An Alignment Method for Noisy Parallel Corpora based on Image Processing Techniques

Jason S. Chang and Mathis H. Chen
Department of Computer Science,
National Tsing Hua University, Taiwan
jschang@cs.nthu.edu.tw mathis@nlplab.cs.nthu.edu.tw
Phone: +886-3-5731069 Fax: +886-3-5723694

Abstract

This paper presents a new approach to bitext correspondence problem (BCP) of noisy bilingual corpora based on image processing (IP) techniques. By using one of several ways of estimating the lexical translation probability (LTP) between pairs of source and target words, we can turn a bitext into a discrete gray-level image. We contend that the BCP, when seen in this light, bears a striking resemblance to the line detection problem in IP. Therefore, BCPs, including sentence and word alignment, can benefit from a wealth of effective, well established IP techniques, including convolution-based filters, texture analysis and Hough transform. This paper describes a new program, *PlotAlign* that produces a word-level bitext map for noisy or non-literal bitext, based on these techniques.

Keywords: alignment, bilingual corpus, image processing

1. Introduction

Aligned corpora have proved very useful in many tasks, including statistical machine translation, bilingual lexicography (Daille, Gaussier and Lange 1993), and word sense disambiguation (Gale, Church and Yarowsky 1992; Chen, Ker, Sheng, and Chang 1997). Several methods have recently been proposed for sentence alignment of the Hansards, an English-French corpus of Canadian parliamentary debates (Brown, Lai and Mercer 1991; Gale and Church 1991a; Simard, Foster and Isabelle 1992; Chen 1993), and for other language pairs such as English-German, English-Chinese, and English-Japanese (Church, Dagan, Gale, Fung, Helfman and Satish 1993; Kay and Röscheisen 1993; Wu 1994).

The statistical approach to machine translation (SMT) can be understood as a word-by-word model consisting of two sub-models: a *language model* for generating a source text segment S and a *translation model* for mapping S to its translation T. Brown et al. (1993) also recommend using a bilingual corpus to train the parameters of $\Pr(S \mid T)$, *translation probability* (TP) in the translation model. In the context of SMT, Brown et al. (1993) present a series of five models of $\Pr(S \mid T)$ for word alignment. The authors propose using an adaptive *Expectation and Maximization* (EM) algorithm to estimate parameters for *lexical translation probability* (LTP) and *distortion probability* (DP), two factors in the TP, from an aligned bitext. The EM algorithm iterates between two phases to estimate LTP and DP until both functions converge.

Church (1993) observes that reliably distinguishing sentence boundaries for a noisy bitext obtained from an OCR device is quite difficult. Dagan, Church and Gale (1993) recommend aligning words directly without the preprocessing phase of sentence alignment. They propose using *char_align* to produce a rough character-level alignment first. The rough alignment provides a basis for estimating the translation probability based on position, as well as limits the range of target words being considered for each source word. *Char_align* (Church 1993) is based on the observation that there are many instances of

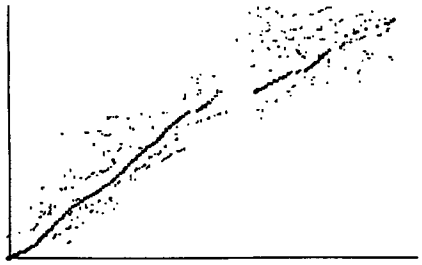

Figure 1. Dotplot. An example of a dotplot of alignment showing only likely dots which lie within a short distance from the diagonal.

cognates among the languages in the Indo-European family. However, Fung and Church (1994) point out that such a constraint does not exist between languages across language groups such as Chinese and English. The authors propose a *K-vec* approach which is based on a *k*-way partition of the bilingual corpus. Fung and McKeown (1994) propose using a similar measure based on Dynamic Time Warping (DTW) between occurrence recency sequences to improve on the *K-vec* method.

The *char-align*, *K-vec* and DTW approaches rely on dynamic programming strategy to reach a rough alignment. As Chen (1993) points out, dynamic programming is particularly susceptible to deletions occurring in one of the two languages. Thus, dynamic programming based sentence alignment algorithms rely on paragraph anchors (Brown et al. 1991) or lexical information, such as cognates (Simard 1992), to maintain a high accuracy rate. These methods are not robust with respect to non-literal translations and large deletions (Simard 1996). This paper presents a new approach based on image processing (IP) techniques, which is immune to such predicaments.

2. BCP as image processing

2.1 Estimation of LTP

A wide variety of ways of LTP estimation have been proposed in the literature of computational linguistics, including Dice coefficient (Kay and Röscheisen 1993), mutual information, ϕ^2 (Gale and Church 1991b), dictionary and thesaurus information (Ker and Chang 1996), cognates (Simard 1992), *K-vec* (Fung and Church 1994), DTW (Fung and McKeown 1994), etc.

Table 1. Linguistic constraints. Linguistic constraints at various level of alignment resolution give rise to different types of image pattern that are susceptible to well established IP techniques.

Constraints	***Image Pattern***	***IP techniques***	***Alignment Resolution***
Structure preserving	Edge	Convolution	Phrase
One-to-one	Texture	Feature extraction	Sentence
Non-crossing	Line	Hough transform	Discourse

Dice coefficient:

$$Dice(s,t) = \frac{2 \cdot prob(s,t)}{prob(s) + prob(t)}$$

mutual information:

$$MI(s,t) = \log \frac{prob(s,t)}{prob(s) \cdot prob(t)}$$

Like the image of a natural scene, the linguistic or statistical estimate of LTP gives rise to signal as well as noise. These signal and noise can be viewed as a gray-level dotplot (Church and Gale 1991), as Figure 1 shows.

We observe that the BCP, when cast as a gray-level image, bears a striking resemblance to IP problems, including edge detection, texture classification, and line detection. Therefore, the BCP can benefit from a wealth of effective, well established IP techniques, including convolution-based filtering, texture analysis, and Hough transform.

2.2 Properties of aligned corpora

The *PlotAlign* algorithms are based on three linguistic constraints that can be observed at different level of alignment resolution, including phrase, sentence, and discourse:

1. **Structure preserving constraint:** The connection target of a word tend to be located next to that of its neighboring words.
2. **One-to-one constraint:** Each source word token connect to at most one target word token.
3. **Non-crossing constraint:** The connection target of a sentence does not come before that of its preceding sentence.

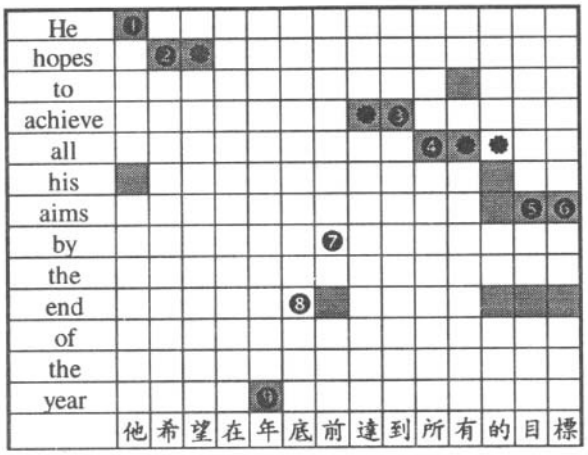

Figure 2. Short edges and textural pattern in a dotplot. The shaded cells are positions where a high LTP value is registered. The cell with a dark dot in it is an alignment connection.

Each of these constraints lead to a specific pattern in the dotplot. The structure preserving constraint means that the connections of adjacent words tend to form short, diagonal edges on the dotplot. For instance, Figure 2 shows that the adjacent words such as "He hopes" and "achieve all" lead to diagonal edges, ❶❷ and ❸❹ in the dotplot. However, edges with different orientation may also appear due to some morphological constraints. For instance, the token "aim" connects to a Mandarin compound "目標," thereby gives rise to the horizontal edge ❺❻. The one-to-one assumption leads to a textural pattern that can be categorized as a region of dense dots distributed much like the 1's in a permutation matrix. For instance, the vicinity of connection dot ❽ (end, 底) is denser than that of a non-connection say (end, 目). Furthermore, the nearby connections ❼, ❽, and ❾, form a texture much like a permutation matrix with roughly one dot per row and per column. The non-crossing assumption means that the connection target of a sentence will not come before that of its preceding sentence. For instance, Figure 1 shows that there are clearly two long lines representing a sequence of sentences where this constraint holds. The gap between these two lines results from the deletion of several sentences in the translation process.

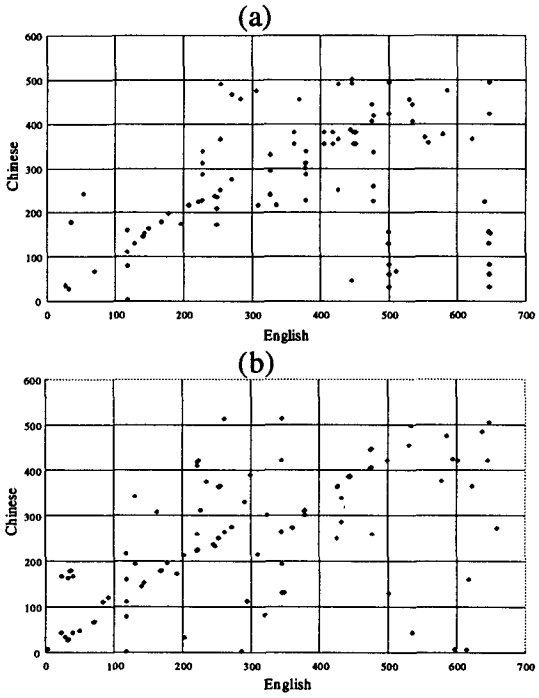

Figure 3. Convolution. (a) LTP dotplot before convolution; and (b) after convolution.

2.3 Convolution and local edge detection

Convolution is the method of choice for enhancing and detecting the edges in an image. For noise or incomplete image, as in the case of LTP dotplot, a discrete convolution-based filter is effective in filling a missing or under-estimated dot which is surrounded by neighboring dots with high LTP value according to the structure preserving constraint. A filtering mask stipulates the relative location of these supporting dots. The filtering can be proceed as follows to obtain $\Pr(s_x, t_y)$, the

translation probability of the position (x, y), from $t(s_{x+i}, t_{y+j})$, the LTP values of itself and neighboring cells:

$$\Pr(s_x, t_y) = \sum_{j=-w}^{w} \sum_{i=-w}^{w} t(s_{x+i}, t_{y+j}) \times mask(i, j)$$

where w is a pre-determined parameter specifying the size of the convolution filter. Connections that fall outside this window are assumed to have no affect on $\Pr(s_x, t_y)$.

For simplicity, two 3×3 filters can be employed to detect and accentuate the signal:

-1	-1	-1
2	2	2
-1	-1	-1

2	-1	-1
-1	2	-1
-1	-1	2

However, a 5 by 5 filter, empirically derived from the data, performs much better.

-0.04	-0.11	-0.20	-0.15	-0.11
0.08	-0.01	-0.25	-0.19	-0.15
-0.13	0.27	1.00	0.27	-013
-0.13	-0.16	-0.22	0.02	0.11
-0.10	-0.14	-0.19	-0.10	-0.02

2.4 Texture analysis

Following the common practice in IP for texture analysis, we propose to extract features to discriminate a connection region in the dotplot from non-connection regions. First, the dotplot should be normalized and binarized, leaving the expected number of dots, in order to reduce complexity and simplify computation. Then, projectional transformation to either or both axes of the languages involved will compress the data further without losing too much information. That further reduces the 2D texture discrimination task to a 1D problem. For instance, Figure 4 shows that the vicinity of a connection (by, 前) is characterized by evenly distributed high LTP values, while that of a non-connection is not. According to the one-to-one constraint, we should be looking for dense and continuous 1D occurrence of dots. A cell with high density and high power density indicate that connections fall on the vicinity of the cell. With this in mind, we proceed as follows to extract features for textural discrimination:

1. Normalize the LTP value row-wise and column-wise.
2. For a window of $n \times m$ cells, set the t (s, t) values of k cells with highest LTP values to 1 and the rest to 0, $k = \mathbf{max}\ (n, m)$.
3. Compute the density and deviation features:

projection:

$$p(x, y) = \frac{1}{\sum_{j=-v}^{v} t(x, y+j)}$$

density:

$$d(x, y) = \frac{\sum_{i=-w}^{w} p(x+i, y)}{2w+1}$$

power density:

$$pd(x, y) = \sum_{i=1}^{c} \sum_{x'=x-w}^{x+w} p(x', y) \cdot p(x'-i, y)$$

where w and v are the width and height of a window for feature extraction, and c is the bound for the resolution of texture. The bound depends on the coverage rate of LTP estimates; 2 or 3 seems to produce satisfactory results.

Since the one-to-one constraint is a sentence level phenomena, the values for w and v should be chosen to correspond to the lengths of average sentences in each of the two languages.

2.5 Hough transform and line detection

The purpose of Hough transform (HT) algorithm, in short, is to map all points of a line in the original space to a single accumulative value in the parameter space. We can describe a line on x-y plane in the form $\rho = x \cdot \mathbf{sin}\theta + y \cdot \mathbf{cos}\theta$. Therefore,

a point (ρ, θ) on the ρ - θ plane describes a line on the x-y plane. Furthermore, HT is insensitive to perturbation in the sense the line of (ρ, θ) is very close to that of $(\rho \pm \Delta\rho, \theta \pm \Delta\theta)$. That enables HT-based line detection algorithm to find high resolution, one-pixel-wide lines, as well as lower-resolution lines.

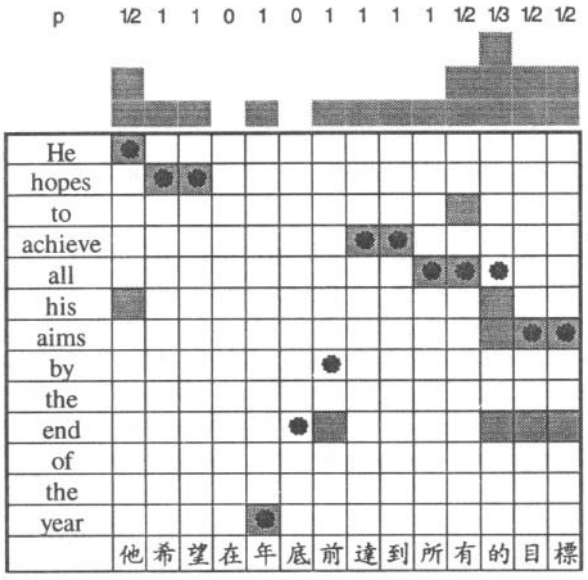

Figure 4. Projection. The histogram of horizontal projection of the data in Figure 2.

As mentioned above, many alignment algorithms rely on anchors, such as cognates, to keep alignment on track. However, that is only possible for bitext of certain language pairs and text genres. For a clean bitext, such as the Hansards, most dynamic programming based algorithms perform well (Simard 1996). To the contrary, a noisy bitext with large deletions, inversions and non-literal translations will appear as disconnected segments on the dotplot. Gaps between these segments may overpower dynamic programming, and lead to a low precision rate. Simard (1996) shows that for the Hansards corpus, most sentence-align algorithms yield a precision rate over 90%. For a noisy corpus, such as literary bitext, the rate drops below 50%. Contrary to the dynamic programming based methods, Hough transform always detect the most apparent line segments even in a noisy dotplot.

Before applying Hough transform, the same processes of normalization and thresholding are performed first. The algorithm is described as follows:

1. Normalize the LTP value row-wise and column-wise.
2. For a window of $n \times m$ cells, set the $t(s, t)$ values of k cells with highest LTP values to 1 and the rest to 0, $k = \mathbf{max}\ (n, m)$.
3. Set **incidence** $(\rho, \theta) = 0$, for all $-k \le \rho \le k$, $-90^\circ \le \theta \le 0^\circ$,
4. For each cell (x, y), $t(x, y) = 1$ and $-90^\circ \le \theta \le 0^\circ$, increment **incidence** $(x \cos \theta + y \sin \theta, \theta)$ by 1.
5. Keep (ρ, θ) pairs that have high incidence value, **incidence** $(\rho, \theta) > \lambda$. Subsequently, filter out dot (x, y) that does not lie on such a line, (ρ, θ) or within a certain distance δ from (ρ, θ).

3. Experiments

To asses the effectiveness of the *PlotAlign* algorithms, we conducted a series of experiments. A novel and its translation was chosen as the test data. For simplicity, we have selected mutual information to estimate LTP. Statistics of mutual information between a source and target words is estimated using an outside source, example sentences and translation in the Longman English-Chinese Dictionary of Contemporary English (LecDOCE, Longman Group, 1992). An additional list of some 3,200 English person names and Chinese translations are used to enhance the coverage of proper nouns in the bitext.

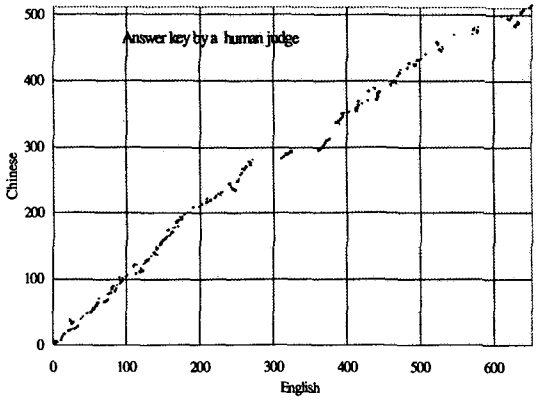

Figure 5. Alignment by a human judge.

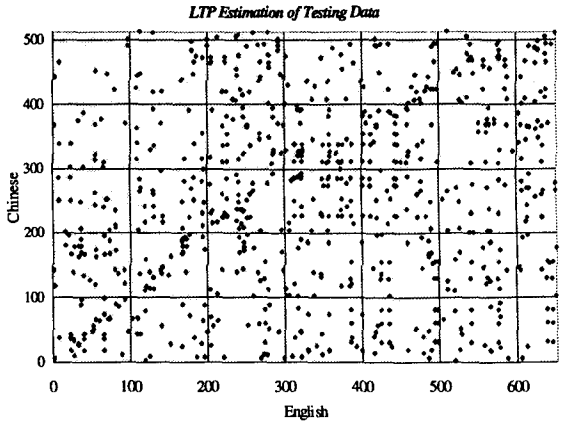

Figure 6. LTP estimation of the test data.

Figure 5 displays the result of word alignment by a human judge. Only 40% of English text and 70% of Chinese text have a connection counterpart. This indicates the translation is not literal and there are many deletions. For instance, the following sentences are freely translated:

1a. It was only a quarter to eleven.
1b. 現在是十點三刻。(10:45.)
2a. She was tall, maybe five ten and a half, but she didn't stoop.
2b. 她有一七五公分以上。 (175cm)
3a. *Larry Cochran tried to keep a discreet distance away.* He knew his quarry was elusive and self-protective: there were few candid pictures of her, which was what would make these valuable. He walked on the opposite side of the street from her; *using a zoom lens, he had already shot a whole roll of film.* When they came to Seventy-ninth Street, he caught a real break when she crossed over to him, and he realised he might be able to squeeze off full-face shots. *Maybe, if it clouded over more, she might take off her dark glasses. That would be a real coup.*
3b. 柯勞瑞不著痕跡地尾隨在後，利用長鏡頭已拍完一捲底片。若是天氣再暗一點，她說不定會摘掉黑鏡，到時後他就可以拍到一張更精彩的相片了。

4. Result and Discussion

Figure 6 shows that the coverage and precision of the LTP estimate is not very high. That is to be expected since the translation is not literal and the mutual information estimate based on an outside source might not be relevant. Nevertheless, *PlotAlign* algorithms seem to be robust enough to produce reasonably high precision that can be seen from Figure 3. Figure 3(a) shows that a normalization and thresholding process based on one-to-one constraints does a good job of filtering out noise. Figure 3(b) shows that convolution-based filtering remove more noise according to the assumption of structure preserving constraint.

Texture analysis does an even better job in noise suppression. Figure 7(a) and 7(b) show that signal-to-noise ratio (SNR) is greatly improved.

The filtering based on Hough Transform, contrary to the other two filtering methods, prefers connection that is consistent with other connections globally. It does a pretty good job of identifying a long line segment. However, isolated, short segments, surrounded by deletions are likely to be missed out. Figure 8(b) shows that filtering based on HT missed out the short line segment appearing near the center of the dotplot shown in Figure 6(b). Nevertheless, this short segment presents most vividly in the result of textural filter, shown in Figure 7(b). By combining filters on all three levels of resolution, we gather as much evidence as possible for optimal result.

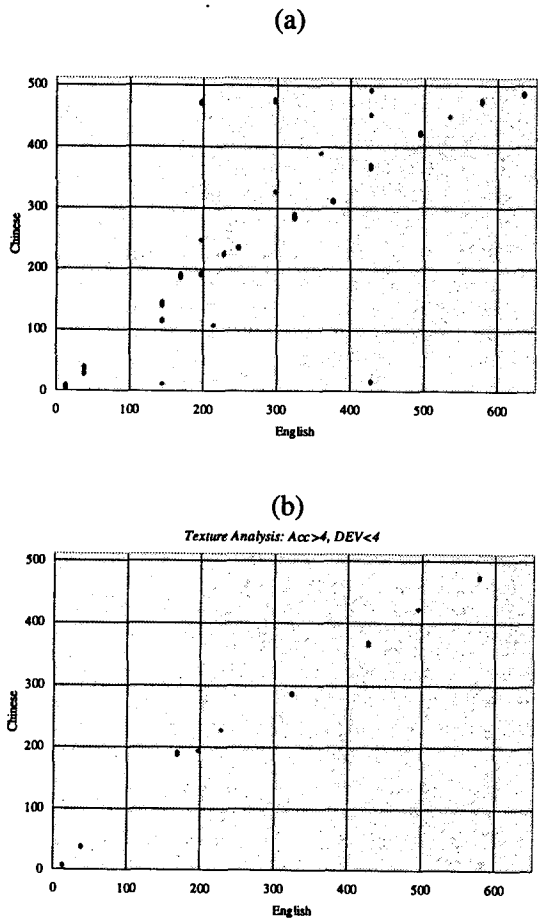

Figure 7. Texture Analysis. (a) Threshold = 3; (b) Threshold = 4.

Table 2. Hough Transform.

ρ	θ	N	ρ	θ	N	ρ	θ	N
5	-42	10	0	-43	6	-61	-56	6
23	0	9	7	-41	6	-83	-60	6
313	0	9	-2	-45	6	113	0	6
387	0	9	-2	-48	6	252	0	6
0	-45	8	-3	-49	6	323	0	6
0	-49	8	-6	-46	6	348	0	6
4	-43	8	-9	-50	6	420	0	6
3	-44	7	32	-1	6	486	0	6
-18	-90	7	46	-31	6	498	0	6
-24	-51	7	-11	-54	6	566	0	6
-38	-53	7	-43	-54	6	-107	-67	6
-39	-53	7	-46	-54	6	-120	-59	6
109	0	7	-53	-57	6	-226	-75	6
226	0	7	-54	-55	6	-353	-90	6

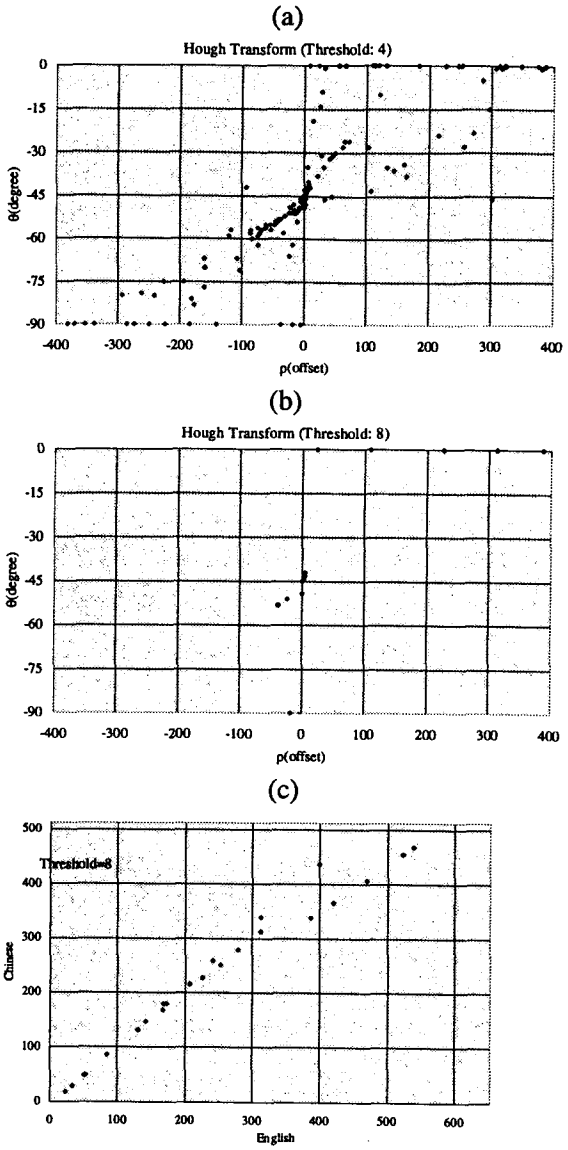

Figure 8. Hough transform of the test data.

5. Conclusion

The algorithm's performance discussed herein can definitely be improved by enhancing the various components of the algorithms, e.g. introducing bilingual dictionaries and thesauri. However, the *PlotAlign* algorithms constitute a functional core for processing noisy bitext. While the evaluation is based on an English-Chinese bitext, the linguistic constraints motivating the algorithms seem to be quite general and, to a large extent, language independent. If that is the case, the algorithms

should be effective to other language pairs. The prospects for English-Japanese or Chinese-Japanese, in particular, seem highly promising. Performing the alignment task as image processing proves to be an effective approach and sheds new light on the bitext correspondence problem. We are currently looking at the possibilities of exploiting powerful and well established IP techniques to attack other problems in natural language processing.

Acknowledgement

This work is supported by National Science Council, Taiwan under contracts NSC-862-745-E007-009 and NSC-862-213-E007-049. And we would like to thank Ling-ling Wang and Jyh-shing Jang for their valuable comments and suggestions.

References

1. Brown, P. F., J. C. Lai and R. L. Mercer, (1991). Aligning Sentences in Parallel Corpora, In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* 169-176, Berkeley, CA, USA.
2. Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics,* 19:2, 263-311.
3. Chen, J. N., J. S. Chang, H. H. Sheng and S. J. Ker, (1997). Word Sense Disambiguation using a Bilingual Machine Readable Dictionary. To appear in *Natural Language Engineering.*
4. Chen, Stanley F., (1993). Aligning Sentences in Bilingual Corpora Using Lexical Information, In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (ACL-91), 9-16, Ohio, USA.
5. Church, K. W., I. Dagan, W. A. Gale, P. Fung, J. Helfman, and B. Satish, (1993). Aligning Parallel Texts: Do Methods Developed for English-French Generalized to Asian Languages? In *Proceedings of the First Pacific Asia Conference on Formal and Computational Linguistics*, 1-12.
6. Church, Kenneth W. (1993), Char_align: A Program for Aligning Parallel Texts at the Character Level, In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics* (ACL-93), Columbus, OH, USA
7. Dagan, I., K. W. Church and W. A. Gale, (1993). Robust Bilingual Word Alignment for Machine Aided Translation, In *Proceedings of the Workshop on Very Large Corpora : Academic and Industrial Perspectives*, 1-8, Columbus, Ohio, USA.
8. Daille, B., E. Gaussier and J.-M. Lange, (1994). Towards Automatic Extraction of Monolingual and Bilingual Terminology, In *Proceedings of the 15th International Conference on Computational Linguistics*, 515-521, Kyoto, Japan.
9. Fung, P. and K. McKeown, (1994). Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping, In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*(AMTA-94), 81-88, Columbia, Maryland, USA.
10. Fung, Pascale and Kenneth W. Church (1994), K-vec: A New Approach for Aligning Parallel Texts, In *Proceedings of the 15th International Conference on Computational Linguistics* (COLING-94), 1096-1140, Kyoto, Japan.
11. Gale, W. A. and K. W. Church, (1991a). A Program for Aligning Sentences in Bilingual Corpora, In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*(ACL-91), 177-184, Berkeley, CA, USA.
12. Gale, W. A. and K. W. Church, (1991b). Identifying Word Correspondences in Parallel Texts, In *Proceedings of the Fourth DARPA Speech and Natural Language Workshop*, 152-157, Pacific Grove, CA, USA.
13. Gale, W. A., K. W. Church and D. Yarowsky, (1992), Using Bilingual Materials to Develop Word Sense Disambiguation Methods, In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation* (TMI-92), 101-112, Montreal, Canada.
14. Kay, M. and M. Röscheisen, (1993). Text-translation Alignment, *Computational Linguistics*, 19:1, 121-142.
15. Ker, Sur J. and Jason S. Chang (1997), Class-based Approach to Word Alignment, to appear in *Computational Linguistics,* 23:2.
16. Longman Group, (1992). *Longman English-Chinese Dictionary of Contemporary English*, Published by Longman Group (Far East) Ltd., Hong Kong.
17. Simard, M., G. F. 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* (TMI-92), 67-81, Montreal, Canada.
18. Simard, Michel and Pierre Plamondon (1996), Bilingual Sentence Alignment: Balancing Robustness and Accuracy, in *Proceedings of the First Conference of the Association for Machine Translation in the Americas* (AMTA-96), 135-144, Montreal, Quebec, Canada.
19. Wu, Dekai (1994), Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria, in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics,* (ACL-94) 80-87, Las Cruces, New Mexicao, USA.