Minimum Bayes Risk Decoding with Enlarged Hypothesis Space in System Combination

Tsuyoshi Okita and Josef van Genabith

Dublin City University, School of Computing, Glasnevin, Dublin 9, Ireland

Abstract. This paper describes a new system combination strategy in Statistical Machine Translation. Tromble et al. (2008) introduced the evidence space into Minimum Bayes Risk decoding in order to quantify the relative performance within lattice or n-best output with regard to the 1-best output. In contrast, our approach is to enlarge the hypothesis space in order to incorporate the combinatorial nature of MBR decoding. In this setting, we perform experiments on three language pairs ES-EN, FR-EN and JP-EN. For ES-EN JRC-Acquis our approach shows 0.50 BLEU points absolute and 1.9% relative improvement obver the standard confusion network-based system combination without hypothesis expansion, and 2.16 BLEU points absolute and 9.2% relative improvement compared to the single best system. For JP-EN NTCIR-8 the improvement is 0.94 points absolute and 3.4% relative, and for FR-EN WMT09 0.30 points absolute and 1.3% relative compared to the single best system, respectively.

1 Introduction

In a sequence prediction task, a max-product algorithm (or Viterbi decoding [29]) is a standard technique to find an approximate solution x which maximizes the joint distribution $p(x)$ (while a sum-product algorithm [23] attempts to find an exact solution x). Max-product is an inference algorithm for a single model in a tree or a chain structure [13]. Suppose that we consider a combination of multiple systems whose model parameters are different. The first problem is that we are required to calibrate the quantities coming from the different models since these quantities are not immediately comparable in general. The second problem is that it is often the case that an increase in the number of participating systems increases the overall computation in a non-linear way; fortunatley, however, it turns out that often a lot of calculations are redundant over systems at the same time. In our particular situation, the number of nodes increases exponentially since the corresponding nodes are searched in a combinatorial manner (even though the overall number of system is small); however, there are a lot of redundancies.

In order to address these problems, this paper imposes practical assumptions limiting our scope but in such a way that our immediate application of Minimum Bayes Risk decoding [14] does not suffer.[1] Our assumptions are that

[1] Note that it is not clear what kind of other applications exist.

A. Gelbukh (Ed.): CICLing 2012, Part II, LNCS 7182, pp. 40–51, 2012.

(i) the model structures are almost identical and that (ii) the probabilities which we compare are indexed and thus can be calibrated locally. Under this assumption, it turned out that we can employ a standard MAP assignment algorithm [13] to calibrate the probabilities arising from different systems, even though the original aim of normalization of MAP assignment is different in that the unnormalized probabilities arise by themselves since MAP assignment partitions variables into E(evidence), Q(query), and H(hidden) variables. Clique tree [24] is a technique to consider only some factors locally, which can be applied here.

With these preparations, we develop a new system combination strategy using Minimum Bayes Risk (MBR) decoding [14] which exploits a larger hypothesis space. A system combination strategy [2,16,6] is a state-of-the-art technique to improve the overall BLEU score. Recently, Tromble et al. [28] attempted to exploit a larger evidence space by using a lattice structure. DeNero et al. [4,5] introduced n-gram expectation, while Arun et al. [1] compared MBR decoding with MAP decoding for general translation tasks in a MERT setting [17].

The remainder of this paper is organized as follows. Section 2 reviews the decoding algorithm in SMT. Section 3 describes our algorithm. In Section 4, our experimental results are presented. We conclude in Section 5.

2 Decoding Algorithm in SMT

There are two popular decoding algorithms in phrase-based SMT: MAP decoding and MBR decoding [10]. MAP decoding is the main appraoch in phrase-based SMT [12], while MBR decoding is mainly used for system combination [2,16,6,28,4]. The MAP decoding algorithm seeks the most likely output sequence, while the MBR decoding seeks the output sequence whose loss is the smallest.

Let E be the target language, F be the source language, A be an alignment which represents the mapping from source to target phrases, and $M(\cdot)$ be an MT system which maps some sequence in the source language F into some sequence in the target language E. MAP decoding can be written as in (1):

$$\hat{E}_{best}^{MAP} = \arg\max_{E} \sum_{A} P(E, A|F) \tag{1}$$

Let $\mathcal{E}$ be the translation outputs of all the MT systems. For a given reference translation E, the decoder performance can be measured by the loss function $L(E, M(F))$. Given such a loss function $L(E, E')$ between an automatic translation E' and the reference E, a set of translation outputs $\mathcal{E}$, and an underlying probability model $P(E|F)$, a MBR decoder is defined as in (2) [14]:

$$\begin{aligned}
\hat{E}_{best}^{MBR} &= \arg\min_{E' \in \mathcal{E}} R(E') \\
&= \arg\min_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} L(E, E') P(E|F) \qquad (2) \\
&= \arg\max_{E' \in \mathcal{E}_H} \sum_{E' \in \mathcal{E}_E} BLEU_E(E') P(E|F) \qquad (3)
\end{aligned}$$

where $R(E')$ denotes the Bayes risk of candidate translation E' under the loss function L, BLEU_E(E') [22] is a function to evaluate a hypothesis E' according to E, $\mathcal{E}_H$ refers to the hypothesis space from which translations are chosen, $\mathcal{E}_E$ refers to the evidence space used for calculating risk. Note that a hypothesis space $\mathcal{E}_H$ and an evidence space $\mathcal{E}_E$ appeared in [9,28,4,1].

The confusion network-based approach [2,16,6] enables us to combine several fragments from different MT outputs. In the first step, we select the sentence-based best single system via a MBR decoder (or single system outputs are often used as the backbone of the confusion network). Note that the backbone determines the general word order of the confusion network. In the second step, based on the backbone which is selected in the first step, we build the confusion network by aligning the hypotheses with the backbone. In this process, we used the TER distance [25] between the backbone and the hypotheses. We do this for all the hypotheses sentence by sentence. Note that in this process, deleted words are substituted as NULL words (or ϵ-arcs). In the third step, the consensus translation is extracted as the best path in the confusion network. The most primitive approach [16] is to select the best word $\hat{e}_k$ by the word posterior probability via voting at each position k in the confusion network, as in (4):

$$\hat{E}_k = \arg\max_{e \in \mathcal{E}} p_k(e|F) \tag{4}$$

Note that this word posterior probability can be used as a measure how confident the model is about this particular word translation [10], as defined in (5):

$$p_i(e|F) = \sum_j \delta(e, e_{j,i}) p(e_j|F) \tag{5}$$

where $e_{j,i}$ denotes the i-th word and $\delta(e, e_{j,i})$ denotes the indicator function which is 1 if the i-th word is e, otherwise 0. However, in practice as is shown by [6,15], the incorporation of a language model in this voting process will improve the quality further. Hence, we use the following features in this voting process: word posterior probability, 4-gram and 5-gram target language model, word length penalty, and NULL word length penalty. Note that Minimum Error-Rate Training (MERT) is used to tune the weights of the confusion network. In the final step, we remove ϵ-arcs, if they exist.

3 Our Algorithm

Tromble et al. [28] introduced a lattice in the evidence space into Minimum Bayes Risk decoding in order to quantify the relative performance within lattice or n-best output with regard to the 1-best output. In contrast, our approach is to enlarge the hypothesis space via different kinds of lattices in order to incorporate the combinatorial nature of MBR decoding.

We first present the motivation for using the enlarged hypothesis space and searching for the optimal subset $\mathcal{E}_0$ among this enlarged hypothesis space $\mathcal{E}$ (where $\mathcal{E}$ is the translation outputs of all the MT systems participating in the

Table 1. Motivating examples. MBR decoding can be schematically described as maximizing the n-gram expectations between the MT output sequence and some sequence, as is described in this table. The left table shows the MT output sequences consisting of 5 systems, while the right table shows the MT output sequences consisting of 4 systems. In this case, the 1-gram expectation of "bbcd" for 4 systems (right table) are better than those for 5 systems (left table). This suggests that it may be better to remove extremely bad MT output from the inputs of system combination.

A	MT outputs	prob	1-gram expectation
1	a a a c	0.30	$\mathbb{E}_A$(aaac)=1.2
2	b b c d	0.20	$\mathbb{E}_A$(bbcd)=2.1
3	b b b d	0.20	$\mathbb{E}_A$(bbbd)=2.0
4	b b c f	0.20	$\mathbb{E}_A$(bbcf)=1.8
5	f f b d	0.10	$\mathbb{E}_A$(ffbd)=1.0

B	MT outputs	prob	1-gram expectation
1	a a a c	0.33	$\mathbb{E}_B$(aaac)=1.32
2	b b c d	0.22	$\mathbb{E}_B$(bbcd)=**2.20**
3	b b b d	0.22	$\mathbb{E}_B$(bbbd)=1.98
4	b b c f	0.22	$\mathbb{E}_B$(bbcf)=1.98
5	- - - -	0.00	

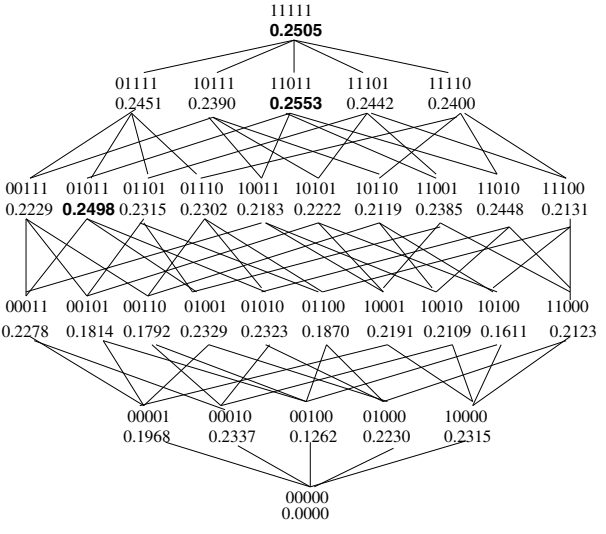

Fig. 1. Figure shows the lattice of five MT output sequences encoded as binary sequences ('11111', '01111', etc) and BLEU scores ('0.2505', '0.2451', etc) for ES-EN JRC-Acquis (Refer Table 2). The top row shows the results using five MT output sequences; the second row uses four MT output sequences; ...; the fourth row uses the individual BLEU scores; the bottom row does not use any MT output sequence (Hence, BLEU score is zero). The observation from this lattice is that the resulting BLEU score is not always between two BLEU scores of adjacent nodes; sometimes the resulting BLEU score is lower than both of them (e.g. '00010' and '10000' resulted in 0.2109.) and it is higher than both of them (e.g. '00011', '01001' and '01010' resulted in 0.2498). The maximal value in the lattice is 0.2553 in the second row in this case.

system combination). The focus is on E of $P(E|F)$ in Eq (2) where E is a set of MT outputs participating in the system combination. That is, if we combine four systems the number of systems, that is $|E|$, is four. A toy example is shown in Table 1. In this example, five MT output sequences "aaac","bbcd","bbbd", "bbcf", and "ffbd" are given. Suppose that we calculate the 1-gram expectation of "bbcd", which constitute the negative quantity in Bayes risk. If we use all the given MT outputs consisting of 5 systems, the expected matches sum to 2.1. If we discard the system producing "ffbd" and only use 4 systems, the 1-gram expectation improves to 2.20. As a conclusion, it is not always the best solution to use the full set of given MT outputs, but removing some bad MT output can be a good strategy. This suggests to consider all possible subsets of the full set of MT outputs, as is shown in (7):

$$\hat{E} = \arg \min_{\mathcal{E}_i \subseteq \mathcal{E}} \sum_{E' \in \mathcal{E}_i} L(E, E') P(E|F) \tag{6}$$

$$= \arg \min_{E' \in \mathcal{E}_{H_i}, \mathcal{E}_{H_i} \subseteq \mathcal{E}} \sum_{E' \in \mathcal{E}_{E_i}} L(E, E') P(E|F) \tag{7}$$

where $\mathcal{E}_{H_i} \subseteq \mathcal{E}$ indicates that we choose $\mathcal{E}_{H_i}$ from all the possible subsets of $\mathcal{E}$ (or a power set of $\mathcal{E}$), $\mathcal{E}_{H_i}$ denotes a i-th hypothesis space, and $\mathcal{E}_{E_i}$ denotes a i-th evidence space corresponding to $\mathcal{E}_{E_i}$.[2]

Now we explain how to formulate an algorithm. As is explained in the latter half of Section 2, a confusion network-based system combination approach takes three steps[3] as follows.

1. Choosing a backbone by a MBR decoder from MT outputs S.
2. Measure the distance between the backbone and each output.
3. Run the decoding algorithm to choose the best path in the confusion network.

Let $|S| = n$. If we consider all the combinations of $|S|$, the simplest algorithm which enumerates all the possibilities requires to repeat these three steps $2^n - 1$ times. However, if we observe this computation we can immediately recognize that there are a considerable number of redundant operations. Hence, our approach is to reduce such redundant operations. First we observe what is changed in these three steps by considering a combinatorial exploration.

– Due to the combinatorial exploration of MT outputs of $|S|$ cases, all the MT outputs can be selected as a backbone for some combination of S in theory. However, if we exclude the combination of using only one or two MT outputs, two cases remain important which have high chances to result in the backbone in most of the cases: the output with the highest BLEU score and that the MBR decoding selects the MT output with highest density (when many MT outputs include the segment).

[2] A power set of $\mathcal{E} = \{1, 2\}$ is $\{\{1, 2\}, \{1\}, \{2\}, \emptyset\}$.

[3] In Section 2, we described the final step. However, this step is just to remove deletion marks and is omitted here.

– Under the combinatorial exploration strategy, what we need to care about is the unnormalized probabilities in the word posterior probabilities. Note that the word posterior probabilities $P(e_j|F)$ in Eq (5) will not vary even if we take the scheme of combinatorial exploration.
– Other quantities, such as language model, word length penalty, and NULL word length penalty will not be changed.

Following on from the second point above, we transform the parallel trees of several MT outputs into a so-called clique tree [13], as is shown in Figure 2. In this clique tree, each clique tree contains the corresponding word pairs in confusion networks. By this transformation, we can reduce the message cost considerably in the third step of decoding to choose the best path in the confusion network, where a message is to connect a node and neighboring node.

Hence, the primitive version which computes all the combinations one by one, takes $O(|S| \times n|T|)$ execution time in the third step where $|T|$ denotes the number of message passing events which is equivalent to the n times the length of the clique tree. Compared to this, the version which uses a clique tree can reduce this message costs from $n|T|$ to $|T|$, hence the overall cost becomes $O(|S| \times |T|)$. If we apply the max-product algorithm, the computation in the clique, which is $O(|S|)$, may be reduced further.

Message passing is done in the clique one by one propagating from the root to the leaf. Let C_i and C_j be the neighboring clique in a clique tree. The value of the message sent from C_i to C_j does not depend on the specific choice of root clique. This argument applies in both directions (p.355 of [13]). Hence, the message from C_i to another clique C_j, denoted as $\delta_{i \to j}$, can be written as (8):

$$\delta_{i \to j} = \max_{C_i - S_{i,j}} \phi_i \prod_{k \in (Nb_i - \{j\})} \delta_{k \to i} \tag{8}$$

where ϕ_i denotes a factor in clique i, and Nb_i denotes the set of indices of cliques that are neighbors of C_i. This message passing process proceeds up the tree. When the root clique has received all messages, it multiplies them with its own initial potential.

4 Heuristic Algorithm

The second algorithm is intended to provide one of the baselines. Suppose we are given 5 translation outputs (the top node marked with '11111' in Fig. 1) and we traverse from this node to the bottom node in a breadth first manner where we only masure the BLEU score on trajectory nodes. Suppose also that we know in advanced each single BLEU score of each translation output ('00001' to '10000'). The first task is to predict which children of '11111' attains the best BLEU score among its siblings ('01111' to '11110'). We choose the combination ('11011') removing a worst single translation output ('00100') will attain the best BLEU score. Then, we measure and compare the actual BLEU score of the parent node and only this child node. (We do not measure the BLEU score of other siblings). If there is an

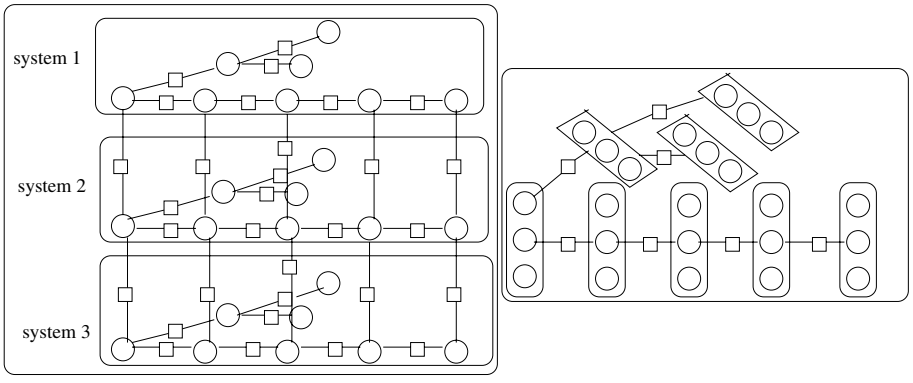

Fig. 2. Figures show a max-product algorithm on multiple systems under two assumptions described in Introduction. In the figure, a circle denote a variable node, a square denote a factor node, and a big rectangle denote a system (in the left figure) and a clique (in the right figure).

Algorithm 1. Heuristic Algorithm

Given: A set of MT devset output $S = \{s_1, \ldots, s_n\}$ and MT testset output $T = \{t_1, \ldots, t_n\}$.
Step 1: Rank devset outputs S according to the performance measure (BLEU, TER, etc) as $S' = \{s'_1, \ldots, s'_n\}$ where $s'_i \prec s'_{i+1}$ (the rank of s'_i is higher than (or the same as) s'_{i+1}).
Step 2: i iteration: Discard the worst system i of S' to make $S'_{(i)}$.
Step 3: Measure the performance of $S'_{(i)}$.
Step 4: If $M(S'_{(i)}) > M(S'_{(i-1)})$ then repeat Step 2.
Step 5: Reply the correspondent MT testset output T with regard to $S'_{(i)}$.

increase, we repeat this process until we reach the bottom node. If we observe decrease, we judge that the parent node attains the best BLEU score. This is shown in Algorithm 1. Although this starts from the full set (of MT systems in a combination) to the empty set (We refer this as Heuristic 1), it is also possible to take the reverse direction which starts from the singleton set to the full set (We refer this as Heuristic 2). There have been no quantitative predictions as far as we are aware.

5 Experiments

We used three different language pairs in our experiments. The first set is ES-EN based on JRC-Acquis [26]; we use the translation outputs of 5 MT systems provided by [7]. The second set is JP-EN provided by NTCIR-8 [8] where translation outputs are prepared by ourselves [20]. The third set is EN-FR

Table 2. Experiment between ES and EN for JRC-Acquis dataset. All the scores are on testset except those marked * (which are on devset). On comparison, we did sampling of three combinations of the single systems, which shows that our results are equivalent to the combination 2. These experimental results validate our motivating results: it is often the case that some radically bad translation output may harm the final output by system combination. In this case, system t3 whose BLEU score is 12.62 has a negative effect on the results of system combination. The best performance was achieved by removing this system, i.e. the combination of systems t1, t2, t4, and t5. The baseline obtained the best score at '01000', the heuristic algorithm obtained at '11011', and our algorithm obtained at '11011'.

	NIST	BLEU	METEOR	WER	PER
system t1 ('10000')	6.3934	0.1968/0.1289*	0.5022487	62.3685	47.3074
system t2 ('01000')	6.3818	0.2337/0.1498*	**0.5732194**	64.7816	49.2348
system t3 ('00100')	4.5648	0.1262/0.0837*	0.4073446	77.6184	63.0546
system t4 ('00010')	6.2136	0.2230/0.1343*	0.5544878	64.9050	50.2139
system t5 ('00001')	6.7082	0.2315/0.1453*	0.5412563	60.6646	45.1949
baseline	6.3818	0.2337	0.5732194	64.7816	49.2348
heuristic 1	6.8419	0.2553	0.5683086	59.9591	44.5357
heuristic 2	6.3818	0.2337	0.5732194	64.7816	49.2348
our algorithm ('11011')	**6.8419**	**0.2553**	0.5683086	**59.9591**	**44.5357**

provided by WMT09 [3]. We use MERT [17] internally to tune the weights and language modeling by SRILM [27].

Tables 2, 3, and 4 include first the BLEU score of individual systems, and then show four results: baseline, heuristic 1 and 2 (Refer Section 4), and our algorithm (Refer Section 3). The baseline is the BLEU score of the best single system.

Table 2 shows our results from ES to EN. The improvement in BLEU was 2.16 points absolute and 9.2% relative compared to the performance of system t2, the single best performing system (we optimized according to BLEU). Except for METEOR, we achieved the best performance in NIST (0.14 points absolute and 2.1% relative), WER (0.71 points absolute and 1.1% relative) and PER (0.64 points absolute and 1.3% relative) as well. However, in this case, Heuristic 1 also achieved the same result. The heuristic algorithm 1 was processed from the point '11011' (BLEU 0.2553) to '11001' (0.2385). The result of heuristic algorithm 1 was 0.2553.

The left half of Table 3 shows our results from JP to EN. The improvement in BLEU was 0.94 points absolute and 3.4% relative compared to the single best performing system. Heuristic 2 and baseline shows the result of system t2. The baseline obtained the result at '01000000000', the heuristic algorithm 1 at '11001111101', the heuristic algorithm 2 at '01000000000', and our algorithm at '11100010101'. The heuristic algorithm 1 was processed from the point '11011111111' (BLEU 0.2202) to '11011111101' (0.2750), '11001111101' (0.2750), and '11001110101' (0.2345). The result of heuristic algorithm 1 was 0.2750. The right half of Table 3 shows the results from EN to FR. The improvement in

Table 3. (Left half) Experiment between JP and EN for NTCIR dataset. The baseline obtained the result at '01000000000', heuristic algorithm 1 was at '11001111101', heuristic algorithm 2 was at '01000000000', and our algorithm obtained at '11100010101'. In this combination, system t3 of BLEU score 0.1243 is included which can be explained that . (Right half) Experiment between EN and FR for WMT 2009 devset. The baseline and the heuristic 2 were at '0000000100000000' and the heuristic 1 was at '0100101110011011'.

JP-EN	NIST	BLEU	METEOR	EN-FR	NIST	BLEU	METEOR
system t1	7.0374	0.2532	0.6083487	system t1	5.6683	0.1652	0.5134530
system t2	7.2992	0.2775	0.6223682	system t2	6.3356	0.2235	0.5765081
system t3	5.1474	0.1243	0.4527874	system t3	5.2992	0.1402	0.4622777
system t4	6.6323	0.1913	0.5590906	system t4	6.0325	0.1945	0.5499950
system t5	6.6682	0.2165	0.5827379	system t5	6.3880	0.2217	0.5579302
system t6	6.8597	0.2428	0.5909936	system t6	5.6773	0.1664	0.5152482
system t7	7.2555	0.2755	0.6193990	system t7	6.2267	0.2170	0.5575926
system t8	6.1250	0.1946	0.6090198	system t8	6.4064	0.2262	0.5614477
system t9	7.2182	0.2529	0.6062563	system t9	6.2788	0.2148	0.5525901
system t10	5.6288	0.1727	0.5141809	system t10	6.0535	0.2034	0.5516885
system t11	7.2625	0.2529	0.6105696	system t11	5.5635	0.1624	0.5137018
				system t12	6.3131	0.2201	0.5574140
				system t13	6.1832	0.2112	0.5514069
				system t14	6.1462	0.2055	0.5582915
				system t15	6.2394	0.2059	0.5303054
				system t16	6.2529	0.2161	0.5567934
baseline	7.2992	0.2775	0.6223682	baseline	6.4064	0.2262	0.5614477
heuristic 1	7.4292	0.2750	0.6228906	heuristic 1	5.5584	0.1799	0.5820681
heuristic 2	7.2992	0.2775	0.6223682	heuristic 2	6.4064	0.2262	0.5614477
our algorithm	**7.5161**	**0.2869**	**0.6305818**	our algorithm	**6.5033**	**0.2292**	**0.5792332**

BLEU was 0.30 points absolute and 1.3% relative compared to the single best performing system. Heuristic 2 and baseline shows the result of system t8. Note that the number of items in the power set (corresponding to the set of all possible sets of MT systems participating in the combination) in ES-EN was 31, JP-EN was 4094, and EN-FR was 65534.

6 Conclusion and Further Studies

This paper investigates the enlarged hypothesis space in MBR decoding in SMT, employing MAP inference on clique tree. This mechanism can substitute the calibration of probabilities with the mechanism of max-product algorithm. First of all, MBR decoding has not been much investigated compared to MAP decoding in SMT, but is rather regarded as a practical tool which achieves state-of-the-art performance for evaluation campaigns. Traditionally, the full set of MT outputs or only to some MT outputs as selected by human beings are employed for MBR decoding. There has been no paper yet to describe the optimization process of this as far as we know (Hence, the search space for the best combination shown

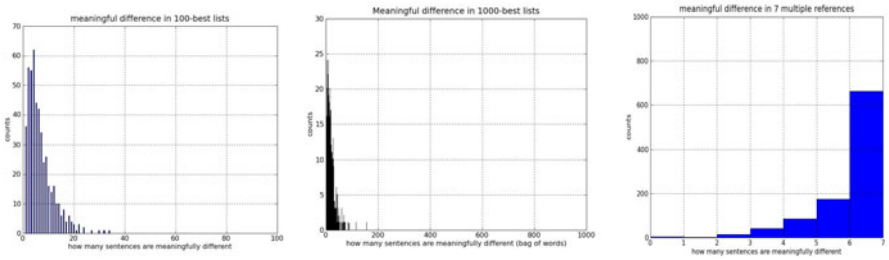

Fig. 3. The left figure shows the count of exact matches among the translation outputs of Moses as a 100-best list after stop-word removal and sorting; We project each sentence in a 100-best list onto a vector space model and count the number of points. The middle figure shows the same quantity for a 1000-best list. The right figure shows the same quantity for a 7-multiple reference (human translation). We use the parallel data of IWSLT 07 JP-EN where we use devset5 (500 sentence pairs) as a development set and devset4 (489 sentence pairs) as a test set; 7-multiple references consist of devset4 and devset5 (989 sentence pairs). For example, the left figure shows that 7% of sentences produce only one really useful translation in a 100-best list and the other 99 sentences in the 100-best list are just reordered versions. In contrast, the right figure of human translation shows that more than 70% of sentences in 7 multiple references are meaningfully different.

in Figure 2 is rarely seen.) Secondly, our algorithm can be successfully applied to the case where the number of participating systems is more than 10, which is the case for the second and the third experiments. Between ES-EN, the improvement was 2.16 BLEU points absolute and 9.2% relative compared to the best single system. Between JP-EN, the improvement was 0.94 points absolute and 3.4% relative. Between FR-EN, the improvement was 0.30 points absolute and 1.3% relative.

There are several avenues for further study. Firstly, to date our experiments involved at most 16 systems. We would like to enlarge the size of the input such as the 1000-best list as in Tromble et al. [28] and DeNero et al. [4], and a general MT translation setting as in Arun et al. [1]. Their improvements are in general quite small compared to the confusion network-based approach. As is shown in Figure 3, the 100-best list and the 1000-best list produced by Moses [11] tend not to be sufficiently different and do not produce useful translation alternatives. As a result, their BLEU score tends to be low compared to the (nearly best) single systems. This means that in our strategy those MT inputs may be better removed rather than employed as a useful source in system combination.

Yet another avenue for further study is to provide prior knowledge into the system combination module. In [19,18,21], we showed that word alignment may include successfuly prior knowledge about alignment links. It would be interesting to incorporate some prior knowledge about system combination, for example, (in)correct words or phrases in some particular translation output.

Acknowledgments. We thank Jinhua Du. This research is supported by the the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the T4ME project (Grant agreement No. 249119).

References

1. Arun, A., Haddow, B., Koehn, P.: A unified approach to minimum risk training and decoding. In: Proceedings of Fifth Workshop on Statistical Machine Translation and MetricsMATR, pp. 365–374 (2010)
2. Bangalore, S., Bordel, G., Riccardi, G.: Computing consensus translation from multiple machine translation systems. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 350–354 (2001)
3. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 workshop on statistical machine translation. In: Proceedings of EACL Workshop on Statistical Machine Translation 2009, pp. 1–28 (2009)
4. DeNero, J., Chiang, D., Knight, K.: Fast consensus decoding over translation forests. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 567–575 (2009)
5. DeNero, J., Kumar, S., Chelba, C., Och, F.: Model combination for machine translation. In: Proceedings of NAACL, pp. 975–983 (2010)
6. Du, J., He, Y., Penkale, S., Way, A.: MaTrEx: the DCU MT System for WMT 2009. In: Proceedings of the Third EACL Workshop on Statistical Machine Translation, pp. 95–99 (2009)
7. Federmann, C.: Ml4hmt workshop challenge at mt summit xiii. In: Proceedings of ML4HMT Workshop, pp. 110–117 (2011)
8. Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H., Shimohata, S.: Overview of the patent translation task at the NTCIR-8 workshop. In: Proceedings of the 8th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access, pp. 293–302 (2010)
9. Goel, V., Byrne, W.: Task dependent loss functions in speech recognition: A-star search over recognition lattices. In: Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH), pp. 51–80 (1999)
10. Koehn, P.: Statistical machine translation. Cambridge University Press (2010)
11. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180 (2007)
12. Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computationa Linguistics (HLT / NAACL 2003), pp. 115–124 (2003)
13. Koller, D., Friedman, N.: Probabilistic graphical models: Principles and techniques. MIT Press (2009)

14. Kumar, S., Byrne, W.: Minimum Bayes-Risk word alignment of bilingual texts. In: Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 140–147 (2002)
15. Leusch, G., Matusov, E., Ney, H.: The rwth system combination system for wmt 2009. In: Fourth EACL Workshop on Statistical Machine Translation (WMT 2009), pp. 56–60 (2009)
16. Matusov, E., Ueffing, N., Ney, H.: Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In: Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 33–40 (2006)
17. Och, F.: Minimum Error Rate Training in Statistical Machine Translation. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 160–167 (2003)
18. Okita, T.: Word alignment and smoothing method in statistical machine translation: Noise, prior knowledge and overfitting. PhD thesis. Dublin City University, pp. 1–130 (2011)
19. Okita, T., Guerra, A.M., Graham, Y., Way, A.: Multi-Word Expression sensitive word alignment. In: Proceedings of the Fourth International Workshop On Cross Lingual Information Access (CLIA 2010, collocated with COLING 2010), Beijing, China, pp. 1–8 (2010)
20. Okita, T., Jiang, J., Haque, R., Al-Maghout, H., Du, J., Naskar, S.K., Way, A.: MaTrEx: the DCU MT System for NTCIR-8. In: Proceedings of the MII Test Collection for IR Systems-8 Meeting (NTCIR-8), Tokyo, pp. 377–383 (2010)
21. Okita, T., Way, A.: Given bilingual terminology in statistical machine translation: Mwe-sensitve word alignment and hierarchical pitman-yor process-based translation model smoothing. In: Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24), pp. 269–274 (2011)
22. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A Method For Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 311–318 (2002)
23. Pearl, J.: Reverend bayes on inference engines: A distributed hierarchical approach. In: Proceedings of the Second National Conference on Artificial Intelligence (AAAI 1982), pp. 133–136 (1983)
24. Shenoy, P., Shafer, G.: Axioms for probability and belief-function propagation. In: Proceedings of the 6th Conference of Uncertainty in Artificial Intelligence (UAI), pp. 169–198 (1990)
25. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proceedings of Association for Machine Translation in the Americas, pp. 223–231 (2006)
26. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., Varga, D.: The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 2142–2147 (2006)
27. Stolcke, A.: SRILM – An extensible language modeling toolkit. In: Proceedings of the International Conference on Spoken Language Processing, pp. 901–904 (2002)
28. Tromble, R., Kumar, S., Och, F., Macherey, W.: Lattice minimum bayes-risk decoding for statistical machine translation. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 620–629 (2008)
29. Viterbi, A.J.: Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13, 260–269 (1967)