# **Research on Intelligent Automatic Translation System in Chinese and English Based on Integration Technology**

Ruijuan Hu

Henan Institute of Education hiruijuan2006@163.com

### Abstract

With the development and revolution of economic globalization, cross-language communication is becoming more and more important. The traditional artificial Chinese and English intelligent automatic translation system is not able to satisfy the growing need for translation. On this basis, through deeply analyzing the deficiencies of current traditional automatic translation system, this paper proposes the use of integration technology to build Chinese and English intelligent automatic translation technology, then by using the log linear model for parameter modulation and decoding of confusion network to obtain better accuracy and better performance of the translation results. Based on the algorithm, this paper constructs experiment for the integration technology in Chinese and English intelligent automatic translation system. Experimental results show that the English and Chinese intelligent automatic translation system based on Integration Technology is capable of obtaining higher precision and better robustness of translation results, with great practical significance.

*Keywords*: Integration technology; confusion network; log-linear model; parameter modulation and decoding; Chinese and English translation system

## **1. Introduction**

With the emergence of the Internet and the formation of the global village, the communication between different countries becomes more and more important, so the demand for language translation is also increasing [1]. Various types of machine translation can be generally divided into two types including rule based translation and corpus based translation. And rule based machine translation system can be divided into transition system, direct translation system and intermediate language system [2] and corpus based machine translation can be divided into statistical machine based translation and example based the machine translation system [3], where the example based machine translation is proposed as a new machine translation system with the rapid growing of bilingual corpus and the improvement of the performance of computer [4].

Different translation models and translation techniques have different theoretical basis, and each of them is proposed to solve different problems. System integration technology is an integration of different translation models and techniques retaining the advantages of various methods while avoiding the weakness of them so as to enhance the quality of machine translation results. The first use of system integration technology is in the field of automatic speech recognition [5]. Considering the fact that the system integration technology significantly reduces the error rate of automatic speech recognition, this method is adopted by the field of machine translation in 2006. Part of the team participated in (WMT) 2008 evaluation of the Machine Translation used the system integration technology and achieved good results. Since 2009 WMT began holding the evaluation of the integration of the system solely, and had held a total of three sessions by 2011. Since 2009, NIST started the evaluation of the integration of the system and had

held four sessions until 2012. It can be seen that system integration technology has caused widespread concern in the world [6].

In addition to effectively enhance the performance of machine translation systems, research on system integration technology can also in turn promote the progress of machine translation technology itself. The difficulties we are facing in using the system integration technology is lexical selection and reordering problem. The exploration on these problems using system integration technology can be directly used for machine translation technology, then realizing the progress of machine translation technology. In this paper, starting from system integration, based on confusion network we construct a Chinese and English intelligent automatic translation system by adding local features to the log-linear model.

# 2. Confusion Network Based Log-Linear Model Machine Translation Method

### 2.1. Traditional MBR Integration Technology

The integration technology of machine translation system is divided into two systems, post processing based system and decoding based multi-model system integration. The decoding based multi-model system integration makes use of the interaction between multiple models in the process of decoding to enhance the consistency of the translation results, and reduce the inconsistent translation results. Currently widely used methods include the method based on Minimum Bias Risk (MBR) method [7] and the method based on general linear model [8]. In the system integration method based on MBR, the minimum expected loss needs to be chose from the combination of the N-best list. The formula is as follows:

$$\hat{e} = \arg\min_{e \in E_h} \sum_{ref \in E_h, ref !=e} L(e', ref) * p(ref \mid f),$$
(2-1)

where  $E_h$  represents the N-best list of the results of multiple machine translation systems. L(e, ref) is the loss function, representing the risk e' of translation results compared to the standard translation ref. The smaller the value of the risk function is, the higher the result of translation quality will be. The widely used risk function is the transformation of automatic machine translation evaluation method, such as BLEU, TER, *etc.* [9]. p(ref | f) is the posterior probability of the translation from source language sentence f into the standard translation ref. Due to the inexistence of the standard translation in actual integration process, each candidate in N-best list is used as ref.

For a typical system integration process, source language sentences may have tens to hundreds of candidate translations and they are from different individual translation system. The integration candidates provided by different translation system are with great differences in overall quality and the integration candidate with low translation quality exists. Using the MBR based sentence level integration method integrates all available candidates. Since sentences with low quality have high similarity in word order and choice of words and the number of them are relatively larger than the number of high quality translation sentences and occupies a very high weight. Therefore, the chose sentences are mostly translated with poor quality. The word level system integration also has similar problem. Since the number of the low quality of translation candidate are larger, for confusion network in a position of all words in it, if a word is derived from the integration candidate of low quality, it will gain very high weights therefore are superior to those from high quality integration candidate words. The final decoding results will be similar to the low quality integration candidates, and then reduce the quality of the integration results of the whole test set.

Aiming at the deficiencies of MBR, this paper constructs the confusion network, and use linear model for parameter modulation and decoding of confusion network to enhance the consistency of the results with the existing integration candidates, so as to improve the accuracy and performance of the integration.

### 2.2. Integration Technology Based on Confusion Network

### 2.2.1. Construct Confusion Network

Common confusion network construction methods adopt the alignment method based on edit distance, mainly through the insertion, deletion, substitution and transposition four kinds of operation to align candidates to the skeleton so the integration candidate word order is same with that of the skeleton. The alignment method based on progressive type [10] has great quality improvement compared to previous alignment. This method can avoid the same word existing in different position of the confusion network, so the words from multiple morphological systems do not appear and thus improve the quality of the translation.

Assuming now there are three integration candidates need to align: "Ilikeballoons", "Ilikebigblueballoons" and "Ilikebluekites". If you use the pair-wise alignment method, you need to first select a skeleton from the 3 integration candidates, then align all the rest of the integration candidates to the skeleton. So the word order of skeleton determines the word order of integration results. Here assuming that the first integration candidate is selected as the skeleton, and then the second and third integration candidates are aligned to the skeleton, using the TER alignment and the results are as follows:

 Table 2-1. The Alignment Result of Pair-wise



Figure 2-1. The Confusion Network Constructed by Pair-Wise

The confusion network created by this alignment is shown in Figure 2-1. From the figure we can find the word "blue" repeated in the confusion network path and the reason is that pair-wise alignment of each integration candidate is not visible, so each alignment optimization can only consider current information, not from the point of view of global optimization.

If you use the incremental alignment method and alignment results are shown in Table 2-2 and the formation of confusion network are shown in Figure 2-2. It is easy to find out from the figure that incremental alignment solves the problem of repeated "blue". Based on progressive alignment method, we first create a confusion network from the skeleton. For a skeleton with N word, a confusion network of N arcs and N + 1 nodes will be created. For the word on the arc, there is a set of system related to the weight which means to vote. When the initial skeleton is created, the weight of each word

corresponding to the slot of the frame system is increased by 1/2, and the weight of the other system is set to 0. Using the greedy algorithm, the rest of the integration candidate are directly aligned to confusion network. The algorithm will constantly try to confuse arbitrary arc between two words in the network as candidates and matching with integration candidate words. And one movement of lexical chunks containing multi words is seen as one single editing operation until no other operation can reduce the edit distance.



 Table 2-2.The Alignment Result of Incremental Algorithm

Figure 2-2. The Confusion Network Constructed by Incremental Algorithm

If the words in integration candidates exact match with the words in confusion network in alignment process, then the weight of the corresponding word will increase 1/(k+1), where k is the ranking in a single translation system N-best of the integration candidates.

(1) If you use the replace operation to replace a by b in the integration candidates., then a new arc labeled a will be created in the corresponding position of the confusion network, and the weights corresponding new arc of the system is set to 1 / (k+1), the rest of the system weights are set to 0.

(2) If you use the delete operation, a new arc labeled NULL will also be created in the corresponding position.

(3) When using the insert operation to align the integration candidates into a confusion network, now a new node is inserted into the corresponding the position, then there will be a new arc labeled the inserted word between new nodes and a node, and the settings of corresponding weights are similar to that of the replacement situation and a new arc labeled NULL will also be created between the new node and the latter node. After adding all the integration candidates, the confidence of all the words between the two nodes need to be normalized, which will be completed by the confusion network after that.

## 2.2.2. Log-Linear Model

The goal of the system is to find the highest scoring path in the confusion network according to the log linear model. For confusion network in the presence of exponential convergence results, search time for a sentence of a medium length is very large, so the actual system fusion decoding process is completed using a heuristic search method, cannot guarantee must find the optimal results, only to find a sufficiently close to the optimal solution of the answer. In the process of decoding, there are two types of errors will lead to the decoding effect is not ideal. The first is the search error, this kind of error is the heuristic search method itself, because the use of heuristic method is not able to completely traverse the entire search space, a good method can reduce this kind of error. The second kind of error is model error, and if the model is wrong, the best result found in this model may not be the true high quality translation.

Log linear model in the field of machine learning is widely used in the field of [11], the linear model in the field of Machine Translation can replace the noise channel model, the problem can be used to express the following formula Machine Translation:

$$e = \arg \max\{\Pr(e \mid f)\}$$
(2-2)

where f represents the source language sentences, e represents the target language sentences. arg max represents the decoding process, both the target language of the sentence, and in all possible sentences to find the maximum probability of A. In the channel model, the formula (2-2) is equivalent to the formula (2-3) by using the Bayesian decision rule. In the new formula, Pr(e') is the target language is the language model, which measures the

probability of the target language sentence, but the translation model  $\Pr(f|e)$ :

$$e = \arg \max\{\Pr(e) * \Pr(f \mid e)\}$$
(2-3)

More widely used in practice, because the use of the formula can be independent of the training of the two knowledge sources, rather than just using a probabilistic model. The noise channel model is trained by using the maximum likelihood method, and the model and the translation model can be used to find the parameters which can be used to maximize the existing data. If the language model depends on the parameters, the translation model depends on the parameters, and the likelihood probability formula of the parallel corpus is as follows:

$$\hat{\theta} = \arg\min x_{\theta} \prod_{s=1}^{s} p_{\theta}(f_s \mid e_s)$$

$$\hat{\gamma} = \arg\min x_{\gamma} \prod_{s=1}^{s} p_{\gamma}(e_s)$$
(2-4)

Using the following decision rules in the actual decoding process:

$$e = \arg\max\{p_{\hat{\theta}}(e') * p_{\hat{\gamma}}(f \mid e')\}$$
(2-5)

Formula (2-5) reach the optimal condition is the estimation of the language model and the translation model and real probability distribution without error, but in the general case according to the formula (2-4) obtained is a very poor on a real distribution estimation, so the language model and the translation model of combinations need to reconsider. In the actual language phenomena exist many dependencies, and this relationship is difficult to be represented by the noisy channel model. In the actual translation process, using  $p_{y}(e^{-|f|})$  in place of  $p_{y}(f|e)$ , the translation model can achieve good results, from the theoretical framework of the noisy channel model is difficult to explain this phenomenon.

The log linear model is used as an improvement of the noise channel model, which avoids the above problems. In the log linear model [12], a set of features  $f_m(e, f), m = 1, 2, ..., M$  are used to model the parameters of each feature. Direct translation probability is expressed in the following formula:

$$\Pr(e \mid f) = p_{\lambda_{1}^{M}}(e \mid f) = \frac{\exp[\sum_{m=1}^{M} \lambda_{m} h_{m}(e, f)]}{\sum_{e} \exp[\sum_{m=1}^{M} \lambda_{m} h_{m}(e', f)]}$$
(2-6)

This modeling method has been widely used in natural language understanding tasks. Due to the denominator in equation (2-6) is the only regular probability, in the actual decision-making is of no effect. Therefore, the following decision rules [13]:

$$e = \arg\max\{\Pr(e' \mid f)\} = \arg\max\{\sum_{m=1}^{M} \lambda_m h_m(e, f)\}$$
(2-7)

It can be found that the noise channel model is a special case of the log linear model, when using the language model and the translation model as the probability, and the corresponding  $\lambda_m$  set to 1, that is, the noise channel model. In the log linear model, the inverse translation model which is difficult to interpret for the noisy channel model can also be easily added to the decoding process as a new feature. Log linear model's greatest strength lies in can be arbitrarily add characteristics of a variety of different types, such as number, a new language model and so on, including used to describe the characteristics of sentence dependent phenomena. Can continue to add new features to improve the performance of Machine Translation, there is a lot of flexibility. The training of log linear model and the gradient descent algorithm is used to find the optimal parameters. Due to its excellent performance, the log linear model is widely used in Natural Language Processing, such as part of speech tagging and other tasks.

### 2.2.3. Log-Linear Model Decoding

The heuristic search method confusion network use is column search (BeamSearch) [14]. This algorithm is also often used in the statistical process of machine translation decoding. Confusion network as a graph has the following characteristics, there exists multiple arcs between every two nodes, any decoding result must pass through all the nodes in the graph, so the decoding process in confuse network is from left to right. Taking the translating Chinese sentences "he is not coming home" to English for example, translation search from the beginning of a null hypothesis of translation from the source language sentence and choose any a word phrase translation position as the first word, such as the choice of the word "he" corresponds to the phrase "he". This will form the translation candidate with only one word, and then choose one from the rest that has not been translated and join into the current integration candidate. This step is referred as the translation hypothesis extension. The translation hypothesis extension can extend recursively until the end of the node, and then from the last node to start backtracking. Then the marked words on the arc can be connected together to get the integration results. The process of translation hypothesis expansion is shown in Figure 2-3.



Figure 2-3. The Decoding Procedure from Source Language Sentence

Since there are too much possible decoding results, the translation hypothesis with lower score need to be deleted continuously in the translation extension process and carry on the translation extension from the higher score translation hypothesis. The first case is called translation hypothesis integration. And the simple case is that the two different translation hypotheses are obtained from different paths, but the corresponding decoding results are the same. As a result, the candidate with lower score is deleted and the candidate with higher score is reserved. The example is shown in Figure 2-4.



Figure 2-4. The First Situation that Translation Assumption Combination

In the other case, N-gram model [15] is used in the process of decoding. The words in the current position of two translation hypothesis are the same, and the past n-1 words are the same too. Then the translation hypothesis with lower scores is discarded and examples can be seen in Figure 2-5. As we can see from the figure, the same translation result "is it" is obtained by two different paths, but the score of one path is higher, so we can directly point the other path to the nearest node. The second is the pruning operation and there are two different types of rules. One is to use a fixed number of pruning and the other is pruning according to the threshold. The rules using a fixed number of pruning is applied to limit the number of the current translation hypothesis that has not been extended, such as 500. If a new translation hypothesis is added, the number of translation hypotheses in the stack is more than 500, then the one with lowest score is discarded. The rules using threshold pruning approach is getting the one with the highest score in the current stack and then the score is multiplied by a certain proportion, such as 0.3. If the score is lower than this threshold, it will be discarded. In the general confusion network decoding process, both methods will be used.



Figure 2-5. The Second Situation that Translation Assumption Combination

### 2.2.4. Log-Linear Model Parameter Modulation

The most commonly used linear model parameter modulation method is the minimum error rate training (MERT) method [16]. Before the MERT method, the goal of machine translation parameter modulation is to optimize the maximum mutual exclusion information, the formula is as follows:

$$\lambda_{1}^{M} = \arg \max_{\lambda_{1}}^{M} \{ \sum_{s=1}^{S} \log p_{\lambda_{1}^{M}}(e_{s} \mid f_{s}) \}$$
(2-8)

where  $\lambda_1^M$  represents the parameters that need to be adjusted, S represents all the bilingual sentences in the training set. And  $\log p_{\lambda_1^M}(e_s | f_s)$  represents the probability

when  $e_s$  is the translation of  $f_s$  according to  $\lambda_1^M$ , a detailed formula for calculating the formula (2-6). Although the optimization objective can guarantee the optimal solution to be globally unique, the formula does not prove that the optimal translation can be obtained when the optimal solution is reached. So the MERT method is proposed the following optimization objective to train the log linear model:

$$\lambda_{1} = \arg \min_{\lambda_{1}}^{M} \{ \sum_{s=1}^{S} E(r_{s} \cdot e(f_{s}; \lambda_{1}^{M})) \}$$
  
= 
$$\arg \min_{\lambda_{1}^{M}} \{ \sum_{s=1}^{S} \sum_{k=1}^{K} E(r_{s} \cdot e_{s,k}) \delta(e(f_{s}; \lambda_{1}^{M}), e_{s,k}) \}$$
  
(2-9)

$$e(f_s; \lambda_1^M) = \arg\max_{e \in C_s} \{\sum_{m=1}^M \lambda_m * h_m(e \mid f_s)\}$$
(2-10)

where the function E(r,e) is used to give the error of the translation result e with respect to the standard translation r. And for any one source language sentence  $f_s$ , there are K different translation results  $C_s = \{e_{s,1}, \dots, e_{s,k}\}$ . From the formula (2-9), it can be seen that the new optimization objective is to reduce the number of translation errors. There are many difficulties in the optimization process.

(1) Formula (2-10) contains the argmax operation, and therefore cannot calculate the gradient. Then the gradient descent method cannot be used to optimize.

(2) There are a large number of local extreme points in the optimization objective function (2-9), and the optimization algorithm must be able to deal with it.

In order to use the gradient descent method, we can use the following smoothing optimization objectives:

$$\lambda_{1}^{M} = \arg\min_{\lambda_{1}}^{M} \{ \sum_{s,k} E(e_{s,k}) \frac{p(e_{s,k} \mid f)^{\alpha}}{\sum_{k} p(e_{s,k} \mid f)^{\alpha}} \}$$
(2-11)

The Powells algorithm can be used to optimize the objective function which is not smooth. In the Powells optimization algorithm, we first optimize a variable and keep the other variables constant, and then optimize the other variables until there are no variables to optimize. Due to the existence of local extreme points, we should choose the best one from multiple optimization results of start points.

# **3.** Chinese and English Machine Translation System Based on Confusion Network

### **3.1. System Construction**

### 3.1.1. System Building Tools

(1) Installation Linux system, use Ubuntu 9.10, GCC version 4.4.1.

(2) Chinese English bilingual corpus. Use 619 sentences in English and Chinese.

(3) Statistical Language Modeling Toolkit V2 CMU-Cambridge. Language model tool for generating a language model, so that to be called by the decoder

(4) GIZA++. Here we use "giza-pp-v1.0.5.tar", which contains the auxiliary tool "mkcls-V2 and" GIZA++-V2 "for the generation of word class.

(5) Word segmentation tool. The Chinese word segmentation is ICTCAL, the English word segmentation is tokenizeE.perl.tmpl in EGYPT.

### 3.1.2 Chinese and English Corpus

We download 1000 English Chinese bilingual corpus sentences from the Internet. In order to distinguish from the sentences used in language model, we used the first 619 sentences as aligned corpus, and the 1000 English sentences as a language model.

### 3.1.3 System framework

According to the above platform and corpus, we put forward the confusion network integration automatic translation system, its system framework is shown in Figure 3-2.



Figure 3-2. The System Frame Diagram of English-Chinese Translation

## **3.2. Experimental Results and Analysis**

In this experiment, we select Chinese translation of English, English translation of Chinese bilingual library as the experimental data set [17], and the BLEU score of a single system on the test set is shown in Table 3-1. The parameter modulation data set is a total of 1003 sentences, a total of 2000 sentences in the test set.

System Id	BLEU Score
Online-B	0.2888
Online-A	0.2822
Rbmt-1	0.2757
Uedin	0.2211
Hyderabad	0.1727
Cu-zeman	0.1557

Table 3-1	Individual	System	of BLEU	Scores
	maiviauai	Oystem		000103

Language model, word structure reliability and the length of sentences are common features of decoding. And ngram-count as local information obtained from the training set and test set, have a great influence on the performance of the integration system, Table 3-2 shows the integration performance comparison of only using language model, word confidence and sentence length features of the fusion result and joining the ngram-count.

The combination feature of	Combination scores of	
system	BLEU	
Comb-without-ngram-count	0.2867	
Comb-ngram-count-2	0.3267	
Comb-ngram-count-3	0.3227	
Comb-ngram-count-4	0.3242	

 Table 3-2. The Combination Ngram-count System of BLEU Scores

From the experimental results, we can find that compared with the traditional single system of BLEU, through the integrating system characteristics, constructing confusion network, and using the log-linear model parameter modulation and decoding, the translation system performance can be effectively improved. From the horizontal comparison, confusion model can steadily improve the accuracy and the performance of the system, and the BLEU score values, to greatly enhance the effect.

# 4. Conclusion

With the continuous progress of times and the rapid development of science and technology, cross-language communication is becoming more and more important, and the traditional Chinese English intelligent automatic translation system has not been able to satisfy the growing demand. With the development of computer technology, the English Chinese translation system based on intelligent technology is capable of achieve automatically cross-language communication and improve the efficiency of communication, and people more and more tend to adopt the automatic translation of cross-cultural communication. On this basis, through deeply analyzing the deficiencies of current traditional automatic translation system, this paper proposes the use of integration technology to build Chinese and English intelligent automatic translation system. First we build a confusion network for integration technology, then by using the log linear model for parameter modulation and decoding of confusion network to obtain better accuracy and better performance of the translation results. At last, based on the algorithm, this paper constructs experiment for the integration technology in Chinese and English intelligent automatic translation system. Experimental results show that the English and Chinese intelligent automatic translation system based on Integration Technology is capable of obtaining higher precision and better robustness of translation results, with great practical significance.

### References

- [1] S. Kashif, T. Cohn, and L. Specia, "A Bayesian non-linear method for feature selection in machine translation quality estimation", Machine Translation, vol. 29, no. 2, (2015), pp. 101-125.
- [2] P. Michael, "Parsing english into abstract meaning representation using syntax-based machine translation", Training, vol. 10, (2015), pp. 218-021.
- [3] G. Caglar, "On using monolingual corpora in neural machine translation", arXiv preprint arXiv:1503.03535, (2015).
- [4] D. Nadir, "The Operation Sequence Model—Combining N-Gram-Based and Phrase-Based Statistical Machine Translation", Computational Linguistics, (2015).
- [5] S. Hendra, "Statistical machine translation features with multitask tensor networks", arXiv preprint arXiv:1506.00698, (2015).
- [6] C. Junyoung, K. Cho, and Y. Bengio, "A Character-level Decoder without Explicit Segmentation for Neural Machine Translation", arXiv preprint arXiv:1603.06147, (2016).
- [7] H. Johannes and U. Hahn, "Adding Multilingual Terminological Resources to Parallel Corpora for Statistical Machine Translation Deteriorates System Performance: A Negative Result from Experiments in the Biomedical Domain", Text, Speech, and Dialogue. Springer International Publishing, (2015).
- [8] G. Caglar, "On using monolingual corpora in neural machine translation", arXiv preprint arXiv:1503.03535, (2015).
- [9] K. Aditi and P. S. Sajja, "A Review of Machine Translation Systems in India and different Translation Evaluation Methodologies", International Journal of Computer Applications, vol. 121, no. 23, (2015).
- [10] H. Xiaodong and K. Toutanova., "Joint optimization for machine translation system combination", U.S. Patent No. 9,201,871. 1, (2015).
- [11] A. Jacob, "On the accuracy of self-normalized log-linear models", Advances in Neural Information Processing Systems, (2015).
- [12] N. Iftekhar and D. Gildea, "Feature-based Decipherment for Large Vocabulary Machine Translation", arXiv preprint arXiv:1508.02142, (2015).
- [13] D. Nadir, "Using Joint Models for Domain Adaptation in Statistical Machine Translation", Proceedings of the Fifteenth Machine Translation Summit (MT Summit XV), Florida, USA. AMTA, (**2015**).
- [14] C. Junyoung, K. Cho and Y. Bengio, "A Character-level Decoder without Explicit Segmentation for Neural Machine Translation", arXiv preprint arXiv:1603.06147, (2016).
- [15] D. Nadir, "The Operation Sequence Model—Combining N-Gram-Based and Phrase-Based Statistical Machine Translation", Computational Linguistics, (2015).
- [16] N. F. G. Preslav and S. Vogel, "Analyzing Optimization for Statistical Machine Translation: MERT Learns Verbosity, PRO Learns Length", CoNLL 2015, (2015), pp. 62.
- [17] G. Caglar, "On using monolingual corpora in neural machine translation", arXiv preprint arXiv:1503.03535, (2015).

### Author



**Ruijuan Hu**, She was born in Nanyang, Henan Province in July 1980. Now he is an assistant professor in Henan Institute of Education. And his main research direction is Applied linguistics, and instructional technology.

International Journal of Multimedia and Ubiquitous Engineering Vol.11, No.12 (2016)