http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (4): 666-672, 2010
ISSN 1812-5638

Arabic-Chinese and Chinese-Arabic Phrase-Based Statistical Machine Translation Systems

[1]Mossa Ghurab, [1]Yueting Zhuang, [1]Jiangqin Wu and [2]Maan Younis Abdullah
[1]College of Computer Science, Zhejiang University, Hangzhou, Zhejiang 310027, People's Republic of China
[2]College of Computer Science, Central South University, Changsha, Hunan,
410083, People's Republic of China

Abstract: Designs for Arabic-to-Chinese and Chinese-to-Arabic translation systems are presented. The core of the system implements standard Phrase-Based Statistical Machine Translation architecture, where Corpus data used for the systems was collected from the United-Nations website and various news engine websites. Here, we focus on its acquisition as it is the training data of Arabic-Chinese and Chinese-Arabic Statistical Machine Translation systems. We trained Statistical Machine Translation systems for two language pairs, which revealed interesting clues into the challenges ahead. Models are then softly integrated into Statistical Machine Translation architecture so they can interact with other models without modifying the basic architecture. As a result, phrase translation probabilities learn directly rather than deriving them heuristically.

Key words: Phrase-based SMT, Corpus collection, Arabic segmentation, Chinese segmentation, Tokenization, Mining

INTRODUCTION

Nowadays, one of the most common paradigms for Machine Translation (MT) is the statistical approach, above all when there is a large amount of parallel texts corpus (plural: corpora) available as is the case of the Arabic-Chinese language pair. The SMT is based on the noisy channel model. Given a foreign-language (e.g., Chinese) input sentence f, it looks for its most likely English translation e:

$$e^* = \arg\max_e P(e|f) = \arg\max_e (P(f|e) \times P(e)) \quad (1)$$

In Eq. 1, P (e) is the target language model; it is trained on monolingual text, e.g., the English side of the training bi-text. The term P(f|e) is the translation model. The noisy channel model is typically extended to a more general log-linear model, where several additional terms are introduced. For each pair of phrases used, there are four terms or components: forward and backward phrase translation probabilities and forward and backward lexicalized phrase translation probabilities. There is also a phrase penalty, which encourages the model to use fewer and thus longer, phrases. A word penalty on the target language side is also included, which controls the overall length of the English output. Finally, the phrase reordering is controlled by a distance-based distortion model.

Under this log-linear model, the most likely English translation e is found as follows:

$$\begin{aligned} e^* &= \arg\max_e P(e|f) \approx \arg\max_{e,s} P(e,s|f) \\ &= \arg\max_{e,s} (P(e)^{\lambda_1} \times \prod_{i=1}^{|s|} P(\bar{f}_i|\bar{e}_i)^{\lambda_2} \times P(\bar{e}_i|\bar{f}_i)^{\lambda_3} \\ &\times P_w(\bar{f}_i|\bar{e}_i)^{\lambda_4} \times P_w(\bar{e}_i|\bar{f}_i)^{\lambda_5} \times d(start_i, end_{i-1})^{\lambda_6} \\ &\times \exp(|\bar{e}_i|)^{\lambda_7} \times \exp(-1)^{\lambda_8}) \end{aligned} \quad (2)$$

In Eq. 2, s is a segmentation of f into phrases. The symbols $\bar{e}_i$ and $\bar{f}_i$ denote an English-foreign translation phrase pair used in the translation and $|s|$ is the number of such pairs under the current segmentation. The terms $P(\bar{e}_i|\bar{f}_i)$ and $P(\bar{f}_i|\bar{e}_i)$ are the phrase-level conditional probabilities and $P_w(\bar{e}_i|\bar{f}_i)$ and $P_w(\bar{f}_i|\bar{e}_i)$ are corresponding lexical weights as described by Koehn *et al*. (2003). The distance-based distortion term $d(start_i, end_{i-1})$ gives the cost for relative reordering of the target phrases at position i and i-1; more complex distortion models are possible, e.g., lexicalized. The remaining two terms $\exp(|\bar{e}_i|)$ and exp(-1) are the word penalty and the phrase penalty, respectively. The parameters λ_i are typically estimated from a tuning set using Minimum Error Rate Training (MERT) as described by Och (2003).

Corpus linguistics is one of the fastest-growing methodologies in contemporary linguistics. It is mainly

Corresponding Author: Mossa Ghurab, College of Computer Science, Zhejiang University, Hangzhou, Zhejiang 310027, People's Republic of China Tel: 008615958183021

classified by two parts: (1) monolingual corpus which refers to text in one language and (2) parallel corpus which is typically used in linguistic circles to refer to texts that are translations of each other. In order to exploit a parallel text, some kind of text alignment that identifies equivalent text segments (approximately sentences) is a prerequisite for analysis.

SYSTEM ARCHITECTURE

We have built two Phrased-based SMT systems, one from Arabic to Chinese translation and the other one from Chinese to Arabic translation. Both systems almost have the same architecture, except that the Chinese has one additional component. This component specifically simplifies the Chinese sentence to words since, Chinese sentences do not have spaces between words. Thus, it is an essential step to pre-process Chinese to prepare it for the translation. Figure 1 shows basic architecture of the translation system which consist of three layers: interface, analyzer and decoder.

Interface: The interface is the part that interacts directly with users or with other services. The users can use their webpage to establish a connection to the translation system to submit their translation requests; services also can access the system as a web service to perform translation.

Analyzer (pre-processing, post-processing): Analyzing input text is a very important step to build any translation system. Here, the analyzer consists of two parts: pre-processing and post-processing. Pre-processing has the following processes: segmenting, tokenizing and managing the input text. We use the ICTCLAS tool as the Chinese segmenter for Chinese text. Similar scripts are used to segment Arabic text. After using these scripts to tokenize text and remove unwanted characters, the text is chunked into sentences and sent in sequence to the decoder. Same steps happened for the post-processes but in reversed order.

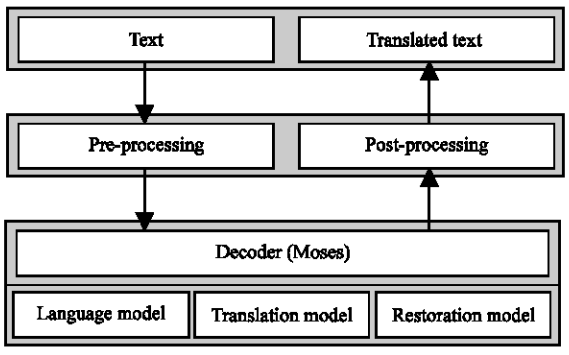

Fig. 1: SMT architecture

Decoder and models: Phrase-based SMT (Koehn *et al*., 2003) has emerged as the dominant paradigm in machine translation research. Such an approach may possibly employ the open-source decoder Moses (Koehn *et al*., 2007). The Moses decoder used an efficient decoding algorithm to calculate the best score of translation. When the foreign text comes as an input sentence to the decoder, it is segmented into many phrases and then uses the models (language model, translation model, restoration model) that are supplied to it to calculate the best score for the source language and then output it to the post-processes step.

CORPUS COLLECTION

Acquisition of a parallel corpus for the use in a SMT system is typically applied to the Arabic-Chinese and Chinese-Arabic translation task. In the following seven steps, corpus data with its pre-processing is described:

- Obtain the raw data (e.g., crawling and mining the web)
- Convert PDF to TEXT (Extracting PDF Documents)
- Extract and map parallel chunks of text (document alignment)
- Break the text into sentences (sentence splitting)
- Segment Chinese sentences into words (Chinese segmentation)
- Prepare the corpus for SMT systems (normalization, tokenization)
- Map sentences in one language to sentences in the other language (sentence alignment)

In the following, we will describe in detail the acquisition of the United Nations corpus from the official website of the United Nations. These proceedings are published in some official languages such as Arabic and Chinese.

Crawling and mining: The website of the United Nations provides the Proceedings of the United Nations in form of PDF files. Each file contains a document ID. The URL for each file contains relevant information for identification, such as its language and the day and number of the thread of discussion.

Crawling this web resource with a web spider is done by starting at an index page and following certain links

based on inclusion and exclusion rules (Fig. 2). Per language, it took several days to obtain enough files for each language.

Besides identifying sources for parallel corpora, we mined some news engines for such data; the mined corpus data is then mixed with the United Nations corpus and used for creating the language model of our SMT systems. We have used an open source tool called VietSpider for such purposes.

The crawler maintains a list of unvisited URLs called the frontier (Liu, 2006). The list is initialized with seed URLs which may be provided by a user or another program. Each crawling loop involves picking the next URL to crawl from the frontier, fetching the page corresponding to the URL through HTTP, parsing the retrieved page to extract the URLs and application specific information and finally adding the unvisited URLs to the frontier. Before the URLs are added to the frontier, they may be assigned a score that represents the estimated benefit of visiting the page corresponding to the URL. The crawling process may be terminated when a certain number of pages have been crawled. If the crawler is ready to crawl another page and the frontier is empty, the situation signals a dead-end for the crawler. The crawler has no new page to fetch and hence it stops.

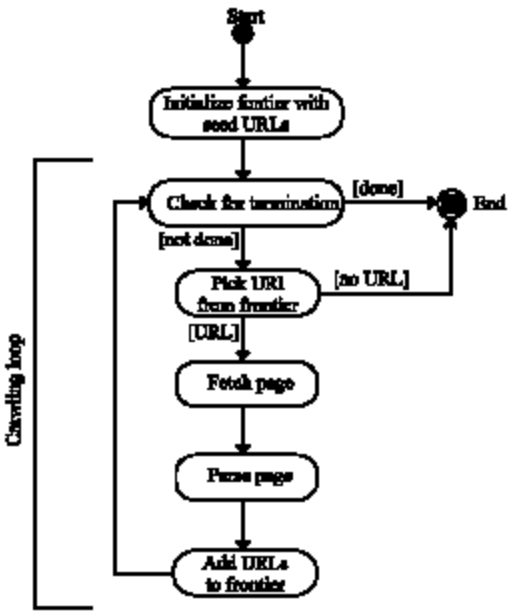

Fig. 2: Flow of a basic sequential crawler

Extracting documents: Because the translations of the Proceedings of the United Nations documents are PDF formatted, we need to extract them in text format. Extracting the corpus in text format from the PDF is the hardest step due to the complexity of the Arabic PDF documents. This process led to two difficulties of extracting the Arabic documents. First, there is no such tool that extracts the Arabic document files in the layout of the original PDF document. We solve this problem by using the following UNIX command:

```
pdftotext -enc UTF-8 ArabicFile.PDF ArabicFile.TXT
```

This method solved the problem of contents extraction, but it leads to another problem of output layout and font display. Figure 3 shows that the output layout of contents has wrong direction in many places in the extracted document and also font display.

We used our scripts to correct the output of text files from abnormalities such as wrong position of tags, incorrect order of words when English words are presented, Indian numbers to Arabic numbers and correction of the document's font display.

After the extraction process of the Arabic PDF files, we save the output to Arabic text files (Fig. 4).

Arabic PDF file

Arabic TEXT file

Fig. 3: Layout and font display problem

In addition, Chinese PDF files were also extracted to text files. Then the Arabic text file was aligned to its Chinese text file translation and both files get an identical prefix ID for later processing.

Document alignment: Each sitting of the United Nations covers a number of topics. A first step is to submit a prefix ID to the Arabic document and to its Chinese document translation. To obtain the maximum amount of data, we match these IDs for the language pair.

Large data collections such as the Proceedings of the United Nations are created over the period of many years, often with changing formatting standards and other sources of error.

The extraction of relevant text from noisy pre-extracted text document is a cumbersome enterprise that requires constant refinement and adaptation. We process the pre-extracted document's data with a Perl program that uses pattern matching to detect and extract the tag identity of the document, as well as the document's date and the document's content.

The documents were automatically aligned on the paragraph level. Each parallel paragraph was taken as a paragraph translation. Based on our goal of creating a translation-corpus, we could afford using an imperfect alignment algorithm, since, we could later use only those paragraphs that were successfully aligned with high confidence. Fortunately, most of the UN documents are divided into paragraphs delimited by the new-line character, so we basically used an alignment algorithm to find as many word-anchors as possible and used them to find matches for each Arabic paragraph. By word-anchor we mean an Arabic word whose equivalent was discovered in the Chinese version of the paragraph. And because Chinese sentences do not have spaces between words, we first segment the document using the ICTCLAS tool before we apply the alignment algorithm.

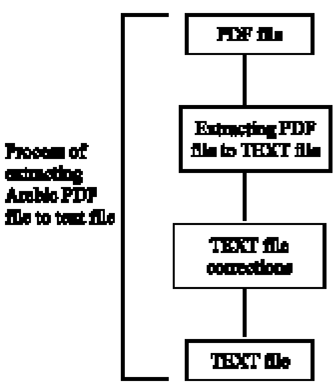

Fig. 4: Converting processes

Data then stored in one file document per language with UTF-8 clear text formatting, as shown in Fig. 5.

We created parallel corpus involving Arabic in this format. Also, we provide corpus in sentence aligned format, which we will describe below. Scripts are provided to generate the other parallel corpus.

The document alignment is done without tokenization and sentence splitting. The motivation behind this is that these are error prone processes for which multiple standards could be applied and we do not want to force any specific standard at this step.

Sentence splitting and tokenization: Sentence splitting and tokenization require specialized tools for each language. One problem of sentence splitting is the ambiguity of the period "." as either an end of sentence marker, or as a marker of meaning (for example, etc.) in the Arabic language.

For training a SMT system, usually all Chinese words are segmented to eliminate the differences between words and sentences (see the following section). In addition, Arabic words are tokenized as it's a high inflectional morphology language.

Issues with tokenization included the Arabic preposition letters that are merged with words such as in والتفاحة (waltufaha / and the apple / 和苹果), which must be treated as two words تفاحة ,و and not a single word. Another issue is the definition ال (Al/the) such as التفاحة (altufaha/the apple/ 苹果) which must be treated as two words تفاحة ال. These processes increase

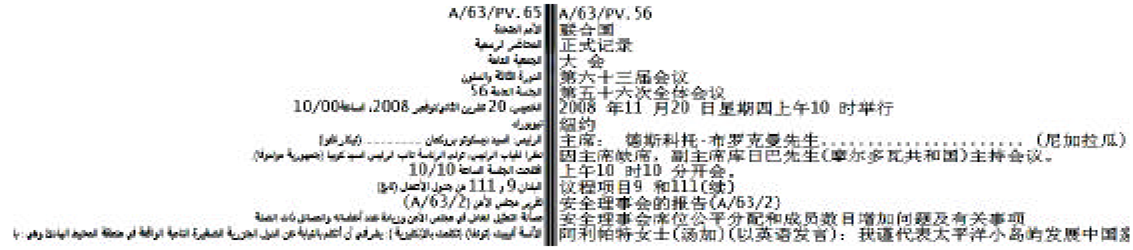

Fig. 5: Format of the released corpus

the probability of the phrases in the translation model during the SMT's training.

As seen in the previous example, the training of a SMT system of Arabic corpus must go through tokenization and stem pre-processing. Then Arabic and Chinese corpus data must be cleared from a non-Arabic and a non-Chinese word existence.

Chinese word segmentation: Chinese word segmentation is one of the pre-processing steps of the Arabic-Chinese and Chinese-Arabic SMT systems. The development of our Arabic-Chinese MT system began in 2008 and this was the first time we faced the issue of Chinese word segmentation. We used the free source codes of ICTCLAS (Zhang *et al*., 2003) which is based on the HHMM-based Chinese lexical analysis; the HHMM-based Chinese lexical analysis comprises five levels: atom segmentation, simple and recursive unknown words recognition, class-based segmentation and POS tagging. In the whole frame, class-based segmentation graph, a directed graph designed for word segmentation, is an essential intermediate data structure that links disambiguation and unknown words recognition with word segmentation and POS tagging.

Atom segmentation, the bottom level of HHMM, is an initial step. Here, an atom is defined to be the minimal segmentation unit that cannot be split in any stage. The atom consists of a Chinese character, punctuation, symbol string, numeric expression and other non-Chinese character string. Any given word is made up of an atom or more. Atom segmentation is to segment original text into an atom sequence and it provides pure and simple source for its parent HMM. For instance, a sentence like 2002.9, ICTCLAS 的自由源码开始发布 (The free source codes of ICTCLAS was distributed in September, 2002) would be segmented as atom sequence 2002.9/, /ICTCLAS/的/自/由/源/码/开/始/发/布/. In this HMM, the original symbol is an observation while the atom is a state.

Our choice to use the ICTCLAS segmenter was based on its performance, accuracy and size of the segmenter. The ICTCLAS stand-alone has a word speed of 996 KB/s, word accuracy of 98.45%, while the API does not exceed 200 KB, has a variety of data compression dictionaries under 12M and is currently the world's best Chinese lexical analyzer. Accuracy and speed test of ICTCLAS shown in Table 1.

Sentence alignment: Sentence alignment is usually a hard problem, but in our case it is simplified by the fact that the texts are already available in paragraph aligned format. Each paragraph consists typically of only 2-5 sentences. If the number of paragraphs of a speaker utterance differs in the two languages, we discard this data for quality reasons. The alignment of sentences in the corpus is done with an implementation of the algorithm by Gale and Church (1994). This algorithms tries to match sentences of similar length and sequence and merges sentences if necessary (e.g., two short sentences in one language to one long sentence in the other language), based on the number of words in the sentence. Since, there are so few sentences per paragraph, alignment quality is very high. There is considerable work on better sentence alignment algorithms. One obvious extension is to not only consider sentence length, but also potential word correspondences within sentence pairs. Work by Melamed (1999) is an example for such an approach.

The sentence aligned data is stored in one file per language, so that lines with the same line number in a file pair are mappings of each other. The markup from the document aligned file is stripped out. Table 2 shows total number of the aligned corpus.

The numbers of sentences and words in the table was taken after the tokenization and sentence-alignment of each other.

Extraction of a common test set: To allow the comparison of machine translation systems, it is necessary not only to define a common training set (as the United Nations corpus), but also a common test set. We suggest to reserve some data as a test set and to use the rest of the corpus as training data.

To be able to compare system performance, we also extracted a set of sentences that are aligned to each other. Figure 6 shows Arabic sentence aligned to its Chinese translation relevant from this collection.

Table 1: Results of an open test of ICTCLAS* (performance, accuracy and speed)

Functional Description**	Open test 1 WS	Open test 2 WS + NER with NW	Open test 3 WS+NER with NW+POS
Test file size (bytes)	4,092,478	4,092,478	4,092,478
Time (sec)	4.094	6.467561	9.094001
Share of the memory core data (Mb)	5.5	7.2	8.9
Speed (KB/s)	999.63	632.77	450.02
Accuracy			
WS (%)	96.56	98.13	98.13
POS (%)			63

*An open test result that motioned on Chinese version of the official document of ICTCLAS2008 tool.** WS: Word segment, NER: Named entity recognition, NW: New word, POS: Part of speech tagging

Table 2: Size of the released corpus

Language	Sentences	Words
Chinese (cn)	907,831	17,781,776
Arabic (ar)	907,831	16,828,090

ARABIC:	ترى من الضروري الحصول على ملاحظات هيئات المعاهدات والمقررين الخاصين، ولا سيما المقرر الخاص المعني بالحق في التعليم، عن تجاربهم ومبادراتهم المتصلة بالتثقيف والتدريب في مجال حقوق الإنسان؛
CHINESE:	认为有必要请条约机构和特别报告员、尤其是教育权问题特别报告员就其人权教育和培训的经验和主动行动提出意见；

Fig. 6: One sentence aligned to its translation

FULL TRANSLATION TASK

Most systems are largely language-independent and building a SMT system for a new language pair is mostly a matter of availability of parallel texts. Our efforts to explore open-domain Arabic-Chinese SMT led us to collecting data from the United Nations. Incidentally, the existence of translations in both languages now enabled us to build translation systems for two language pairs.

The translation task of these systems is focused on the Arabic and Chinese language pair. Translation quality was evaluated by using automatic evaluation metrics (Papineni *et al*., 2002) and NIST (Doddington, 2002).

Parallel corpus data go through preparation steps to integrate into the system: (1) segment Chinese corpus part to have spaces between the words using ICTCALS tool that was mentioned earlier; (2) segment Arabic corpus part to separate words from ال/Al/The/和, preposition letters and pronouns; (3) Tokenize parallel corpus from unwanted characters and (4) clean parallel corpus from long and short sentences on both sides.

When corpus data is ready to use, we build the language model of 5-gram for both languages using SRILM tool. Then, widely used alignment tool GIZA++ was used to extract phrases and build the alignment files. This step took 8 days in a Linux system (2.83 GHz CPU, 1 GB memory, 100GB HDD) with a parallel corpus of 907,831 sentence pairs.

The next step is to use a script that comes with the Moses decoder to train the translation model and restoration model. After that, system configuration file is tuned to use some reserved data to get the best translation result.

The systems are now ready to use. To check its accuracy, we used 1000 sentence pairs of Arabic and Chinese as a development test. We then translate from Arabic to Chinese and from Chinese to Arabic. After that, we evaluate the systems with automatic evaluation metrics (NIST, BLEU). The evaluation test shows a good result.

RESULTS AND DISCUSSION

The current system was developed under the CADAL project. Once the system was reasonably stable, we improved the system based on a small development set of data. For development, we used a test set that we reserved from during the corpus collection. Average sentence length was approximately 20 words. Development consisted primarily of translation from Arabic to Chinese; the other test was from Chinese to Arabic. Given the limited resources, we found the results to be highly encouraging. For many of the development input sentences, translations are reasonably comprehensible.

To quantitatively evaluate the results achieved so far and to comparatively assess the performance of our automatically acquired Phrased-based SMT, the results were evaluated using several automatic metrics for MT evaluation. These automatic metrics compare the translations with human-produced reference translations for the test sentences. For this test set, two reference translations were obtained. As mentioned previously, we used the BLEU (Papineni *et al*., 2002) and NIST (Doddington, 2002) automatic metrics for MT evaluation. The scores for the Arabic-Chinese System are displayed in Table 3 and 4. The heights score is 7.9905 NIST score and a 0.4916 BLEU score, whereas the scores for the Chinese-Arabic System is displayed in Table 5 and 6. The heights score is a 7.0643 NIST score and a 0.4678 BLEU.

Table 3: NIST, BLEU individual scores of the Arabic-Chinese Phrase-Based SMT system

	Arabic to Chinese translation score, individual N-gram scoring				
	1-gram	2-gram	3-gram	4-gram	5-gram
NIST	6.2310	1.5060	0.2012	0.0398	0.0125
BLEU	0.8053	0.5635	0.4427	0.3576	0.2937

Table 4: NIST, BLEU cumulative scores of the Arabic-Chinese Phrase-Based SMT system

	Arabic to Chinese translation score, cumulative N-gram scoring				
	1-gram	2-gram	3-gram	4-gram	5-gram
NIST	6.2310	7.7370	7.9382	7.9780	7.9905
BLEU	0.7647	0.6397	0.5561	0.4916	0.4389

Table 5: NIST, BLEU individual scores of the Chinese-Arabic Phrase-Based SMT system

	Chinese to Arabic translation score, individual N-gram scoring				
	1-gram	2-gram	3-gram	4-gram	5-gram
NIST	5.9354	0.9961	0.1080	0.0197	0.0050
BLEU	0.6965	0.5059	0.4102	0.3403	0.2859

Table 6: NIST, BLEU cumulative scores of the Chinese-Arabic Phrase-Based SMT system

	Chinese to Arabic translation score, cumulative N-gram scoring				
	1-gram	2-gram	3-gram	4-gram	5-gram
NIST	5.9354	6.9316	7.0396	7.0593	7.0643
BLEU	0.6919	0.5897	0.5213	0.4678	0.4234

CONCLUSION

The main focus of our study was the acquisition of discriminative learning techniques that would enhance the overall linguistic proficiency and to improve the Arabic and Chinese machine translation precisely. The goal of our system was to allow us to be able to write documents and predict good sentences using trained phrases based on high probability phrases. To be able to write documents accurately using the machine rather than human involvement was also essential. In addition, we described the acquisition of the United Nations corpus and used it to build SMT systems for two language pairs, the first serious effort at building such a system for Chinese-Arabic and Arabic-Chinese translation systems. The widely ranging quality of the different SMT systems for the different language pairs demonstrate the many different challenges for SMT research, which we have only touched upon.

REFERENCES

Doddington, G., 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. Proceedings of the 2nd International Conference on Human Language Technology Research, March 24-27, San Diego, California, pp: 138-145.

Gale, W.A. and K.W. Church, 1994. A program for aligning sentences in bilingual corpora. Comput. Linguistics, 19: 75-102.

Koehn, P., F.J. Och and D. Marcu, 2003. Statistical phrase-based translation. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, May 27-June 01, Edmonton, Canada, pp: 48-54.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch and M. Federico *et al*., 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, July 2007, Prague, Czech Republic, pp: 177-180.

Liu, B., 2006. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data (Data-Centric Systems and Applications). Springer, New York, pp: 274-321.

Melamed, I.D., 1999. Bitext maps and alignment via pattern recognition. Comput. Linguistics, 25: 107-130.

Och, F.J., 2003. Minimum error rate training in statistical machine translation. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, July 07-12, Sapporo, Japan, pp: 160-167.

Papineni, K., S. Roukos, T. Ward and W.J. Zhu, 2002. BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, July 07-12, Philadelphia, Pennsylvania, pp: 311-318.

Zhang, H.P., H.K. Yu, D.Y. Xiong and Q. Liu, 2003. HHMM-based Chinese lexical analyzer ICTCLAS. Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, July 11-12, Sapporo, Japan, pp: 184-187.