Regenerating Hypotheses for Statistical Machine Translation

Boxing Chen, Min Zhang, Aiti Aw and **Haizhou Li**
Department of Human Language Technology
Institute for Infocomm Research
21 Heng Mui Keng Terrace, 119613, Singapore
{bxchen, mzhang, aaiti, hli}@i2r.a-star.edu.sg

Abstract

This paper studies three techniques that improve the quality of N-best hypotheses through additional regeneration process. Unlike the multi-system consensus approach where multiple translation systems are used, our improvement is achieved through the expansion of the N-best hypotheses from a single system. We explore three different methods to implement the regeneration process: re-decoding, n-gram expansion, and confusion network-based regeneration. Experiments on Chinese-to-English NIST and IWSLT tasks show that all three methods obtain consistent improvements. Moreover, the combination of the three strategies achieves further improvements and outperforms the baseline by 0.81 BLEU-score on IWSLT'06, 0.57 on NIST'03, 0.61 on NIST'05 test set respectively.

1 Introduction

State-of-the-art Statistical Machine Translation (SMT) systems usually adopt a two-pass search strategy (Och, 2003; Koehn, et al., 2003) as shown in Figure 1. In the first pass, a decoding algorithm is applied to generate an N-best list of translation hypotheses, while in the second pass, the final translation is selected by rescoring and re-ranking the N-best translations through additional feature functions. The fundamental assumption behind using a second pass is that the generated N-best list may contain better translations than the best choice found by the decoder. Therefore, the performance of a two-pass SMT system can be improved from two aspects, i.e. scoring models and the quality of the N-best hypotheses.

Rescoring pass improves the performance of machine translation by enhancing the scoring models with more global sophisticated and discriminative feature functions. The idea for applying two passes instead of one is that some global feature functions cannot be easily decomposed into local scores and computed during decoding. Furthermore, rescoring allows some feature functions, such as word and n-gram posterior probabilities, to be estimated on the N-best list (Ueffing, 2003; Chen et al., 2005; Zens and Ney, 2006).

In this two-pass method, translation performance hinges on the N-best hypotheses that are generated in the first pass (since rescoring occurs on these), so adding the translation candidates generated by other MT systems to these hypotheses could potentially improve the performance. This technique is called system combination (Bangalore et al., 2001; Matusov et al., 2006; Sim et al., 2007; Rosti et al., 2007a; Rosti et al., 2007b).

We have instead chosen to regenerate new hypotheses from the original N-best list, a technique which we call *regeneration*. Regeneration is an intermediate pass between decoding and rescoring as depicted in Figure 2. Given the original N-best list (N-best1) generated by the decoder, this regeneration pass creates new translation hypotheses from this list to form another N-best list (N-best2). These two N-best lists are then combined and given to the rescoring pass to derive the best translation.

We implement three methods to regenerate new hypotheses: re-decoding, n-gram expansion and confusion network. Re-decoding (Rosti et al., 2007a) based regeneration re-decodes the source sentence using original LM as well as new trans-

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 105–112
Manchester, August 2008

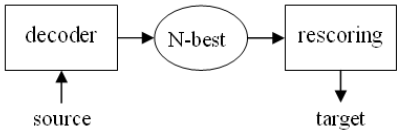

Figure 1: Structure of a typical two-pass machine translation system. N-best translations are generated by the decoder and the 1-best translation is returned after rescored with additional feature functions.

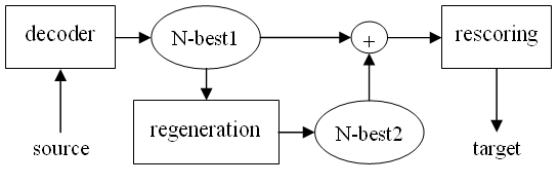

Figure 2: Structure of a three-pass machine translation system with the new regeneration pass. The original N-best translations list (N-best1) is expanded to generate a new N-best translations list (N-best2) before the rescoring pass.

lation and reordering models that are trained on the source-to-target N-best translations generated in the first pass. N-gram expansion (Chen et al., 2007) regenerates more hypotheses by continuously expanding the partial hypotheses through an n-gram language model trained on the original N-best translations. And confusion network generates new hypotheses based on confusion network decoding (Matusov et al., 2006), where the confusion network is built on the original N-best translations.

Confusion network and re-decoding have been well studied in the combination of different MT systems (Bangalore et al., 2001; Matusov et al., 2006; Sim et al., 2007; Rosti et al., 2007a; Rosti et al., 2007b). Researchers have used confusion network to compute consensus translations from the outputs of different MT systems and improve the performance over each single systems. (Rosti et al., 2007a) also used re-decoding to do system combination by extracting sentence-specific phrase translation tables from the outputs of different MT systems and running a phrase-based decoding with this new translation table. Finally, N-gram expansion method (Chen et al., 2007) collects sub-strings occurring in the N-best list to produce alternative translations.

This work demonstrates that a state-of-the-art MT system can be further improved by means of *regeneration* which expands its own N-best translations other than taking the translation candidates from the other MT systems.

2 SMT Process

Phrase-based statistical machine translation systems are usually modeled through a log-linear framework (Och and Ney, 2002). By introducing the hidden word alignment variable a (Brown et al., 1993), the optimal translation can be searched for based on the following criterion:

$$\tilde{e}^* = \arg\max_{e,a}\left(\sum_{m=1}^{M}\lambda_m h_m(\tilde{e}, \tilde{f}, a)\right) \quad (1)$$

where $\tilde{e}$ is a string of phrases in the target language, $\tilde{f}$ is the source language string of phrases, $h_m(\tilde{e}, \tilde{f}, a)$ are feature functions, weights λ_m are typically optimized to maximize the scoring function (Och, 2003).

Our MT baseline system is based on Moses decoder (Koehn et al., 2007) with word alignment obtained from GIZA++ (Och et al., 2003). The translation model (TM), lexicalized word reordering model (RM) are trained using the tools provided in the open source Moses package. Language model (LM) is trained with SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing method (Chen and Goodman, 1998).

3 Regeneration Methods

Given the original N-best translations, regeneration pass is to generate M new target translations which are not seen in the original N-best choices.

3.1 Regeneration with Re-decoding

One way of regeneration is by running the decoding again to obtain new hypotheses through a re-decoding process (Rosti et al., 2007a). In this work, the same decoder (Moses) is used to produce the new M-best translations using a new translation model and reordering model trained over the word-aligned source input and original N-best target hypotheses. Although the target-to-source phrase alignments are available in the original N-best hypotheses, to enlarge the difference between the new M-best translations and the original N-best translations, we re-align the words using GIZA++.

Weights of the decoder are re-optimized by the tool in the Moses package over the development set. The process of such a re-decoding is summarized as follows:

1. Run GIZA++ to align the words between the source input and target N-best translations;
2. Train translation and reordering model;
3. Optimize the weights of the decoder with the new models;
4. Decode the source input by using new models and new weights to generate N+M distinct translations ("distinct" here refers to the target language string only, not considering the phrase segmentation, etc.);
5. Output M-best translations which are not seen in the original N-best translations.

Re-decoding on test set follows the same steps, but without the tuning step, step 3.

3.2 Regeneration with N-gram Expansion

N-gram expansion (Chen et al., 2007) combines the sub-strings occurred in the original N-best translations to generate new hypotheses. Firstly, all n-grams from the original N-best translations are collected. Then the partial hypotheses are continuously expanded by appending a word through the n-grams collected in the first step. We explain this method in more detail using the following example.

Suppose we have four original hypotheses shown in Figure 3. Firstly, we collect all the 3-grams from the original hypotheses. The first n-grams of all original entries in the N-best list are set as the initial partial hypotheses. They are: *it's 5 minutes*, *it is 5*, *it's about 5* and *i walk 5*. Then the expansion of a partial hypothesis starts by computing the set of n-grams matching its last n-1 words. As shown in Figure 4, the n-gram *5 minutes on* matches the last two words of the partial hypothesis *it's about 5 minutes*. So the hypothesis is expanded to *it's about 5 minutes on*. The expansion continues until the partial hypothesis ends with a special end-of-sentence symbol that occurs at the end of all N-best strings.

Figure 5 shows some new hypotheses that are generated from the example in Figure 3. This is an example excerpted from our development data. One reference is also given in Figure 5; the first new generated hypothesis is equal to this reference. But unfortunately, there is no such hypothesis in the original N-best translations.

During the new hypotheses generation, the translation outputs of a given source sentence are computed through a beam-search algorithm with a log-linear combination of the feature functions. In addition to n-gram frequency and n-gram posterior probability which have been used in (Chen et al., 2007), we also used language model, direct/inverse IBM model 1, and word penalty in this work. The size of the beam is set to N+M, to ensure more than M new hypotheses are generated.

Original hypotheses	1. it's 5 minutes on foot . 2. it is 5 minutes on foot . 3. it's about 5 minutes' to walk . 4. i walk 5 minutes .
n-grams	it's 5 minutes, 5 minutes on, …… on foot ., about 5 minutes …… 5 minutes .

Figure 3: Example of original hypotheses and 3-grams collected from them.

partial hyp.	it's	about	5	minutes	
n-gram	+		5	minutes	on
new partial hyp.	it's	about	5	minutes	on

Figure 4: Expanding a partial hypothesis via a matching n-gram.

New hypotheses	it's about 5 minutes on foot . it's 5 minutes . i walk 5 minutes on foot . ……
Reference	it's about five minutes on foot .

Figure 5: New generated hypotheses through n-gram expansion and one reference.

3.3 Regeneration with Confusion Network

Confusion network based regeneration builds a confusion network over the original N-best hypotheses, and then extracts M-best hypotheses from it. The word order in the N-best translations could be very different, so we need to choose a hypothesis with the "most correct" word order as the confusion network skeleton (alignment reference), then align and reorder other hypotheses in this word order.

Some previous work compute the consensus translation under MT system combination, which differ from ours in the way of choosing the skeleton and aligning the words. Matusov et al. (2006) let every hypothesis play the role of the skeleton once and used GIZA++ to get word alignment. Bangalore et al. (2001), Sim et al. (2007), Rosti et al. (2007a), and Rosti et al. (2007b) chose the hypothesis that best agrees with other hypotheses on average as the skeleton. Bangalore et al. (2001) used a WER based alignment and Sim et al. (2007), Rosti et al. (2007a), and Rosti et al. (2007b) used minimum Translation Error Rate

(TER) based alignment to build the confusion network.

Choosing alignment reference: Since the N-best translations are ranked, choosing the first best hypothesis as the skeleton is straightforward in our work.

Aligning words: As a confusion network can be easily built from a one-to-one alignment, we develop our algorithm based on the one-to-one assumption and use competitive linking algorithm (Melamed, 2000) for our word alignment. Firstly, an association score is computed for every possible word pair from the skeleton and sentence to be aligned. Then a greedy algorithm is applied to select the best word-alignment. In this paper, we use a linear combination of multiple association scores, as suggested in (Kraif and Chen, 2004). As the two sentences to be aligned are in the same language, the association scores are computed on the following four clues. They are cognate (S_1), word class (S_2), synonyms (S_3), and position difference (S_4). The four scores are linearly combined with empirically determined weights as shown is Equation 2.

$$S(f_j, e_i) = \sum_{k=1}^{4} \lambda_k \times S_k \qquad (2)$$

Reordering words: After word alignment, the words in all other hypotheses are reordered to match the word order of the skeleton. The aligned words are reordered according to their alignment indices. The unaligned words are reordered in two strategies: moved with its previous word or next word. In this work, additional experiments suggested that moving the unaligned word with its previous word achieve better performance. In the case that the first word is unaligned, it will be moved with its next word. Each word is assigned a score based on a simple voting scheme. Figure 6 shows an example of creating a confusion network.

Extracting M-best translations: New translations are extracted from the confusion network. We again use beam-search algorithm to derive new hypotheses. The same feature functions proposed in Section 3.2 are used to score the partial hypotheses. Moreover, we also use position based word probability (i.e. in Figure 6, the words in position 5, "on" scored a probability of 0.5, and "ε" scored a probability of 0.25) as a feature function. Figure 6 shows some examples of new hypotheses generated through confusion network regeneration.

Original hypotheses	**1. it's 5 minutes on foot .** 2. it is 5 minutes on foot . 3. it's about 5 minutes' to walk . 4. i walk 5 minutes .
Alignments	**it's 5 minutes on foot . ε** it 5 minutes on foot . is it's 5 minutes' to walk . about i 5 minutes ε walk .
Confusion network	**it's ε 5 minutes on foot .** it is 5 minutes on foot . it's about 5 minutes' to walk . i ε 5 minutes ε walk .
New hypotheses	1. it's about five minutes on foot . 2. it about five minutes on foot . 3. it's about five minutes on walk . 4. i about 5 minutes to work .

Figure 6: Example of creating a confusion network from the word alignments, and new hypotheses generated through the confusion network. The sentence in bold is the alignment reference.

4 Rescoring model

Since the final N+M-best hypotheses are produced either from different methods or same decoder with different models, local feature functions of each hypothesis are not directly comparable, and thus inadequate for rescoring. We hence exploit rich global feature functions in the rescoring models to compensate the loss of local feature functions. We apply the following 10 feature functions and optimize the weight of each feature function using the tool in Moses package.

- direct and inverse IBM model 1 and 3
- association score, i.e. hyper-geometric distribution probabilities and mutual information
- lexicalized word/block reordering rules (Chen et al., 2006)
- 6-gram target LM
- 8-gram target word-class based LM, word-classes are clustered by GIZA++
- length ratio between source and target sentence
- question feature (Chen et al., 2005)
- linear sum of n-grams relative frequencies within N-best translations (Chen et al., 2005)
- n-gram posterior probabilities within the N-best translations (Zens and Ney, 2006)
- sentence length posterior probabilities (Zens and Ney, 2006)

5 Experiments

5.1 Tasks

We carried out two sets of experiments on two different datasets. One is in spoken language domain while the other is on newswire corpus. Both experiments are on Chinese-to-English translation.

Experiments on spoken language domain were carried out on the *Basic Traveling Expression Corpus* (BTEC) (Takezawa et al., 2002) Chinese- to-English data augmented with *HIT-corpus*[1]. BTEC is a multilingual speech corpus which contains sentences spoken by tourists. 40K sentence-pairs are used in our experiment. *HIT-corpus* is a balanced corpus and has 500K sentence-pairs in total. We selected 360K sentence-pairs that are more similar to BTEC data according to its sub-topic. Additionally, the English sentences of *Tanaka* corpus[2] were also used to train our LM. We ran experiments on an IWSLT[3] *challenge track* which uses IWSLT-2006[4] DEV clean text set as development set and IWSLT-2006 TEST clean text as test set. Table 1 summarizes the statistics of the training, dev and test data for IWSLT task.

Experiments on newswire domain were carried out on the FBIS[5] corpus. We used NIST[6] 2002 MT evaluation test set as our development set, and the NIST 2003, 2005 test sets as our test sets. Table 2 summarizes the statistics of the training, dev and test data for NIST task.

data		Chinese	English
Train	Sentences	406,122	
	Words	4,443K	4,591K
	Vocabulary	69,989	61,087
Dev.	Sentences	489	489×7
	Words	5,896	45,449
Test	Sentences	500	500×7
	Words	6,296	51,227
Additional target data	Sentences	-	155K
	Words	-	1.7M

Table 1: Statistics of training, development and test data for IWSLT task.

[1] http://mitlab.hit.edu.cn/
[2] http://www.csse.monash.edu.au/~jwb/tanakacorpus.html
[3] International Workshop for Spoken Language Translation
[4] http:// www.slc.atr.jp/IWSLT2006/
[5] LDC2003E14
[6] http://www.nist.gov/speech/tests/mt/

data		Chinese	English
Train	Sentences	238,761	
	Words	7.0M	8.9M
	Vocabulary	56,223	63,941
NIST 02 (dev)	Sentences	878	878×4
	Words	23,248	108,616
NIST 03 (test)	Sentences	919	919×4
	Words	25,820	116,547
NIST 05 (test)	Sentences	1,082	1,082×4
	Words	30,544	141,915
Additional target data	Sentences	-	2.2M
	Words	-	61.5M

Table 2: Statistics of training, development and test data for NIST task.

System	#hypo	Dev set		Test set	
		BLEU	NIST	BLEU	NIST
1-best	-	29.98	7.468	29.10	7.103
RESC1	1,200	31.60	7.657	30.42	7.165
RD	1,200	32.46	7.664	30.95	7.175
NE	1,200	32.58	7.660	31.02	7.178
CN	1,200	32.33	7.671	30.82	7.200
RESC2	2,000	31.72	7.659	30.55	7.166
COMB	2,000	**32.98**	**7.673**	**31.36**	**7.202**

Table 3: Translation performances (BLEU% and NIST scores) of IWSLT task: decoder (1-best), rescoring on original 1,200 N-best (RESC1) and 2,000 N-best hypotheses (RESC2), re-decoding (RD), n-gram expansion (NE), confusion network (CN) and combination of all hypotheses (COMB).

5.2 Results

We set N = 800 and M = 400 for IWSLT task, i.e. 800 distinct translations for each source input are extracted from the decoder and used for regeneration; and 400 new hypotheses are generated for each regeneration system: re-decoding (RD), n-gram expansion (NE) and confusion network (CN). System COMB combines the original N-best and the three regenerated M-best hypotheses lists (totally, 2,000 distinct hypotheses: 800 + 3×400). Then each system computes the 1-best translation through rescoring and re-ranking its hypotheses list. For comparison purpose, the performance of rescoring on two sets of original N-best translations are also computed and they are applied based on 1,200 (RESC1) and 2,000 (RESC2) distinct hypotheses extracted from the decoder. For NIST task, we set N = 1,600, and M = 800, thus, RESC2 and COMB compute 1-

System	#hypo	NIST'02 (dev)		NIST'03 (test)		NIST'05 (test)	
		BLEU	NIST	BLEU	NIST	BLEU	NIST
1-best	1	27.67	8.498	26.68	8.271	24.82	7.856
RESC1	2,400	28.13	8.519	27.09	8.312	25.29	7.868
RD	2,400	28.46	8.518	27.34	8.320	25.54	7.897
NE	2,400	28.52	8.539	27.47	8.329	25.65	7.907
CN	2,400	28.40	8.545	27.30	8.332	25.54	7.913
RESC2	4,000	28.27	8.522	27.21	8.320	25.43	7.875
COMB	4,000	**28.92**	**8.602**	**27.78**	**8.401**	**26.04**	**7.994**

Table 4: Translation performances (BLEU% and NIST scores) of NIST task: decoder (1-best), rescoring on original 2,400 N-best (RESC1) and 4,000 N-best hypotheses (RESC2), re-decoding (RD), n-gram expansion (NE), confusion network (CN) and combination of all hypotheses (COMB).

1	Reference	No tax is needed for this item . Thank you .
	RESC2	you don't have to do not need to pay duty on this . thank you .
	COMB (RD)	not need to pay duty on this . thank you .
2	Reference	Certainly . The fitting room is over there . Please come with me .
	RESC2	the fitting room is over there . can you come with me .
	COMB (NE)	yes , you can . the fitting room is over there . please come with me .
3	Reference	OK . I will bring it to you in five minutes .
	RESC2	a good five minutes , we will give you .
	COMB (CN)	ok . after five minutes , i will give it to you .

Table 5: Translations output by system RESC2 and COMB on IWSLT task (case-insensitive).

best from 4,000 ($1{,}600 + 3 \times 800$) distinct hypotheses.

Our evaluation metrics are BLEU (Papineni et al., 2002) and NIST, which are to perform case-insensitive matching of *n*-grams up to $n = 4$. The translation performance of IWSLT task and NIST task is reported in Tables 3 and 4 respectively. The row "1-best" reports the scores of the translations produced by the decoder. The column "#hypo" means the size of the N-best hypotheses involved in rescoring. Note that on top of the same global feature functions as mentioned in Section 4, the local feature functions used during decoding were also involved in rescoring RESC1 and RESC2.

First of all, we note that both BLEU and NIST scores of the first decoding step were improved through rescoring. If rescoring was applied after regeneration on the N+M best lists, additional improvements were gained for all the development and test sets on all three regeneration systems. Absolute improvement on BLEU score of 0.4-0.6 on IWSLT'06 test set and 0.25-0.35 on NIST test sets were obtained when compared with system RESC1. Comparing the performance of three regeneration methods, we can see that re-decoding and confusion network based method achieved very similar improvement; while n-gram expansion based regeneration obtained slightly better improvement than the other two methods. Combining all regenerated hypotheses with the original hypotheses further increased the scores on both tasks. Compared with RESC2, system COMB obtained absolute improvement of 0.81 (31.36 – 30.55) BLEU score on IWSLT'06 test set, 0.57 (27.28 – 27.21) BLEU score on NIST'03 and 0.61 (26.04 – 25.43) BLEU score on NIST'05 respectively.

We further illustrate the effectiveness of the regeneration mechanism using some translation examples obtained from system RESC2 and COMB as shown in Table 5.

6 Discussion

To better interpret the performance improvement; first let us check if the regeneration pass has produced better hypotheses. We computed the oracle scores on all four 1,200-best lists in IWSLT task. The oracle chooses the translation with the lowest word error rate (WER) with respect to the references in all cases. The results are reported in Table 6. It is worth noticing that the first 800-best (original N-best) hypotheses are the same in

all four lists, with differences found only in the remaining 400 hypotheses (M-best). The consistent improvement of oracle scores shows that the translation candidates have been really improved.

From another viewpoint, Table 7 shows the number of translations generated by each method in the final translation output (translations of COMB). After re-ranking N+3M entries, it is observed that more than 25% (e.g. for IWSLT'06 test set, (50+74+39)/500=32.6%; NIST'03 test set, (77+85+68)/919=25.1%; NIST'05 test set, (95+110+82)/1082=26.5%) of best scored outputs were generated by the regeneration pass, showing that new generated translations are quite often the rescoring winner. This also proved that the new-generated hypotheses contain better ones than the original ones.

List		BLEU	NIST	WER	PER
Dev.	Moses	46.10	8.765	36.29	30.94
	RD	46.91	8.764	35.29	30.62
	NE	46.95	8.811	36.05	30.72
	CN	46.85	8.769	36.17	30.83
Test	Moses	45.09	8.403	37.07	32.04
	RD	45.67	8.418	36.50	31.82
	NE	45.82	8.481	36.44	31.70
	CN	45.68	8.471	36.55	31.81

Table 6: Oracle scores (BLEU%, NIST, WER% and PER%) on IWSLT task 1,200-best lists of four systems: decoder (Moses), re-decoding (RD), n-gram expansion (NE) and confusion network (CN).

		# sentence				
	Set	Tot.	Orig.	RD	NE	CN
IWSLT	Dev	489	325	52	76	36
	Test	500	337	50	74	39
NIST	NIST 02	878	613	92	100	73
	NIST 03	919	689	77	85	68
	NIST 05	1082	795	95	110	82

Table 7: Number of translations generated by each method in the final translation output of system COMB: decoder (Orig.), re-decoding (RD), n-gram expansion (NE) and confusion network (CN). "Tot." is the size of the dev/test set.

Then, let us consider each single regeneration method to understand why regeneration can produce better hypotheses. Re-decoding may introduce new and better phrase-pairs which are extracted from the N-best hypotheses to the translation model thus generate better hypotheses. N-gram expansion can (almost) fully exploit the search space of target strings, which can be generated by an n-gram LM. As a result, it can produce alternative translations which contain word re-orderings and phrase structures not considered by the search algorithm of the decoder (Chen, et al., 2007). Confusion network based regeneration reinforces the word choice by considering the posterior probabilities of words occur in the N-best translations.

7 Conclusions

In this paper, we proposed a novel three-pass SMT framework against the typical two-pass system. This framework enhanced the quality of the translation candidates generated by our proposed regeneration pass and improved the final translation performance. Three regeneration methods were introduced, namely, re-decoding, word-based n-gram expansion, and confusion network based regeneration.

Experiments were based on the state-of-the-art phrase-based decoder and carried out on the IWSLT and NIST Chinese-to-English task. We showed that all three methods improved the performance with the n-gram expansion method achieving the greatest improvement. Moreover, the combination of the three methods further improves the performance.

We conclude that translation performance can be improved by increasing the potential of translation candidates to contain better translations. We have presented an alternative solution to ameliorate the quality of translation candidates in a way that differs from system combination which takes translations from other MT systems. We demonstrated that the translation performance could be self-boosted by expanding the N-best list through hypotheses regeneration.

References

S. Bangalore, G. Bordel, and G. Riccardi. 2001. Computing consensus translation from multiple machine translation systems. In *Proceeding of IEEE workshop on Automatic Speech Recognition and Understanding*, pages 351–354. Madonna di Campiglio, Italy.

P. F. Brown, V. J. Della Pietra, S. A. Della Pietra & R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2) 263-312.

B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo and M. Federico. 2005. The ITC-irst SMT System for

IWSLT-2005. In *Proceeding of IWSLT-2005*, pp.98-104, Pittsburgh, USA, October.

B. Chen, M. Cettolo and M. Federico. 2006. Reordering Rules for Phrase-based Statistical Machine Translation. In *Proceeding of IWSLT-2006*, Kyoto, Japan.

B. Chen, M. Federico and M. Cettolo. 2007. Better N-best Translations through Generative n-gram Language Models. In *Proceeding of MT Summit XI*. Copenhagen, Denmark.

S. F. Chen and J. T. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. *Technical Report TR-10-98*, Computer Science Group, Harvard University.

P. Koehn, F. J. Och and D. Marcu. 2003. Statistical Phrase-based Translation. In *Proceedings of HLT/NAACL*, pp 127-133, Edmonton, Canada.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceeding of ACL-2007*, pp. 177-180, Prague, Czech Republic.

O. Kraif, B. Chen. 2004. Combining clues for lexical level aligning using the Null hypothesis approach. In *Proceeding of COLING-2004*, Geneva, pp. 1261-1264.

E. Matusov, N. Ueffing, and H. Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceeding of EACL-2006*, Trento, Italy.

I. D. Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2), pp. 221-249.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL-2003*. Sapporo, Japan.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1), pp. 19-51.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceeding of ACL-2002*.

A. Rosti, N. F. Ayan, B. Xiang, S. Matsoukas, R. Schwartz and B. Dorr. 2007a. Combining Outputs from Multiple Machine Translation Systems. In *Proceeding of NAACL-HLT-2007*, pp. 228-235. Rochester, NY.

A. Rosti, S. Matsoukas and R. Schwartz. 2007b. Improved Word-Level System Combination for Machine Translation. In *Proceeding of ACL-2007*, Prague.

K. C. Sim, W. J. Byrne, M. J.F. Gales, H. Sahbi, and P. C. Woodland. 2007. Consensus network decoding for statistical machine translation system combination. In *Proceeding of ICASSP-2007*.

A. Stolcke. 2002. SRILM - an extensible language modelling toolkit. In *Proceeding of ICSLP-2002*. 901-904.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceeding of LREC-2002*, Las Palmas de Gran Canaria, Spain.

R. Zens and H. Ney. 2006. N-gram Posterior Probabilities for Statistical Machine Translation. In *Proceeding of HLT-NAACL Workshop on SMT*, pp. 72-77, NY.