

# Text Alignment in the Real World: Improving Alignments of Noisy Translations Using Common Lexical Features, String Matching Strategies and N-Gram Comparisons<sup>1</sup>

Mark W. Davis, Ted E. Dunning and William C. Ogden

Computing Research Laboratory

New Mexico State University

Box 30001/3CRL

Las Cruces, New Mexico

USA

{madavis,ted,ogden}@crl.nmsu.edu

## Abstract

Alignment methods based on byte-length comparisons of alignment blocks have been remarkably successful for aligning good translations from legislative transcriptions. For noisy translations in which the parallel text of a document has significant structural differences, byte-alignment methods often do not perform well. The Pan American Health Organization (PAHO) corpus is a series of articles that were first translated by machine methods and then improved by professional translators. Many of the Spanish PAHO texts do not share formatting conventions with the corresponding English documents, refer to tables in stylistically different ways and contain extraneous information. A method based on a dynamic programming framework, but using a decision criterion derived from a combination of byte-length ratio measures, hard matching of numbers, string comparisons and *n-gram* co-occurrence matching substantially improves the performance of the alignment process.

## 1 Introduction

Given texts in two languages that are to some degree translations of one another, an alignment of the texts associates sentences, paragraphs or phrases in one document with their translations in the other. Successful approaches to alignment can be divided into two primary types: those that use comparisons of lexical elements between the documents (Wu, 1994; Chen 1993; Catizone, Russell and Warwick, 1989), and

those that use a statistical decision process derived from byte-length ratios between alignment blocks (Wu, 1994; Church, 1993; Gale and Church, 1991). Methods vary for the former approach, but in the latter approach, a dynamic programming framework is used to sequentially align blocks as the alignment process proceeds. Under this model, blocks are compared only with nearby blocks as the alignment proceeds, substantially reducing the computational overhead [ $O(n^2) \ll O(n!)$ ] of the alignment process.

In the primary literature on alignment, the texts are typically well-behaved. In byte-length ratio approaches, the presence of long stretches of blocks that have roughly similar lengths can be problematic, and some improvement can be achieved by augmenting the byte-length measure by scores derived from lexical feature matching (Wu, 1994). When combined with radical formatting departures between documents that often arise in text translations, the difficulties of producing good alignments are exacerbated by the presence of untranslated segments, textual rearrangements and other problematic text features. The dynamic programming framework makes long runs of segments that have no translation in their parallel text difficult to ignore because the limited window size prevents passing over those segments to reach appropriate areas of the document further downstream. Taken together, these difficulties can be catastrophic to the alignment process. Our experience shows that the fraction of correct alignments can drop to less than 5%.

Noisy translations of this sort do reflect human error and the preferences of translators, and they are probably much more prevalent than alignment work on legislative transcriptions has indicated. The purpose of this research was to ascertain what types of information contained in a document could be used to improve the alignment process, while not making gross assumptions about the source text format con-

---

<sup>1</sup>This research was funded under DoD Contract #MDA 904-94-C-E086

ventions and peculiarities. The Pan American Health Organization (PAHO) corpus was used as a test corpus for evaluating the performance of the modified alignment algorithm. The PAHO texts are a series of documents on Latin American health issues and our test segment consisted of 180 documents that ranged from 20 to 3825 lines in length. From these documents, several of the more problematic texts were hand aligned for analysis and comparison with the results of automatic alignment methods.

## 2 A General Approach

The byte-length ratio methods are very general in that they rely only upon a heuristic segmentation procedure to divide a text into sentence-level chunks. Although determining sentence boundaries can be problematic across languages, simple assumptions appear to work well even for comparisons between European and Oriental languages, primarily because

the segmentation heuristic is uniformly applied to each document and therefore an "undersegmented" section can combine together to match with single blocks in the opposite language as necessary.

Less general would be a method that relied on deep analysis of the source texts to determine appropriate boundaries for alignment blocks. A model that accounted for all of the formatting discrepancies, comparative rescalings of sentence or phrase length due to the economy of the language expression, and other properties that may define a corpus, will not necessarily be appropriate to other corpora or to text in general.

We chose to remain as general as possible with our investigations of alignment methods. In particular, the heuristics for text segmentation regarded periods followed by a space, a newline (paragraph boundaries) or a tab as a sentence boundary for both English and Spanish texts (Figure 1). Multiple periods separated by spaces were ignored for alignment segmentation to

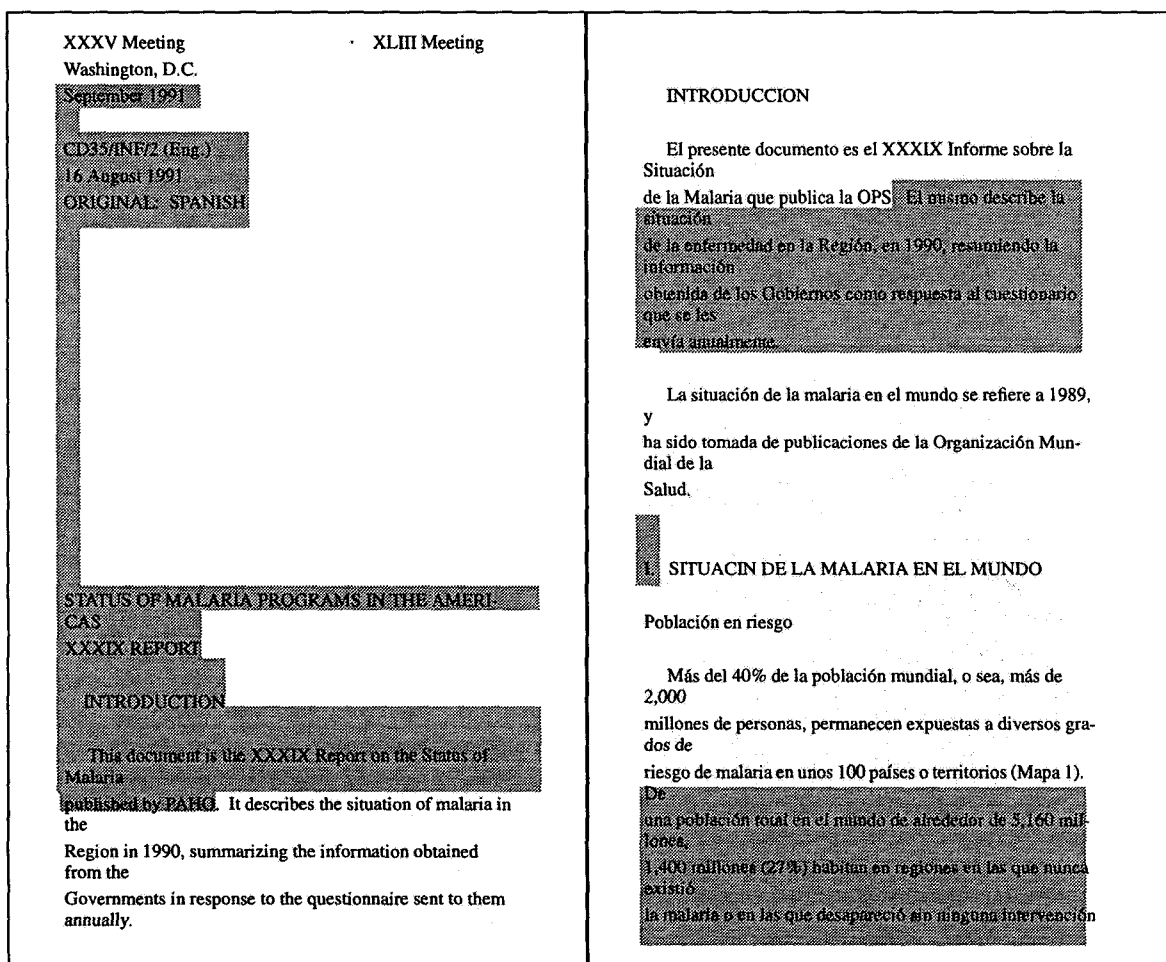


Figure 1: Sample English and Spanish texts with contiguous grey areas indicating alignment blocks.

allow for ellipsis. This approach did not, therefore, regard many abbreviations as a unique class of textual event. The end result was an extremely simplistic segmentation.

### 3 The PAHO Corpus: Noisy, Problematic Texts

The PAHO texts serve as an important counterpart to our translator's workstation, Norm (Ogden, 1993). During the translation process, translators can access many different resources including a variety of on-line dictionaries, reference works and parallel texts. The parallel texts include examples of translations that different translators have compiled in the past and serve as a series of examples of how to translate words and phrases for a particular context. The PAHO texts also serve as a basis for our multi-lingual information retrieval system (Davis and Dunning, 1995; Dunning and Davis, 1993a, 1993b). The need for robust strategies to process and align large parallel corpora automatically is therefore a critical component of our ongoing research.

In the PAHO corpus, many of the texts are well-behaved, with similar tokenization at the boundaries delineating paragraphs. But some are extremely noisy, with added text in the English or Spanish document that lacks a counterpart in the parallel document. Formatting conventions differ in many cases, with multiple periods delimiting contents listings in one, while spaces serve a similar role in the other, or tables and reference formats differing between the two texts. Another formatting problem is the addition of significant runs of whitespace and newlines that simply do not occur in the parallel text. The document pair shown in Figure 1 is representative of the quality of the PAHO texts.

### 4 Features and Alignment

One of the most striking features of English-Spanish translations is the fact that native English speakers with little knowledge of Spanish appear able to identify parallel texts with remarkable accuracy. The reason appears to be the large number of cognate terms that Spanish and English translations share, especially technical terms, and other lexical features such as numbers and proper names that may appear with similar placement and frequency across two parallel texts. The work by Simard, Foster and Isabelle (1993) as well as Church (1993) demonstrated that cognate-matching strategies can be highly effective in aligning text. Native English

speakers with limited Spanish appear to be capable of aligning even noisy texts like many of the PAHO documents, with difficulty causing a decrease in speed of alignment, rather than decreased accuracy of the alignment. From these observations, we examined five different sources of information for alignment discrimination:

- Byte-length ratios
- Unordered character *n*-gram comparisons
- Simple ordered string-matching
- Number matching
- Bilingual-dictionary translations

The analyses of each of these information sources are presented in sections 5.1 through 5.5.

For each method, a hand-aligned document from PAHO corpus that was problematic for byte-ratio methods was used for evaluation, first for comparing the method's score distribution between random blocks and the hand-aligned set, then for performing realignments of the documents. The document was quite long for the PAHO set, containing about 1400 lines of text and 360 alignment blocks in the English document and 1000 lines and 297 blocks in the Spanish text. In these particular documents, the English text had nearly 400 lines of extraneous data abutted to the end of it that was not in the Spanish document, increasing the error potential for byte-length methods.

### 5 Improving Alignments

We used a modified and extended version of Gale and Church's byte-ratio algorithm (1991) as a basis for an improved alignment algorithm. The standard algorithm derives a penalty from a ratio of the byte lengths of two potential aligned blocks and augments the penalty with a factor based on the frequency of matches between blocks in one language that equate to a block or blocks in the second language. The byte-ratio penalty is the measurement-conditioned or *a posteriori* probability of a match while the frequency of block matches gives the *a priori* probability of the same match. Our version of the basic algorithm differs in the mechanics of memory management (we use skip-lists to improve performance of the dynamic programming, for example), includes both positive and negative information about the probability of a given match and fuses multiple sources of information for evaluating alignment probabilities.

For two documents,  $D_1$  and  $D_2$ , consisting of  $n$  and  $m$  alignment blocks, respectively,  $a_{i \leq n}$  and  $b_{j \leq m}$ , an alignment,  $A$ , is a set consisting of  $a_i \dots a_{i+1} \leftrightarrow b_j \dots b_{j+p}$  pairs. For compactness, we will write this as  $\alpha_{i,l} \leftrightarrow \beta_{j,p}$ .

$$P(A|D_1, D_2) \approx \prod_{(\alpha_{i,l} \leftrightarrow \beta_{j,p}) \in A} P(\alpha_{i,l} \leftrightarrow \beta_{j,p} | \delta_1, \delta_2, \dots, \delta_k) \quad (\text{Eq 1})$$

$$P(\alpha_{i,l} \leftrightarrow \beta_{j,p} | \delta_1, \delta_2, \dots, \delta_k) = \frac{P(\delta_1, \delta_2, \dots, \delta_k | \alpha_{i,l} \leftrightarrow \beta_{j,p}) P(\alpha_{i,l} \leftrightarrow \beta_{j,p})}{P(\delta_1, \delta_2, \dots, \delta_k)} \quad (\text{Eq 2})$$

$$P(\alpha_{i,l} \leftrightarrow \beta_{j,p} | \delta_1, \delta_2, \dots, \delta_k) \quad (\text{Eq 3})$$

$$= \frac{P(\delta_1, \delta_2, \dots, \delta_k | \alpha_{i,l} \leftrightarrow \beta_{j,p}) P(\alpha_{i,l} \leftrightarrow \beta_{j,p})}{P(\delta_1, \delta_2, \dots, \delta_k | \alpha_{i,l} \leftrightarrow \beta_{j,p}) P(\alpha_{i,l} \leftrightarrow \beta_{j,p}) + P(\delta_1, \delta_2, \dots, \delta_k | \neg(\alpha_{i,l} \leftrightarrow \beta_{j,p})) P(\neg(\alpha_{i,l} \leftrightarrow \beta_{j,p}))}$$

$$P(\delta_1, \delta_2, \dots, \delta_k | \alpha_{i,l} \leftrightarrow \beta_{j,p}) \approx \prod_k P(\delta_k | \alpha_{i,l} \leftrightarrow \beta_{j,p}) \quad (\text{Eq 4})$$

$$\vartheta_1 = \log P(\delta_k | \alpha_{i,l} \leftrightarrow \beta_{j,p}) \quad (\text{Eq 5})$$

$$\vartheta_2 = \log [P(\delta_k | \alpha_{i,l} \leftrightarrow \beta_{j,p}) P(\alpha_{i,l} \leftrightarrow \beta_{j,p}) + P(\delta_k | \neg(\alpha_{i,l} \leftrightarrow \beta_{j,p})) P(\neg(\alpha_{i,l} \leftrightarrow \beta_{j,p}))] \quad (\text{Eq 6})$$

$$\vartheta_3 = \log P(\alpha_{i,l} \leftrightarrow \beta_{j,p}) \quad (\text{Eq 7})$$

$$\arg \max_A P(A|D_1, D_2) = \sum_{(\alpha_{i,l} \leftrightarrow \beta_{j,p}) \in A} \sum_k -\vartheta_1 + \vartheta_2 - \vartheta_3 \quad (\text{Eq 8})$$

## Figure 2. Equations.

Following the Gale and Church approach, we choose an alignment that maximizes the probability over all possible alignments:

$$\arg \max_A [P(A|D_1, D_2)]$$

If we assume that the probabilities of individually aligned block pairs in an alignment are independent, the above equation becomes:

$$P(A|D_1, D_2) \approx \prod_{(\alpha_{i,l} \leftrightarrow \beta_{j,p}) \in A} P(\alpha_{i,l} \leftrightarrow \beta_{j,p} | D_1, D_2)$$

Further assuming that the individual probabilities of aligning two blocks,  $P(\alpha_{i,l} \leftrightarrow \beta_{j,p} | D_1, D_2)$ , are dependent on features in the text described by a series of feature scores,  $\delta_k$ , the above equations expands into Equation 1 in figure 2.

Now, for each of the feature scoring functions, the *a posteriori* probabilities can be calculated from Bayes' Rule as shown in Equation 2, Figure 2 which, given an approximation of the joint *a posteriori* probabilities by assuming independence, produces Equation 3, Figure 2.

Note that the term in the denominator of Equation 3 reflects both the statistics of the positive and negative information for the alignment. In Gale and Church's

original work, the denominator term was assumed to be a constant over the range of  $\delta$ , and therefore could be safely ignored during the maximization of probabilities over the alignment set. In reality, this assumption is true only in the case of a uniform distribution of  $P(\delta_k | \neg(\alpha_{i,l} \leftrightarrow \beta_{j,p}))$ , and is perhaps not even true in that case due to the scaling properties of the logarithm when the maximization problem above is converted to a minimization problem (below).

In any case, the probability of a given value of  $\delta$  occurring is not merely dependent on the probability of that score in the hand-aligned set, but is dependent on the comparative probabilities of the score for the hand aligned set and a set of randomly chosen alignment blocks. Clearly, if a value of  $\delta$  is equally likely for both the hand aligned and random sets, then the measurement cannot contribute to the decision process. Equation 3 presents a very general approach to the fusion of multiple sources of information about alignment probabilities. Each of the sources contributes to the overall probability of an alignment, but is in turn scaled by the total probability of a given score occurring over the entire set of possible alignments.

We can convert the maximization of probabilities into a minimization of penalties taking the negative logarithm of Equation 2 and substituting Equation 3,

where  $\vartheta_1$ ,  $\vartheta_2$  and  $\vartheta_3$  are as given in figure 2, Equations 5, 6 and 7. Equation 8 in the same figure is the result.

The feature functions,  $\delta_k$ , are derived from estimates of the probability of byte length differences, number matching score probabilities and string match score probabilities in our approach.

The Bayesian prior,  $P(\alpha_{i,l} \leftrightarrow \beta_{j,p})$ , can be estimated as per Gale and Church (1991) by assuming that it is equal to the frequency of distinct  $n$ - $m$  matches in the training set.

### 5.1 Byte-length Ratios, $\delta_1$

The probability of an alignment based on byte-length ratios is  $P(\delta_1 | \alpha_{i,l} \leftrightarrow \beta_{j,p}) = P(\delta_1(l(\alpha_{i,l}), l(\beta_{j,p})))$ , where  $l()$  is the byte-length function. The distribution is assumed to be a Gaussian random variable derived from the block length differences in the hand aligned set. Following Gale and Church (1991), the slope of the average of the length differences describes the average number of Spanish characters generated per English character. Assuming that the distribution is approximately Gaussian, we can normalize it to mean 0 and variance 1, resulting in:

$$\delta_1(l(\alpha_{i,l}), l(\beta_{j,p})) = \frac{l(\beta_{j,p}) - l(\alpha_{i,l})c}{\sqrt{l(\alpha_{i,l})\sigma^2}}$$

where  $c = E(l(\beta_{j,p})/l(\alpha_{i,l})) \approx 0.99$  and  $\sigma^2 \approx 0.16$  is the observed variance. The histogram in Figure 3a shows the actual distribution of the hand-aligned data set. The shape of the histogram is approximately Gaussian. The distribution of the corresponding random

segments is shown in Figure 3b. Note that the distribution of the random set has significantly higher standard deviation than the corresponding hand aligned set. This diagram, as well as Figure 4 for the  $n$ -gram approach on the following page, indicate the statistical quality of the information provided by the scores. Good sources of information would produce a marked difference between the two distributions. For comparatively poor sources of information, the distributions would show little or no differences.

### 5.2 4-gram Matching, $\delta_2$

Cognates in English and Spanish often have short runs of letters in common. A measure that counts the number of matching  $n$ -grams in two strings is an unordered comparison of similarities within the strings. In this way, runs of letters in common between cognate terms are measured. We used an efficient  $n$ -gram matching algorithm that requires a single scan of each string, followed by two sorts and a linear-time list comparison to count matches. The resulting score was normalized by the total number of  $n$ -grams between the strings. Formally, for two strings  $e_1e_2\dots e_p$  and  $s_1s_2\dots s_q$ , the  $n$ -gram match count,  $K_n$ , is given by:

$$K_n = \frac{1}{pq} \sum_{i < p} \sum_{j < q} m(e_i e_{i+1} \dots e_{i+n}, s_j s_{j+1} \dots s_{j+n})$$

where  $m()$  is the matching function. The function,  $m()$ , is equal to 1 only for equivalent  $n$ -grams, else it is 0.

We chose to use 4-gram scores for the alignment algorithm,  $\delta_2 = K_4$ . The distributions of the 4-gram

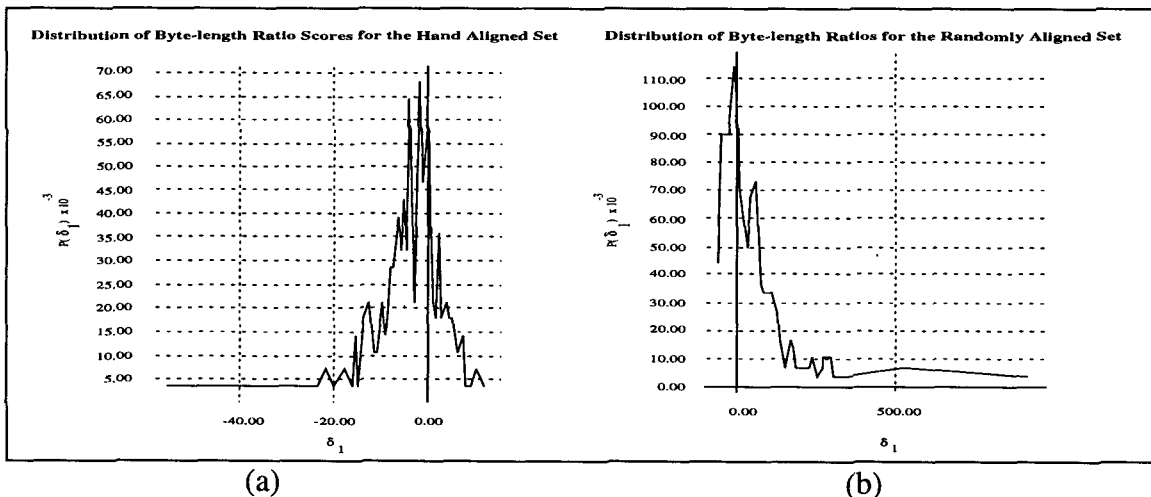


Figure 3. Distribution of  $\delta_1$  for (a) hand aligned and (b) randomly aligned blocks

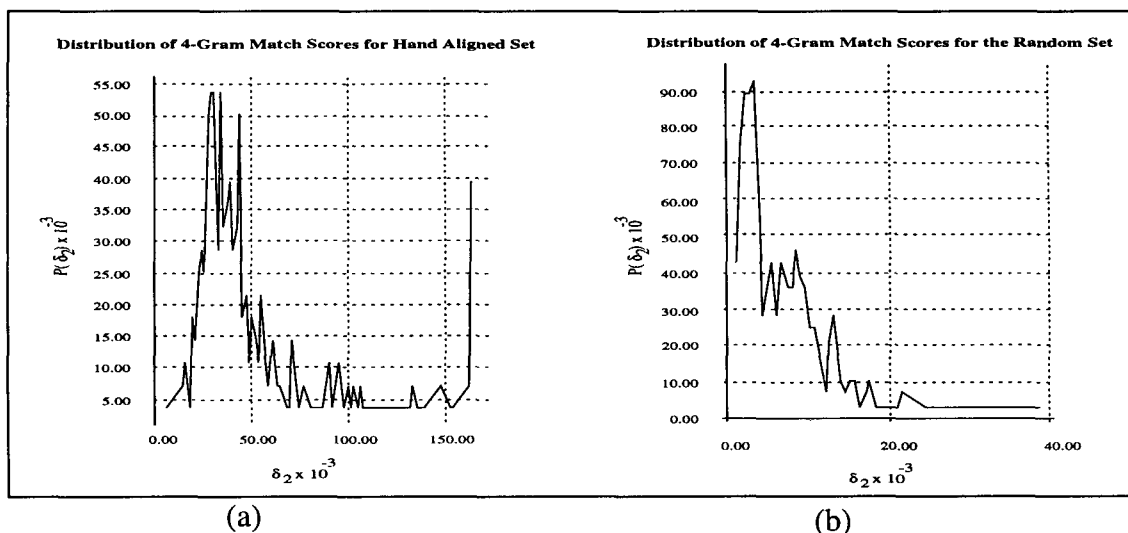


Figure 4. 4-gram matching score distributions,  $\delta_2$ , for (a) hand aligned and (b) randomly aligned blocks.

counts were computed for both the hand-aligned and random alignment blocks. Figure 4 shows the resulting distributions. The results suggest that, on the whole, the use of  $n$ -gram methods should be considered for improving alignments that contain lexically similar cognates. Being unordered comparisons, however, they cannot exploit any intrinsic sequencing of lexical elements.

### 5.3 Ordered String Comparisons, $\delta_3$

The value of unordered comparisons like the  $n$ -gram matching may be enhanced by ordered comparisons. An ordered comparison can reduce the noise associated with matching unrelated  $n$ -grams at opposite ends of parallel alignment blocks. We chose to evaluate a simple string-matching scheme as a possible method for improving alignment performance. The scheme compares the two alignment blocks character-by-character, skipping over sections in one block that do not match in the opposite, thus primarily penalizing the inclusion of dissimilar text segments in either block. The resulting sum of the matches is scaled by the sum of the lengths of the two blocks. In comparison with the random block scoring, the distribution of the hand aligned data set had a greater number of matches with high string-match scores.

### 5.4 Number Matching, $\delta_4$

The PAHO texts are distinguished by a number of textual features, especially the fact that they are all in some way related to Latin American health issues. The preponderance of the documents are technical reports on epidemiology, proceedings from meetings and conferences, and compendiums of resources and

citations. Within these documents, numbers occur regularly. The string-matching technique suggested that if a class of lexical distinction could be matched directly, the alignments might be significantly improved. Numbers are sufficiently general that we felt we were not violating the spirit of the restriction on generality by using a number-matching scheme.

For each alignment block pair, the number matching algorithm extracted all numbers. The total number of exact matches between the number sets from each alignment block was then normalized by the sizes of both sets of numbers. This approach has several drawbacks, such as the differences in the format of numbers between Spanish and English. In Spanish, for example, commas are used instead of decimal points. These distinctions were ignored, however, to preserve the generality of the algorithm. This generality will potentially extend to other languages, including Asiatic languages, which tend to use Arabic numerals to represent numbers. The distributions of both the hand and random block scoring both showed a substantial mass of very low scores.

It should be noted that numbers are simply a special case of cognates and certainly contribute to the  $n$ -gram scores. Adding in number matching strategies therefore only enhances the  $n$ -gram results.

### 5.5 Translation Residues

Despite the fact that non-Spanish speakers can often achieve success at aligning English documents with Spanish texts, the added knowledge of someone with both Spanish and English language understanding is an added benefit and should facilitate alignment. To evaluate the role of translation-based alignment scor-

**Table 1. Performance comparisons between byte-length ratio methods and the improved algorithm.**

	Byte-length Ratio Method			Improved Method		
	#HAND	#FOUND	#CORRECT	#HAND	#FOUND	#CORRECT
Document 1	281	196	65	281	222	138
Document 2	787	440	3	787	614	553

ing, the Collins Spanish-English and English-Spanish bilingual dictionaries were used to produce a score equal to the *residue* from a translation attempt of the terms in potential aligning blocks.

Given a set of English terms,  $e_i$ , Spanish terms,  $s_j$ , from two blocks, the *translation operation*,  $T(l)$ , generates a set of terms in the opposite language by stemming each term and retrieving the terms that the stemmed word translates to in Collins. The *residue*,  $R$ , is then a penalty equal to the (normalized) number of terms in each translation set that do not have a match in the opposite translation set:

$$R = 1 - \frac{\sum_k \|T(l_k) \cap l_k\|}{\sum_k \|T(l_k) \cup l_k\|}$$

In comparison test, the distributions of scores between random Spanish blocks and English blocks, and between the hand-aligned sets, were surprisingly similar, making a statistical discrimination of proper alignments difficult. We believe that dictionary-based discrimination performs poorly primarily due to the noisy nature of the dictionary we used. It was initially thought that subsenses and usage patterns for each term would be an aid to discrimination by providing a stronger basis for matches between true parallel blocks. The added terms beyond the critical primary sense in the dictionary had high hit rates with usage terms throughout the dictionary. The result was a noisy translation set that robbed the residue measure of discriminatory power. The results discouraged us from including the  $R$  measure in the error function for the dynamic programming system, although we suspect that improved dictionaries may ultimately provide better discrimination. It may also be possible to apply a kill list to the dictionary to reduce the number of high frequency terms in each definition, increasing the relevancy of the overall residue measure.

## 6 Implementation

The fact that our formulation of the alignment probability for two blocks is dependent on both the positive and negative information about the alignment probability means that the probability density functions can be used directly in the algorithm. Specifically, the distributions shown in Figures 3 and 4, as well as the distributions for ordered string comparisons and number comparisons, were loaded into the algorithm as histograms. During the dynamic programming operation, probability scores were determined by direct look-up of the  $\delta$  scores in the appropriate histogram, with some weighted averaging performed for values between the boundaries of the histogram bars for smoothing. This approach eliminated the necessity of estimating a distribution function for the rather non-Gaussian functions that are assumed to underlay the experimental data. Using this approach, the byte-length ratios could be simplified by not assuming a Gaussian-like distribution and directly using the histograms of byte-length probabilities. For comparison, however, we chose to use the Gale and Church derivation without modifying  $\delta_1$ .

## 7 Performance

Table 1 shows the performance of the original alignment algorithm compared to the improved algorithm. The results are for two documents. #HAND is the number of alignment blocks found in the hand aligned set. #FOUND is the number of alignment blocks found by the algorithms. Values of #FOUND lower than the value of #HAND indicates that alignment blocks that contain multiple segments have been found by the algorithm (e.g., a 3-3 match has supplanted three 1-1 matches). #CORRECT is the number of the found blocks which *exactly* match blocks in the hand aligned set. Note that the number of exact matches is a conservative estimate of the number of

acceptable alignments, as different translators may, for example, differ about whether a 2-2 match can take the place of two 1-1 matches and still be considered aligned.

In general, the performance of the improved alignment algorithm was very good, improving the hit rates from 23% to 49% on Document 1 and from 0.00381% to 70% on Document 2. The abysmal performance of the byte-length method on Document 2 can be attributed to the massive amounts of header information, significant added whitespace and inconsistent table and list formats that occurred in one document but not the other. The algorithm encountered only 1 hit (the document start) in the first quarter of the document. The training texts for these runs were the texts themselves, and therefore the results must be reviewed with care. The statistics of just two documents, applied directly to those two documents for evaluation does not necessarily provide a direct estimate of the same statistics to a broader spectrum of documents.

## 8 Conclusions

In the real world, poor-quality translations are common due to the preferences of individual translators, lack of formal format guidelines for translations and outright mistakes. Our method combines four feature scores into a simple measure of the probability of two textual segments aligning. The algorithm is fairly general in that all of the feature scores used are more or less applicable to a wide range of Spanish and English translations and are also applicable to a degree to other European languages. It is further likely that the methods we used can improve alignments between many non-European languages by exploiting the increasingly common English phrases and Arabic number occurrences in professional and public communications throughout the world.

Our alignment algorithm presents a new formulation of Bayesian methods combined with a direct approach to data fusion for multiple sources of information. This approach should work well with a wide range of data sources, including direct comparisons of co-occurrence probabilities for specific classes of lexical elements.

## References

CATIZONE, ROBERTA, GRAHAM RUSSELL & SUSAN WARWICK. 1989. Deriving Translation Data from Bilingual Texts. In *Proceedings of the First International Acquisition Workshop*, Detroit, MI.

CHEN, STANLEY F. 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information. In *Proceedings of the 31st Annual Conference of the Association of Computational Linguistics*, 9-16, Columbus, OH.

CHURCH, KENNETH W. 1993. Char-align: A Program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Conference of the Association of Computational Linguistics*, 1-8, Columbus, OH.

DAVIS, MARK W. & TED E. DUNNING. 1995. Query Translation using Evolutionary Programming for Multi-lingual Information Retrieval. To appear in *Proceedings of the Fourth Annual Conference on Evolutionary Programming*, San Diego, CA.

DUNNING, TED E. & MARK W. DAVIS. 1993a. A Single Language Evaluation of a Multi-Lingual Text Retrieval System. *NIST Special Publication 500-207: The First Text Retrieval Conference (TREC-1)*, D.K. Harman, Ed., Computer Systems Laboratory, NIST.

DUNNING, TED E. & MARK W. DAVIS. 1993b. Multi-Lingual Information Retrieval. *Memoranda in Computer and Cognitive Science*, MCCS-93-252, Computing Research Laboratory, New Mexico State University.

GALE, WILLIAM A. & KENNETH W. CHURCH. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Conference of the Association of Computational Linguistics*, 177-184, Berkeley, CA.

OGDEN, WILLIAM C. 1993. Norm - A System for Translators. Demonstration at *ARPA Workshop on Human Language Technology*, Merrill-Lynch Conference Center, Plainsboro, NJ.

SIMARD, M., G. FOSTER & P. ISABELLE. 1992. Using Cognates to Align Sentences in Bilingual Corpora. *Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*. Montreal Canada.

WU, DEKAI. 1994. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics*, 80-87, Las Cruces, NM.