HLT-NAACL 2003 Workshop: Building and Using Parallel Texts
Data Driven Machine Translation and Beyond , pp. 81-87
Edmonton, May-June 2003

Efficient Optimization for Bilingual Sentence Alignment Based on Linear Regression

Bing Zhao
Language Technologies Institute
Carnegie Mellon University
bzhao@cs.cmu.edu

Klaus Zechner
Educational Testing Service
Rosedale Road, Princeton, NJ 08541
kzechner@ets.org

Stephan Vogel
Language Technologies Institute
Carnegie Mellon University
vogel+@cs.cmu.edu

Alex Waibel
Language Technologies Institute
Carnegie Mellon University
ahw@cs.cmu.edu

Abstract

This paper presents a study on optimizing sentence pair alignment scores of a bilingual sentence alignment module. Five candidate scores based on perplexity and sentence length are introduced and tested. Then a linear regression model based on those candidates is proposed and trained to predict sentence pairs' alignment quality scores solicited from human subjects. Experiments are carried out on data automatically collected from Internet. The correlation between the scores generated by the linear regression model and the scores from human subjects is in the range of the inter-subject agreement score correlations. Pearson's correlation ranges from 0.53 up to 0.72 in our experiments.

1 Introduction

In many instances, multilingual natural language systems like machine translation systems are developed and trained on parallel corpora. When faced with a different, unseen text genre, however, translation performance usually drops noticeably. One way to remedy this situation is to adapt and retrain the system parameters based on bilingual data from the same source or at least a closely related source. A bilingual sentence alignment program (Gale and Church, 1991, and Brown et al., 1991) is the crucial part in this adaptation procedure, in that it collects bilingual document pairs from the Internet, and identifies sentence pairs, which should have a high likelihood of being correct translations of each other. The set of identified bilingual parallel sentence pairs is then added to the training set for parameter re-estimation.

As is well known, text mined from the Internet is very noisy. Even after careful html parsing and filtering for text size and language, the text from comparable html-page pairs still contains mismatches of content or non-parallel junk text, and the sentence order can be too different to be aligned. Together with a large mismatch of vocabulary, the aligned sentence pairs, which are extracted from these collected comparable html-page pairs, contain a number of low translation quality alignments. These need to be removed before the re-training of the MT system.

In this paper, we present an approach to automatically optimizing the alignment scores of such a bilingual sentence alignment program. The alignment score is a combination (by linear regression) of two word translation lexicon scores and three sentence length scores and predicts the translation quality scores from a set of human annotators. We also present experiments analyzing how many different human scorers are needed for good prediction and also how many sentence pairs should be scored per human annotator.

The paper is structured as follows: in section 2, the text mining system is briefly described. In section 3, five sentence alignment models based on lexical information and sentence length are explained. In section 4, a regression model is proposed to combine the five models to get further improvement in predicting alignment quality. We describe alignment experiments in section 5, focusing on the correlation between the alignment scores predicted by the sentence alignment models and by humans. Conclusions are given in section 6.

2 System of Mining Parallel Text

One crucial component of statistical machine translation (SMT) system is the parallel text mining from

Internet. Several processing modules are applied to collect, extract, convert, and clean the text from Internet. The components in our system include:

- A web crawler, which collects potential parallel html documents based on link information following (Philip Resnik 1999);
- A bilingual html parser (based on flex for efficiency), which is designed for both Chinese and English html documents. The paragraphs' boundaries within the html structure are kept.
- A character encoding detector, which judges if the Chinese html document is GB2312 encoding or BIG5 encoding.
- An encoding converter, which converts the BIG5 documents to GB2312 encoding.
- A language identifier to ensure that source and target documents are both of the proper language. (Noord's Implementation).
- A Chinese word segmenter, which parses the Chinese strings into Chinese words.
- A document alignment program, which judges if the document pair is close translation candidates, and filters out those non-translation pairs.
- A sentence boundary detector, which is based on punctuation and capitalized characters;
- And the key component, a sentence alignment program, which aligns and extracts potential parallel sentence pairs from the candidate document pairs.

After sentence alignment, each candidate of a parallel sentence pair is then re-scored by the regression models (to be described in section 5). These scores are used to judge the quality of the aligned sentences. Thus one can select the aligned sentence pairs, which have high alignment quality scores, to re-estimate the system's parameters.

2.1 Sentence Alignment

Our sentence alignment program uses IBM *Model-1* based perplexity (section 2.2) to calculate the similarity of each sentence pair. Dynamic programming is applied to find Viterbi path for sentence alignments of the bilingual comparable document pair. In our dynamic programming implementation, we allow for seven alignment types between English and Chinese sentences:

- 1:1 – exact match, where one sentence is the translation of the other one;
- 2:2 – the break point between two sentences in the source document is different from the segmentation in the target document. E.g. part of sentence one in the source might be translated as part of the second sentence in the target;
- 2:1, 1:2, and 3:1 – these cases are similar to the case before: they handle differences in how a text is split into sentences. The case 1:3 has not been used in the final configuration of the system, as this type did not occur in any significant number;
- 1:0 (deletion) and (0:1) insertion – a sentence in the source document is missing in the translation or vice versa.

The deletion and insertion types are discarded, and the remaining types are extracted to be used as potential parallel data. In general, one Chinese sentence corresponds to several English sentences. In (Bing and Stephan, 2002), experiments on a 10-year XinHua news story collection from the Linguistic Data Consortium (LDC) show that alignment types like (2:1) and (3:1) are common, and this 7-type alignment is shown to be reliable for English-Chinese sentence alignment. However, only a small part of the whole 10-year collection was pre-aligned (Xiaoyi, 1999) and extracted for sentence alignment.

The picture can be very different when directly mining the data from Internet. Due to the mismatch between the training data and the data collected from Internet, the vocabulary coverage can be very low; the data is very noisy; and the data aligned is not strictly parallel. The percentage of alignment types of insertion (0:1) and deletion (1:0) become very high as shown in section 5. The aligned sentence pairs are subject to many alignment errors. The alignment errors are not desired in the re-training of the system, and need to be removed.

Though the sentence alignment outputs a score from Viterbi path for each of the aligned sentence pairs, this score is only a rough estimation of the alignment quality. A more reliable re-scoring of the data is desirable to estimate the alignment quality as a post processing step to filter out the errors and noise from the aligned data.

2.2 Statistical Translation Lexicon

We use a statistical translation lexicon known as IBM *Model-1* in (Brown et al., 1993) for both efficiency and simplicity.

In our approach, *Model-1* is the conditional probability that a word *f* in the source language is translated given word *e* in the target language, *t(f|e)*. This probability can be reliably estimated using the expectation-maximization (EM) algorithm (Cavnar, W. B. and J. M. Trenkle, 1994).

Given training data consisting of parallel sentences: $\{(f^{(i)}, e^{(i)}), i = 1..S\}$, our *Model-1* training for *t(f|e)* is as follows:

$$t(f \mid e) = \lambda_e^{-1} \sum_{s=1}^{S} c(f \mid e; f^{(s)}, e^{(s)})$$

Where λ_e^{-1} is a normalization factor such that

$$\sum_j t(f_j \mid e) = 1.0$$

$c(f \mid e; f^{(s)}, e^{(s)})$ denotes the expected number of times that word *e* connects to word *f*.

$$c(f \mid e; f^{(s)}, e^{(s)}) = \frac{t(f \mid e)}{\sum_{k=1}^{l} t(f \mid e_k)} \sum_{j=1}^{m} \delta(f, f_j) \sum_{i=1}^{l} \delta(e, e_i)$$

With the conditional probability t(f|e), the probability for an alignment of foreign string F given English string E is in (1):

$$P(F \mid E) = \frac{1}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{n} t(f_j \mid e_i) \qquad (1)$$

The probability of alignment F given E: $P(F \mid E)$ is shown to achieve the global maximum under this EM framework as stated in (Brown et al.,1993).

In our approach, equation (1) is further normalized so that the probability for different lengths of F is comparable at the word level:

$$\bar{P}(F \mid E) = \left[\frac{1}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{n} t(f_j \mid e_i) \right]^{1/m} \qquad (2)$$

The alignment models described in (Brown et al., 1993) are all based on the notion that an alignment aligns each source word to exactly one target word. This makes this type of alignment models asymmetric.

Thus by using the conditional probability t(e|f) translation lexicon trained from English (source) to Chinese (target), different aspects of the bilingual lexical information can be captured. A similar probability to (2) can be defined based on this reverse translation lexicon:

$$\bar{P}(E \mid F) = \left[\frac{1}{(l+1)^m} \prod_{i=1}^{m} \sum_{j=0}^{n} t(e_i \mid f_j) \right]^{1/n} \qquad (3)$$

Starting from the Hong Kong news corpora provided by LDC, we trained the translation lexicons to be used in the parallel sentence alignment. Each sentence pair has a perplexity, which is calculated using the minus log of the probability eg. equation (2).

3 Alignment Models

The alignment model is aimed at automatically predicting the alignment scores of a bilingual sentence alignment program. By scoring the alignment quality of the sentence pairs, we can filter out those mis-aligned sentence pairs, and save our SMT system from being corrupted by mis-aligned data.

3.1 Lexicon Based Models

It is necessary to include lexical features in the aligned quality evaluation. One way is to use the translation lexicon based perplexity as in our sentence alignment program.

For each of the aligned sentence pairs, the sentence alignment generated a score, which is solely based on equation (2). Using this score only, we can do a simple filtering by setting a threshold of perplexity. The sentence pairs which have a higher perplexity than the threshold will be removed. However the perplexity based on (2) is definitely not discriminative enough to evaluate the quality of aligned sentence pairs.

In our experiment, it showed that perplexity (3) has more discriminative power in judging the quality of the aligned sentence pairs for Chinese-English sentence alignment. It is also possible that equation (2) is more suitable for other language pairs. Both (2) and (3) are applied in our sentence alignment quality judgment, which is to be explained in section 4.

3.2 Sentence Length Models

As was shown in the sentence alignment literature (Church, K.W. 1993), the sentence length ratio is also a very good indication of the alignment of a sentence pair for languages from a similar family such as French and English. For language pairs from very different families such as Chinese and English, the sentence length ratio is also a good indication of alignment quality as shown in our experiments.

For the language pair of Chinese and English, the sentence length can be defined in several different ways.

3.2.1 Sentence Length

In general, a Chinese sentence does not have word boundary information; so one way to define Chinese sentence length is to count the number of bytes of the sentence. Another way is to first segment the Chinese sentence into words (section 3.2.2) and count how many words are in the sentence. For English sentences, we can similarly define the length in bytes and in words.

The length ratio is assumed to be a Gaussian distribution. The mean and variance are calculated from the parallel training corpus, which, in our case, is the Hong Kong parallel corpus with 290K parallel sentence pairs.

3.2.2 A Chinese Word Segmenter

The word segmenter for Chinese is to parse the Chinese string into words. Different word segmenters can generate different numbers of words for the same Chinese sentence.

There are many word segmenters publicly available. In our experiments, we applied a two-pass strategy to segment the word according to the dictionary of the LDC bilingual dictionary of Chinese-English. The two-pass started first from left to right, and then from right back to left, to calculate the maximum word frequency and select one best path to segment the words.

In general, the sentence length is not sensitive to the segmenters used. But for reliability, we want each seg-

mented word can have an English translation, thus we used the LDC bilingual dictionary as a reference word list for segmentation.

3.2.3 Sentence Length Model

Assume the alignment probability of $P(A \mid s,t)$ is only related to the length of source sentence s and target sentence t:

$$\begin{aligned} P(A \mid s,t) &= P(\mid s \mid \leftrightarrow \mid t \parallel s,t) \\ &\cong P(\mid s \mid \leftrightarrow \mid t \parallel\parallel s \mid, \mid t \parallel) \\ &\cong P(\mid s \mid - \mid t \mid) \\ &\cong P(\delta(\mid s \mid, \mid t \mid)) \end{aligned}$$

where $\mid s \mid$ and $\mid t \mid$ are the sentence lengths of *s* and *t*.

The difference of the length $\delta(\mid s \mid, \mid t \mid)$ is assumed to be a Gaussian distribution (Church, K.W. 1993) and can be normalized as follows:

$$\bar{\delta} = \frac{\mid t \mid - \mid s \mid c}{\sqrt{(\mid s \mid +1)\sigma^2}} \sim N(0,1) \qquad (4)$$

where *c* is a constant indicating the mean length ratios between source and target sentences and σ^2 is the variance of the length ratios.

In our case, we applied three length models described in the following Table 1:

Table 1. *Three Length Models description*

L-1	Both English and Chinese sentence are measured in *bytes*
L-2	Both English and Chinese sentence are measured in *words*
L-3	English sentence is measured in *words* and Chinese sentence is measured in *bytes*

The means and σ^2 of the length ratios for each of the length models are calculated from Hong Kong news parallel corpus. The statistics of the three sentence length models are shown in Table 2.

Table 2. *Sentence length ratio statistics*

	L-1	L-2	L-3:
Mean	1.59	1.01	0.33
Var	3.82	0.79	0.71

In general, the smaller the variance, the better the sentence length model can be. From Table 2 we observe that the bytes based length ratio model has significantly larger variance (3.82) than the other two models (L-2: 0.79, L-3: 0.71). This means L1 is not as reliable as L2 and L3. Both L2 and L3 have similar variance, which indicates measuring English sentences in words will entail smaller variance in length model; measuring Chinese sentences in bytes or words entails only a slight difference in variance. This also indicates that the length model is not so sensitive to the Chinese word segmenter applied. L-1, L-2 and L-3 capture the length relationship of parallel sentence in different views. Their modeling power has overlap, but they also compensate each other in capturing the parallel characteristics of good translation quality. A combination of these models can potentially bring further improvement, which is shown in our experiment in section 6.

4 Regression Model

Rather than doing a binary decision (classification) that the aligned sentence pair is either good or not, the regression can give a confidence score indicating how good the alignment can be, thus offering more flexibility in decisions. Predicting the alignment quality using the candidate models is considered as a regression problem in that different scores are combined together.

There are many ways such as genetic programming, to combine the candidate models, and regression is one of the straight forward and efficient ones. So in this work, we explored linear regression.

4.1 Candidate Models

We have five candidate models described in section 3. They are: PP1, the perplexity based on the word pair conditional probability *p(f|e)* in equation (2); PP2, the perplexity based on the reverse word pair conditional probability *p(e|f)* in equation (3); L-1, Length ratio model measured in bytes (mean=1.59, var=3.82); L-2, length ratio model measured in words (mean=1.01, var=0.79); L-3, length ratio model, where the English sentence is measured in words and the Chinese sentence is measured in bytes (mean=0.33, var=0.71). These five models capture different aspects of the aligned quality of the sentence pair. The idea is to combine these five models together to get better prediction of the aligned quality.

Linear regression is applied to combine these five models. It is trained from the observation of the five models together with the label of human judgment on a training set.

4.2 Regression Model Training

The linear regression model tries to discover the equation for a line that most nearly fits the given data (Trevor Hastie et al. 2001). That linear equation is then used to predict values for the data.

Now given human subject judgment of the aligned translation quality of sentence pairs, we can train a regression model based on the five models we described in section 4.1 under the objective of least square errors.

The human evaluation is measures translation quality of aligned pairs on a discrete 6-point scale between 1 (very bad) and 5 (perfect translation). The score 0 was used for alignments that were not genuine translation e.g., both sentences were from the same language. We will use *n* for the number of total sentence pairs labeled by humans and used in training.

Let A= [PP1, PP2, L-1, L-2, L-3] be the machine-generated scores for each of the sentence pairs. In our case, A is a $n \times 5$ matrix.

Let H= [Human-Judgment-Score] be the human evaluation of the sentence pairs on a 6-point scale. In our case, H is a $n \times 1$ matrix.

In linear regression modeling, a linear transformation matrix W should satisfy the least square error criterion:

$$W^* = \min_w \{ \| AW - H \| \} \quad (5)$$

where W is in fact a 5x1 weight matrix. The equation can be solved as:

$$W^* = (A^T A)^{-1} A^T H \quad (6)$$

The inverse of matrix $A^T A$ is usually calculated using singular vector decomposition (SVD). After W is calculated, the predicted score from the regression model is:

$$H' = AW^* \quad (7)$$

where H' is the final predicted alignment quality score of the regression model. We can also view H' as a weighted sum of the five models shown in section 4.1. The calculation of H' reduces to a linear weighted summation, which is very efficient to compute.

5 Experiments

1500 pairs of comparable html document pairs were obtained from bilingual web pages crawled from Internet. After preprocessing, filtering, and sentence alignment, the alignment types were distributed as shown in Table 3. Ignoring the alignment type of insertion (0:1) and deletion (1:0), we extracted around 5941 parallel sentences.

Table 3. *Alignment types' distribution of mined data from noisy web data crawled*

	1:0	0:1	1:1	2:1	1:2	2:2	3:1
%	23.7	41.9	29.4	1.99	0.01	0.02	2.79

From Table 3, we see the data is very noisy, containing a large portion of insertions (23.7%) and deletions (41.9%). This is very different from the LDC XinHua pre-aligned collection provided by LDC, which is relatively clean.

For this set of English-Chinese bilingual sentences, we randomly selected 200 sentence pairs, focusing on Viterbi alignment scores below 12.0 from sentence alignment, which was an empirically determined threshold (The alignment scores here were purely reflecting the *Model-1* parameters using equation (2)). Three human subjects then had to score the 'translation quality' of every sentence pair, using a 6 point scale described in section 4.2. We further excluded very short sentences from consideration and evaluated 168 remaining sentences.

Pearson R correlation is applied to calculate the magnitude of the association between two variables (human-human or human-machine in our case) that are on an interval or ratio scale. The correlation coefficients (Pearson R) between human subjects were in Table 4 (all are statistically significant):

Table 4. *Correlation between Human Subjects*

	H2	H3
H1	0.786	0.615
H2	----	0.568

Overall, more than 2/3 of the human scores are identical or differ by only 1 (between subjects).

For the automatic score prediction, the five component scores described in section 4.1 are used, which are then combined using a standard Linear Regression as described in section 4.2. Table 5 shows the correlation between alignment scores based on Model X and human subjects' predicted quality scores:

Table 5. *Correlation between optimization models and human subjects*

Model	human-1	human -2	human -3
PP-1	.57	.53	.32
PP-2	.60	.58	.46
L-1	.42	.41	.30
L-2	.46	.41	.40
L-3	.40	.38	.29
Naïve	.58	.56	.38
Regression	**.72**	**.68**	**.53**

The data we used in our training of the lexicon is Hong Kong news parallel data from LDC. There are 290K parallel sentence pairs, with 7 million words of English and 7.3 million Chinese words after segmentation. The IBM *Model-1* for PP-1 and PP-2 are both trained using 5 EM iterations. The other three length models are also calculated from the same 290K sentence pairs. Punctuation is removed before the calculation of all automatic score prediction models.

The regression model here is the standard linear regression using the observations from three human subjects as described in section 4.1. The average performance of the regression model is shown in the bottom line of the above Table 5. The average correla-

tion varies from 0.53 upto 0.72, which shows that the regression model has a very strong positive correlation with the human judgment.

Also from Table 5, we see both lexicon based models: PP-1 and PP-2 are better than the length models in term of correlation with human scorer. Model PP-2 has the largest correlation, and is slightly better than PP-1. PP-2 is based on the conditional probability of *p(e|f)*, which models the generation of an English word from a Chinese word. The vocabulary size of Chinese is usually smaller than English vocabulary size, so this model can be more reliably estimated than the reverse direction of *p(f|e)*. This explains why PP-2 is slightly better than PP-1.

For sentence length models, we see L-2, for which the lengths of both the English sentence and the Chinese sentence are measured in words, has the best performance among the three settings of a sentence length model. This indicates that the length model measured in words is more reliable.

Also shown in Table 5, the naïve interpolation of these different models, i.e. just using each model with equal weight, resulted in lower correlation than the best single alignment model.

We also performed correlation experiments with varied numbers of training sentences from either Human-1/Human-2/Human-3 or from all of the three human subjects. We picked the first 30/60/90/120 labeled sentence pairs for training and saved the last 48 sentence pairs for testing. The average performance of the regression model is as follows:

Table 6. *Correlation between different training set sizes and human scorers.*

Training set size	Human-1	Human -2	Human –3
30	.686	.639	.447
60	.750	.707	.452
90	.765	.721	.456
120	.760	.721	.464

The average correlation of the regression models showed here increased noticeably when the training set was increased from 30 sentence pairs to 90 sentence pairs. More sentence pairs caused no or only marginal improvements (esp. for the third human subject).

Figure 1 shows a scatter plot, which illustrates a good correlation (here: Pearson R=0.74) between our regression model predictors and the human scorers.

6 Conclusion

In this paper, we have demonstrated ways to efficiently optimize a sentence alignment module, such that it is able to select aligned sentence pairs of high translation quality automatically. This procedure of alignment score optimization requires (a) a small number of human subjects who annotate a set of about 100 sentence pairs each for translation quality; and (b) a set of alignment scores, based on perplexity and sentence length ratio, to be able to learn to predict the human scores. Based on the learned predictions, by means of linear regression, the alignment program can choose the best sentence pair candidates to be included in the training data for the SMT system re-estimation.

Our experiments showed that, for Chinese-English language pair, perplexity based on the reverse word pair conditional probability *p(e|f)* (PP-2) gives the most reliable prediction among the five models proposed in this paper; the regression model, which combines those five models, give the best correlation between human score and automatic predictions. Our approach needs only a fairly limited number of human labeled sentences pairs, and is an efficient optimization of the sentence alignment system.

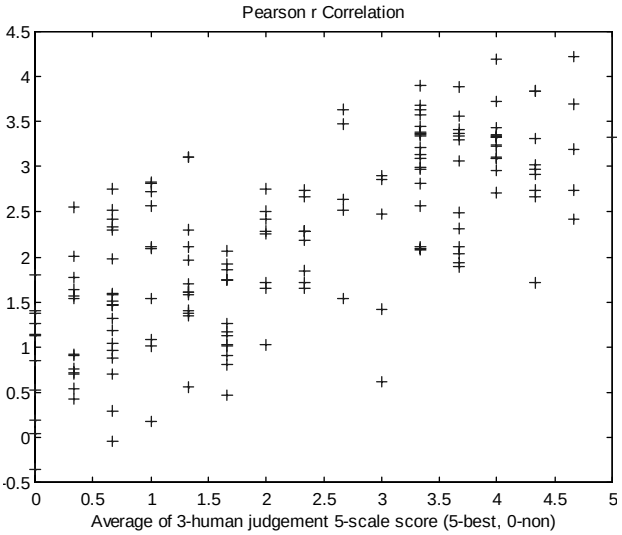

Figure 1. Correlation between regression model and human scorers, Pearson R=0.74.

References

Bing Zhao, Stephan Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. *IEEE International Conference on Data Mining (ICDM 02)* , pp. 745-748. Japan.

Brown, P., Lai, J. C., and Mercer, R. 1991. Aligning Sentences in Parallel Corpora. In *Proceedings of ACL-91*, Berkeley CA. 1991

Cavnar, W. B. and J. M. Trenkle. 1994. N-Gram-Based Text Categorization. *Proceedings of Third Annual Symposium on Document Analysis and Information*

Retrieval, Las Vegas, NV, UNLV Publications/Reprographics, pp. 161-175, 11-13.

Stanley Chen. 1993. Aligning sentences in Bilingual corpora using lexical information. In *proceedings of the 31st Annual Conference of the Association for computational linguistics*, pages 9-16, Columbus, Ohio, June 1993

Church, K. W. 1993. Char_align: A Program for Aligning Parallel Texts at the Character Level. *Proceedings of ACL-93*, Columbus OH.

Gale, W. A. and Church, K. W. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of ACL-91*, Berkeley CA. 1991.

Melamed, I.D. 1996. A Geometric Approach to Mapping Bitext Correspondence. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA. 1996

Noord's Implementation of Textcat: http://odur.let.rug.nl/~vannoord/TextCat/index.html

Peter F. Brown, Stephan A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathmatics of Statistical Machine Translation: Parameter estimation. *Computational Linguistics*, vol 19, no.2 , pp.263-311.

Philip Resnik. 1999. Mining the Web for Bilingual Text. *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, University of Maryland, College Park, Maryland.

Trevor Hastie, Robert Tibshirani, Jerome Friedman. 2001. The Elements of Statistical Learning: Data Mining, Inference and Prediction. *Springer Publisher*.

Xiaoyi Ma, Mark Y. Liberman, "BITS: A Method for Bilingual Text Search over the Web". *Machine Translation Summit VII*, 1999