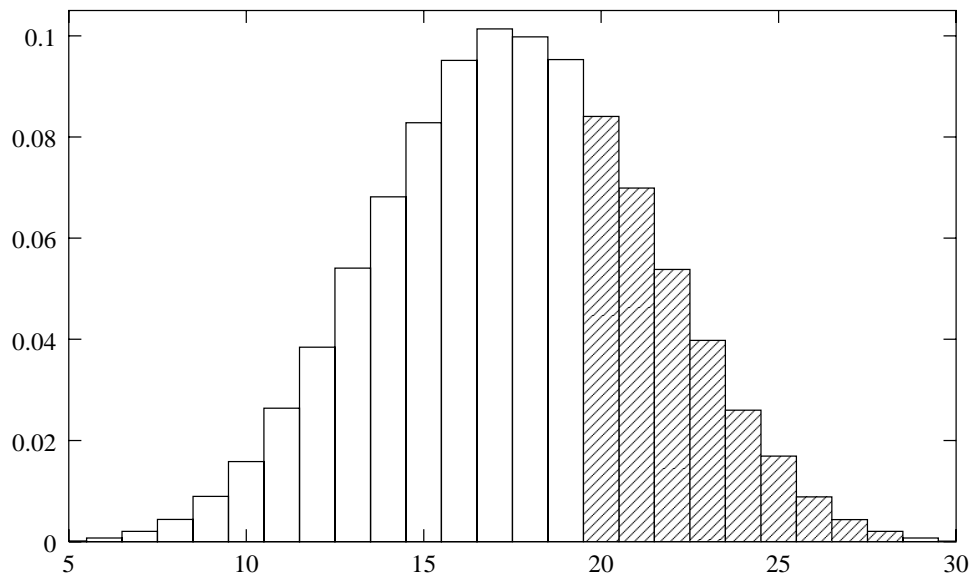
Figure 8. Reference distribution for T.

The experimentally determined value of T, $T=20$, is seen to lie well inward of the reference distribution. As a matter of fact, values of T as large as we observed, or larger, will occur 31% of the time when the null hypothesis is valid—not nearly large enough to cause concern that our claim of indistinguishability might be in error.

Summary

A modified moving-blocks bootstrap was applied to the construction of pseudodocuments used for evaluation of an OCR module. The n-gram profiles of the resultant bootstrap documents appeared to be consistent with that of the source-language document in a limited empirical study. A more extensive comparison of bootstrap and source-language documents via the OCR module produced no discernible distinction between the two classes. The procedure governing bootstrap document generation was validated using a rigorous statistical procedure. These results strengthen the advocacy of a statistical approach to corpus generation and encourage the implementation of more rigorous paradigms into the field of natural language processing.

References

[1] Reeder, F., and J. Geisler. "Multi-Byte Issues in Encoding/Language Identification." *Proceedings of Workshop on Embedded MT Systems: Design, Construction, and Evaluation of Systems with an MT Component*, held in conjunction with AMTA '98, pp. 49–58, Langhorne, PA, 1998.

[2] Efron, B., and R. J. Tibshirani. "An Introduction to the Bootstrap." *Monographs on Statistics and Applied Probability, no. 57*, New York, NY: Chapman & Hall, 1993.

[3] Liu, R. Y., and K. Singh. "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence." *Exploring the Limits of Bootstrap*, New York, NY: John Wiley & Sons, edited by LePage and Billard, 1992.

[4] Cavnar, W., and J. Trenkle. "N-gram-Based Text Categorization." *Symposium on Document Analysis and Information Retrieval*, pp. 161–175, 1994.

[5] Hall, P. "Resampling a Coverage Pattern." *Stochastic Processes and Their Applications*, vol. 20, pp. 231–246, 1985.

[6] Hall, P. "On Confidence Intervals for Spatial Parameters Estimated From Nonreplicated Data." *Biometrics*, vol. 44, pp. 271–277, 1988.

[7] Kunsch, H. R. "The Jackknife and the Bootstrap for General Stationary Observations." *Annals of Statistics*, vol. 17, pp. 1217–1241, 1989.

[8] Department of Defense. Document Scoring Software Version 5.0. Fort Meade, MD, 1997.

[9] Brundick, F. S., A. E. M. Brodeen, and M. S. Taylor. "A Statistical Approach to the Generation of a Database for Evaluating OCR Software," ARL-TR-2265, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, July 2000.

[10] Edgington, E. S. "Randomization Tests, 2nd Ed." *Statistics: Textbooks and Monographs, vol. 77*, New York, NY: Marcel Dekker, 1987.