Sentence alignment algorithm based on bi-variate population model

Scott Songlin Piao
Email: s.piao@dcs.shef.ac.uk

Tony McEnery
Email: mcenery@comp.lancs.ac.uk

Abstract

In the past decade, a number of algorithms of sentence alignment have been proposed and tested. Employing various approaches, including sentence-length based, cognates based and dictionary/linguistic knowledge based algorithms, those algorithms have obtained promising success on some language pairs. However, most of the algorithms depend on linguistic knowledge base or a-priori statistical parameters extracted from manually aligned sample. Such dependencies make it difficult to port the algorithms to other language pairs or domain/genres. In this paper, we present a new statistical algorithm, which is based on a statistical bivariate population model and less dependent on domain/genre. This algorithm was evaluated on an English-Chinese parallel corpus and obtained success rates between 91% -99% on data from different domains. Although it has bot been tested, possibility exists that our algorithm is language independent to some extent.

1. Introduction

In the past decade, numerous algorithms of sentence alignment have been suggested and evaluated (Brown *et al*., 1991; Simard *et al*., 1992; Gale and Church, 1993; Johansson *et al*., 1993; Kay & Roscheisen, 1993; McEnery *et al*., 1994; Wu, 1994; Fung, 1994; Hofland, 1995; Haruno & Yamazaki, 1996; Melamed, 1997; Collier *et al*., 1998; Simard, Foster, Hannan, Macklovitch, Plamondon, 2000; Oakes & McEnery, 2000). In particular Gale and Church's algorithm, which is based on sentence length in terms of character number, often with some modification and enhancement, has been widely applied to practical corpus alignment. This algorithm has obtained a remarkable success on some European language pairs. However, its performance, to a large extent, depends on *a-priori* statistical parameters: mean ratio of sentence lengths and its variance, and *a-priori* likelihood of different types of alignments such as 1:1, 1:2 and 1:3 alignments. When these parameters fit the corpus to be aligned a high success rate can be expected. Otherwise, the algorithm suffers from a higher error rate.

In this paper, we propose a new algorithm of sentence alignment which alleviates the dependency on a-priory parameters. Based on a statistical bivariate population model, instead of relying on a-priori parameters extracted from pre-aligned sample, this algorithm dynamically extracts local parameters from the corpus to be aligned and obtains an optimal local alignment. The algorithm is evaluated on Chinese-English parallel corpus (CEPC) built in Lancaster University. In our experiment this algorithm obtained a rather stable performance on samples from different domains.

2. A new sentence-alignment algorithm based on bivariate population model

In this section, we describe a new sentence alignment algorithm designed based on a statistical bivariate population model. The *regression line*, *correlation coefficient*, and *standard variation* of bivariate population are coordinated to achieve an optimum local alignment. In the following section, the bivariate statistical model will be described to provide a setting to the algorithm.

2.1. Statistical bivariate population model

In statistics, a bivariate population refers to a pair of corresponding variable sets. Alder & Roessler (1972: 195) defines it as follows:

> If for every measurement of a variable *X* we know a corresponding value of a second variable *Y*, the resulting set of pairs of variates is called a *bivariate population*.

If a bivariate population is plotted onto a scatter diagram, we can see points clustering around a notional line as shown in Figure 2.1 below (for the convenience of observation the axis is plotted explicitly). In the chart, the horizontal and vertical axes denote lengths of the corresponding English and Chinese sentences respectively.

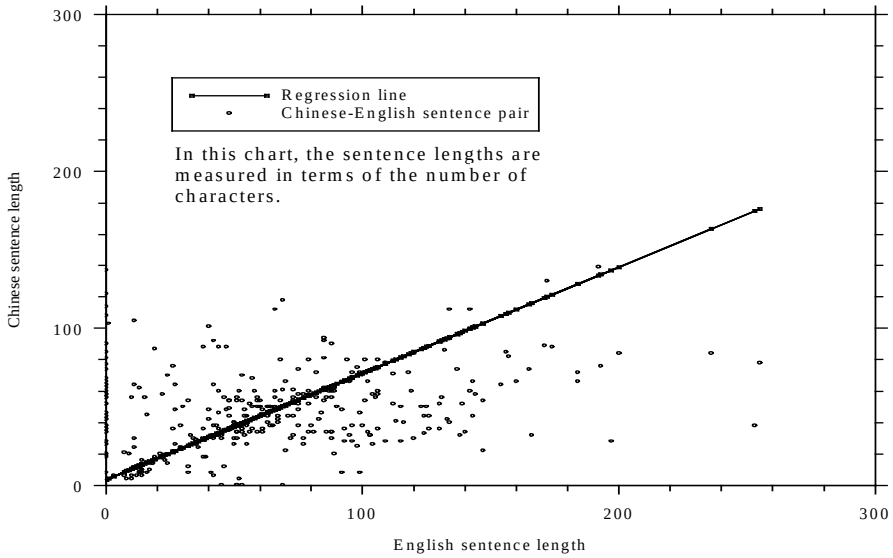

Figure 2.1: A sample chart plotting a bivariate population of English-Chinese sentence lengths

In figure 2.1, the line represents a certain linear relationship between the two groups of variables. Mathematically, this is a line to which the sum of distances from the points is the minimal. The line is called the *line of regression of prediction for Y*. Its mathematical expression can be obtained as follows.

In mathematics, a line can be expressed by a linear equation in the form of

(2-1) $$Y = a + bX,$$

where *X* and *Y* are variables and *a* and *b* are constants. A specific line is determined by the values of *a* and *b*. In order to let the equation (2-1) represent the regression line, *a* and *b* need to be evaluated in such a way that the points in the diagram lie as close to the line as possible.

For each *X*, a *Y*-value is calculated by the equation (2-1), which is an estimation of the actual *Y*-value. If we denote the estimated *Y*-value by Y_e, and the actual *Y*-value by *Y*, then the equation can be rewritten as,

(2-2) $$Y_e = a + bX,$$

where *X* and Y_e are called independent and dependent variables respectively. Then the absolute value $|Y - Y_e|$ represents the error entailed by using the estimated Y_e instead of the

actual *Y*-value. The regression line is the line that corresponds to the minimal value of the sum of these errors $\sum|Y-Y_e|$. In differential calculus, *a* and *b* must be the solution of the following simultaneous linear equations:

(2-3)
$$an+b\sum X=\sum Y,$$
$$a\sum X+b\sum X^2=\sum XY.$$

Hence, the value of *a* and *b* can be calculated as follows:

(2-4)
$$a=\frac{\sum X^2\sum Y-\sum X\sum XY}{n\sum X^2-(\sum X)^2},$$

(2-5)
$$b=\frac{n\sum XY-\sum X\sum Y}{n\sum X^2-(\sum X)^2}.$$

For a given bivariate population, the regression line is determined by formulae (2-4) and (2-5). In other words, the relationship between two variable groups of a bivariate population can be established with these two formulae.

For a pair of bivariate population, the level of correlation between them can be measured as follows. With the regression line determined, we can measure the level of spread of a set of points around the regression line by standard deviation, or standard error ($\hat{s}_e$), which can be calculated by the formula:

(2-6)
$$\hat{s}_e=\sqrt{\frac{\sum(Y-Y_e)^2}{n}}.$$

This standard error has the property that, under certain conditions, about 95% of the points will fall within the region $[Y_e-2\hat{s}_e, Y_e+2\hat{s}_e]$.

Another measure of the correlation of the bivariate population is the correlation coefficient, which is denoted as *r* and defined as

(2-7)
$$r=\frac{\sum xy}{\sqrt{\sum x^2\sum y^2}} \quad (|r|\le 1),$$

where

$$x=X-\overline{X},$$
$$y=Y-\overline{Y},$$
$$\overline{X}=\frac{\sum X}{n},$$
$$\overline{Y}=\frac{\sum Y}{n}.$$

If the *r*-value close to zero, that means there is a weak correlation between the two sets of variables; if the *r*-value close to ±1, that indicates there is a very strong linear relationship between them.

The standard error and correlation coefficient provide the means of estimating and measuring the correlation of bivariate populations. They provide the basic metrics based on which the new sentence alignment algorithm is built. However, before we proceed to describe the algorithm, let's present a case study to assess the reliability of the metrics.

2.2. A case study: do the correlation metrics reflect sentence alignment?

In order to apply the correlation metrics in sentence alignment, we first need to test whether the metrics reflect sentence alignment. In other words, do the correlation coefficient (denoted by r) and standard deviation (denoted by $\hat{s}_e$) really represent the degree of correlation of translated sentence pairs in a corpus? Below, this relationship is examined with a sample from the CEPC.

First, a pair of English and Chinese paragraphs were selected, which contain five English sentences and six Chinese sentences respectively, as shown in figure 2.2.

English paragraph:
 <s> Various heating products are available to help you keep warm: check in your local shops or gas or electricity showroom. </s> <s> Some electricity suppliers operate 'Budget Warmth' schemes designed to guarantee one warm room. </s> <s> Ask at your local electricity showroom or office for details. </s> <s> Staff at your local gas showroom can give advice on appropriate gas heaters or ring the British Gas or Electricity company energy efficiency free phone number for advice. </s> <s> You may prefer to contact your Local Energy Advice Centre for independent advice. </s>

Chinese paragraph:
 <s> 有 各种 各样 的 取暖 产品 可以 助 你 保持 温暖 。 </s> <s> 到 你 当地 的 商店 或 煤气 或 电力公司 陈列室 查看 。 </s> <s> 有些 电力 供应商 设有 省钱 保暖 方安 ， 目的 是 要 保证 有 一 间 房子 温暖 。 </s> <s> 向 你 当地 的 电力公司 陈列室 或者 办事处 查询 详情 。 </s> <s> 你 当地 的 煤气公司 陈列室 的 工作 人员 可 就 适当 的 便宜 的 煤气 暖 提供 建议 或 打 电话 给 英国 煤气公司 或 电力 公司 的 免费 节能 专线 得到 谘询 。 </s> <s> 你 可能 更 愿意 打 电话 给 你 当地 的 能源谘询中心 得到 个别 的 建议 。 </s>

Figure 2.2: A pair of English-Chinese sample paragraphs

When the sentence lengths are counted and listed, we obtain Table 2.1. In this table, the second and third columns show English and Chinese sentence lengths in terms of token (words plus punctuation marks) numbers while the last column shows the differences of lengths. As shown in Figure 2.2, none of the sentence pairs in the corresponding positions are true translations of each other. After the sentences are manually aligned, Table 2.1 becomes Table 2.2.

Paragraph

No.	Toknumb[E]	Toknumb[C]	Diff[E-C]
1	22	12	10
2	16	12	4
3	11	18	-7
4	30	11	19
5	14	34	-20
6	0	16	-16

Table 2.1: Chinese-English sentence length distribution in a sample paragraph

Paragraph

No.	Toknumb[E]	Toknumb[C]	Diff[E-C]
1	22	12+12	-2
2	16	18	-2
3	11	11	0
4	30	34	-4
5	14	16	-2
6	0	0	0

Table 2.2: Chinese-English sentence length distribution in the sample paragraph after alignment

In order to apply the bivariate population model, the English and Chinese sentence lengths are taken as a set of bivariate population. The former is taken as the independent variables and the latter the dependent variables. Our aim is to examine whether the paragraph produces distinct values of correlation metrics before and after the alignment.

From Table 2.1 and Table 2.2 r and $\hat{s}_e$ were calculated respectively. As a result, the former produced r-score of 0.310 and $\hat{s}_e$-score of 7.535 while the latter produced r-score of 0.999 and –score of 0.585. As shown in this sample, it is clear that there is a relationship between the value of these two metrics and the success of alignment (English and Chinese in this case).

Of course, such a small test is insufficient in itself to test the relationship. For a further examination, the same process was repeated on a larger sample containing 363 English-Chinese sentence pairs. The sample was divided into two groups: a) English-Chinese paragraph pairs containing the same number of (termed *Group I*), b) the other paragraph pairs (termed *Group II*). After the division, *Group I* contained 247 sentence pairs while *Group II* contained 116 sentence pairs. Here the sentences were paired simply by their locations in the corresponding paragraphs, including 1:0 or 0:1 pairs. After the paring, it was found that most sentence pairs in *Group I* are true translations of each other; the minor mismatches were manually corrected. On the other hand, 73 pairs in *Group II* were mismatches.

In this experiment, *Group I* and *Group II* are assumed to represent correctly aligned and mis-aligned corpora respectively. The two probabilistic values were calculated for the whole sample, then for *Group I* and finally for *Group II*. The results are shown in Table 3.7 below.

Probabilistic value	Whole sample	*Group I*	*Group II*
$\hat{s}_e$	6.344812	3.732527	7.721320
r	0.775947	0.938085	0.351723

Table 2.3: Standard deviation and correlation coefficient of a sample corpus (containing 363 sentences)

As shown in the table, *Group I* yields a high *r*-score and a small standard deviation ($\hat{s}_e$), which reflects a strong correlation between English and Chinese sentence length; *Group II* yields a low *r*-score and a greater $\hat{s}_e$, which reflects the weak correlation in *Group II* caused by numerous alignment mismatches. The *r*-score and $\hat{s}_e$ for the whole sample lie between these two groups, as one would expect.

The experiments presented above support my assumption that $\hat{s}_e$ and *r*-score can be used to estimate and measure sentence alignment. The test showed that these scores truly reflect the case of sentence alignment for a text and roughly correspond to expected error rates. The conclusion of my test is: *In the English-Chinese parallel corpora the degree of alignment in a local context is generally in direct proportion to the r-score.* However, it is open to question whether this is true for all language pairs and genres.

Based on this conclusion, a sentence alignment algorithm for Chinese-English parallel corpora is based on the above conclusion, as will be described in the following sections.

2.3. A boundary for distinguishing between true and false sentence alignment

In the previous section, we have shown that the correlation coefficient *r* and standard deviation $\hat{s}_e$ can be indicators for how likely a pair of sentences are true alignments. In this section, we explore the possibility of determining correctly matched and mismatched sentence pairs based on the correlation score.

We assume that this issue can be explored by considering whether a pair of candidate sentence alignments is subject to correlation or not. Suppose we have a pre-aligned representative sample corpus. Using this sample, we can extract the regression line and standard error $\hat{s}_e$ from the sample, which is representative of the corpus to be aligned. If we denote the source and target language sentences with the independent variate *X* and the dependent variate *Y* respectively, the range $[\, Y_e - 2\hat{s}_e, Y_e + 2\hat{s}_e \,]$ defines a region in which about 95% of the correctly matched sentence pairs in the pre-aligned sample fall (see section 2.1). Therefore, if the equation of the regression line is applied to the entire corpus, most correctly aligned sentence pairs in the whole corpus can be expected to fall within this region. On the other hand, if a sentence pair falls outside the region, it is likely to be a mismatched pair. In short, most true alignments should satisfy the following condition:

(2-8) $$Y \in [\, Y_e - 2\hat{s}_e, Y_e + 2\hat{s}_e \,].$$

For the sake of simplicity, as well as for a clearer graphical interpretation of the range of variance, the expression in (3-13) is converted into the following equivalent form: $|Y - Y_e| \leq 2\hat{s}_e$. However, this expression is not flexible enough to cope with sentences of different lengths. Greater differences of sentence length were observed for longer sentence translation pairs than for shorter ones, as one would expect. This indicates that standard error $\hat{s}_e$ should be relaxed for longer sentences. To reflect this, standard error $\hat{s}_e$ is weighted as follows.

Let *ML* be the mean sentence length of the corpus, let *LW* be the length weight for $\hat{s}_e$, and let *X*(*i*) and *Y*(*i*) be a pair of English-Chinese sentence lengths, then

(2 - 9) $$LW = \begin{cases} 1 & \text{if } X(i) \leq ML \text{ and } Y(i) \leq ML, \\ \sqrt{\dfrac{Max(X(i), Y(i))}{ML}} & \text{if } X(i) > ML \text{ and } Y(i) > ML. \end{cases}$$

In formula (2-9), the value of $\sqrt{\frac{Max(X(i),Y(i))}{ML}}$ is always greater than 1. Finally, expression 2-8 becomes

(2-10) $$|Y - Y_e| \leq 2\,\hat{s}_e * LW.$$

The expression (2-10) defines a range in which most correct sentence alignments are expected to fall. Figure 2.2 illustrates this range for a sample from the CEPC containing about 2,221 sentence pairs. In this experiment, the English and Chinese sentences in corresponding positions in the matched paragraphs are paired. In this chart, each dot represents a pair of matched English-Chinese sentences. The horizontal axis represents English sentence lengths *X* (as the source language) while the vertical axis represents the variance of actual Chinese sentence lengths *Y* (as the target language) from the expected length Y_e. The vertical dotted line denotes the mean sentence length. The other dotted lines and curves enclose an area defined by formula 2-10.

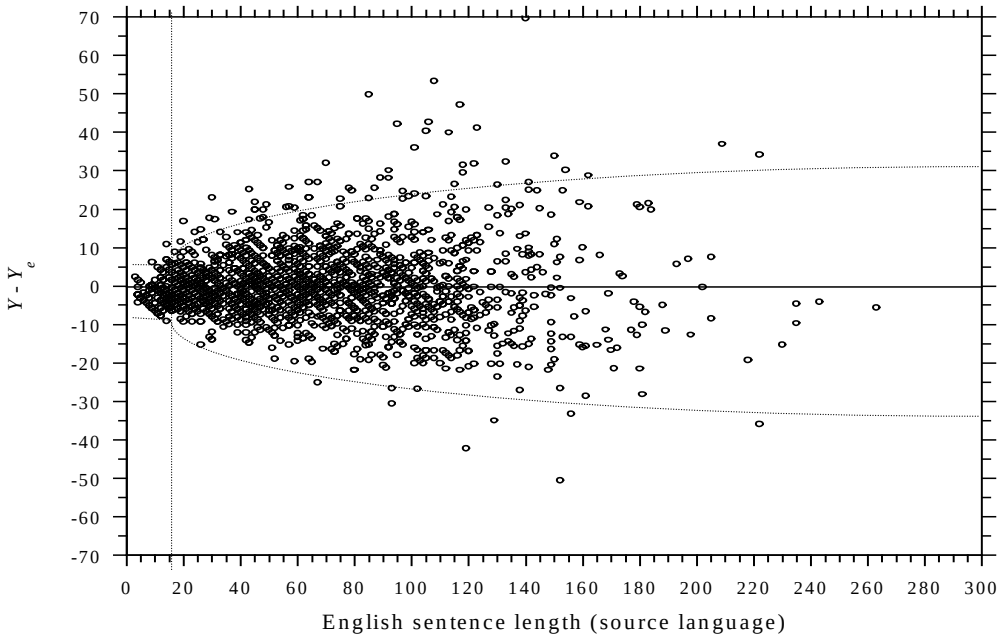

Figure 2.2: Illustration of the range of $|Y - Y_e| \leq 2\hat{s}_e * LW$ ($\hat{s}_e$=6.8)

In figure 2.2, for a given pair of Chinese-English candidate alignments, if the actual Chinese sentence length *Y* is equivalent to the expected length Y_e, the point representing the sentence pair will fall on the solid line Diff[Y-Ye] = 0. Otherwise, if the actual Chinese sentence length is longer or shorter than the expected length, the point will float above or fall below the solid line. The area enclosed by dotted lines denotes the region $[-2\hat{s}_e * LW, +2\hat{s}_e * LW]$. Here $\hat{s}_e$ is weighted for those sentences longer than mean length (15.9 in this case. Depicting the weighting, after the point of mean length, the dotted lines depart wider as the sentence length grows. In the figure, if a given pair of candidate sentence alignments are true translations of each other, the points representing them are likely to fall within the range defined by the dotted lines. If the pair is a false match, then the points representing them are likely to fall outside the range defined by the dotted lines.

We assume there exists a **critical value** (CV henceforth) near the $\hat{s}_e$-value which defines an optimal region [-2CV*LM, +2CV*LM] that covers as many true alignments as possible. The $\hat{s}_e$-value obtained from the pre-aligned sample approximates the CV. If a regression line and the CV determined for a corpus to be aligned, such an optimal region can be defined.

2.4. A sentence alignment algorithm

As shown in the previous section, a region of standard deviation can be derived from a pre-aligned representative sample which can be used to distinguish most true sentence alignments from false ones. The standard deviation $\hat{s}_e$, when weighted properly, approximates a critical value CV which can be used to judge if a pair of candidate sentence alignments are true or false matches.

However, in order to apply this approach to aligning sentences, there are two issues must be solved. Firstly, it is not desirable to prepare a pre-aligned sample each time we align a corpus. Secondly, it can be argued that the regression line and standard deviation extracted from a pre-aligned sample are not necessary accurate for a given corpus to be aligned.

In the following sub-sections, three sub-algorithms will be described. They address the issues mentioned above and comprise a complete algorithm to align sentences automatically. In details, these three sub-algorithms implement the following three tasks:

1) Automatic extraction of a local regression line and a standard deviation for the true sentence alignments in a given division[1] of a parallel corpus (section 3.4.4.1).
2) A core algorithm for aligning sentences with a given regression line and standard deviation (3.4.4.2).
3) Automatic extraction of the CV which produces an approximation to optimum alignment within a given division of the corpus (3.4.4.3).

2.4.1. Automatic extraction of the local regression line and standard deviation for sentence alignment

For a statistical sentence alignment algorithm, the necessity of pre-aligned sample for extracting a-priori parameter causes problems in practical alignment. For example, Gale and Church (1993) suggested that universal a-priori parameters, the mean ratio of sentence lengths and its variance, and a-priori likelihood of different types of alignments, are applicable for most European languages. However, this assumption has been questioned by controversial results.

For example, in the CRATER Project, McEnery *et al*. (1996) reported a success rate of 75% on English-German sentence alignment in contrast to 98% on English-French sentence alignment. Later, they applied the same algorithm on four different genres of English-Polish corpus, the algorithm yielded widely different success rates: 100%, 64.4%, 66.7% and 85.2% on fiction, fiction, medical and financial genres respectively (McEnery *et al*. (1997: 221). These results suggest that the parameters used in Gale et al.'s algorithm might be dependent not only on different languages but also on different genres.

One possible solution to the problem is to pre-align sample corpus and extract local a-priori parameters from it each time aligning a corpus. Obviously it is not desirable, if not impossible in practical work. An alternative approach we suggest here is to extract local a-priori parameters, regression line (see formula 2-1, 2-3 and 2-4) and standard deviation (see formula 2-6), from the paragraph pairs containing the same numbers of sentences in the corpus to be aligned, termed *Group I* (see section 2.2).

This approach is based on our assumption that, in *Group I*, the majority of sentences in the corresponding positions are true translations of each other. If it is true, it is possible to obtain correctly aligned sample by simply pairing sentences in positional sequence within each paragraph. Although minor mismatches might occur in the sample, the impact of the errors

[1] Here division refers to a part of a corpus which is kept in a computer file.

can be alleviated by filtering out some, is not all, of the mismatches. Because such samples derive from the corpus, often a division of it, to be aligned, they are accurately representative of the local context.

In order to test our assumption, three samples were selected from the CEPC: *health*, *education* and *zaobao*. The first sample is from government leaflets and the text is translated highly literally. The second sample is from booklets and the texts are less literally translated comparing to the former. The last sample is from a Chinese-English bilingual column of Singaporean online newspaper *Lianhe Zaobao*, and more creative translation is involved. In details, first the paragraph pairs containing the same numbers of sentences were collected. Second, the sentences are sequentially paired within each paragraph. Lastly the sentence pairs were manually checked to find mismatches pairs. Table 2.4 shows the result.

Genre of sample	No. of sentence pairs	Mismatches	Percentage of correct matches
health	1,869	2	99.89%
education	455	19	95.82%
A sample of *zaobao*	359	28	92.20%

Table 2.4: Percentage of true alignments among sequence-paired sentences in *Group I*

As shown in table 2.4, although the percentage of true alignments decrease as freer translation is involved, in average 95.97% of the sentence pairs are true alignments. Therefore, we assume that, after a pruning, *Group I* can be used as an approximation of pre-aligned sample. The Group I data is pruned as follows.

It is found that the length difference between mis-aligned sentences is tend to be greater than that of true alignments. A score *ft* based on the length difference is used to determine and filter out mismatches. It is calculated as follows:

(2-11) $$ft = \frac{|l_c - l_e|}{l_c + l_e},$$

where l_e and l_c denote the length of English and Chinese sentences respectively. This formula takes the sum of the Chinese and English sentence lengths into account, allowing greater length variation for lengthy matching sentences. For example, suppose we have two pairs of sentences, *pair* 1 and *pair* 2, as follows:

pair 1: $l_{c1} = 5$, $l_{e1} = 14$
pair 2: $l_{c2}=30$, $l_{e2} = 39$

Then,
$ft(l_{c1}, l_{e1}) = |5\text{-}14|/(5+14) = 0.47$,
$ft(l_{c2}, l_{e2}) = |30\text{-}39|/(30+39) = 0.13$.

Figure 2.3: Calculation of *ft* scores

As shown in Fig. 2.3, although the length difference of both of the pairs is 9, the *ft* score for the first pair (0.47) is greater than that for the other pair, which is longer (0.13). If a threshold of 0.4 is used, the first pair would be filtered out while the second pair would be accepted.

When extracting the regression line and standard deviation from *Group I* of a given corpus, every pair of sentences involved was tested for *ft*-score. If *ft* is greater than a given threshold, the sentence pair is excluded from the calculation. The threshold was determined empirically from the observation of a contrastive data of *ft* score *vs*. the sum of l_c+l_e. Figure 3.9 and Figure 3.10 illustrates the pruning the *health* and *education* subcorpora.

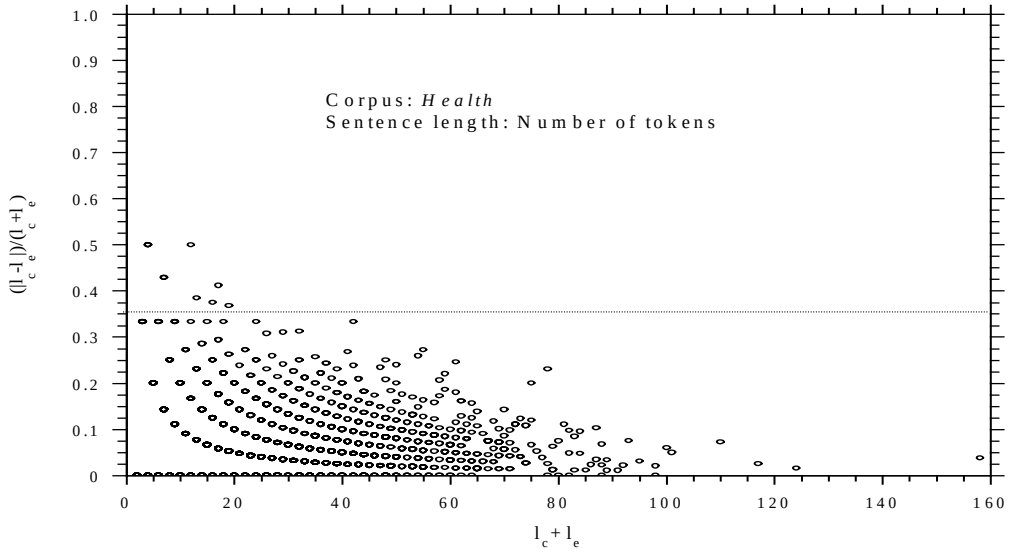

Figure 2.4: Chart of *ft vs.* l_c+l_e of *Group I* of *health* (1,869 sentence pairs included)

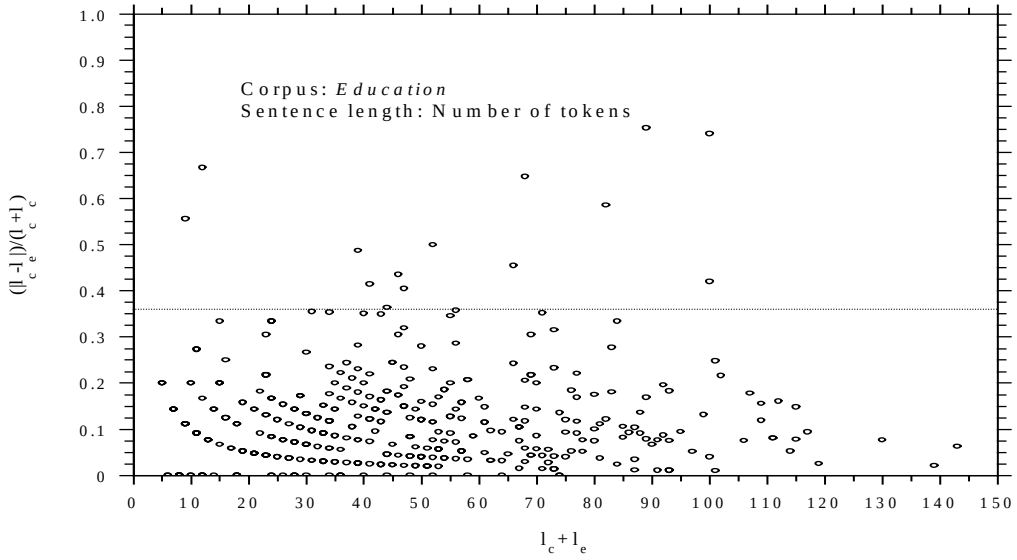

Figure 2.5: Chart of *ft* vs. l_c+l_e of *Group I* of *education* (455 sentence pairs included)

As shown in figures 2.4 and 2.5, the majority of sentence pairs occur in the lower part of the charts. In this specific case a *ft* threshold of 0.37 was used. This threshold filtered out 13 sentence pairs from *Group I* of the *education* subcorpus, among which 10 were mismatches. Note that there were 19 mismatches in the total of 455 sentence pairs in the same sample (see table 2.4). Therefore, the filter removed 52.63% of the mismatches, improving the percentage of true alignments from 95.82% to 97.96%. As the experiment shows, although this threshold failed to remove every mismatch while filtering out some true alignments, it clearly improves the precision of the regression line as it increases the percentage of true alignments in *Group I*.

The algorithm presented in this section makes it possible to automatically extract a regression line and standard deviation for the corpus to be aligned. Although minor mismatches may be included in the process, a close approximation can be expected.

2.4.2. The core algorithm for sentence alignment

In the previous section, an algorithm was presented for automatically extracting the regression line and standard deviation $\hat{s}_e$ for a given corpus. With these parameters available, it is possible to apply the bivariate population model (see section 2.1) to the sentence alignment. Here $\hat{s}_e$ is an approximation of the critical value CV. The extraction of true CV will be addressed in section 2.4.3. In this section, supposing the CV equals to $\hat{s}_e$, a core algorithm will be presented for aligning sentences.

The algorithm works as follows. For a given regression line and a CV, sentences in a given division are aligned as follows (also refer to Fig. 3.11):

a) For a given source sentence length *Xi* (the independent variable), calculate the expected target sentence length Y_{ie} with the equation of the regression line (see formula 2-1, 2-3 and 2-4) established from *Group I* of the corpus to be aligned.
b) Compare the difference between the actual and expected target sentence lengths $|Y_i - Y_{ie}|$ against the CV. If the difference falls within the range of [-2CV**LW*, +2CV**LW*], judge the pair to be a true alignment, then re-start step a) for the next sentence pair. Otherwise judge the pair to be a mismatch and proceed to step c).
c) If a pair of sentences X_i, Y_i (i = 1, ..., m, where m refers to the number of source sentences in a given paragraph) is judged to be a mismatch, the following sentence is taken into consideration. If $Y_i > Y_{ie}$, then X_{i+1} will be added to X_i to form a new candidate X_{i2} (2:1 alignment); if $Y_i < Y_{ie}$, then Y_{i+1} will be combined with Y_i to form a new candidate Y_{i2} (1:2 alignment). This step can be applied recursively to X_{i2} and Y_{i2}, leading to X_{i3} or Y_{i3} (1:3 or 3:1 alignment)
d) For the updated sentence pair, (X_{i2}, Y_{i2}) or (X_{i3}, Y_{i3}), repeat the process from step a) to c).
e) The procedure from step a) to step d) will be repeated until the pair is judged to be a true alignment in step b) or the k in X_{ik} or Y_{ik} exceeds 3 (exceed 1:3 or 3:1 alignment). Then a new cycle of the whole procedure begins for the next candidate sentence pair.

It was found that a sentence from a shorter paragraph (in terms of the number of sentences) is more likely to be translated into multiple sentences in the longer counterpart paragraphs. Based on this observation, an optional substitute algorithm for step c) was set up, as shown below:

c') Suppose m and n refer to the numbers of source and target sentences in a given paragraph respectively, then if $m>n$, X_i is given *a-priori*ty to combine with X_{i+1} to form a new candidate X_{i2}; if $m<n$, Y_i is given *a-priori*ty to be combined with Y_{i+1} to form a new candidate Y_{i2}.

Figure 2.6 describes the algorithm in quasi-program expression. In the figure, the double equals mark "==" denotes an alignment, $X(i)$ denotes the length of the ith source language sentence, $Y(i)$ denotes the length of the ith target language sentence, and $Y_e(i)$ denotes the expected length of the ith target language sentence. In lines 7 and 14, the two pairs of differential equations divided by slashes represent two options provided by the algorithms in c) and c').

SUPPOSE $Y = aX + b$ is the regression line, and m and n denote numbers of source and target language sentences, ML denotes the mean sentence length of the corpus, and LW denotes the length weight for the CV.

THEN

(1) $ML = \dfrac{\sum_{i=1}^{m} X(i) + \sum_{j=1}^{n} Y(i)}{m+n}$;

(2) **IF**($Max(X(i), Y(i)) > ML$) **THEN** $LW(i) = \sqrt{\dfrac{Max(X(i), Y(i))}{ML}}$;

(3) **ELSE** $LW(i) = 1$;

(4) $Y_e(i) = a + bX(i)$;

(5) **IF** $|Y(i) - Y_e(i)| \le (2CV*LW(i))$ **THEN** $X(i)==Y(i)$;
(6) **ELSE {**
(7) **IF** $(Y(i) < Y_e(i)) / (m < n)$ **{**
(8) **IF** $| Y(i) + Y(i+1) - Y_e(i)| \le (2CV*LW(i))$ **THEN** $X(i) == Y(i) + Y(i+1)$;
(9) **ELSE {**
(10) **IF** $((Y(i) + Y(i+1)) < Y_e(i))$ **THEN** $X(i) == Y(i) + Y(i+1) + Y(i+2)$;
(11) **IF** $((Y(i) + Y(i+1)) > Y_e(i))$ **THEN** $X(i) + X(i+1) == Y(i) + Y(i+1)$;
(12) **}**
(13) **}**
(14) **IF** $(Y(i) > Y_e(i)) / (m > n)$ **{**
(15) $Y_e(i+1) = a + b(X(i) + X(i+1)$;
(16) **IF** $|Y(i) - Y_e(i+1)| \le (2CV*LW(i))$ **THEN** $X(i) + X(i+1) == Y(i)$;
(17) **ELSE {**
(18) **IF** $(Y(i) > Y_e(i+1))$ **THEN** $X(i) + X(i+1) + X(i+2) == Y(i)$;
(19) **IF** $(Y(i) < Y_e(i+1))$ **THEN** $X(i) + X(i+1) == Y(i) + Y(i+1)$;
(20) **}**
(21) **}**
(22) **}**

Figure 2.6: Core algorithm of sentence alignment

This core algorithm works upon one paragraph each time. Given that the majority of sentences in *Group I* of the CEPC are 1:1 alignments (see section 2.4.1), when the core algorithm is applied to *Group I*, 1:1 alignment is given priority by adding a priority weight (PW) to the threshold 2CV. Thus, formula 2-10, the basic condition for sentence alignment, becomes

$$|Y - Y_e| \le 2CV+PW, \tag{2-11}$$

where CV = $\hat{s}_e$. The PW was empirically set to 5. In addition, it was found that very few 1:3 or 3:1 alignments exist in *Group I* (see Table 4.2 in section 4.2.1), therefore the multiple alignment is constrained to 1:2/2:1 alignments for *Group I*.

The suggested algorithm deals with various alignments: 1:1 (line 5), 1:*n* (lines 7-10), *n*:1 (lines 14-18) and 2:2 (lines 11 and 19), where n = 1, 2, 3. However, the sub-algorithms in lines 11 and 19 failed to detect any 2:2 alignments in my data, even though four such alignments were present in the experiment. As the problem is marginal and it more likely make mistakes than aligning correctly, it was decided to exclude these two lines from the program, thus precluding any chance of identifying 2:2 alignments.

Note that the core algorithm described in this section is based on the CV which provides the algorithm with the means of distinguishing true alignments from mismatches. It was supposed that CV equals to $\hat{s}_e$, which is not necessarily true. An algorithm will be presented in the following section for determining the CV automatically.

2.4.3. Automatic extraction of the CV

In section 2.3, we assumed that there exists a value CV near $\hat{s}_e$ which defines a range $|Y - Y_e| \leq 2CV+PW$ that optimally distinguishes true alignments from false ones. Later in section 2.4.2, we let $\hat{s}_e$ extracted from the *Group I* of the corpus be the value of CV. In practice, however, although $\hat{s}_e$ may approximate the actual CV, it is not necessarily the CV itself. Therefore, an algorithm is needed to determine the CV for a given $\hat{s}_e$.

In order to automatically determine the CV, a recursive algorithm is designed which incorporates the correlation coefficient and standard deviation. The procedure of this algorithm is:

a) A rough range is determined, within which the CV is assumed to fall.
b) Each value within the search range is assumed to be a candidate CV. For each candidate CV, the core alignment algorithm described in figure 2.6 is carried out on the data. The correlation coefficient is calculated for each output of the algorithm.
c) The alignment which produces the maximum correlation coefficient is searched for. This candidate CV is then taken to be a true CV.

In section 2.2, it was shown that the correlation coefficient *r* can predict the degree of alignment. Here, the correlation coefficient is used as an indicator of the level of sentence alignment in the corpus under consideration. Note that r-score is generally in direct ratio to the level of sentence alignment of a corpus. Therefore, the CV, which corresponds to the highest correlation coefficient score, should result in a close approximation to optimum sentence alignment in the corpus to be aligned. This algorithm is described in details below.

First, a search range for the CV is determined. In section 2.2, it was found that *Group I* (paragraph pairs which contain the same number of sentences) of the CEPC sample yielded a small standard deviation (denoted by $\hat{s}_e(1)$) while *Group II* (paragraph pairs containing different numbers of sentences) yielded a greater standard deviation (denoted as $\hat{s}_e(2)$). The standard deviation extracted from *Group II* tends to be big. Therefore, it is reasonable to assume that the CV for the whole corpus (*Group I* + *Group II*) is less than $\hat{s}_e(2)$, and lies somewhere between [0, $\hat{s}_e(2)$] (0 is the minimal value for CV). In the case of highly literal translations, the CV is expected to be close to $\hat{s}_e(1)$, therefore the search range can be narrowed down to [$\hat{s}_e(1)-\delta$, $\hat{s}_e(2)$], where δ is a positive constant. In fact, the former range is a specific case of the latter range where $\delta = \hat{s}_e(1)$. If this assumption is true, then when we plot the candidate CVs and corresponding correlation coefficients (denoted by an *r*-score) in a curve chart, the global peak *r*-score will occur within either one of two search ranges. The sustainability of the search ranges [0, $\hat{s}_e(2)$] and [$\hat{s}_e(1)-\delta$, $\hat{s}_e(2)$] will examined shortly.

With regards to the simulation of the sentence alignment algorithm (see Figure 2.6), first, all of the relevant statistical data and information, including sentence lengths, were extracted from the corpus. Given a regression line equation that is extracted from *Group I* (see section 2.4.1) and a candidate CV, the core sentence alignment algorithm was carried out. For each output, the correlation coefficient was extracted and recorded as a pair with the corresponding candidate CV.

The algorithm is repeated for all available candidate CVs within the search range introduced previously, resulting in a set of pairs of candidate CVs and their corresponding correlation

coefficients. The candidate CV with the highest correlation coefficient was selected as the true CV[2].

In order to test the algorithm suggested so far and check the sustainability of the [0, $\hat{s}_e(2)$] and [$\hat{s}_e(1)$–δ, $\hat{s}_e(2)$] ranges, the search algorithm was tested on nine divisions of the CEPC. In order to examine whether the global peak *r*-score really occurs within the search range, either [0, $\hat{s}_e(2)$] or [$\hat{s}_e(1)$–δ, $\hat{s}_e(2)$], the search range in the experiment was extended to [0, 20]. The nine charts in Figure 2.7 illustrate the search processes for the nine divisions. In this figure, the horizontal axis denotes the candidate CV while the vertical axis denotes the correlation coefficient.

As shown in figure 2.7, if we denote [$\hat{s}_e(1)$–δ, $\hat{s}_e(2)$] (δ = 0.5) with range(1) and [0, $\hat{s}_e(2)$] with range(2), in the five divisions of the *health* subcorpus which represent a literal translation, the peak *r*-score was occurred within range(1) (see charts 1-5). However, in the four divisions of the *education* subcorpus, which is translated less literally, range(1) failed to include the global peak r-score (see charts 6-9). This shows that range(1) does not guarantee the inclusion of the CV within the search range for various genres. On the other hand, range(2) contains the CV corresponding to the global peak *r*-score in all of the charts in Fig. 3.12. It shows that $\hat{s}_e(2)$ effectively defines the maximum range of the CV. Therefore range (2) is selected as the default search range for the CV in this thesis. In fact, the only difference the different search ranges cause is a slight difference in speed, with range(1) slightly faster than the other.

This algorithm for the extraction of the CV enables the sentence alignment algorithm to be implemented automatically, as will be shown in the following section.

[2] Because of the small intervals used in the algorithm, such as 0.1, the OCV actually represents a range of values. For example, in chart 1 of Fig. 2.7, the OCV is any value between [3.7, 4.0].

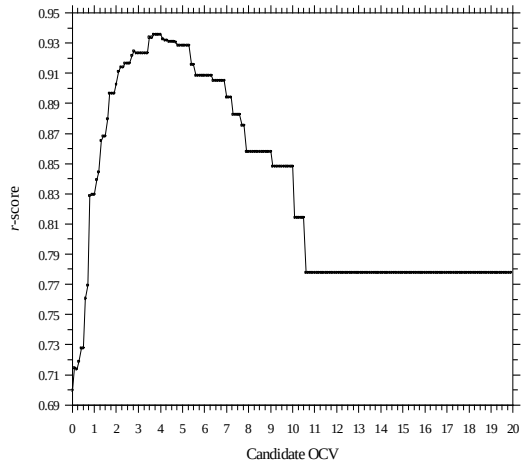

Chart 1: *hlth1* ($\hat{s}_e(1)$=3.67, $\hat{s}_e(2)$=7.68)

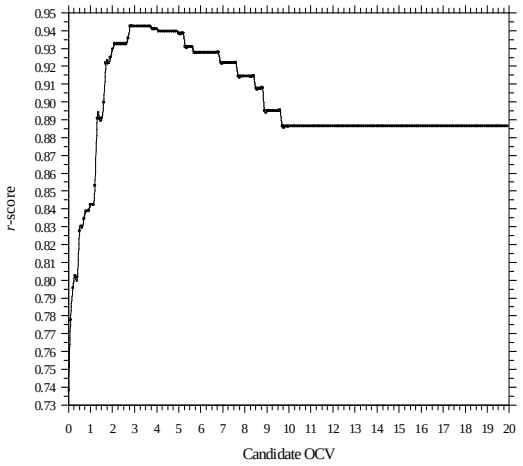

Chart 2: *hlth2* ($\hat{s}_e(1)$=3.12, $\hat{s}_e(2)$=5.80)

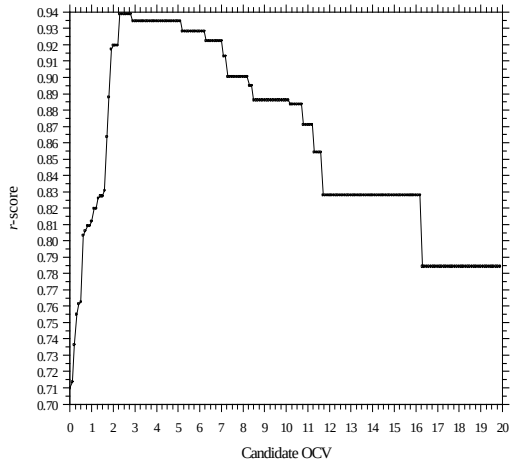

Chart 3: *hlth3* ($\hat{s}_e(1)$=3.63, $\hat{s}_e(2)$=11.64)

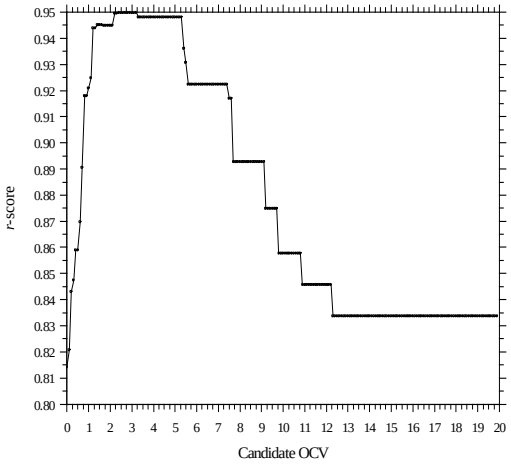

Chart 4: *hlth4* ($\hat{s}_e(1)$=2.83, $\hat{s}_e(2)$=7.55)

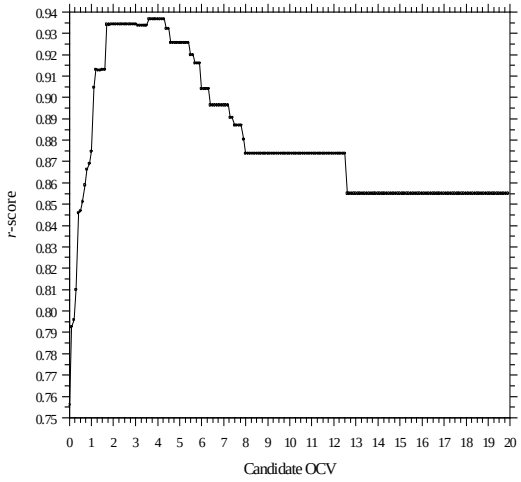

Chart 5: *hlth5* ($\hat{s}_e(1)$=2.92, $\hat{s}_e(2)$=7.10)

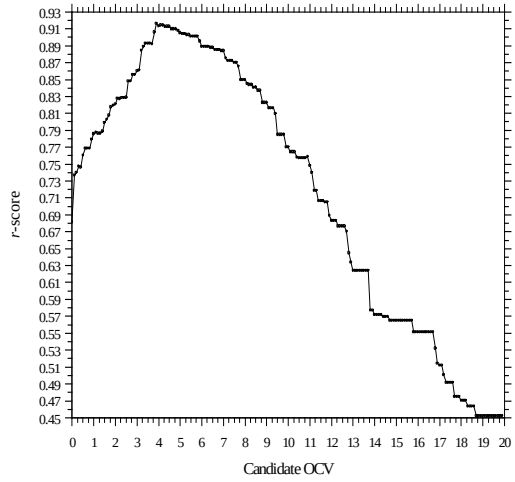

Chart 6: *education1* ($\hat{s}_e(1)$=8.40, $\hat{s}_e(2)$=18.44)

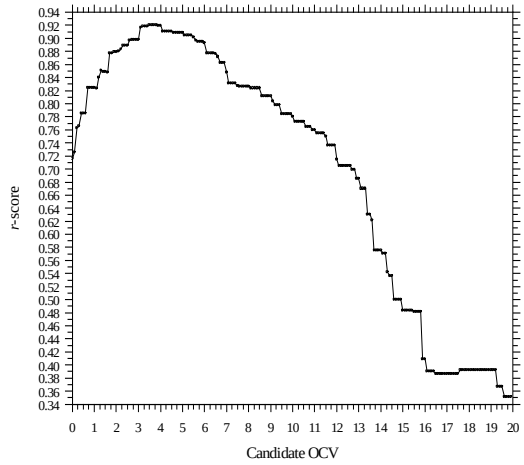

Chart 7: *education2* ($\hat{s}_e(1)$=5.31, $\hat{s}_e(2)$=20.70)

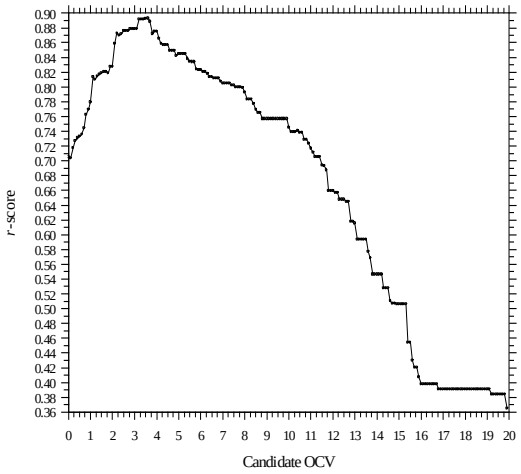

Chart 8: *education3* ($\hat{s}_e(1)$=6.52, $\hat{s}_e(2)$=19.14)

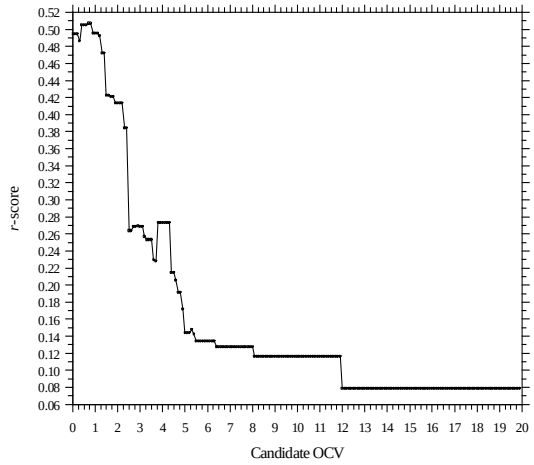

Chart 9: *education4* ($\hat{s}_e(1)$=15.14, $\hat{s}_e(2)$=23.66)

Fig. 2.7: Search for the CV

2.5. A statistical system of sentence alignment

The algorithms described in the previous sections are incorporated into an automatic sentence alignment system. The system works in three steps:

a) Extract relevant statistical data and positional information regarding titles, paragraphs and sentences.

b) Extract the CV using the algorithms developed in section 2.4.3, producing a simulated output of sentence alignment containing positional information on headings, paragraphs and sentences.

c) Retrieve tokens for appropriate positions designated by the positional information in the simulated output.

Note that in section 2.4.1, we argued that *a-priori* parameters, such as the **mean ratio** of sentence lengths c, **variance** s^2 and the ***a-priori* likelihood** of different types of alignments in Gale *et al.*'s algorithm, are dependent on the language pair and the genre/domain. Similarly,

we argue that the regression line, standard deviation and CV in our sentence alignment system are language pair- and genre/domain-dependent, too. Therefore, the system recalculates these parameters based on the local context of each given division[3] of a corpus. The sentences were then aligned in the division using the local parameters. As will be shown in section ???, local parameters produce better alignment results than a global parameter of a larger corpus.

In terms of speed, although the system includes a rather heavy computation of statistical scores and a recursive simulation in the search for the CV, it is quite efficient. It takes 18 seconds in *SunOS 4.1.3_U1* workstation to align the sentences in the CEPC, which contains 61,534 English words and 197,074 bytes of Chinese characters.

Figure 2.8 shows a sample output of the sentence alignment system. In the sample, the source of the aligned regions is shown in one of two forms:

<*L*[1/2]_*Title_articlenumber*> and

<*L*[1/2]_*number1_number2_number3_number4*>.

In both, *L1* and *L2* denote the source language and target language respectively. In the first case, *Title* means the following sentence is a heading, and *articlenumber* indicates the location of the text in the corpus. For example, <L2_Title_20> means that the following item is a target language heading of the twentieth article in the corpus.

In the second case, *number1* indicates the position of the whole article/text in the corpus, *number2* indicates the position of the paragraph in the article/text, *number3* indicates the position of the sentence in the paragraph, and *number4* indicates the type of alignment. *Number4* has five possible values - 11, 12, 13, 21, 31 -which represent 1:1, 1:2, 1:3, 2:1, 3:1 alignments respectively. Also, each word is linked to its POS tag and, in the case of English, its lemma.

[3] Usually the computer files which comprise a corpus are considered as the divisions.

```
**************************************************************************
<L1_Title_20> THE_AT_the HEALTH_NN1_health OF_IO_of THE_AT_the
                NATION_NN1_nation AND_CC_and YOU_PPY_you
<L2_Title_20> 国民_NN0 以及 _CJ0 您_PN0 的_DE1 健康_NN0
**************************************************************************

***Paragraph [1]***
***============================================================***
<L1_20_1_1_11> Introduction_NN1_introduction
<L2_20_1_1_11> 介召_NN0
***------------------------------------------------------------***

***Paragraph [2]***
***============================================================***
<L1_20_2_1_11> No_PNQS31_no matter_PNQS32_matter who_PNQS33_who we_PPIS2_we
                are_VBR_be or_CC_or what_DDQ_what we_PPIS2_we do_VD0_do
                ,_,_PUNC we_PPIS2_we all_DB_all know_VV0_know how_RGQ_how
                precious_JJ_precious good_JJ_good health_NN1_health is_VBZ_be
                ._._PUNC
<L2_20_2_1_11> 不管_CJ0 我们_PN0 是_VS 谁_PN0 □_XX0 或者_CJ0 做_VT0 什麽_PN0
                □_XX0 大家_PN0 都_AV0 知道_VT0 健康_NN0 是_VS 多麽_AV0
                的_DE1 宝贵_AJ0 °_XX0
***------------------------------------------------------------***
<L1_20_2_2_12> After_CS_after several_DA2_several months_NNT2_month of_IO_of
                wide_JJ_wide consultation_NN1_consultation ,_,_PUNC the_AT_the
                Government_NNJ_government has_VHZ_have produced_VVN_produce
                "_"_PUNC The_AT_the Health_NN1_health of_IO_of the_AT_the
                Nation_NN1_nation "_"_PUNC ,_,_PUNC a_AT1_a strategic_JJ_strategic
                plan_NN1_plan aimed_VVD_aim at_II_at achieving_VVG_achieve
                better_JJR_good health_NN1_health for_IF_for everyone_PN1_everyone
                in_II_in England_NP1_england ._._PUNC
<L2_20_2_2_12> 经过_VT0 数_MC 个_NMW 月_NN0 广泛_AJ0 的_DE1 征求_VT0
                意见_NN0 □_XX0 政府_NN0 推出_VT0 了_AUX 这_PN0 本_NMW
                “_XX0 国民_NN0 健康_NN0 “_XX0 °_XX0 这_PN0 一_MC 策略_NN0
                性_ELM 计划_NN0 的_DE1 目的_NN0 在于_VI0 使_VI0 每_PN0 一_MC
                位_NMW 在_PRP 英国_NN0 居住_VI0 的_DE1 人_NN0 都_AV0
                获得_VT0 更_AV0 好_AJ0 的_DE1 健康_NN0 °_XX0
***------------------------------------------------------------***
<L1_20_2_3_11> The_AT_the plan_NN1_plan sets_VVZ_set out_RP_out targets_NN2_target
                for_IF_for the_AT_the nation_NN1_nation 's_VBZ_be health_NN1_health
                ._._PUNC
<L2_20_2_3_11> 该_PND 计划_NN0 提出_VT0 国民_NN0 健康_NN0 的_DE1 目标_NN0
                □_XX0
***------------------------------------------------------------***
<L1_20_2_4_11> This_DD1_this is_VBZ_be the_AT_the first_MD_first time_NNT1_time
                that_CST_that such_DA_such targets_NN2_target have_VH0_have
                been_VBN_be set_VVN_set for_IF_for England_NP1_england ._._PUNC
<L2_20_2_4_11> 这_PN0 是_VS 英国_NN0 首次_AV0 制定_VI0 的_DE1 类似_AJ0
                目标_NN0 °_XX0
***------------------------------------------------------------***
```

Fig. 2.8: A sample of the Chinese-English sentence alignment

3. Evaluation

The algorithm described in the previous sections was implemented in a C-program and tested on the CEPC, which contains 61,534 English words and 98,537 Chinese characters. The aligned sentences were checked manually. For a comparative study on texts from different domains, the statistics on the performance of the program were collected separately for two sections of the corpus, *health* and *education*. Also, in order to examine the influence of corpus size on the performance of the algorithm, the *health* section was also divide into five parts[4] and aligned separately. Furthermore, in order to examine the affect of different measures of sentence length on sentence alignment, i.e. number of characters versus number of tokens, the *health* section was aligned with these two measures separately and for a comparison. Note that in section 3.1, the sentence length is measured with the number of tokens while in section 3.2 the sentence length is measures with the number of characters.

3.1. Statistical analysis on the performance of the algorithm

First, the results from the *health* section were examined. Because English is the source language in this section, it is taken as the independent variate. Table 3.1 shows the overall success rates for the corpus and its five divisions. The first five rows describe the program's performance for the five divisions, the sixth row shows the mean accuracy for the five cases, and the seventh row shows the case in which the *health* section was aligned as a whole.

[Independent variate: English]

Divisions	Aligned Pairs	Mismatches	Error Rate	Accuracy
hlth1	351	17	4.84%	95.16%
hlth2	446	0	0.00%	100.00%
hlth3	435	1	0.23%	99.77%
hlth4	417	0	0.00%	100.00%
hlth5	533	3	0.56%	99.44%
*Sum of div.s	2182	21	0.96%	99.04%
Whole Corpus	2182	25	1.15%	98.85%

Table 3.1: Statistics of the sentence alignment of the *health* section, with token-number approach

As shown in table 3.1, a total of 2,182 alignment pairs were extracted. As all sentences in the *health* section were considered and aligned, recall is 100%. When this section was aligned as a whole, a success rate of 98.85% was achieved. Slightly higher success rates, with an average of 99.04%, were obtained when the divisions were aligned separately. This indicates that the statistical *a-priori* parameters reflect local situations more accurately than a wider scale of corpus. This result shows that the performance of the system is stable on this sample.

Table 3.2 illustrated the performance of program from a different point of view. It shows the system's performance on different types of alignments. As in table 3.1, the statistics in this table were also collected separately for the two cases: a) the section aligned division by division and, b) the section aligned as a whole.

Predictably, the algorithm works best on 1:1 alignments. The results trail off as the table proceeds. The only exception is the high accuracy rate - 100% - for 1:3 alignments. While the score is encouraging, there are too few cases of 1:3 alignments in the section to allow for a

[4] The corpus is kept in a number of computer files each of which contains some articles. In this case, these files served as the divisions of the corpus.

general claim to be made about the efficiency of the program in dealing with this type of alignments.

[Independent variate: English]

Corpus	Type(Eng:Chi)	Align. Pairs	Mismatches	Error Rate	Accuracy
Aligned division by division	1 : 1	2128	13	0.61%	99.39%
	1 : 2	47	6	12.77%	87.23%
	2 : 1	5	1	20.00%	80.00%
	1 : 3	2	0	0.00%	100.00%
	2 : 2	0	~	~	~
Aligned as a whole	1 : 1	2128	18	0.85%	99.15%
	1 : 2	47	6	12.77%	87.23%
	2 : 1	5	1	20.00%	80.00%
	1 : 3	2	0	0.00%	100.00%
	2 : 2	0	~	~	~

Table 3.2: Analysis of the precision of sentence alignment types by token-number approach (*health*)

As shown in table 3.2, about 97.53% of the alignments in the *health* section are one-to-one alignments, and the average sentence length for both English and Chinese is relatively short (E = 15.925875, C = 15.945430). These factors may have contributed to the high precision score of the program. In order to assess the influence of such factors on the performance of the program, its performance on another section, *education*, is examined.

The *education* section has some distinct features in comparison to the *health* section. Firstly, the average sentence length in the former (E = 25.22, C = 27.79) is much longer than that in the latter. Consequently, despite their similar sizes (31,044 English words and 87,794 bytes of Chinese translation in *education* versus 30,490 and 109,280 in *health*), the former contains fewer sentences than the latter (1,031 and 1,297 Chinese and English sentences *vs*. 2,214 and 2,170 respectively).

Secondly, the data in the *education* section was translated with more complicated patterns, including two 1:5 Chinese-to-English translations, than those in the *health* section. Numerous non-one-to-one translations occurred in the former, making up 20.67% of its data (see table 3.3 below). The higher proportion of non-one-to-one translations in the *education* section caused a greater mean variance of difference between English-Chinese sentence lengths than was the case for the *health* section, as shown in tables 3.3 and 3.4.

[sentence length: number of tokens]

Alignment type (English : Chinese)	Average English sentence length	Average Chinese sentence length	Variance of difference
Overall	25.22	27.79	8.28
1:1	25.48	22.31	6.23
1:2	38.80	39.47	3.35
2:1	50.97	42.95	6.99

Table 3.3: Average sentence lengths of various types of sentence alignment and their difference variances in the *education* section

[sentence length: number of tokens]

Alignment type (English : Chinese)	English average length	Chinese average length	Variance of difference
Overall	15.93	15.95	2.34
1:1	15.95	15.93	2.34
1:2	25.02	26.30	3.57
2:1	24.20	22.40	2.17

Table 3.4: Average sentence lengths of various types of sentence alignment and their difference variances in the *health* section

All these factors present difficulties for the statistical algorithm. As a result, the algorithm produced less accurate sentence alignment on the *education* section. Table 3.5 shows the result. As shown in the table, the corpus yielded 1,132 aligned Chinese-English sentence pairs[5]. Of them, 20.67% are non-one-to-one alignments (by comparison to 2.54% in the *health*), and consequently the precision drops to 92.93%.

[Independent variate: Chinese]

Type(Chi:Eng)	Alignment Pairs	Mismatch	Error Rate	Success Rate
1 : 1	898	44	4.90%	95.10%
1 : 2	166	16	9.64%	90.36%
2 : 1	30	6	20.00%	80.00%
2 : 2	4	4	100%	0.00%
1 : 3	31	8	25.81%	74.19%
3 : 1	1	0	0.00%	100.00%
1 : 4	0	0	~	~
1 : 5	2	2	100%	0.00%
Sum	1132	80	7.07%	92.93%

Table 3.5: Statistics of the sentence alignment in the *education* sections using a token-number approach

When we compare tables 3.2 to 3.3 and 3.5 to 3.4, the precision of the system is in inverse proportion to the variance. However, the affect is less noticeable for 1:2 and 2:1 alignments. The system obtained 87.23% and 80.00% precision on 1:2 and 2:1 in the *health* section and 90.36% and 80.00% precision for 1:2 and 2:1 alignment respectively in the *education* section. Nonetheless, greater variance is likely to yield a lower precision score.

Factors in the *education* section, such as greater overall difference variance of English-Chinese sentence lengths and a higher portion of non-one-to-one alignments, explain the lower precision. In this section, the system had to deal with much more complex alignments than it did in the *health* section. For example, as shown in table 4.3, 166 and 30 cases of 1:2 and 2:1 alignments occurred respectively in the *education* section in comparison to 47 and 5 in the *health* section. In the *education* section, the system dealt with non-one-to-one alignments about 20.67% of the time compared to only 2.54% of the time in the *health* section.

The above discussion shows that the variance of sentence length and translation style affect the performance of the algorithm. To cope with this problem, this system needs to be enhanced with linguistic knowledge, e.g. electronic English-Chinese bilingual dictionary.

[5] The number of pairs were counted automatically. Because some sentences are merged in one-to-multiple alignment and some mismatches are included in the counting as well, the number of sentences given here is different from the original number.

3.2 Character-measuring versus token-measuring of sentence length

One interesting issue in sentence alignment is how to measure sentence length. Generally, the sentence length can be measures either in terms of number of words/tokens or number of characters. For example, Brown *et al*. (1991) measured sentence length using the number of tokens, while Gale *et al*. (1991) used the number of characters as the measure of sentence lengths. Gale *et al*. (1993: 89-90) argue, that

> It might seem that a word is a more natural linguistic unit than a character. However, we have found that words do not perform as well .... We believe that characters are better because there are more of them, and therefore there is less uncertainty.

In order to examine the influence of different measures of sentence length on the performance of the algorithm, the *health* section was re-aligned using the character measurement of the sentence length and the result was compared against that of the token measurement approach. Tables 3.6 and 3.7 show statistics of the results. Table 3.6 shows the overall accuracy rate of the algorithm on five divisions of the section and table 3.7 classifies the performance of the algorithm for different types of alignments (compare to tables 3.1 and 3.2).

[Independent variate: English]

Division	Aligned Pairs	Mismatches	Error Rate	Success Rate
hlth1	350	19	5.43%	94.57%
hlth2	446	11	2.47%	97.53%
hlth3	433	3	0.69%	99.31%
hlth4	417	6	1.44%	98.56%
hlth5	530	3	0.57%	99.43%
Sum	2176	42	1.93%	98.07%

Table 3.6: Result of sentence alignment in the *health* section using character measure of sentence length

[Independent variate: English]

Manner	Type	Aligned Pairs	Mismatches	Error Rate	Success Rate
Aligned division by division	1 : 1	2122	32	1.51%	98.49%
	1 : 2	47	8	17.02%	82.98%
	2 : 1	5	1	20.00%	80.00%
	1 : 3	2	1	50.00%	50.00%
	2 : 2	0	~	~	~

Table 3.7: Result classification for various types of sentence alignment in the *health* section using the character measure of sentence length

A comparison of table 3.1 to 3.6 reveals a slightly lower accuracy rate for character measure. The average success rate drops from 99.04 to 98.07. When we compare table 3.2 to table 3.7, many more errors are found when the character measure is used. In particular, the error rate for 1:1 alignments, which form the majority of cases, is doubled with the character measure.

Suspecting that a weaker correlation of sentence lengths in terms of characters might have caused the success rate to decrease, the correlation coefficients of English to Chinese sentence lengths using the token measure and the character measure were examined. Table 3.8 shows the correlation coefficients in the five aligned divisions of the *health* section.

Sample	Corr_Coef. (token measure)	Corr_Coef. (char. measure)
hlth1	0.935412	0.932190
hlth2	0.942526	0.928577
hlth3	0.938964	0.934824
hlth4	0.949761	0.950201
hlth5	0.934259	0.940396
Average	0.940184	0.937237

Table 3.8: Comparison of post-alignment correlation coefficients in five divisions of the *health* section, using token *vs*. character measures of sentence length

As shown in table 3.8, in three out of five divisions of the section, the correlation is stronger when the token measurement is used. On the other hand, in the other two divisions, the character measurements resulted in a stronger sentence length correlation. Yet, on average, the character measurement corresponds to a slightly lower correlation coefficient than the token measurement: 0.937237 *vs*. 0.940184. This finding helps explain why the character measurement yielded a slightly lower accuracy rate for sentence alignment than the token measurement.

Due to the minute scale of the differences observed in table 3.8, it is difficult to draw a definite conclusion from it. The influence of the two different measures on sentence length correlation between Chinese and English in the *health* section is surprisingly trivial. Nevertheless, according to the comparative study of the token and character measurements so far, measuring sentence length in terms of the number of tokens seems preferable for Chinese-English alignment algorithm.

4. Conclusion

In this paper, we described a new statistical sentence alignment algorithm based on a statistical bivariate model. Avoiding the need for universal *a-priori* parameters, which not necessarily is language/genre independent, this algorithm extracts local *a-priori* information from a division of the corpus to be aligned. We argued that such local *a-priori* parameters are more accurate than universal *a-priori* parameters.

As shown in the experiment presented previously, the parameters, such as regression line and correlation coefficient, provide locally accurate *a-priori* information for sentence alignment in a given division of the CEPC. Our algorithm makes use of this property of local parameters to achieve an optimum sentence alignment within a given section of the CEPC. By doing so, it avoids the need for universal *a-priori* parameters determined empirically from pre-aligned samples, which become inappropriate when the genre/domain of the corpus changes.

Our algorithm automates the whole process of sentence alignment by extracting an approximate regression line from the paragraph pairs which contain the same number of sentences. Although it is an approximation of the true regression line, with a filtering algorithm to exclude mismatches from the calculation, it is presumably accurate enough to provide a means of estimating the expected ratio of sentence lengths between the languages to be aligned. The dynamic search algorithm for the CV ensures optimum sentence alignment. The high accuracy rate of the algorithm on the CEPC testifies to the reliability of this algorithm. Nevertheless, the assumption that Group I data alone is significant to establish the regression line is a possible limitation.

The algorithm aligns a division of the CEPC at a time. The size of the division could be another factor that affects the performance of the algorithm. From out experiment presented, a

division size of between 400 to 600 sentences yields the highest success rates. If the data of a corpus are highly homogenous, we assume that the size of the division is less of a concern.

Translation style is another factor that may affect the performance of the algorithm. In the experiment, we saw more errors in the alignment of the *education* section, which is less literally translated, than the *health* section. Some complex translations, such as 1:5 alignments, are beyond the capability of a purely statistical algorithm. We assume that a higher success rate from a literally translated corpus and a lower success rate from freely translated corpora can be expected.

Different measures of sentence length – number of characters versus number of tokens – do affect the performance of the algorithm. But our experiment shows that such influence is trivial. Considering the difficulty of word segmentation in some languages, such as Chinese and Japanese, the character measure of sentence length is preferable to the word/token measure despite a slight sacrifice of accuracy.

As a purely statistical algorithm, our algorithm can not deal with excessively complex translations such as 1:4 translations. To cope with such complex translation, this algorithm needs to be enhanced with cognate approaches and language knowledge, such as machine-readable bilingual dictionary, thesaurus, etc.

References:

Alder, Henry L. & Edward B. Roessler (1972), 'Chapter 12: Regression and correlation' in *Introduction to Probability and Statistics* (*fifth edition*), W. H. Freeman and Company, San Francisco.

Brown, P. F., J. C. Lai and R. L. Mercer (1991), 'Aligning sentences in parallel corpora', in *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, USA. pp. 169-176.

Chen, Kuang-Hua and Hsin-Hsi Chen (1994), 'A part-of-speech-based alignment algorithm, in *Proceedings of COLING-94*, vol.1, Kyoto, Japan, pp. 166-171.

Collier, Nigel, Kenji Ono and Hideki Hirakawa (1998), 'An experiment in hybrid dictionary and statistical sentence alignment', in *Proceedings of COLING-ACL '98*, Montreal, Canada, vol.1; pp. 268-267.

Collier, Nigel, Hideki Hirakawa and Akira Kumano (1998), 'Machine translation vs. dictionary term translation - a comparison for English-Japanese news article alignment', in *Proceedings of COLING-ACL '98*, Montreal, Canada, vol.1; pp. 263-267.

Fung, Pascale and Kenneth Ward Church (1994), 'K-vec: a new approach for aligning parallel texts', in *Proceedings of COLING '94*, Kyoto, Japan, Vol. III, pp. 1996-2001.

Gale, William A. and Kenneth W. Church (1993), 'A program for aligning sentences in bilingual corpus', in *COLING 91*, pp. 177-184.

Hofland, K. (1995), 'A program for aligning English and Norwegian sentences', in *Proceedings of the ACH/ALLC Conference*, Santa Barbara, USA.

Kay, M., and M. Roscheisen (1993), 'Text-Translation Alignment', *Computational Linguistics*, 19:1. pp. 121-142.

Haruno, Masahiko and Takefumi Yamazaki (1996), 'High-performance bilingual text alignment using statistical and dictionary information', in *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics*, June, California, USA, pp. 131-138.

McEnery, A. and M. P. Oakes (1995), 'Sentence and word alignment in the CRATER project: methods and assessment', in S. Warwick-Armstrong (ed.), *Proceedings of the Association for Computational Linguistics Workshop SIG-DAT Workshop*, Dublin.

McEnery, A., M. P. Oakes and R. G. Garside (1994), 'The use of approximate string matching techniques in the alignment of sentences in parallel corpora', in A. Vella (ed.), *The Proceedings of MT - 10 Years On*, Cranfield, UK.

McEnery, Tony and Michael Oakes (1996), 'Sentence and word alignment in the CRATER project', in Jenny Thomas and Mick Short (eds.), *Using Corpora for Language Research*, Longman, pp. 211-231.

Melamed, I. Dan (1999), 'Bitext maps and alignment via pattern recognition', in *Computational Linguistics*, March 1999, Vol. 25, No. 1, MIT Press, pp. 107-130.

Simard, M., G. Foster and P. Isabelle (1992), 'Using cognates to align sentences in bilingual corpora', in *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI92)*, Montreal, Canada. pp. 67-81.

Simard, M., George Foster, Marie-Louise Hannan, Elliott Macklovitch and Pierre Plamondon (2000), 'Bilingual text alignment: where do we draw the line?', in Simon Philip Botley, Anthony Mark McEenry and Andrew Wilson (eds.) *Multilingual Corpora in Teaching and Research*, Amsterdam - Atlanta, GA. pp. 38-64.

Vogal, Stephan, Hermann Ney and Christoph Tillmann (1996), 'HMM-based word alignment in statistical translation', in *Proceedings of COLING - 96*, Vol. 1, pp. 836-841.

Wu, Dekai (1994), 'Aligning a parallel English-Chinese corpus statistically with lexical criteria', in *32nd Annual Meeting of the Association for Computational Linguistics*, New Mexical, USA, pp. 80-87.