Pattern Analysis and Applications manuscript No.
(will be inserted by the editor)

Minimum Bayes' Risk Subsequence Combination for Machine Translation

Jesús González-Rubio · Francisco Casacuberta

Received: date / Accepted: date

Abstract System combination has proved to be a successful technique in the pattern recognition field. However, several difficulties arise when combining the outputs of tasks, e.g. machine translation, that generate structured patterns. So far, machine translation system combination approaches either implement sophisticated classifiers to select one of the provided translations, or generate new sentences by combining the "best" subsequences of the provided translations. We present minimum Bayes' risk system combination (MBRSC), a system combination method for machine translation that gathers together the advantages of sentence-selection and subsequence-combination methods. MBRSC is able to detect and utilize the "best" subsequences of the provided translations to generate the optimal consensus translation with respect to a particular performance metric. Experiments show that MBRSC yields significant improvements in translation quality.

Keywords minimum Bayes' risk · system combination · statistical machine translation

1 Introduction

Machine translation (MT) is a fundamental technology that is emerging as a core component of language processing systems. However, after a major development boost in the early nineties, MT technology seems to have reached a technical plateau nowadays [31, 6]. The combination of multiple MT systems is a promising

Jesús González-Rubio (✉) Francisco Casacuberta
D. Sistemas Informáticos y Computación
Universitat Politècnica de València
C/ de Vera s/n
46021, Valencia, Spain
Tel.: +34-96-3877069
Fax: +34-96-3877239
E-mail: jegonzalez@dsic.upv.es

Francisco Casacuberta
E-mail: fcn@dsic.upv.es

research direction to overcome this stagnation. The key idea of system combination [12] is that it is often very difficult to find the real best system for the task at hand, while different systems (for instance, trained on different data or using different learning paradigms) can exhibit complementary strengths and limitations. Therefore, a proper combination of various systems could be more effective than using a single monolithic system.

The combination of outputs from multiple systems have been found to improve performance in a number of classification task such as part-of-speech tagging [39], text categorization [27] and speech recognition [16]. However, unlike part-of-speech tagging or text categorization where the classes are atomic units (either a part-of-speech or a category), classes in a translation task are sequences (sentences of words). When combining MT systems, we can consider either the full sentence or the individual words as the atomic classes, which leads to two different MT system combination approaches.

MT system combination methods that consider the full sentences as the classification classes implement the so-called sentence-selection approach. The decision on the consensus translation is taken as a selection of one of the translation provided by the individual MT systems [5, 32, 36, 11, 13]. Their main limitation is that they cannot generate new translations that include "good" subsequences from different individual sentences. In exchange, they can implement sophisticated classifiers such as minimum Bayes' risk classifiers [14], which constitutes their main virtue.

In contrast, MT system combination methods that consider the individual words as the classification classes implement the so-called subsequence-combination approach. These methods detect which subsequences of words in the individual translations are "correct", and combine them to generate a consensus translation with reduced error [16]. Unfortunately, the translations provided by the individual systems can be of different length or have a different word order. Therefore, a synchronization (alignment) step is required to detect which is the correspondence between the subsequences of the different translations. The consensus translation is given by the highest scoring path throughout the graph, the so-called confusion network, defined by the computed alignment [1, 21, 37, 29, 19]. These methods have one obvious advantage over sentence-selection: they can generate new consensus translations that potentially contain the "best" subsequences of the individual translations. However, they have to deal with the challenging word alignment problem that has a substantial effect on combination performance [19]. Moreover, these methods also require additional data to train complex search models that score the paths throughout the consensus network, which hinders their application to languages with scarce resources.

We present minimum Bayes' risk system combination (MBRSC), a method designed to gather together the advantages of sentence-selection and subsequence-combination methods. MBRSC can detect the "best" subsequences of the provided translations, and combine them into a new consensus translation which is optimal with respect to a particular performance measure. We choose the BLEU score [35] as our performance measure of interest. BLEU considers a sentence as a vector of n-gram[1] occurrences rather than a word sequence. Therefore, BLEU can compare sentences without the need of a word alignment between them. Additionally, BLEU

[1] We will refer as n-gram to a sequence of n consecutive words in a sentence.

is the standard performance measure for MT, thus, by using loss function based on BLEU, we are optimizing our system towards the most widespread translation quality measure.

Compared with sentence-selection methods, MBSRC also implements a sophisticated classifier, and, additionally, it is able to generate new consensus translations that include the "best" subsequences from different individual translations. Regarding subsequence-combination methods, MBRSC has several advantages over the dominant confusion network approach:

- Translations do not have to be synchronized which avoids the limitations imposed by the alignment.
- The full target language is explored in the search for the consensus translation.
- A minimum Bayes' risk classifier is implemented. Thus, the consensus translations are optimal with respect to the final evaluation measure.
- The consensus translation is computed directly from the translations of the individual systems. I.e., no additional data is required to train graph-search models which allows the effective application of MBRSC to languages with scarce resources.

The basic concept of MBRSC has been previously described in a conference publication [17]. Since then, the process to obtain the consensus translation have been substantially improved. We describe a novel dynamic programming beam search algorithm [22] that efficiently explores the full output language, outperforming the previously used gradient ascent algorithm.

The rest of the article is organized as follows. Section 2 reviews the basics of Bayesian decision theory and introduces minimum Bayes' risk classifiers. Section 3 presents our system combination algorithm, MBRSC, in detail. Experimental results are presented in section 4. Finally, we conclude with a summary in section 5.

2 Minimum Bayes' risk classifiers

Let $\mathbf{x} \in \mathcal{X}$ be a domain of objects, and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}$ a set of classes. A classification system is defined by a classification function $(C : \mathcal{X} \rightarrow \mathcal{Y})$ that maps each object to one class [14]. Given a loss function $L(C(\mathbf{x}), \mathbf{y}')$ that measures the error of classifying object $\mathbf{x}$ into class $C(\mathbf{x})$ knowing that the correct class is $\mathbf{y}'$, the performance of a classification function is measured through the Bayes' risk[2]:

$$R(C(\mathbf{x})) = \mathbb{E}_{Pr(\mathbf{y}\,|\,\mathbf{x})}[L(C(\mathbf{x}), \mathbf{y}')] \tag{1}$$

The optimal classification function $\widehat{C}(\cdot)$ minimizes the Bayes' risk for each object [4], the so-called minimum Bayes' risk (MBR) classifier:

$$\hat{\mathbf{y}} = \widehat{C}(\mathbf{x}) = \underset{\mathbf{y}\in\mathcal{Y}}{\arg\min} \sum_{\mathbf{y}'\in\mathcal{Y}} Pr(\mathbf{y}' \,|\, \mathbf{x}) \cdot L(\mathbf{y}, \mathbf{y}') \tag{2}$$

MBR classifiers usually are computationally costly, particularly, when applied to tasks (e.g. MT) where the number of classes $|\,\mathcal{Y}\,|$ is very large or even infinite.

[2] $Pr(\cdot)$ denotes general probability distributions, $P(\cdot)$ denotes model-based distributions, and $\mathbb{E}_{Pr(X)}[X]$ denotes the expected value of a random variable X under distribution $Pr(X)$.

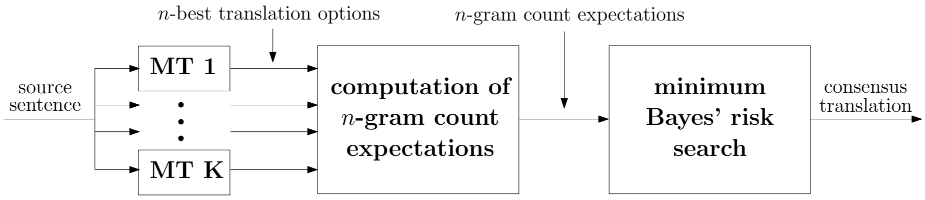

Fig. 1 Overview of the process followed by the proposed MBRSC method to generate a consensus translation.

However, this complexity can be greatly reduced if we consider linear loss functions of the form $L(\mathbf{y}, \mathbf{y}') = \sum_d \theta_d(\mathbf{y}) \cdot \phi_d(\mathbf{y}')$, where $\phi_d(\mathbf{y}')$ is a real-valued feature of the reference class $\mathbf{y}'$, and $\theta_d(\mathbf{y})$ is the count of that feature in the candidate class $\mathbf{y}$. Then, the MBR classifier in Equation (2) can be re-written as:

$$
\begin{aligned}
\hat{\mathbf{y}} &= \operatorname*{arg\,min}_{\mathbf{y} \in \mathcal{Y}} \sum_{\mathbf{y}' \in \mathcal{Y}} Pr(\mathbf{y}' \,|\, \mathbf{x}) \cdot \sum_d \theta_d(\mathbf{y}) \cdot \phi_d(\mathbf{y}') \\
&= \operatorname*{arg\,min}_{\mathbf{y} \in \mathcal{Y}} \sum_d \theta_d(\mathbf{y}) \cdot \sum_{\mathbf{y}' \in \mathcal{Y}} Pr(\mathbf{y}' \,|\, \mathbf{x}) \cdot \phi_d(\mathbf{y}') \\
&= \operatorname*{arg\,min}_{\mathbf{y} \in \mathcal{Y}} \sum_d \theta_d(\mathbf{y}) \cdot \mathbb{E}_{Pr(\mathbf{y}' \,|\, \mathbf{x})}[\phi_d(\mathbf{y}')] \qquad (3)
\end{aligned}
$$

Unfortunately, many loss functions of interest (e.g. BLEU) are nonlinear, and so Equation (3) does not apply. However, this loss functions usually are functions of features of $\mathbf{y}'$. That is, they can be expressed as $\widetilde{L}(\mathbf{y}; \Phi(\mathbf{y}'))$ for a feature mapping $\Phi : \mathcal{Y} \rightarrow \mathbb{R}^n$. Based on this observation, DeNero *et al.* [10] proposed to follow the structure of Equation (3) also for nonlinear functions, choosing a class $\mathbf{y}$ based on the feature expectations of $\mathbf{y}'$:

$$
\hat{\mathbf{y}} \approx \operatorname*{arg\,min}_{\mathbf{y} \in \mathcal{Y}} \widetilde{L}(\mathbf{y}, \mathbb{E}_{Pr(\mathbf{y}' \,|\, \mathbf{x})}[\Phi(\mathbf{y}')]) \qquad (4)
$$

Note that for nonlinear loss functions, this MBR classifier over features differs from the exact MBR classifier in Equation (2), but MT system combination results reported in [10] showed that there were no significant difference in performance between the two approaches.

The main advantage of the MBR formulation over features in Equation (4) is that the computation of the Bayes' risk is independent of the number of classes which largely simplifies its implementation. The main computational challenge that remains is the well-studied search problem ($\arg\min_{\mathbf{y} \in \mathcal{Y}}$ operation). The exact formulation of the search problem depends of the particular loss function under consideration, but it can be solved through several general purpose techniques such as dynamic programming [3] or branch-and-bound [26], that additionally can implement beam search [22] to improve their efficiency.

3 Minimum Bayes' risk system combination

We now present the details of the proposed method: minimum Bayes' risk system combination (MBRSC). Section 3.1 presents the probabilistic translation model of MBRSC and its MBR formulation for BLEU. Section 3.2 describes the process to train the free parameters of the model. Finally, section 3.3 describes the search algorithm that generates the consensus translation. Figure 1 gives an overview of the process followed by MBRSC to generate a consensus translation.

3.1 MBRSC model

Let $\{C_1, \ldots, C_k, \ldots, C_K\}$ denote K individual MT systems. Under the assumption that the systems are statistically independent, we model the multi-system classifier as a weighted ensemble of systems [23]:

$$P(\mathbf{y} \mid \mathbf{x}) = \sum_{k=1}^{K} \alpha_k \cdot P_k(\mathbf{y} \mid \mathbf{x}) \tag{5}$$

where $P_k(\mathbf{y} \mid \mathbf{x})$ denotes the probability distribution over translations modelled by system C_k. Free parameters $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K\}$ are scaling factors that can be interpreted as a measure of the importance of each individual system ($\sum_{k=1}^{K} \alpha_k = 1$). The optimal classification function for the ensemble model in Equation (5) is an instance of the MBR classifier in Equation (2):

$$\hat{\mathbf{y}} = \arg\min_{\mathbf{y} \in \mathcal{Y}} \sum_{\mathbf{y}' \in \mathcal{Y}} \left(\sum_{k=1}^{K} \alpha_k \cdot P_k(\mathbf{y}' \mid \mathbf{x}) \right) \cdot L(\mathbf{y}, \mathbf{y}') \tag{6}$$

We choose the widespread BLEU [35] score as loss function. BLEU computes the geometric mean of the precision of n-grams of various lengths between a candidate and a reference translation. This geometric average is multiplied by a factor that penalizes translations shorter than the reference, namely the brevity penalty. Following the standard BLEU implementation, we consider $n = 4$ as the maximum n-gram length. Formally, the BLEU score between a candidate $\mathbf{y}$ and a reference translation $\mathbf{y}'$ is given by:

$$\mathrm{BLEU}(\mathbf{y}, \mathbf{y}') = \left(\prod_{n=1}^{4} p_n(\mathbf{y}, \mathbf{y}') \right)^{\frac{1}{4}} \cdot \mathrm{BP}(\mathbf{y}, \mathbf{y}') \tag{7}$$

where the n-gram precisions $p_n(\mathbf{y}, \mathbf{y}')$ and the brevity penalty $\mathrm{BP}(\mathbf{y}, \mathbf{y}')$ are computed as:

$$p_n(\mathbf{y}, \mathbf{y}') = \frac{\displaystyle\sum_{\mathbf{w} \in \mathcal{W}_n(\mathbf{y})} \min(\#_{\mathbf{w}}(\mathbf{y}), \#_{\mathbf{w}}(\mathbf{y}'))}{\displaystyle\sum_{\mathbf{w} \in \mathcal{W}_n(\mathbf{y})} \#_{\mathbf{w}}(\mathbf{y})} \tag{8}$$

$$\mathrm{BP}(\mathbf{y}, \mathbf{y}') = \min\left(\exp\left(1 - \frac{|\,\mathbf{y}'\,|}{|\,\mathbf{y}\,|} \right), 1 \right) \tag{9}$$

where $\mathcal{W}_n(\mathbf{y})$ is the set of n-grams of size n in $\mathbf{y}$, $\#_{\mathbf{w}}(\mathbf{y})$ represents the count of n-gram $\mathbf{w}$ in translation $\mathbf{y}$ and $|\,\mathbf{y}\,|$ denotes its length.

BLEU is a percentage with a value of one denoting an exact match between $\mathbf{y}$ and $\mathbf{y}'$. Thus, the $\arg\min_{\mathbf{y}\in\mathcal{Y}}$ in Equation (6) is substituted by an $\arg\max_{\mathbf{y}\in\mathcal{Y}}$. Finally, the BLEU-based MBR classifier for the ensemble is formulated as:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \sum_{\mathbf{y}'\in\mathcal{Y}} \sum_{k=1}^{K} \alpha_k \cdot P_k(\mathbf{y}' \,|\, \mathbf{x}) \cdot \mathrm{BLEU}(\mathbf{y}, \mathbf{y}') \tag{10}$$

This MBR classifier has a high temporal complexity in $O(|\,\mathcal{Y}\,|^2 \cdot K \cdot I)$, where $|\,\mathcal{Y}\,|$ denotes the number of possible target language sentences, and I represents the maximum sentence length given that $\mathrm{BLEU}(\mathbf{y}, \mathbf{y}')$ has a computational complexity in $O(\max(|\,\mathbf{y}\,|, |\,\mathbf{y}'\,|))$.

Since $\mathrm{BLEU}(\mathbf{y}, \mathbf{y}')$ references $\mathbf{y}'$ only via its n-gram counts[3], we can follow [10] and approximate Equation (10) by choosing a translation $\mathbf{y}$ based on n-gram count expectations:

$$\begin{aligned} \hat{\mathbf{y}} &= \arg\max_{\mathbf{y}\in\mathcal{Y}} \widetilde{\mathrm{BLEU}}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[\Phi(\mathbf{y}')]) \\ &= \arg\max_{\mathbf{y}\in\mathcal{Y}} \left(\prod_{n=1}^{4} \widetilde{p_n}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[\Phi(\mathbf{y}')])\right)^{\frac{1}{4}} \cdot \widetilde{\mathrm{BP}}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[\Phi(\mathbf{y}')]) \end{aligned} \tag{11}$$

where $\mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[\Phi(\mathbf{y}')]$ are the expected n-gram counts according to the probability distribution $P(\mathbf{y}'\,|\,\mathbf{x})$ of the ensemble model in Equation (5). We reformulate $p_n(\mathbf{y}, \mathbf{y}')$ and $\mathrm{BP}(\mathbf{y}, \mathbf{y}')$ as functions of expected n-gram counts:

$$\widetilde{p_n}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[\Phi(\mathbf{y}')]) = \frac{\displaystyle\sum_{\mathbf{w}\in\mathcal{W}_n(\mathbf{y}')} \min(\#_{\mathbf{w}}(\mathbf{y}), \mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[\#_{\mathbf{w}}(\mathbf{y}')])}{\displaystyle\sum_{\mathbf{w}\in\mathcal{W}_n(\mathbf{y}')} \#_{\mathbf{w}}(\mathbf{y})} \tag{12}$$

$$\widetilde{\mathrm{BP}}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[\Phi(\mathbf{y}')]) = \min\left(\exp\left(1 - \frac{\mathbb{E}_{P(\mathbf{y}'\,|\,\mathbf{x})}[|\,\mathbf{y}'\,|]}{|\,\mathbf{y}\,|}\right), 1\right) \tag{13}$$

the n-gram count expectations can be computed in advance, thus Equation (11) has a computational complexity in $O(|\,\mathcal{Y}\,| \cdot I)$.

To compute the expected n-gram counts, all systems should share the same candidate translations. However, due to differences in generative capabilities, training data selection, and various pruning techniques, the domain of translations of the different systems are always not identical in practice. Our approach is to compute the count expectations individually for each system[4] and combine these counts according to the ensemble weights $\boldsymbol{\alpha}$. If a probability distribution over translations is not available, e.g. translations generated by non-statistical MT systems, we can use a uniform distribution or assign a rank-based probability [37] to each translation.

[3] The brevity penalty is also a function of n-gram counts: $|\,\mathbf{y}'\,| = \sum_{\mathbf{w}\in\mathcal{W}_1(\mathbf{y}')} \#_{\mathbf{w}}(\mathbf{y}')$.

[4] This can be done straightforwardly if the domain of translations is represented as a list. For more complex graph-based representations, we can use the algorithms proposed in [25, 10, 11].

$P(\mathbf{y} \mid \mathbf{x})$	$\mathbf{y}$
0.35	we are certainly faced with enormous challenges .
0.25	certainly we must tackle enormous challenges .
0.40	we are faced with enormous challenges .

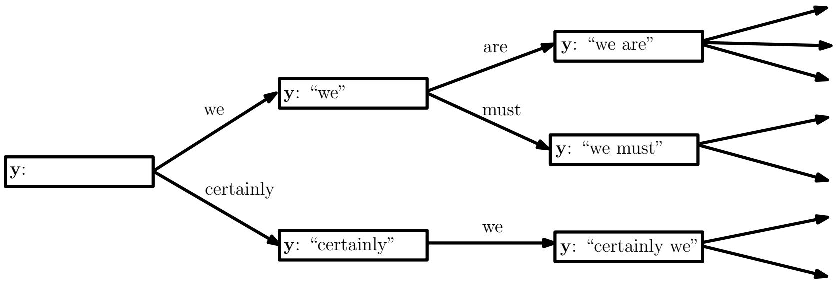

Fig. 2 Example of the search-graph explored by MBRSC when combining three translations.

3.2 MBRSC training

The objective of the training procedure is to obtain suitable values for the free parameters $\boldsymbol{\alpha}$ of the ensemble model. By "suitable", we mean parameter values that yield good translation quality on unseen data, the so-called minimum error rate training (MERT) [34]. Given a function $Q(\mathbf{y}, \mathbf{y}')$ that measures the quality of a translation $\mathbf{y}$ with respect to a reference translation $\mathbf{y}'$, our goal is to obtain the parameter values that maximize the translation quality of the consensus translations generated by MBRSC for a representative training set $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_S, \mathbf{y}_S)\}$:

$$\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \sum_{s=1}^{S} Q(C(\mathbf{x}_s; \boldsymbol{\alpha}), \mathbf{y}_s) \tag{14}$$

where function $C(\mathbf{x}_s; \boldsymbol{\alpha})$ returns the consensus translation for source sentence $\mathbf{x}_s$ given by the MBRSC decision function (Equation (11)) using parameter values $\boldsymbol{\alpha}$. We solve this optimization problem with the downhill-simplex algorithm [30] using BLEU as quality function.

3.3 MBRSC search

We now address the search problem also referred to as generation or decoding. Its goal is to solve Equation (11) which involves to find for a given source sentence the translation of maximum expected BLEU score among all possible target language sentences. The main difficulty in the computation of Equation (11) is the potentially infinite number of target language sentences $\mathbf{y} \in \mathcal{Y}$ that have to be considered as candidate translations during the search process. A similar search problem also arises in conventional MT models which has been demonstrated to be an NP-complete problem [24, 42], so we cannot expect to develop efficient algorithms to perform an exact search.

We formalize the MBRSC search as a dynamic programming problem [3]. Search is then interpreted as a sequence of decisions that incrementally generate new translation hypotheses $\mathbf{y}'$. Starting with an empty hypothesis, each decision expand a hypotheses of size $i-1$ with one new target vocabulary word $y \in \Sigma$ to create a hypothesis of size i. This search space can be represented as a directed acyclic graph where the states denote partial hypotheses and the edges are labelled with expansion words. Figure 2 shows an example of the two first expansions in the search graph when combining three sentences. We avoid repeated computations by traversing the search graph in a topological order, thus performing a breadth-first exploration of the search space. In other words, before we process a node, i.e. expand a hypothesis, we have to make sure that we have visited all predecessor states. We can easily guarantee the topological order by processing the nodes according to the size of the partial hypotheses.

Each possible expansion of a partial hypothesis will be assigned a score representing its expected BLEU score. Among all possible paths of the search graph, we are interested in that of the highest score. As have been explained above, a state of the graph represents a partial hypothesis, however only the n-grams counts of the partial hypothesis are required to compute its score. Two partial hypotheses sharing the same n-grams are indistinguishable, and we are only interested in the hypothesis of higher score. According to these considerations, each state of the graph can be represented by a specific bag (namely a specific multiset) $\mathcal{N}$ of n-grams. We define $Q(\mathcal{N}) = \{q, \mathbf{y}\}$, where q is the maximum score of a path leading from the initial state to the state $(\mathcal{N})$, and $\mathbf{y}$ is the highest-scoring hypothesis in the state. The usage of $\mathcal{N}$ and $\mathbf{y}$ may seem redundant, however, while $\mathcal{N}$ allows to distinguish between hypotheses, the actual ordered sequence of words $\mathbf{y}$ is required to generate the subsequent expanded hypothesis. We also define $\widehat{Q} = \{\hat{q}, \hat{\mathbf{y}}\}$ as the final state of the optimal translation $\hat{\mathbf{y}}$. Finally, we obtain the following dynamic programming recursion equations:

$$
\begin{aligned}
Q(\emptyset) &= \{0, ""\} \\
Q(\mathcal{N}_e) &= \left\{ \max_{\substack{y \in \Sigma, \{\cdot, \mathbf{y}_p\} = Q(\mathcal{N}_p), \\ \mathbf{y}_e = \mathbf{y}_p\, y, \mathcal{N}_e = \mathcal{N}_p \cup \Theta(\mathbf{y}_p, y)}} \widetilde{\mathrm{BLEU}}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\Phi(\mathbf{y}')]),\ \mathbf{y}_e \right\} \\
\widehat{Q} &= \left\{ \max_{\substack{\{\cdot, \mathbf{y}_p\} = Q(\mathcal{N}_p) \\ \hat{\mathbf{y}} = \mathbf{y}_p\, \$}} \widetilde{\mathrm{BLEU}}(\hat{\mathbf{y}}, \mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\Phi(\mathbf{y}')]),\ \hat{\mathbf{y}} \right\}
\end{aligned}
$$

where \$ is the end-of-sentence symbol that denotes a complete translation, and $\Theta(\mathbf{y}, y)$ returns the new n-grams generated when expanding hypothesis $\mathbf{y}$ with word y. For example, given the hypothesis $\mathbf{y}_p$= "we are faced with" and the expansion word y= "enormous", the expanded hypothesis $\mathbf{y}_e$= "we are faced with enormous" contains four[5] n-grams more than $\mathbf{y}_p$: "enormous", "with enormous", "faced with enormous", and "are faced with enormous".

As defined in the dynamic programming equations, every target language word is a potential expansion option for each partial translation. However, not all word sequences form correct natural language sentences. E.g., given the partial translation $\mathbf{y}_p$= "we are faced with", it is clear that word y= "enormous" can be a valid

[5] Following the definition of the BLEU score (see previous section), we take into consideration n-grams up to size four.

input : $\mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\#_{\mathbf{w}}(\mathbf{y}')]$ (n-gram count expectations)
N (pruning parameter),
I (maximum translation length)
output : $\widehat{Q}$ (optimal translation along with its score)
auxiliary: $\Theta(\mathbf{y}, y)$ (new n-grams after expanding hypothesis $\mathbf{y}$ with word y),
$\Delta(\mathbf{y})$ (set of expansion words for $\mathbf{y}$),
$\overline{S}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\#_{\mathbf{w}}(\mathbf{y}')])$ (returns the complete score of $\mathbf{y}$),
$\Pi(i, N)$ (prunes out low-scoring hypotheses of size i)

begin
  $Q(\cdot) \leftarrow \{0, ""\}$; $\widehat{Q} \leftarrow \{0, ""\}$;
  for $i = 1$ **to** I **do**
    forall $\mathcal{N}_p : Q(\mathcal{N}_p) = \{\cdot, \mathbf{y}_p\} \wedge |\,\mathbf{y}_p\,| == i - 1$ **do**
      $\{\cdot, \mathbf{y}_p\} \leftarrow Q(\mathcal{N}_p)$;
      forall $y \in \Delta(\mathbf{y}_p)$ **do**
        $\mathbf{y}_e \leftarrow \mathbf{y}_p\, y$;
        $q_e \leftarrow \overline{S}(\mathbf{y}_e, \mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\#_{\mathbf{w}}(\mathbf{y}')])$;
        if $y == \$$ **then**
          $\{\hat{q}, \cdot\} \leftarrow \widehat{Q}$;
          if $q_e > \hat{q}$ **then**
            $\widehat{Q} \leftarrow \{q_e, \mathbf{y}_e\}$;
        else
          $\mathcal{N}_e \leftarrow \mathcal{N}_p \bigcup \Theta(\mathbf{y}_p, y)$;
          $\{q, \cdot\} \leftarrow Q(\mathcal{N}_e)$;
          if $q_e > q$ **then**
            $Q(\mathcal{N}_e) \leftarrow \{q_e, \mathbf{y}_e\}$;
    $\Pi(i, N)$;
end

Algorithm 1: Pseudocode of the dynamic programming beam search algorithm with pruning.

expansion option while word y="with" cannot. Thus, we consider $y \in \Sigma \cup \{\$\}$ as a valid expansion word for partial hypothesis $\mathbf{y}_p$ only if at least one of the new n-grams in the resulting expanded hypothesis $\mathbf{y}_e = \mathbf{y}_p\, y$ has a expected n-gram count above zero. Formally, the set of expansion words $\Delta(\mathbf{y}_p)$ for a partial hypothesis $\mathbf{y}_p$ is given by:

$$\Delta(\mathbf{y}_p) = \left\{ y \mid \exists\, \mathbf{w} \in \Theta(\mathbf{y}_p, y) \wedge \mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\#_{\mathbf{w}}(\mathbf{y}')] > 0 \right\}$$

This dynamic programming search is optimal. Unfortunately, due to the exponential number of states[6], we cannot expect to efficiently obtain the optimal consensus translation. To speed up the search, we use a beam search algorithm [22] with pruning. Specifically, for each size i, we keep only the N best-scoring hypotheses and discard the rest of them. To assure a fair competition between hypotheses, the score of each of them is given by a combination of its score so far, and an estimate of the rest score to complete the translation. Following [18], we apply a light search process (considering at each step the single best expansion) to estimate the

[6] The number is computed by the multiset coefficient [41] and it is exponential in the size of the target vocabulary.

score of the complete translation that can be obtained from the hypothesis. This score is then used as the complete score of the hypothesis.

Algorithm 1 shows the pseudocode of the dynamic programming beam search algorithm with pruning. It takes as input the set of n-gram count expectations ($\mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\#_{\mathbf{w}}(\mathbf{y}')]$), the number of hypotheses to keep after pruning (N), and the maximum translation length under consideration (I). We use some auxiliary functions: $\Theta(\mathbf{y}, y)$ returns the set of new n-grams generated in the expansion of hypothesis $\mathbf{y}$ with word y, $\Delta(\mathbf{y})$ returns the set of valid expansion words for $\mathbf{y}$, $\overline{S}(\mathbf{y}, \mathbb{E}_{P(\mathbf{y}' \mid \mathbf{x})}[\#_{\mathbf{w}}(\mathbf{y}')])$ returns the complete score (current score plus rest score estimation) of $\mathbf{y}$, and $\Pi(i, N)$ is a function that prunes out search states that represent partial hypotheses of size i keeping only the N best-scoring ones for future expansions.

The first loop in Algorithm 1 assures that the search graph is traversed in topological order. Additionally, it introduces an upper bound to the maximum translation size under consideration, and thus, to the number of iterations of the algorithm. At each iteration, line 4 loops over the states that remain from the previous iteration, i.e., non-pruned states that store a translation of size $i-1$. For each of these predecessor states, line 6 loops over the corresponding expansion words. Given a predecessor state ($\mathcal{N}_p$) that stores a hypothesis $\mathbf{y}_p$, and a valid expansion word y, we compute the complete score (current score plus rest score estimation) q_e of the expanded hypothesis $\mathbf{y}_e = \mathbf{y}_p\, y$ (line 8). Then, if the expansion word is the end-of-sentence symbol ($y == \$$), the expanded hypothesis is a complete translation, and if it improves the score ($\hat{q}$) of the best consensus translation so far, we update this optimal translation (lines 9–12). For any other expansion words, we first compute the bag of n-grams $\mathcal{N}_e$ of the expanded hypothesis (line 14). Then, if the score q_e of the expanded hypothesis improves the score stored in the corresponding successor state ($\mathcal{N}_e$) (line 16), we update the state. Finally, we prune out states that represent low-scoring hypotheses of current size i (line 18).

This beam search algorithm with pruning has a computational complexity in $O(I^2 \cdot N \cdot D)$, where N denotes the pruning parameter that controls the number of predecessor states in line 4, D denotes the maximum number of expansion words in line 6, and I is the maximum translation size in line 3. The extra $O(I)$ factor is given by the score computation[7] in line 8. Note that the computational complexity of Algorithm 1 does not depend on the number of translations provided by the individual systems.

4 Experiments

We now describe the experiments performed to study the soundness of the proposed system combination method. First, we describe the evaluation criteria used in the experimentation. Then, we present results for several comparative experiments between different setups of MBRSC. Finally, we compare MBRSC with several other state-of-the-art system combination algorithms.

[7] The BLEU-based score cannot be computed incrementally due to the $\min(\cdot)$ functions in its formulation.

Table 1 Average number of translation options provided, and case insensitive BLEU scores for the single best translation of each system.

System	**#trans_opts**	**BLEU [%]**
A	13	24.8
B	9	25.2
C	41	25.8
D	263	25.8
E	126	26.4

4.1 Evaluation criteria

4.1.1 Translation quality measures

We used two well-established automatic measures to evaluate the quality of the consensus translations: BLEU [35], and TER [40]. BLEU measures the geometric average of the n-gram precisions multiplied by a factor that penalize short translations, see Equation (7). TER measures the percentage of words that must be edited to convert the candidate translation into the reference translation; valid edit operations are: deletion, insertion and substitution of single words and shift of word sequences. Each measure assumes a different definition of "translation quality". BLEU is a percentage that measures to which extent the candidate translation contains the same information as the reference translation. A 100% BLEU value denotes a candidate translation equal to the reference. In contrast, TER aims at measuring the amount of work needed to fix a candidate translation. Thus, TER is an error measure where a 0% denotes a perfect matching between the candidate translation and the reference. Since MBRSC is designed to optimize BLEU, we expect translation quality improvements in BLEU to be particularly important. We also report TER scores to independently assess BLEU results.

4.1.2 Statistical Significance

We apply statistical significance testing to establish that an observed performance difference between two methods is in fact significant, and has not just arisen by chance. The usual approach is to state as null hypothesis: "Methods A and B do not differ with respect to the evaluation measure of interest". Then, we determine the probability, namely the p-value, that an observed difference has arisen by chance given the null hypothesis. If the p-value is lower than a predefined significance level (usually $p < 0.01$, or $p < 0.05$) we can reject the null hypothesis. To do that, we use randomization tests [33], specifically a randomization version of the paired t-test based on [9]:

1. Collect the absolute difference in evaluation measure $Q(\cdot)$ for methods A and B
   $|Q(A) - Q(B)|$
2. Shuffle N times ($N = 9999$ in our experiments)
3. Count the number of times ($N^{\geq}$) that
   $|Q(A') - Q(B')| \geq |Q(A) - Q(B)|$
4. The estimate of the p-value is $\frac{N^{\geq}+1}{N+1}$
   (1 is added to achieve an unbiased estimate)

Table 2 Influence of individual MBRSC components on the quality of the generated consensus translations.

System	BLEU[%]	TER[%]
worst single system	24.8	60.4
best single system	26.4	56.0
sentence selection baseline [15]	27.4	55.5
MBRSC system combination translation:		
sentence selection (feature expectations loss)	27.4	55.5
gradient ascent search [17]	27.7	55.4
beam search (uniform weights)	27.8	55.1
+ automatic parameter optimization	28.0	54.9
oracle (beam search + reference n-gram counts)	43.3	42.2

Initially, we use an evaluation measure $Q(\cdot)$ (e.g. BLEU) to determine the absolute difference between the original outcomes of methods A and B. Then, we repeatedly create shuffled versions A' and B' of the original outcomes, determine the absolute difference between their evaluation metrics, and count the number of times $N^{\geq}$ that this difference is equal or larger than the original difference. To create the shuffled versions of the data sets, we iterate over each data point in the original outcomes and decide based on a simulated coin-flip whether data points should be exchanged between A and B. The p-value is the proportion of iterations in which the absolute difference in evaluation metric was indeed larger for the shuffled version (corrected to achieve an unbiased estimate).

4.2 Comparative experiments

First, we performed comparative experiments to evaluate the influence of the different features of MBRSC on MT quality. This experiments were performed on French-English, from the translation task of the 2009 Workshop on Statistical Machine Translation[8] [7]. We combined the outputs of the five statistical MT systems that submitted lists of n-best translation options to the task. Table 1 shows the average number of translation options for each source sentence, and case insensitive BLEU scores for the single best translation of each system. System outputs were tokenized and lower-cased before performing the combination. We report case-insensitive evaluation results to factor out the effect of true-casing of the English words from the effect of computing a consensus translation.

Table 2 displays case-insensitive BLEU and TER results for the computed consensus translations. We used different setups of MBRSC to generate consensus translations that combine all the translation options provided by the five individual systems. On average, for each source sentence we combined about 450 translations. We also report results for the best and worst individual systems, and for an oracle experiment where the expected n-gram counts were computed directly from the reference translations.

As a baseline, we present results for a conventional sentence-selection MBR classifier [15] for the ensemble model in Equation (5). The risk of each candidate translation was computed by exhaustively calculating its BLEU score with respect to the rest of the translations (Equation (10)). Results in Table 2 show that this

[8] `http://statmt.org/wmt09/translation-task.html`

Table 3 Examples of translation quality improvements resulting from system combination.

single MT	no aircraft universal also today is that the telephone .
MBRSC	no current apparatus is as universal as the telephone .
reference	no contemporary machine is as universal as the telephone .
single MT	no confirmation was able to be obtained from aig .
MBRSC	no confirmation could be obtained from aig .
reference	no confirmation could be obtained from aig .
single MT	simply complete , through the usb connector , the device of music from the computer .
MBRSC	it is enough to fill , through the usb connector , the music from the computer .
reference	it 's enough to fill the device with music using the usb from the computer .
single MT	for their operation to be effective , they have indeed need much less clients as a classic operator .
MBRSC	for their operation to be effective , they need far fewer customers than a classic operator .
reference	in order to function effectively , they require many fewer customers than a classic operator does .

baseline already resulted in a substantial improvement over the best individual system: +1.0 BLEU points and −0.5 TER points.

We replicated this baseline sentence-selection experiment using n-gram count expectations to compute the loss (Equation (11)) and obtained the same BLEU and TER scores than the baseline. These results indicate that MBR over feature expectations is an accurate approximation to the exact MBR classifier even for nonlinear loss functions such as BLEU, a finding consistent with prior research [10].

Then, we generated consensus translations using the beam search algorithm described in section 3.3; a pruning parameter value $N = 100$ was used. Results showed a slight performance improvement: +0.4 BLEU points over sentence-selection search, and +0.1 BLEU points over the gradient ascent search algorithm described in [17]. Since we used the same n-gram count expectations in all three experiments, these BLEU improvements imply that beam search was able to generate better translations than the translations already provided by the individual systems (sentence-selection search), and that it explores a broader search space than the gradient ascent search.

Finally, we automatically optimized the values of free parameters $\boldsymbol{\alpha}$ in a separate development set which further improved performance of MBRSC: +0.2 BLEU points and −0.2 TER points. This scarce improvements are rather surprising given that a much larger improvement, +1 BLEU points, was obtained in the development set. We hypothesize that this is due to overfitting: in fact, the optimized weight for one of the systems was very small. This can happen if the quality of a system varies between datasets. In this case, the importance of these systems in determining the consensus translation may be underestimated. Nevertheless, this final experiment showed a statistically significant improvement ($p = 0.0003$) of +0.6 BLEU points and −0.6 TER points over baseline. Table 3 shows examples of how the translation quality can be improved with system combination. Here, the consensus translation is compared with the translation of the best individual system, as well as with a human reference translation.

We performed one last comparative experiment (oracle) to measure the upper bound for the performance of MBRSC. Instead of expected counts, we generated

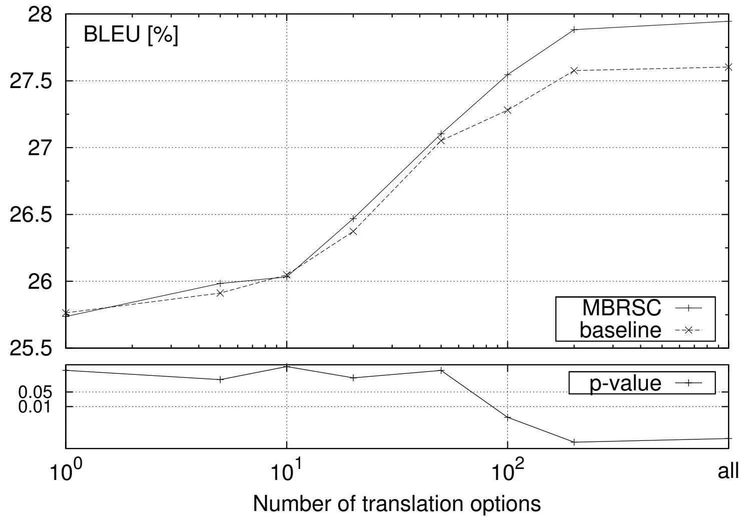

Fig. 3 Performance, and significance testing against baseline, of MBRSC as a function of the number of translation options combined. Parameter optimization was performed for both methods.

consensus translations using n-gram counts computed directly from the reference translations. Naturally, oracle results showed a huge improvement in performance over the best individual system. Since MBRSC barely explores one tenth of this potential, we conclude that refinements in the estimation of n-gram count expectations could have the potential to boost translation quality.

Additionally, we evaluated the performance of MBRSC as a function of the number of translation options combined. For each source sentence only a subset of translation options are combined to generate the consensus translation, namely the top scoring ones. Figure 3 compares MBRSC against the baseline sentence-selection search algorithm. Additionally, we report significance level of the difference in performance between them[9]. We mark two standard levels of significance, 0.01 and 0.05, for reference. MBRSC consistently outperformed baseline, although these differences were not statistically significant below 100 translation options. This is not surprising since the search space for the baseline sentence-selection method grows linearly with the number of translation options while for MBRSC it grows exponentially. Thus, as more translation options were used MBRSC was able to explore a broader space which involved a statistically significant difference in performance when combining 100 translation options or more.

4.3 Comparison with state-of-the-art system combination methods

We now compare MBRSC against several state-of-the-art subsequence system combination techniques. This experiments were performed on the official evaluation

[9] Similarly as done in [2], we give p-values on a logarithmic scale.Note that 10^{-4} is the smallest possible p-value that can be computed with 9999 shuffles in the randomized test.

Table 4 BLEU [%] scores of MBRSC in comparison with the best-performing system combination methods presented in the system combination task of the 2011 workshop on statistical machine translation.

System	cz→en	en→cz	de→en	en→de
MBRSC	29.5	20.8	25.2	18.4
BBN [38]	29.9	–	26.5	–
CMU [20]	28.7	20.1	25.1	17.6
JHU [43]	29.4	–	24.9	–
RTWH [28]	–	–	25.4	–

sets from the system combination task [10] of the 2011 Workshop on Statistical Machine Translation [8]. Consensus translations were generated for both translation directions of the following language pairs: Czech–English (cz–en), German–English (de–en), Spanish–English (es–en) and French–English (fr–en). For each translation direction, we combined the outputs of all the system that submit translations to the translation task. In contrast to the previous experiments, for each source sentence only single best translations were provided by each individual system. Thus, each experiment combined only about 10 translations.

Table 4 compares the performance of MBRSC with respect to the various systems that participate in the system combination task. For the sake of simplicity, we show results only for the four (out of ten) best-performing systems. All these system combination methods align the provided translations to build a consensus network, and compute the consensus translation as the highest-scoring path through the network in the style of [16]. They differ in the alignment method and the path-scoring models used. We report results only for cz↔en and de↔en translation directions. Experiments for other directions lead to similar conclusions.

It is important to note that the experimental conditions of this task favored consensus network methods. On the one hand, only single-best translations were available so the n-gram count expectations could not be smoothly estimated and were biased to those single translations. On the other hand, organizers allowed the use of any additional data which permits network methods to train their complex search models. However, we found that even in this pessimistic setting MBRSC was the best performer for en→cz and en→de, and was between the top-performing systems for the rest of translation directions.

Not surprisingly, MBRSC scored particularly high for those translation directions (cz and de) whose target language had scarcer resources. For these languages, network-based systems simply did not had enough data to train their complex network search models. In fact, many participants submitted consensus translations for only a limited number of translation directions. In contrast, MBRSC does not require any additional data. Since the consensus translation is directly computed from the provided translation options, MBRSC obtained competitive results in all translation directions. These results confirm the soundness and generality of the proposed system combination technique.

[10] http://www.statmt.org/wmt11/system-combination-task.html

5 Conclusion

We have described minimum Bayes' risk system combination (MBRSC) a new subsequence system combination approach for MT. MBRSC is able to detect and combine the "best" parts of the provided translations to generate the optimal consensus translation with respect to the BLEU score.

Despite its simplicity, MBRSC provides strong performance by leveraging different modelling, training and search techniques. We have performed a thorough analysis of how individual features of the algorithm influence the translation quality, and have compared the overall performance with the upper bound achievable by the algorithm. These comparative experiments showed that MBRSC significantly outperforms MBR sentence-selection techniques. Additionally, we compared MBRSC with several state-of-the-art subsequence combination systems in the system combination task of the 2011 workshop on statistical machine translation. Experiments show that even in this pessimistic setting, better suited for the dominant network-based techniques, MBRSC obtained competitive results specially for languages with scarce resources.

Acknowledgements Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), the iTrans2 (TIN2009-14511) project, the UPV under grant 20091027, the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

1. Bangalore, S.: Computing consensus translation from multiple machine translation systems. In: IEEE Automatic Speech Recognition and Understanding Workshop, pp. 351–354 (2001)
2. Becker, M.A.: Active learning - an explicit treatment of unreliable parameters. Ph.D. thesis, University of Edinburgh (2008)
3. Bellman, R.: Dynamic Programming. Princeton University Press,Princeton, NJ (1957)
4. Bickel, P.J., Doksum, K.A.: Mathematical statistics : basic ideas and selected topics. Holden-Day, San Francisco (1977)
5. Callison-burch, C., Flournoy, R.S.: A program for automatically selecting the best output from multiple machine translation engines. In: Proceedings of the VIII Machine Translation Summit, pp. 63–66 (2001)
6. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: Further meta-evaluation of machine translation. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 70–106. Association for Computational Linguistics (2008)
7. Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, pp. 1–28. Association for Computational Linguistics, Athens, Greece (2009)
8. Callison-Burch, C., Koehn, P., Monz, C., Zaidan, O.F. (eds.): Proceedings of the 6th Workshop on Statistical Machine Translation. Association for Computational Linguistics, Edinburgh, Scotland (2011)

9. Chinchor, N.: The statistical significance of the muc-4 results. In: Proceedings of the Conference on Message Understanding, pp. 30–50 (1992)
10. DeNero, J., Chiang, D., Knight, K.: Fast consensus decoding over translation forests. In: Proceedings of the 47th annual meeting of the Association for Computational Linguistics, pp. 567–575. Association for Computational Linguistics (2009)
11. DeNero, J., Kumar, S., Chelba, C., Och, F.: Model combination for machine translation. In: Proceedings of the 11th conference of the North American chapter of the Association for Computational Linguistics, pp. 975–983. Association for Computational Linguistics (2010)
12. Dietterich, T.G.: Ensemble methods in machine learning. In: Proceedings of the First International Workshop on Multiple Classifier Systems, MCS '00, pp. 1–15. Springer-Verlag (2000)
13. Duan, N., Li, M., Zhang, D., Zhou, M.: Mixture model-based minimum bayes risk decoding using multiple machine translation systems. In: Proceedings of the 23rd conference on Computational Linguistics, pp. 313–321 (2010)
14. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification (2nd Edition), 2 edn. Wiley-Interscience (2001)
15. Ehling, N., Zens, R., Ney, H.: Minimum bayes risk decoding for bleu. In: Proceedings of the 45th annual aeeting of the Association for Computational Linguistics, pp. 101–104. Association for Computational Linguistics (2007)
16. Fiscus, J.G.: A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). In: Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 347–352 (1997)
17. González-Rubio, J., Juan, A., Casacuberta, F.: Minimum bayes-risk system combination. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics, pp. 1268–1277 (2011)
18. He, X., Toutanova, K.: Joint optimization for machine translation system combination. In: Proceedings of the 2009 conference on Empirical Methods in Natural Language Processing, pp. 1202–1211. Association for Computational Linguistics (2009)
19. He, X., Yang, M., Gao, J., Nguyen, P., Moore, R.: Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In: Proceedings of the 2008 conference on Empirical Methods in Natural Language Processing, pp. 98–107. Association for Computational Linguistics (2008)
20. Heafield, K., Lavie, A.: Cmu system combination in wmt 2011. In: Proceedings of the 6th workshop on Statistical Machine Translation, pp. 145–151. Association for Computational Linguistics, Edinburgh, Scotland (2011)
21. Jayaraman, S., Lavie, A.: Multi-engine machine translation guided by explicit word matching. In: Proceeding of the 10th conference of the European Association for Machine Translation, pp. 143–152 (2005)
22. Jelinek, F.: Statistical methods for speech recognition. MIT Press, Cambridge, MA, USA (1997)
23. Kittler, J., Hatef, M., Duin, R.P.W., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analisis and Machine Intelligence **20**, 226–239 (1998). DOI 10.1109/34.667881
24. Knight, K.: Decoding complexity in word-replacement translation models. Computational Linguistics **25**(4), 607–615 (1999). URL

http://dl.acm.org/citation.cfm?id=973226.973232
25. Kumar, S., Macherey, W., Dyer, C., Och, F.: Efficient minimum error rate training and minimum bayes-risk decoding for translation hypergraphs and lattices. In: Proceedings of the 47th annual meeting of the Association for Computational Linguistics, pp. 163–171. Association for Computational Linguistics (2009)
26. Land, A.H., Doig, A.G.: An automatic method of solving discrete programming problems. Econometrica **28**(3), 497–520 (1960)
27. Larkey, L.S., Croft, B.W.: Combining classifiers in text categorization. In: H.P. Frei, D. Harman, P. Schäuble, R. Wilkinson (eds.) Proceedings of the 19th ACM International Conference on Research and Development in Information Retrieval, pp. 289–297. ACM Press, New York, US (1996)
28. Leusch, G., Freitag, M., Ney, H.: The rwth system combination system for wmt 2011. In: Proceedings of the 6th workshop on Statistical Machine Translation, pp. 152–158. Association for Computational Linguistics, Edinburgh, Scotland (2011)
29. Matusov, E., Leusch, G., Banchs, R.E., Bertoldi, N., Dechelotte, D., Federico, M., Kolss, M., suk Lee, Y., no, J.B.M., Paulik, M., Roukos, S., Schwenk, H., Ney, H.: System combination for machine translation of spoken and written language. IEEE Transactions on Audio, Speech and Language Processing **16**, 1222–1237 (2008)
30. Nelder, J.A., Mead, R.: A Simplex Method for Function Minimization. The Computer Journal **7**(4), 308–313 (1965)
31. NIST: NIST 2006 machine translation evaluation official results. http://www.itl.nist.gov/iad/mig/tests/mt/ (2006)
32. Nomoto, T.: Multi-engine machine translation with voted language model. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics, pp. 494–501. Association for Computational Linguistics (2004)
33. Noreen, E.: Computer-intensive methods for testing hypotheses: an introduction. A Wiley Interscience publication. Wiley (1989)
34. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st annual meeting on Association for Computational Linguistics, pp. 160–167. Association for Computational Linguistics (2003)
35. 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 on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
36. Paul, M., Doi, T., Hwang, Y., Imamura, K., Okuma, H., Sumita, E.: Nobody is perfect: Atr's hybrid approach to spoken language translation. In: Proceedings of the 2005 International Workshop on Spoken Language Translation, pp. 55–62 (2005)
37. Rosti, A., Ayan, N.F., Xiang, B., Matsoukas, S., Schwartz, R., Dorr, B.: Combining outputs from multiple machine translation systems. In: Proceedings of the 6th conference of the North American Chapter of the Association for Computational Linguistics, pp. 228–235. Association for Computational Linguistics (2007)
38. Rosti, A., Zhang, B., Matsoukas, S., Schwartz, R.: Expected bleu training for graphs: Bbn system description for wmt11 system combination task. In: Proceedings of the 6th workshop on Statistical Machine Translation, pp. 159–

165. Association for Computational Linguistics (2011)
39. Roth, D., Zelenko, D.: Part of speech tagging using a network of linear separators. In: Proceedings of the 17th international conference on Computational linguistics - Volume 2, COLING '98, pp. 1136–1142. Association for Computational Linguistics (1998)
40. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Weischedel, R.: A study of translation error rate with targeted human annotation. In: Proceedings of the 7th conference of the Association for Machine Transaltion in the Americas, pp. 223–231 (2006)
41. Stanley, R.: Enumerative combinatorics. Cambridge studies in advanced mathematics. Cambridge University Press (2002)
42. Udupa, R., Maji, H.K.: Computational complexity of statistical machine translation. In: D. McCarthy, S. Wintner (eds.) Proceedings of the European Chapter of the Association for Computational Linguistics. The Association for Computer Linguistics (2006). URL http://acl.ldc.upenn.edu/E/E06/E06-1004.pdf
43. Xu, D., Cao, Y., Karakos, D.: Description of the jhu system combination scheme for wmt 2011. In: Proceedings of the 6th workshop on Statistical Machine Translation, pp. 171–176. Association for Computational Linguistics (2011)