Improving Alignment of System Combination by Using Multi-objective Optimization

Tian Xia[+], **Zongcheng Ji**[*], **Shaodan Zhai**[+], **Yidong Chen**[++], **Qun Liu**[*], **Shaojun Wang**[+]
[++] Xiamen University, Xiamen 361005, P.R. China
[+] Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, USA
[*] Institute of Computing Technology, Chinese Academy of Sciences
P.O. Box 2704, Beijing 100190, China
{jizongcheng, liuqun}@ict.ac.cn **and** ydchen@xmu.edu.cn
{xia.7, zhai.6, shaojun.wang}@wright.edu

Abstract

This paper proposes a multi-objective optimization framework which supports heterogeneous information sources to improve alignment in machine translation system combination techniques. In this area, most of techniques usually utilize confusion networks (CN) as their central data structure to compact an exponential number of an potential hypotheses, and because better hypothesis alignment may benefit constructing better quality confusion networks, it is natural to add more useful information to improve alignment results. However, these information may be heterogeneous, so the widely-used Viterbi algorithm for searching the best alignment may not apply here. In the multi-objective optimization framework, each information source is viewed as an independent objective, and a new goal of improving all objectives can be searched by mature algorithms. The solutions from this framework, termed Pareto optimal solutions, are then combined to construct confusion networks. Experiments on two Chinese-to-English translation datasets show significant improvements, 0.97 and 1.06 BLEU points over a strong Indirected Hidden Markov Model-based (IHMM) system, and 4.75 and 3.53 points over the best single machine translation systems.

1 Introduction

System combination (SC) techniques have the power of boosting translation quality in BLEU by several percent over the best among all input machine translation systems (Bangalore et al., 2001; Matusov et al., 2006; Sim et al., 2007; Rosti et al., 2007b; Rosti et al., 2007a; Huang and Papineni, 2007; He et al., 2008; Rosti et al., 2008; He and Toutanova, 2009; Li et al., 2009; Feng et al., 2009; Pauls et al., 2009). A central data structure in the SC is the confusion network, and its quality greatly affects the final performance. He et al. (2008) proposed a new hypothesis alignment algorithm for constructing high-quality confusion networks called Indirect Hidden Markov Model (IHMM), which does better in synonym matching compared with the classic translation edit rate (TER) based algorithm (Rosti et al., 2007b; Rosti et al., 2008; Sim et al., 2007). Now, current state-of-the-art SC systems have been using IHMM or variants in their alignment algorithms more or less (Li et al., 2009; Feng et al., 2009).

Our motivation derives from an observation that in an ideal alignment of a pair of sentences, many-to-many alignments often exist. For instance, “be about to” has the same meaning with “be on the point of”. Because Hidden Markov Model based alignment algorithms, e.g. IHMM for system combination, HMM in GIZA++ software for statistical machine translation (SMT) (Och and Ney, 2000; Koehn et al., 2003), are designed for one-to-many alignment, and running GIZA++ from two directions to gain better performance turns into a standard operation in SMT, therefore we are seeking a way to empower IHMM by introducing bi-directional information.

However, it appears to be intractable in an IHMM model to search the optimal solution by simply defining a new goal as a product of probabilities

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 535–544,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

from two directions. To bypass this problem, Liang et al. (2006) adopts a simple and effective variational inference algorithm.

Further, different alignment algorithms capture different information and linguistic phenomena for a pair of sentences, hence more information would be expected to benefit the final alignment. Liang's method may not be suitable for this expected outcome.

We propose to adopt multi-objective optimization framework to support heterogeneous information sources which may induce difficulties in a conventional search algorithm. In this framework, there exist a variety of matured multi-objective optimization algorithms, e.g. evolutionary algorithm (Deb et al., 2000; Deb et al., 2002), Tabu search (Hansen, 1997), ants colony (Engelbrecht, 2005), and simulated annealing (Serafini, 1994). In this work, we select the multi-objective evolutionary algorithm because of its public open source software (http://www.iitk.ac.in/kangal/codes.shtml). On the other hand, this framework is also totally unsupervised. It prevents weights of a linearly combined goal from training even if all information is homogeneous and applicable in a Viterbi search (Forney Jr, 1973). This framework views any useful information benefiting alignment as an independent objective, and researchers just need to write short codes for objective definitions. The search algorithm seeks for potentially better solutions which are no worse than the current solution set. The output from multi-objective optimization algorithms includes a set of solutions, called *Pareto optimal solutions*, each one being a many-to-many alignment. We then combine and normalize them into a unique one-to-one alignment to perform confusion network construction (Section 3.3).

Our work is conducted on the classic pipeline which has three modules, pair-wise hypothesis alignment, confusion network construction, and training. Now many work integrates neighboring modules to avoid propagated errors to gain improved performance. For example, Rosti et al. (2008), and Li et al. (2009) combine the first and the second module, and He and Toutanova (2009) combine all modules into one directly. Nevertheless, the classic structure also owns its merits. Because of the independence between modules, a system is relatively simple to maintain, and improvements on each module might contribute to final performance additively. Based on our work, lattice-based minimum error rate training (lattice-MERT) and minimum bayes risk training techniques (Kumar et al., 2009) could be adopted on the third module. And Feng et al. (2009) in the second module adopts a different data structure called lattice which could directly use our better many-to-many alignment for construction.

Experiments on the Chinese-to-English task on two datasets use four objectives, IHMM probability (Section 3.2.1), and alignment probability from GIZA++ (Section 3.2.2) from two directions. Results show multi-objective optimization framework efficiently integrates different information to gain approximately 1 BLEU point improvement over a strong baseline.

2 Background

We briefly give an introduction to confusion networks, and because the IHMM based alignment is an important objective in our multi-objective framework, here we also provide detailed definition of formulas for completeness of content.

2.1 Confusion Network

Table 1 shows hypotheses h_1 and h_2 are aligned to selected backbone h_0. When alignment algorithm obtains good enough results, the expected output *"he prefers apples"* is included in its corresponding confusion network in Figure 1. This suggests developing better alignment algorithm may help creating high-quality confusion networks. This also motivates us to use the BLEU of oracle hypotheses to approximately measure the quality of a set of CNs. We hereafter call it an oracle BLEU of a CN. See more in Section 5.1.

h_0 :he	feels	like	apples
h_1 :he	prefer	ε	apples
h_2 :him	prefers	to	apples

Table 1: A toy example of hypothesis alignment, where h_0 is the backbone hypothesis. $h_1$and h_2 are aligned to the backbone separately. The resulting confusion network is in Figure 1.

A confusion network $G = (V, E)$ is a directed acyclic graph with a unique source and sink vertex,

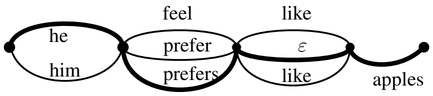

Figure 1: A classic confusion network, and the bold path the expected output.

formally a weighted finite state automation (FSA), where V is the set of nodes and E is the set of edges. Each edge is restricted to attach to a single word as well as an associated probability. A special mark ε is a place-holder denoting no word here.

2.2 IHMM-based Alignment

Indirected Hidden Markov Model (IHMM) was firstly proposed by He et. al (2008). Compared with TER-based alignment performing literal matching, IHMM supports synonym comparison in redefining emission probabilities in an IHMM model.

Let $\boldsymbol{f}^I = (f_1, \ldots f_I)$ be a backbone hypothesis, and $\boldsymbol{e}^J = (e_1, \ldots e_J)$ be a hypothesis aligned to the backbone, both being English sentences in our experiments. Let $\boldsymbol{a}^J = \{a_1, \ldots a_j\}$ be an alignment. Suppose the a_jth word in $\boldsymbol{f}^I$ is aligned to jth word in $\boldsymbol{e}^J$, and the conditional probability that the hypothesis is generated by the backbone, shown in the upper graph of Figure 3, is given by

$$p(\boldsymbol{f}^I, \boldsymbol{e}^J) = \sum_{a^J} \prod_{j=1}^{J} \{p^t(a_j|a_{j-1}, I) p^o(e_j|f_{a_j})\} \tag{1}$$

The *distortion probability* $p^t(a_j|a_{j-1}, I)$ from position a_{j-1} to a_j, relies on jumped distance, which is computed as follows:

$$p^t(i^{'}|i, I) = \frac{c(i^{'} - i)}{\sum_{t=1}^{I} c(t - i)} \tag{2}$$

The *distortion parameters* $c(d)$ are grouped into 11 buckets, $c(\leq -4)$,$c(-3)$,$c(-2)$…$c(5)$,$c(\geq 6)$. Because all the hypotheses in system combination are in the same language, the IHMM model would support more monotonic alignments, and non-monotonic alignments will be penalized.

$$c(d) = (1 + |d - 1|)^{-K}, d = -4 \ldots 6 \tag{3}$$

where K is tuned on held-out data.

Let p_0 be the probability of jumping to a $null$ word state, which is also tuned on held-out data, and the accurate transition probability becomes:

$$p^t(i^{'}|i, I) = \begin{cases} p_0 & \text{if } i^{'} = null \\ (1 - p_0) p^t(i^{'}|i, I) & \text{otherwise} \end{cases} \tag{4}$$

The *output probability* $p^o(e|f)$ from the state word f to the observation word e, also called translation probability, is a linear interpolation of semantic similarity $p^{sem}(e|f)$ and surface similarity $p^{sur}(e|f)$, and α is the interpolation factor:

$$p^o(e|f) = \alpha p^{sem}(e|f) + (1 - \alpha) p^{sur}(e|f) \tag{5}$$

When calculating *semantic similarity* $p^{sem}(e|f)$, source sentence src is needed, and a bilingual probabilistic dictionary $p^{dic}(w_1|w_2)$ is necessary.

$$p^{sem}(e|f) \approx \sum_{c \in src} p^{dic}(c|f) \cdot p^{dic}(e|c) \tag{6}$$

Note that $p^{sem}(e|f)$ has been updated with different source sentences.

The *surface similarity* $p^{sur}(e|f)$ is measured by the literal matching rate:

$$p^{sur}(e, f) = \exp\{\rho[\frac{\text{LMP}(f, e)}{\max(|f|, |e|)} - 1]\} \tag{7}$$

where LMP(f, e) is the length of the longest matched prefix, and ρ is a smoothing parameter.

3 Multi-objective Optimization

Many decision making problems in the real world consider more than one objective. One natural way is to scalarize multiple objectives into one by assigning it with a weight vector. This method allows a simple optimization algorithm in many cases, while in system combination, it would cause problems.

In the first module, in order to train suitable weights of objectives, extra labeled data is needed, besides that, the efficient Viterbi algorithm for searching the optimal alignment would not work for the alignment objectives in this work. More, the parameter training in the third module relies on the CNs constructed from the output of the first module, which increases the instability of the whole system. Therefore, an unsupervised multi-objective algorithm may be a good choice allowing for more alignment information.

There exist other alternative optimization algorithms in the multi-objective optimization framework, though the evolutionary algorithm is adopted here, we only introduce some general concepts.

3.1 Pareto Optimal Solutions

A general multi-objective optimization problem consists of a number of objectives and is associated with a number of constraints. Mathematically, the problem can be written as follows (Deb, 2001)

$$\begin{array}{rll} \text{Maximize} & f_i(x) & i = 1 \ldots M \\ \text{s.t.} & g_j(x) \leq 0 & j = 1 \ldots N \\ & h_k(x) = 0 & k = 1 \ldots K \end{array}$$

where x denotes a potential solution, its structure relying on different problems, and the number of constraints M, N, K depend on different problems. All the functions f_i, g_j, h_k map a solution x into a scalar. We will explain them in terms of system combination.

In this work, we refer to $x = \{x_{i,j} | x_{i,j} \in \{0, 1\}\}$ as a potential alignment of a pair of hypotheses, where $x_{i,j}$ is a boolean value to denote *whether the ith word in the first hypothesis is aligned to the jth word in the second hypothesis*. Here the definition of $\boldsymbol{x}$ seems different from that of $\boldsymbol{a}$ in Formula 1, and they could convert to each other. Using a line-based access style, a matrix can be unfolded as a vector. We refer to f as IHMM alignment probability (He et al., 2008) and GIZA++ alignment probability (Chen et al., 2009), total four objectives from two directions, and the larger the objectives, the better. The g_js and h_ks serve as the role of checking if $\boldsymbol{x}$ represents a legal alignment. For instance, the subscripts of $x_{i,j}$ are not in bounds.

Definition 1. *Let x, x' be two potential alignments. If $f_i(x) \geq f_i(x')$ holds for all i, we call the alignment x dominates the alignment x'. If there does not exist any alignment x'' to dominate x, we call the alignment x to be non-dominated.*

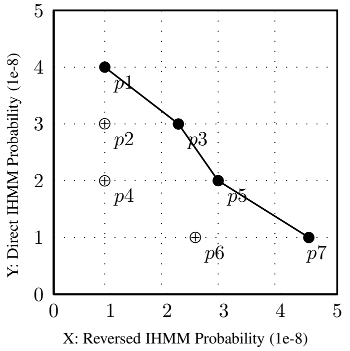

Figure 2: Sample solutions with only two objectives. Pareto Optimal Solutions p_1, p_3, p_5, p_7. Other points p_2, p_4, p_6 are dominated by at least one point in the Pareto optimal solutions.

Definition 2. *A alignment x is said to be Pareto optimal if there is no other alignment x' found to dominate x.*

In Figure 2, p_1 dominates p_2, and p_2 dominates p_4. To summarize, a point is dominated by the ones on its upper and right side with ties. In this example, p_1, p_3, p_5, p_7 are Pareto optimal.

In some cases, Pareto optimal solutions can be used for good candidate solutions. Considering the IHMM model, maximizing Y axis, the top-4 best alignments are p_1, p_2, p_3, p_4. But from the view of Pareto optimal, the top-4 alignments would be p_1, p_3, p_5, p_7 without order, which considers a greater range than a single optimization model. In our method, we just combine these Pareto optimal solutions equally into a unique alignment (Section 3.3).

Our adopted multi-objective optimization searching algorithm is the non-dominated sorting genetic algorithm II (NSGA-II) (Deb et al., 2000; Deb et al., 2002) with an open source software (http://www.iitk.ac.in/kangal/codes.shtml). NSGA-II has a complexity of O(mn^2), where m is the number of objectives and n is the population size in an evolutionary algorithm.

3.2 Objectives in Evolutionary Algorithm

The optimization objectives in our experiments can be categorized as an IHMM alignment probability (He et al., 2008) and GIZA++ alignment probability

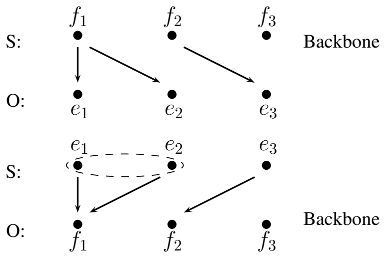

Figure 3: The same alignment $(f_1, e_1)(f_1, e_2)(f_2, e_3)$ in two IHMM models. The upper one is a typical example in IHMM, and in the bottom one, because any word in the observation is required not to correspond to two statuses, it has a minor trouble. S: status sequence, O: observation sequence.

(Chen et al., 2009), total four from two directions.

3.2.1 IHMM Probability

A typical IHMM alignment is demonstrated in the upper graph of Figure 3, where a backbone is acting the role of a status sequence. The unnormalized conditional alignment probability is $[p^t(1|null)] \cdot [p^t(1|1)p^t(2|1)] \cdot [p^o(e_1|f_1)p^o(e_2|f_1)p^o(e_3|f_2)]$. However, the same alignment $(f_1, e_1)(f_1, e_2)(f_2, e_3)$, if we change the alignment direction, the backbone being observations, would be a bit different. We offer a minor modification to Formula 1.

Look at the bottom graph of Figure 3, the observation f_1 has two statuses, e_1 and e_2 at the same time, it becomes ambiguous to compute the transitional probability between $p^t(3|1)$ and $p^t(3|2)$. This is because IHMM algorithm deals with one-to-many alignments, and MOEA permits many-to-many alignments.

We hence empirically modify the IHMM model to support many-to-many alignments. A new status is defined, rather than a single position $p^t(j|i)$, but as a set of positions $p^t(\{j\}|\{i\})$. The positions in one status need not to be adjacent to each other.

The redefined *transitional probability*

$$p^t(\{j\}|\{i\}) = \frac{1}{|\{j\}| \cdot |\{i\}|} \sum_{i,j} p^t(j|i)$$

The redefined *emission probability*

$$p^o(j|\{i\}) = \prod_i p^o(j|i)$$

We need to note that there is no guarantee on the closed property of probabilities, though these approximations prove to be effective in a practical sense. Straightforwardly, when there is only one position in a new status, the expanded IHMM degenerates to the standard IHMM.

Let us return to the second IHMM example. The new probability becomes $[p^t(1|\text{null})p^t(2|\text{null})] \cdot [\frac{1}{2}p^t(3|1)p^t(3|2) \cdot p^t(\text{null}|3)] \cdot [p^o(f_1|e_1)p^o(f_1|e_2)p^o(f_2|e_3)p^o(f_3|\text{null})]$.

3.2.2 Alignment Probability

GIZA++ considers very different and more information in alignment, we attempt to utilize them. All probabilities appearing in below formulas can be looked up in GIZA++.

Given a pair of hypotheses $\boldsymbol{f}^I = (f_1, \ldots f_I)$, $\boldsymbol{e}^J = (e_1, \ldots e_J)$, and their alignment $\boldsymbol{a}$, the alignment probability could be calculated as follows

$$p^{Giza}(\boldsymbol{e}^J|\boldsymbol{f}^I, \boldsymbol{a}) = \prod_{e_i} T(e_i|\boldsymbol{f}^I, \boldsymbol{a})$$

$$T(e_i|\boldsymbol{f}^I, \boldsymbol{a}) = \begin{cases} n(\phi_i|e_i)\sum_{(j,i)\in\boldsymbol{a}} t(e_i|f_j)a(j|i)/\phi_i & \text{if } \phi_i \neq 0 \\ n(0|e_i)t(e_i|\text{null})a(0|i) & \text{otherwise} \end{cases}$$

$$\phi_i = |\{j|(i,j) \in \boldsymbol{a}\}|$$

where ϕ_i is the fertility number, $t(e|c)$ *the translation probability* for the word pair, $z(j|i)$ *alignment probability* to show how likely a target word at position i could be translated into a source word at position j, and $n(\phi|e)$ is the *fertility probability* to show how likely a given target word e is translated into ϕ source words.

In order to increase the coverage of words, we collect all the hypothesis pairs in both the tuning set and the test set and feed them into GIZA++. This is an off-line operation, which makes it not suitable for an online translation system. In some circumstances, users submit a pile of documents in the hope of high-quality translations, thus more useful knowledge sources would be helpful. In our experiments, a pure GIZA++ based system combination does not perform as well as IHMM based, but does benefit the final translation quality if combined in our multi-objective optimization framework.

3.3 Configuration of Evolutionary Algorithm

3.3.1 Encoding

Given a sentence pair <$\boldsymbol{f}^I$, $\boldsymbol{e}^J$>, we define a two-dimensional matrix $\boldsymbol{x} = \{z_{i,j} | z_{ij} \in \{0, 1\}\}$ to encode a set of possible alignments. Using a line-based access style, the matrix could be unfolded as a vector with $|I| \cdot |J|$ bits of length.

3.3.2 Initialization

Because in NSGA-II software the initial population are generated at random. In order to make NSGA-II more consistent and flexible, better initial seeds should be fed with, thus we combine an existing word alignment results as input. Here we use together two N-best lists generated from directional HMM and reversed HMM respectively for initialization.

3.3.3 Normalization of Pareto Optimal Solutions

Multi-objective optimization algorithms do not pose weights on objectives, thus they output a set of so-called Pareto optimal solutions, each of which is a many-to-many alignment. We can understand them as an N-best alignment list without explicit preferences. We also empirically compare it with the idea that directly cuts an N-best list from the IHMM based alignment.

We describe a two-stage strategy for normalization. Firstly, we use a simple and effective voting strategy to combine a set of many-to-many alignments into a single many-to-many alignment, and Secondly we normalize it into a one-to-one alignment for confusion network construction. In the first stage, we count the number of word-to-word alignments on each position pair (i, j). If there is more than a half number of alignments, then we output 1, otherwise 0. In the second stage, if any word relates to more than one word alignment, the one with the highest posterior probability is selected (He et al., 2008; Feng et al., 2009). The posterior probabilities can be computed in a classic forward-backward procedure in IHMM (He et al., 2008).

4 Training and Decoding

Our work does not change the classic pipeline, thus the model and features are nearly identical to the ones in (Rosti et al., 2007b; He et al., 2008), which are modeled in a log-linear fashion in Eq. 8. Translation on a CN is just a concatenation of edges traversed, on which 4 categories of features are defined.

1. word posterior probabilities. In Eq. 8, $p(w|sys, span)$ are word confidence scores. If the word w comes from the kth hypothesis of thesys-th system, the raw score should be $\frac{1}{k+1}$, and then it would be normalized by the same sys and $span$. The same word coming from different systems owns a different score, so there are sys system weights λ_{sys}.

2. logarithm of language model score, $L(h)$.

3. number of $null$ edge, Num_{null}.

4. number of words, Num_w.

$$\begin{array}{rcl} log(h) & = & \sum_{span} \log(\sum_{sys} \lambda_{sys} p(w|sys, span)) \\ & + & w_0 L(h) + w_1 Num_{null} + w_2 Num_w \end{array} \tag{8}$$

Decoding a confusion network is straightforward, traversing each node from left to right, and the beam search algorithm will retain for each node an N-best list. The final N-best can be acquired following (Huang and Chiang, 2005).

The training process follows minimum error rate training (MERT) described in (Och, 2003; Koehn et al., 2003). In each iteration, the Powell algorithm would attempt to predict the optimal parameters on the cumulative N-best list.

5 Experiments

We evaluate our method in two datasets in the Chinese-to-English task. In the first one, NIST MT 2002 and 2005 are used for tuning and testing respectively, and in the second, the newswire part of MT 2006 and 2008 are for tuning and testing. A 5-gram language model is trained on the Xinhua portion of the Gigaword corpus. We report the case-sensitive NIST-BLEU score.

Four single machine translation systems participating in the system combination consist of a BTG-based system using a Max-Entropy based reordering model, a hierarchical phrase-based system, a Moses decoder and a syntax-based system. 10-best unique hypotheses from a single system on the development

SYSTEM	MT 2005	MT 2008(news)
best single	0.3207	0.3016
IHMM*	0.3585(+3.78%)	0.3263(+2.47%)
IncIHMM	0.3639(+4.32%)	0.3320(+3.04%)
GIZA++	0.3438(+2.31%)	0.3166(+1.50%)
PPBD	0.3619(+4.10%)	0.3306(+2.90%)
N-best IHMM	0.3590(+3.83%)	0.3270(+2.54%)
dH+rH	0.3604	0.3284
dH+dT	0.3610	0.3290
dH+rH+dT	0.3609	0.3289
dH+rH+rT	0.3630*(+4.27%)	0.3320∗(+3.04%)
dH+rH+dT+rT	**0.3682****(+4.75%)	**0.3369****(+3.53%)

Table 2: PPBD is a posterior probabilistic-based decoding (section 5.3). N-best IHMM simulates the Pareto optimal solutions in our method (section 5.3). The last five systems adopt different objective combinations. The improvement percents in parentheses are compared to the best single. dH: directed IHMM, rH: reversed IHMM, dT: directed translation probability, rT: reversed translation probability. ** significance at 0.01 level, and * significance at 0.05 level over the IHMM model.

and test sets are collected as the input of the system combination.

Our baseline systems are described as follows. Two main baseline systems are IHMM based and incremental IHMM (Li et al., 2009). The first system differs from our method just in hypothesis alignment algorithm, and the second combines the first and second module of the system combination pipeline.

Because our method utilizes bidirectional information, we also provide another two alternative systems for comparison, which are GIZA++ based alignment and the posterior probability based alignment (Liang et al., 2006). Finally, we also provide an N-best alignment IHMM system, which combines an N-best alignment list to simulate the Pareto optimal solutions in our method.

The method that linearly combines all objectives is not listed as our baseline like (Duh et al., 2012) does, because their algorithm finds the best weighted solution in a fixed and small solution set, while in our problem, the solution space is a trellis-style structure consisting of an exponential number of solutions, and no efficient algorithms apply here.

The IHMM based alignment utilizes typical settings (He et al., 2008; Feng et al., 2009). The smoothing factor for the surface similarity model, and $\rho = 3$ the controlling factor for the distortion model, $K = 2$. The bilingual probabilistic dictionary is trained in the FBIS corpus which includes about 230k parallel sentence pairs. GIZA++ based system is to run GIZA++ from two directions to align all the hypotheses, and make the intersection using grow-diag-final heuristics (Koehn et al., 2003). The many-to-many alignments are normalized with the same method with ours. Our system employs NSGA-II software to realize the MOEA algorithm. The main parameters, generation number, cross probability and mutation probability, and population size, are empirically set as 100, 0.9, 0.001 and 40, and we examine the influence of difference populations sizes in the full system combination.

5.1 The Quality of Confusion Networks

This experiment shows the relationship between hypothesis alignment and confusion network. Intuitively, we expect a better hypothesis alignment would reduce the error in constructing confusion networks, and then improve the final translation quality.

We first use the *alignment error rate* (AER) (Och and Ney, 2000), which is widely used to measure the quality of hypothesis alignment. The smaller, the better. For convenience, we only examine exact literal matching. IHMM based alignment reaches around 0.15 in AER, and our method 0.145.

As the AER may not vividly reflect the relations between alignment and the final BLEU of systems, and the quality of confusion network is hard to measure directly, we assume that the quality of confusion networks could be measured by the oracle hypotheses that could be generated from them. We test the BLEU of the oracle hypotheses.

From this angle, we demonstrate several oracle BLEU of CNs generated from some conventional alignment algorithms. The results are shown in Table 3.

We find the confusion network from IHMM based alignment (He et al., 2008) is better than that from TER based alignment (Rosti et al., 2007b) by about 1 point in both two datasets. These quantities agree with the final improvements in the BLEU score in (He et al., 2008). As confusion networks from MOEA based alignment also show superiority over

alignment	MT02	MT05
GIZA++	0.5690	0.5228
TER	0.5720	0.5270
IHMM	0.5883	0.5382
IncIHMM	0.5931	0.5453
MOEA	**0.6017**	**0.5526**

Table 3: Oracle BLEUs of CNs. GIZA++: invoking GIZA++ software. TER: minimum translation edit rate. IHMM: indirect hidden markov model. IncIHMM: incremental indirect hidden markov model. MOEA: multi-objective evolution algorithm.

that from IHMM based in the oracle BLEU, we expect our final translation quality would be improved.

In Table 3, GIZA++ and TER perform similarly, because the former is more capable of tackling many-to-many alignments over the latter, while latter based might obtain relatively more precise alignment information. Both of the two do not consider synonym matching compared to IHMM.

Our method and IncIHMM overpass IHMM on this metric due to different strategies. Obtaining better hypothesis alignment or better construction of confusion networks benefit the quality of CNs.

5.2 Different Objective Combinations

As our framework is convenient to support different alignment information, we test the influence of different objective combinations to the final translation quality. We adopt four objectives to depict the candidate alignment, directed IHMM probability (dH), reversed IHMM probability (rH), directed alignment probability (dT), and reversed alignment probability (rT). Table 2 demonstrates all the results.

We can see that the IHMM based system outperforms the GIZA++ based system by about 1-1.5 points in BLEU, which agrees with the difference of oracle BLEU in Table 1. From (He et al., 2008), the IHMM based system outperforms the TER based by 1 point, which also agrees with our results in Table 1. Our system, using dH + rH + dT + rT, improves BLEU score by about 1 points over the IHMM based system. This comparison verifies our assumption, improving the quality of the confusion network does improve system performance.

The different feature combinations exhibit interesting results. The system with dH + rH + dT is 0.05 point better than the system with dH + rH, and the system dH + rH + rT is 0.3 point better than system with dH + rH, so the contributions of feature dT and rT are 0.05 and 0.3 respectively. While the two features are used together in the fourth system, the contribution is about 0.8 point, rather than 0.35. This phenomenon also proves the correlations between different features.

Our method explores a way to integrate GIZA++ and IHMM, and is supportive of useful features. Compared to the classic and powerful IHMM based system, we obtained an improvement of 0.97 points on MT 05 and 1.06 points on news of MT 2008, and equivalently over the best single system by 4.75 points and 3.53 points respectively. More, compared with the incremental IHMM, our system also shows moderate improvement, though not much. We hope these two ideas could be effectively combined in the future work.

5.3 Comparison with Other Bi-directional Alignment Methods

Our method introduces multiple alignment information into system combination to obtain improvements, thus it would be interesting to explore other alternative methods for utilizing this information. We provide three alternative methods similar to our motivations, and they fall into two categories.

The first category is from the angle of bi-directional alignment. We use GiZA++ alignment and the posterior probability decoding-based alignment for comparison. The basic idea for the latter is setting a word-to-word alignment $x_{i,j}$ as 1, if its approximate posterior marginal probability $q(x_{i,j}, x) = p_d(x_{i,j}|x, \theta_d) \cdot p_r(x_{i,j}|x, \theta_r)$ is greater than a threshold δ, where p_d and p_r are posterior marginal probabilities from directed and reversed IHMM models, which could be conveniently computed with a forward-backward algorithm, and the δ is tuned on a validation-set optimized data. We just list some δ values to examine its best performance shown in Table 4.

The second class is because our method combines the Pareto optimal solutions that consist of several candidate alignments, thus for fairness we also use a 100-best outputs from the directed IHMM model and conduct the same normalization technique.

The general results are shown in Table 2. We can

δ	**MT 2005**	**MT 2008**
IHMM	0.3585	0.3263
0.15	0.3556	**0.3391**
0.2	**0.3619**	**0.3306**
0.25	0.3575	0.3278
0.3	0.3608	0.3259

Table 4: Posterior decoding. When threshold δ are set to suitable values, simple bi-directional alignment could overpass the baseline.

see that, GIZA++ leads to the worst performance, which can be explained as GIZA++ does not support synonym matching like IHMM. The N-best IHMM has a minor improvement over the IHMM method. We found differences in the N-best list are not obvious enough. In comparison, the posterior decoding method brings relatively significant improvements on both datasets. However, the threshold δ must be selected suitably. Table 4 lists the ideal results, which will be hampered when tuning on a validation set.

All of the three candidate methods can not conveniently support extra alignment information, and a linear model poses restrictions on features to get an efficient decoding, the multi-objective optimization may be a good selection as an inference algorithm in many circumstances.

5.4 Population Size

We test the influence of final translation quality and time consumed by different population size.

population	**BLEU**
size	MT 2005
20	0.3597
40	0.3682
60	0.3655

Table 5: Big population size consumes more CPU time. In our experiments, we use a multi-thread technique to speed up the alignment, and choose 40 as the parameter to leverage the time and BLEU.

We expect enlarging the population size would improve the translation quality, but the BLEU in population size set as 60 does not overpass when set as 40. We conjecture that, in our code, if the N-best size from IHMM (we set as 50-best) does not reach the population size, we would use randomly generated seeds, which may hamper the performance of MOEA. We also tried a larger population in MOEA, but did not receive obvious improvement on performance.

We exerted a hard restriction on the genes in evolutionary algorithm, that is many-to-many discontiguous alignment is forbidden. This trick speeds up running by about 20 times, and does not harm system performance. Now our method runs about 0.9 seconds to align a pair of hypotheses. In practice, we utilize multi-thread to speed up.

6 Conclusion

In this paper, we explore a multi-objective framework to conveniently support more useful alignment objectives to improve the hypothesis alignment. By a minor modification of the first module in the classic pipeline, we successfully combine GIZA++ and IHMM to obtain significant improvement over a powerful and state-of-the-art IHMM based system. In comparison with another genre of improving system combination by combing adjacent modules of the pipeline, more powerful incremental IHMM here, our system also show moderate improvement. Though, our best system may not overpass He and Toutanova (2009) who combine all the modules into a unified training procedure, we believe our method could boost many work on the higher modules of the pipeline to obtain a further improvement to match their work.

7 Acknowledgement

This research is partially supported by Air Force Office of Scientific Research under grant FA9550-10-1-0335, the National Science Foundation under grant IIS RI-small 1218863 and a Google research award. We thank the anonymous reviewers for their insightful comments.

References

B Bangalore, German Bordel, and Giuseppe Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Automatic Speech Recognition and Understanding*.

Yidong Chen, Xiaodong Shi, Changle Zhou, and Qingyang Hong. 2009. A word alignment model based on multiobjective evolutionary algorithms. *Computers and Mathematics with Applications*, 57.

Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and Tanaka Meyarivan. 2000. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. *Lecture notes in computer science*, 1917:849–858.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *Evolutionary Computation, IEEE Transactions on*, 6(2):182–197.

Kalyanmoy Deb. 2001. Multi-objective optimization. *Multi-objective optimization using evolutionary algorithms*, pages 13–46.

John DeNero, Shankar Kumar, Ciprian Chelba, and Franz Och. 2010. Model combination for machine translation. In *Proc. of NAACL*, pages 975–983.

Kevin Duh, Katsuhito Sudoh, Xianchao Wu, Hajime Tsukada, and Masaaki Nagata. 2012. Learning to translate with multiple objectives. In *Proc. of ACL*, pages 1–10.

Andries P Engelbrecht. 2005. *Fundamentals of computational swarm intelligence*, volume 1. Wiley Chichester.

Yang Feng, Yang Liu, Haitao Mi, Qun Liu, and Yajuan Lü. 2009. Lattice-based system combination for statistical machine translation. In *Proc. of EMNLP*, EMNLP '09.

G David Forney Jr. 1973. The viterbi algorithm. *Proc. of the IEEE*, 61(3):268–278.

Michael Pilegaard Hansen. 1997. Tabu search for multiobjective optimization: Mots. In *Proc. of Multiple Criteria Decision Making*, pages 574–586.

Xiaodong He and Kristina Toutanova. 2009. Joint optimization for machine translation system combination. In *Proc. of EMNLP*.

Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In *Proc. of EMNLP*.

Liang Huang and David Chiang. 2005. Better k-best parsing. In *Proc. of IWPT*.

Fei Huang and Kishore Papineni. 2007. Hierarchical system combination for machine translation. In *Proc. of EMNLP-CoNLL*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL: Poster*, pages 177–180.

Shankar Kumar, Wolfgang Macherey, Chris Dyer, and Franz Och. 2009. Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices. In *Proc. of Joint ACL and AFNLP*.

Zhifei Li and Sanjeev Khudanpur. 2009. Forest reranking for machine translation with the perceptron algorithm. *GALE book chapter on MT From Text*.

Chi-Ho Li, Xiaodong He, Yupeng Liu, and Ning Xi. 2009. Incremental hmm alignment for mt system combination. In *Proc. of Joint ACL and AFNLP*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proc. of NAACL*.

Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proc. of EACL*.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. *Proc. of ACL-08: HLT*, pages 192–199.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. pages 440–447, October.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167.

Adam Pauls, John DeNero, and Dan Klein. 2009. Consensus training for consensus decoding in machine translation. In *Proc. of EMNLP*.

Antti-Veikko I Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz, and Bonnie Dorr. 2007a. Combining outputs from multiple machine translation systems. In *Proc. of NAACL-HLT*.

Antti-Veikko I Rosti, Spyros Matsoukas, and Richard Schwartz. 2007b. Improved word-level system combination for machine translation. In *Proc. of ACL*, volume 45.

Antti-Veikko I Rosti, Bing Zhang, Spyros Matsoukas, and Richard Schwartz. 2008. Incremental hypothesis alignment for building confusion networks with application to machine translation system combination. In *Proc. of WSMT*.

Paolo Serafini. 1994. Simulated annealing for multi objective optimization problems. In *Proc. of Multiple Criteria Decision Making*, pages 283–292. Springer.

Khe Chai Sim, William J Byrne, Mark JF Gales, Hichem Sahbi, and Phil C Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *Proc. of ICASSP*, volume 4.

Yong Zhao and Xiaodong He. 2009. Using n-gram based features for machine translation system combination. In *Proc. of NAACL: Short Papers*, pages 205–208.