The RWTH Aachen Speech Recognition and Machine Translation System for IWSLT 2012

Stephan Peitz, Saab Mansour, Markus Freitag, Minwei Feng, Matthias Huck
Joern Wuebker, Malte Nuhn, Markus Nußbaum-Thom and Hermann Ney

Human Language Technology and Pattern Recognition Group
Computer Science Department
RWTH Aachen University
Aachen, Germany
<surname>@cs.rwth-aachen.de

Abstract

In this paper, the automatic speech recognition (ASR) and statistical machine translation (SMT) systems of RWTH Aachen University developed for the evaluation campaign of the *International Workshop on Spoken Language Translation* (IWSLT) 2012 are presented. We participated in the ASR (English), MT (English-French, Arabic-English, Chinese-English, German-English) and SLT (English-French) tracks. For the MT track both hierarchical and phrase-based SMT decoders are applied. A number of different techniques are evaluated in the MT and SLT tracks, including domain adaptation via data selection, translation model interpolation, phrase training for hierarchical and phrase-based systems, additional reordering model, word class language model, various Arabic and Chinese segmentation methods, postprocessing of speech recognition output with an SMT system, and system combination. By application of these methods we can show considerable improvements over the respective baseline systems.

1. Introduction

This work describes the automatic speech recognition (ASR) and statistical machine translation (SMT) systems developed by RWTH Aachen University for the evaluation campaign of IWSLT 2012 [1]. We participated in the ASR track, machine translation (MT) track for the language pairs English-French, Arabic-English, Chinese-English, German-English and the spoken language translation (SLT) track. State-of-the-art ASR, phrase-based and hierarchical machine translation systems serve as baseline systems. To improve the MT baselines, we evaluated several different methods in terms of translation performance. We show that phrase training for the phrase-based (forced alignment) as well as for hierarchical approach (forced derivation) can reduce the phrase table size while even improving translation quality. In addition, different word segmentation methods are tested for both Arabic and Chinese as source language. For English as source language, we perform a part-of-speech-based adjective reordering as preprocessing step. System combination is employed in three language pairs of the MT track to improve the translation quality further. Moreover, we investigate the use of the Google Books n-grams. For the SLT track, an SMT system is applied to perform a postprocessing of the given ASR output. This paper is organized as follows. In Section 2 and 3 we describe our ASR system and baseline translation systems. Sections 4 and 5 give an account of the phrase training procedure for the hierarchical phrase-based system and the system combination applied in several MT tasks. Our experiments for each track are summarized in Section 6. We conclude in Section 7.

2. ASR System

The ASR system is based on our English speech recognition system that we also successfully applied in Quaero evaluations [2].

In the acoustic feature extraction, the system computes Mel-frequency cepstral coefficients (MFCC) from the audio signal, which are transformed with a vocal tract length normalization (VTLN). In addition, a voicedness feature is computed. Acoustic context is incorporated by concatenating nine feature vectors in a sliding window. The resulting feature vector is reduced to 45 dimensions by means of a linear discriminant analysis (LDA). Furthermore, bottleneck features derived from a multilayer perceptron (MLP) are concatenated with the feature vector.

The acoustic model is based on hidden Markov models (HMMs) with Gaussian mixture models (GMMs) as emission probabilities. The GMM has a pooled, diagonal covariance matrix. It models 4500 generalized triphones which are derived by a hierarchical clustering procedure (CART). The parameters of the GMM are estimated with the expectation-maximization (EM) algorithm with a splitting procedure according to the maximum likelihood criterion.

The language model is a Kneser-Ney smoothed 4-gram. Several language models are trained on different datasets. The final language model is obtained by linear interpolation.

Table 1: Acoustic training data of ASR system

Corpus	Amount of data [hours]
quaero-2011	268h
hub4+tdt4	393h
epps	102h

Table 2: Language model training data of ASR system

Corpus	Amount of data [running words]
Gigaword 4	2.6B
TED	2.7M
Acoustic transcriptions	5M

The vocabulary of the recognition lexicon is obtained by applying a count-cut-off on the language model data. Each word in the lexicon can have multiple pronunciations. Missing pronunciations are derived with a grapheme-to-phoneme tool.

The recognition is structured in three passes, In the first pass, a speaker independent model is used. The recognition result of the first pass is used for estimating feature transformations for speaker adaptation (CMLLR). The second pass uses the CMLLR transformed features. Finally, a confusion network decoding is performed on the word lattices obtained from the second pass.

The acoustic model of the ASR system is trained on 793 hours of transcribed acoustic data in total, see Table 1. The acoustic training data consists of American broadcast news data (hub4+tdt4), European parliament speeches (epps), and British broadcast conversations (quaero). The MLP is trained on the 268 hours of the quaero corpus only. We use 4500 triphone states and perform eight EM splits, resulting in a GMM with roughly 1.1 million mixture components.

The language model is trained on a large amount of news data (Gigaword), the transcriptions of the audio training data, and a small amount of in-domain data (TED), see Table 2. The recognition lexicon consists of 150k words.

3. Baseline SMT Systems

For the IWSLT 2012 evaluation RWTH utilized state-of-the-art phrase-based and hierarchical translation systems as well as our in-house system combination framework. GIZA++ [3] was employed to train word alignments, all LMs were created with the SRILM toolkit [4] and are standard 4-gram LMs with interpolated modified Kneser-Ney smoothing, unless stated otherwise. We evaluate in truecase, using the BLEU [5] and TER [6] measures.

3.1. Phrase-based Systems

For the phrase-based SMT systems, we used in this work both an in-house implementation of the state-of-the-art MT decoder (PBT) described in [7] and the implementation of the decoder based on [8] (SCSS) which is part of RWTH's open-source SMT toolkit Jane 2.1 [1]. We use the standard set of models with phrase translation probabilities and lexical smoothing in both directions, word and phrase penalty, distance-based reordering model, an *n*-gram target language model and three binary count features. The parameter weights are optimized with MERT [9] (SCSS, HPBT) or the downhill simplex algorithm [10] (PBT).

3.2. Hierarchical Phrase-based System

For our hierarchical setups, we employed the open source translation toolkit Jane [11], which has been developed at RWTH and is freely available for non-commercial use. In hierarchical phrase-based translation [12], a weighted synchronous context-free grammar is induced from parallel text. In addition to contiguous *lexical* phrases, *hierarchical* phrases with up to two gaps are extracted. The search is carried out with a parsing-based procedure. The standard models integrated into our Jane systems are: phrase translation probabilities and lexical smoothing probabilities in both translation directions, word and phrase penalty, binary features marking hierarchical phrases, glue rule, and rules with non-terminals at the boundaries, four binary count features, phrase length ratios and an *n*-gram language model. Optional additional models are IBM model 1 [13], discriminative word lexicon (DWL) models, triplet lexicon models [14], a discriminative reordering model [15] and several syntactic enhancements like preference grammars and string-to-dependency features [16]. We utilize the cube pruning algorithm [17] for decoding and optimize the model weights with standard MERT [9] on 100-best lists.

4. Forced Derivation

As proposed in [18], an alternative to the heuristic phrase extraction from word-aligned data is to train the phrase table with an EM-inspired algorithm. Since in [18] a phrase table for a phrase-based system was learned, we employed the idea of force-aligning the training data on a hierarchical phrase-based setup [19]. Instead of applying a modified version of the decoder, a synchronous parsing algorithm based on two successive monolingual parses is performed. The idea of the two-parse algorithm is to first parse the source sentence. Then, phrases extracted from the source parse tree are used to parse the target sentence. After parsing, we apply the inside-outside algorithm on the generated target parse tree to compute expected counts for each applied phrase. Using the expected counts, we update the phrase probabilities and apply a threshold pruning on the phrase table. Leave-one-out

[1] http://www-i6.informatik.rwth-aachen.de/jane/

Table 3: Forced Derivation (FD) results for the MT task English-French including phrase table (PT) size.

system	dev		test		PT size
	BLEU	TER	BLEU	TER	# phrases
baseline	27.4	56.9	30.4	51.2	72M
FD	27.6	56.6	30.5	51.3	8.7M

is applied to counteract over-fitting effects. We tested this procedure on the English-French MT task. The results are shown in Table 3. The phrase table size was reduced by 88% without hurting performance.

5. System Combination

System combination is used to produce consensus translations from multiple hypotheses generated with different translation engines. System combination can be divided into two steps. The first step produces a word to word alignment for the given single system hypotheses. In a second step a confusion network is constructed. Then, the hypothesis with the highest probability is extracted from this confusion network. For the alignment procedure, we have to choose one of the given single system hypotheses as primary system. To this primary system all other hypotheses are aligned and thus the primary system defines the word order. In Figure 1 a system combination of four different system is shown. We select the bold hypothesis as primary hypothesis. The other hypotheses are aligned to the primary using the METEOR [20] alignment. The resulting hypotheses have different word lengths and thus it is possible to align a word to an empty word marked as $. Once the alignment is given, we are able to built a confusion network. As the hypotheses consist of different words and may have different sentence length, the unaligned words could produce incorrect arcs. To fix the incorrect arcs, we introduce a reordering model based on the language model scores of the given adjacent incorrect arcs. For unaligned parts, we take the hypothesis with the highest language model score and align the unaligned parts of all hypotheses to that one. As result we get a more meaningful confusion network. In Figure 1 different confusion networks with and without the reordering model are shown. A more compact representation of the confusion network is given in Figure 2.

As choosing a primary hypothesis is a hard decision, we build for each hypothesis as primary system one confusion network. To combine these different networks, we just use the Union operation from the automata theory. The next step is to extract the most probably translation from the confusion network. Each arc in the confusion network is rescored with different statistical models as word or phrase counts of the single systems, a language model score, a word penalty and a binary feature which marks the primary system of the partial confusion network. We give each model a weight and

system hypotheses	**this is it**						
	that was future						
	this is in the future						
	future is this						
alignment	that|**this** was|**is** $\|**it** future\|**$**						
	this|**this** is|**is** $\|**it** in\|**$** the|**$** future\|**$**						
	future|**$** this|**this** is|**is** $|**it**						
confusion network	**$**	**this**	**is**	**it**	**$**	**$**	**$**
	$	that	was	$	future	$	$
	$	this	is	$	in	the	future
	future	that	is	$	$	$	$
reordering of unaligned words	**$**	**this**	**is**	**it**	**$**	**$**	**$**
	$	that	was	$	$	$	future
	$	this	is	$	in	the	future
	$	this	is	$	$	$	future

Figure 1: Example for system combination of four different hypotheses.

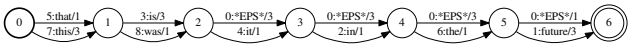

Figure 2: Confusion network of four different hypotheses.

Table 4: System combination results for the MT tasks English-French (en-fr), Arabic-English (ar-en) and Chinese-English (zh-en).

system		tst2010	
		BLEU	TER
en-fr	best single system	32.0	50.1
	system combination	32.9	42.9
ar-en	best single system	27.1	54.4
	system combination	28.0	53.4
zh-en	best single system	14.7	74.5
	system combination	15.4	74.1

combine them in a log-linear model. The weights can be optimized with MERT and the translation with the best score within the lattice is the consensus translation.

By applying system combination in the English-French, Arabic-English and Chinese-English MT task, we achieve improvements of up to +0.9 points in BLEU and up to -1.0 points in TER.

6. Experimental Evaluation

6.1. Automatic Speech Recognition

In Table 5 we compare the word error rate (WER) of the three different passes. A lower WER indicates a better recognition quality. We achieve an improvement of 2.5 points in WER by applying the second pass. Furthermore, the confusion network decoding improves the recognition by 0.2 points.

Table 5: Results of the English ASR task. Our ASR system is incrementally improved with each pass.

	dev2010	**tst2010**
pass 1	20.0	18.4
pass 2	17.5	15.9
cn-decoding	17.3	15.7

6.2. English-French

For the English-French task, RWTH employed both phrase-based decoders (SCSS, PBT), different hierarchical phrase-based systems (HPBT) and a system combination of the best setups. All experimental results are given in Table 6.

The SCSS baseline system is trained on the in-domain data (TED) [21]. For this baseline, we achieve the biggest improvement by training an additional translation model on the available out-of-domain data (+1.1% BLEU). The system is further improved by applying part-of-speech-based adjective reordering rules as preprocessing step [22] (+0.3% BLEU) and a 7-gram word class language model (+0.3% BLEU).

For the PBT setups, the baseline is a system trained on all available data (allData). By adding phrase-level discriminative word lexicons [14] (DWL) and a reordering model, which distinguishes monotone, swap, and discontinuous phrase orientations [23, 24] (MSD-RO), the baseline system is improved by 0.9 points in BLEU and 0.7 points in TER.

The HPBT baseline is trained on the in-domain data. By limiting the recursion depth for the hierarchical rules with a shallow-1 grammar [25], we achieve an improvement of 0.6 points in BLEU. The bigger language model is trained on the target part of the bilingual corpus, the Shuffled News data and the 10^9 and French Gigaword corpora. As for the SCSS system, we trained an additional phrase table on the out-of-domain data. All in all, we are able to improve the HPBT baseline by +2.3% BLEU and -1.8% TER.

To increase the translation quality further, we employed system combination as described in Section 5 on several systems including the last year's primary submission (HPBT.2011). We gain an enhancement of 0.9 points in BLEU and 0.7 points in TER compared to the best single system. Compared to the last year's submission on the 2011 evaluation set, we could improve our best single system by 1.6 points in BLEU and 1.8 points in TER and further 1.0% BLEU with system combination (Table 7).

6.2.1. Google Books n-grams

For the English-French translation task we also investigated upon using the Google Books n-grams [26] which is a collection of n-gram counts extracted from digitized books. These counts are categorized by language and publication year of the books containing the n-grams. Selecting a range of years and using the vanilla n-grams resulted in language models with very high perplexities: The preprocessing steps applied to the underlying corpus do not match the preprocessing used in our system. By adapting the vanilla n-grams reasonable perplexities were obtained. We could further improve the language model by selecting only n-grams from books published in the last few years.

Our final language model uses 4-grams obtained from the

Table 6: Results for the English-French MT task. The open-source phrase-based decoder (SCSS) is incrementally augmented with a second translation model trained on out-of-domain data (*oodDataTM*), adjective-reordering as preprocessing step (*adj-reordering*) and a word class language model (*WordClassLM*). The in-house phrase-based decoder (PBT) is trained on all available bilingual data (allData) and incrementally augmented with a discriminative word lexicon (*DWL*) and an additional reordering model (*MSD-RO*). The hierarchical phrase-based decoder (HPBT) is incrementally augmented with a shallow-1 grammar (*shallow*), a bigger language model (*bigLM*), an alternative lexical smoothing (*IBM-1*), forced derivation (*FD*) and a second translation model trained on out-of-domain data (*oodDataTM*). The primary submission is a system combination of all systems marked with *.

system	**dev2010**		**tst2010**		
	BLEU	TER	BLEU	TER	
SCSS TED	25.9	58.3	29.3	52.1	
+oodDataTM	28.2	56.1	31.4	50.9	
+adj-reordering	28.2	56.4	31.7	50.5	*
+WordClassLM	28.3	56.0	32.0	50.1	*
PBT allData	27.9	55.8	30.9	50.6	*
+DWL	28.0	56.1	31.6	50.3	*
+MSD-RO	28.1	55.8	31.8	49.9	*
HPBT TED	25.7	58.6	29.0	52.8	
+shallow	26.6	57.8	29.6	52.0	
+bigLM	26.8	57.6	30.2	51.7	
+IBM-1	27.4	56.9	30.4	51.2	*
+FD	27.6	56.6	30.5	51.3	*
+oodDataTM	27.7	56.5	31.3	51.0	*
HPBT.2011	27.4	57.0	31.1	50.7	*
system combination	29.5	54.9	32.9	49.2	

Table 7: Comparison of 2011 and 2012 English-French task submission on tst2011.

submission	**tst2011**	
	BLEU	TER
2011 (single system)	36.1	43.8
2012 (best single system)	37.7	42.0
2012 (system combination)	38.7	40.9

Google Books n-grams which are mixed with our previously described language model. The resulting language model has a perplexity of 81.4 on our development set which compares to a perplexity of 85.0 of the original language model. However, we did not use the improved language model in our final system since very small to no increase in translation quality was observed whereas the language model size was increased. We believe that the combination of mismatch in preprocessing, OCR errors and the very broad domain of the Google Books n-grams lead to the rather small improvements. It should be noted that a newer version of the Google Books n-grams [27] is available that was not available during the time of work.

6.3. Arabic-English

RWTH participated last year in the Arabic-English TED task, achieving the best automatic results in the evaluation. This year, the architecture of the Arabic-English system is similar to last year, where a system combination is performed over different systems with differing Arabic segmentation methods. The differences from last year include: larger bilingual in-domain training data (130K versus 90K last year), the inclusion of the English Gigaword for language-modeling, and phrase table interpolation. We experimented with linear phrase table interpolation, where the phrase probabilities in both directions are interpolated linearly with a fixed weight optimized on the development set. We created two phrase tables, one using the TED in-domain and the other using the UN corpus, and interpolated them with a weight of 0.9 for the TED phrase table. The interpolation resulted in 1% BLEU improvement over a system using a phrase table trained over the full data.

The different segmentation methods are similar to last year, and include:

FST A finite state transducer-based approach introduced and implemented by [28]. The segmentation rules are encoded within an FST framework.

SVM A reimplementation of [29], where an SVM framework is used to classify each character whether it marks the beginning of a new segment or not.

CRF An implementation of a CRF classifier similar to the SVM counterpart. We use CRF++[2] to implement the method.

MorphTagger An HMM-based Part-Of-Speech (POS) tagger implemented upon the SRILM toolkit [30].

MADA v3.1 An off-the-shelf tool for Arabic segmentation [31]. We use the following schemes: D1,D2,D3 and ATB (TB), which differ by the granularity of the segmentation.

[2]http://crfpp.sourceforge.net/

Table 8: Arabic-English results on the test set (tst2010) for different segmentations, comparing 2011 and 2012 systems. *MADA-TB ALL* is a system using unfiltered bilingual data. The primary submission is a system combination of all the listed systems.

system	**2011**		**2012**	
	BLEU	TER	BLEU	TER
FST	25.1	57.0	26.5	55.8
SVM	25.4	57.4	26.6	54.4
HMM	25.7	56.9	26.9	55.1
CRF	25.7	56.7	26.9	54.5
MADA-D1	24.7	57.1	26.3	55.4
MADA-D2	25.2	57.1	26.9	54.7
MADA-D3	25.4	57.1	27.0	54.0
MADA-TB	26.1	56.4	-	-
MADA-TB ALL	26.1	56.6	27.1	54.4
system combination	27.0	54.7	28.0	53.4

As in last year, adaptation using filtering is done for both LM training and TM training. To build the LM, we use a mixture of all available English corpora, where News Shuffle, giga-fren.en and the English Gigaword are filtered. For translation model filtering, we use the combined IBM-1 and LM cross-entropy scores. We perform filtering for the MultiUN corpus, selecting $\frac{1}{16}$ of the sentences (400K). Due to the different Arabic segmentations we utilize, we performed the sentence selection only once over the MADA-TB method, and used the same selection for all other setups.

We trained phrase-based systems for all different segmentation schemes using the interpolation of TED and the 400K selected portion of the UN corpus. Additionally, one system was trained on all available data, preprocessed with MADA-TB. The results are summarized in Table 8. The table includes a comparison between the 2011 and 2012 systems on the test set. This year systems clearly improves over last year, with improvements ranging from 1% up-to 1.7% BLEU. The single system *MADA-TB ALL* of 2012 performs similarly to the system-combination submission of 2011. The final system combination improves over last year submission with +1% BLEU and -1.3% TER.

6.4. Chinese-English

Results of Chinese-English systems are given in Table 9. The system combination in Table 9 is RWTH's primary submission. The system combination was done as follows. We use both a phrase-based decoder [7] and a hierarchical phrase-based decoder Jane [11]. For each of the two decoders we do a bi-directional translation, which means the system performs standard direction decoding (left-to-right) and reverse direction decoding (right-to-left). We thereby obtain a total of four different translations.

Table 9: Chinese-English results on the dev test set for different segmentations. The primary submission is a system combination of all the listed systems.

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
PBT	12.2	80.0	14.2	73.7
PBT-reverse	11.9	79.6	13.7	74.3
HPBT	12.7	80.0	14.7	74.5
HPBT-reverse	12.8	81.0	14.5	76.2
HPBT-withUN-a	12.1	81.4	14.1	76.0
HPBT-withUN-b	12.5	80.4	14.0	75.5
system combination	13.7	78.9	15.4	74.1

To build the reverse direction system, we used exactly the same data as the standard direction system and simply reversed the word order of the bilingual corpora. For example, the bilingual sentence pair "今天␣是␣星期天␣。 ||Today␣is␣Sunday␣." is now transformed to "。␣星期天␣是␣今天||.␣Sunday␣is␣Today". With the reversed corpora, we then trained the alignment, the language model and our translation systems in the exactly same way as the normal direction system. For decoding, the test corpus is also reversed. The idea of utilizing right-to-left decoding has been proposed by [32] and [33] where they try to combine the advantages of both of the left-to-right and right-to-left decoding with a bidirectional decoding method. We also try to gain benefits from two-direction decoding, however, we use a system combination to achieve this goal.

In Table 9, first four systems do not use UN data. For *HPBT-withUN-a* and *HPBT-withUN-b* we additionally select 800k bilingual sentences from UN. *HPBT-withUN-a* and *HPBT-withUN-b* are built using the same setup but with differently optimized feature weights. PBT-reverse is the reverse system of PBT. HPBT-reverse is the reverse system of HPBT. *HPBT-withUN-a* and *HPBT-withUN-b* are trained with normal the left-to-right direction. From the results we draw the conclusions: HPBT performs better than PBT; UN data does not help; system combination of the six systems gets the best result.

6.5. German-English

For the German-English task, RWTH submitted a phrase-based system which is extended by several state-of-the-art improvements. In a preprocessing step, the German source is decompounded [34] and part-of-speech-based long-range verb reordering rules [22] are applied. The baseline uses a 4-gram language model trained on the target side of the bilingual data. When using additional monolingual data, we perform data selection as described in [35].

The results are given in Table 10. We created two baselines, one trained on all available bilingual data, one trained on the in-domain TED data only. The pure in-domain system clearly outperforms the general system on the TED data sets. This baseline is improved by forced-alignment phrase training (+0.1% BLEU) [18], adding $\frac{1}{4}$ of the Shuffled News data (+0.7% BLEU), a 7-gram word class language model (+0.6% BLEU), a second translation model trained on all available out-of-domain data (+0.5% BLEU) and finally by adding $\frac{1}{8}$ of each of the 10^9 and Gigaword corpora to the LM training data (+0.5% BLEU).

Table 10: Results for the German-English MT task. The phrase-based decoder (SCSS) trained on TED data is incrementally augmented with forced alignment phrase training (FA), additional monolingual data (ShuffledNews, Gigaword), a word class language model (WordClassLM) and a second translation model trained on out-of-domain data (oodDataTM).

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
SCSS allData	29.0	49.5	27.5	51.6
SCSS TED	29.9	48.4	28.4	50.3
+FA	30.3	47.7	28.5	49.9
+ShuffledNews	31.1	47.9	29.2	50.2
+WordClassLM	31.2	47.8	29.8	49.7
+oodDataTM	31.9	47.4	30.3	49.3
+Gigaword	32.6	46.4	30.8	48.6

6.6. Spoken Language Translation (SLT)

The input for the translation systems in the SLT track is the automatic transcription provided by the automatic speech recognition track. In this work, we used the recognitions of our ASR system described in Section 2. Due to the fact that the output of the ASR system does not provide punctuation marks or case information and contains recognition errors, we have to adapt the standard text translation system used in the English-French MT track.

Firstly, as described in [36], we trained a translation system on data without punctuation marks and case information in the source language, but including punctuation and casing in the target language. By translating ASR output with such a system, punctuation and case information are predicted during the translation process. We denote this as IMPLICIT.

As a second approach an SMT system was trained on a corpus with ASR output as source language data and the corresponding manual transcription as target language data, i.e. we interpret the postprocessing of the ASR output as machine translation [37]. We denote this as POSTPROCESSING. In order to built such a corpus we recognized the provided talks with our ASR system. On this corpus a standard phrase-based SMT was trained. During the translation of the ASR output punctuation and case information are restored. The output of this SMT system is the input of a standard text

translation system.

Table 11: Comparison between the methods IMPLICIT and POSTPROCESSING on the SLT task English-French (IWSLT 2012).

system	dev2010		tst2010	
	BLEU	TER	BLEU	TER
IMPLICIT	19.2	67.8	22.5	61.6
POSTPROCESSING	20.1	67.2	23.4	60.7

In Table 11, we compare the IMPLICIT method with our second approach (POSTPROCESSING). Note, for the experiments we utilized the best single system of the MT English-French track. POSTPROCESSING outperforms IMPLICIT and we achieve an improvement of 0.9 points in BLEU and 0.9 points in TER.

7. Conclusion

RWTH participated in ASR, MT (English-French, Arabic-English, Chinese-English, German-English) and SLT tracks of the IWSLT 2012 evaluation campaign.

Considerable improvements over respective baseline systems were achieved by applying several different techniques.

For the MT track, among these are phrase training for the phrase-based as well as for the hierarchical system, an additional reordering model, word class language model, data filtering techniques, phrase table interpolation, and different Arabic and Chinese segmentation tools. To improve the SLT system, postprocessing of the ASR output is modelled as machine translation. By system combination, additional improvements of the best single system were achieved.

8. Acknowledgements

This work was partly achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation. The research leading to these results has also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287658.

9. References

[1] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK, December 2012.

[2] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A. El-Desoky Mousa, S. Hahn, D. Nolden, R. Schlüter, and H. Ney, "The RWTH 2010 Quaero ASR evaluation system for English, French, and German," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP*, 2011, pp. 2212–2215.

[3] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, Mar. 2003.

[4] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, vol. 2, Denver, CO, Sept. 2002, pp. 901–904.

[5] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318.

[6] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, Massachusetts, USA, August 2006, pp. 223–231.

[7] R. Zens and H. Ney, "Improvements in Dynamic Programming Beam Search for Phrase-based Statistical Machine Translation," in *International Workshop on Spoken Language Translation*, Honolulu, Hawaii, Oct. 2008, pp. 195–205.

[8] J. Wuebker, M. Huck, S. Peitz, M. Nuhn, M. Freitag, J.-T. Peter, S. Mansour, and H. Ney, "Jane 2: Open source phrase-based and hierarchical statistical machine translation," in *International Conference on Computational Linguistics*, Mumbai, India, Dec. 2012, to appear.

[9] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.

[10] J. A. Nelder and R. Mead, "A Simplex Method for Function Minimization," *The Computer Journal*, vol. 7, pp. 308–313, 1965.

[11] D. Vilar, D. Stein, M. Huck, and H. Ney, "Jane: Open source hierarchical translation, extended with reordering and lexicon models," in *ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, Sweden, July 2010, pp. 262–270.

[12] D. Chiang, "Hierarchical Phrase-Based Translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[13] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, June 1993.

[14] A. Mauser, S. Hasan, and H. Ney, "Extending statistical machine translation with discriminative and trigger-based lexicon models," in *Conference on Empirical Methods in Natural Language Processing*, Singapore, Aug. 2009, pp. 210–217.

[15] R. Zens and H. Ney, "Discriminative Reordering Models for Statistical Machine Translation," in *Human Language Technology Conf. (HLT-NAACL): Proc. Workshop on Statistical Machine Translation*, New York City, NY, June 2006, pp. 55–63.

[16] D. Stein, S. Peitz, D. Vilar, and H. Ney, "A Cocktail of Deep Syntactic Features for Hierarchical Machine Translation," in *Conf. of the Association for Machine Translation in the Americas (AMTA)*, Denver, CO, Oct./Nov. 2010.

[17] L. Huang and D. Chiang, "Forest Rescoring: Faster Decoding with Integrated Language Models," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague, Czech Republic, June 2007, pp. 144–151.

[18] J. Wuebker, A. Mauser, and H. Ney, "Training phrase translation models with leaving-one-out," in *Proceedings of the 48th Annual Meeting of the Assoc. for Computational Linguistics*, Uppsala, Sweden, July 2010, pp. 475–484.

[19] S. Peitz, A. Mauser, J. Wuebker, and H. Ney, "Forced derivations for hierarchical machine translation," in *International Conference on Computational Linguistics*, Mumbai, India, Dec. 2012, to appear.

[20] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," Prague, Czech Republic, June 2007, pp. 228–231.

[21] M. Cettolo, C. Girardi, and M. Federico, "Wit^3: Web inventory of transcribed and translated talks," in *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[22] M. Popović and H. Ney, "POS-based Word Reorderings for Statistical Machine Translation," in *International Conference on Language Resources and Evaluation*, 2006, pp. 1278–1283.

[23] C. Tillmann, "A Unigram Orientation Model for Statistical Machine Translation," in *Proc. of the HLT-NAACL: Short Papers*, 2004, pp. 101–104.

[24] 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. Constantine, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," Prague, Czech Republic, June 2007, pp. 177–180.

[25] A. de Gispert, G. Iglesias, G. Blackwood, E. R. Banga, and W. Byrne, "Hierarchical Phrase-Based Translation with Weighted Finite-State Transducers and Shallow-n Grammars," *Computational Linguistics*, vol. 36, no. 3, pp. 505—-533, 2010.

[26] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. Nowak, and E. Lieberman-Aiden, "Quantitative analysis of culture using millions of digitized books," *Science*, vol. 331, pp. 176–182, 2011.

[27] Y. Lin, J.-B. Michel, E. Aiden Lieberman, J. Orwant, W. Brockman, and S. Petrov, "Syntactic annotations for the google books ngram corpus," in *Proceedings of the ACL 2012 System Demonstrations*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 169–174.

[28] A. El Isbihani, S. Khadivi, O. Bender, and H. Ney, "Morpho-syntactic Arabic Preprocessing for Arabic to English Statistical Machine Translation," in *Proceedings on the Workshop on Statistical Machine Translation*, New York City, June 2006, pp. 15–22.

[29] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," in *HLT-NAACL 2004: Short Papers*, D. M. S. Dumais and S. Roukos, Eds., Boston, Massachusetts, USA, May 2 - May 7 2004, pp. 149–152.

[30] S. Mansour, "Morphtagger: Hmm-based arabic segmentation for statistical machine translation," in *International Workshop on Spoken Language Translation*, Paris, France, December 2010, pp. 321–327.

[31] R. Roth, O. Rambow, N. Habash, M. Diab, and C. Rudin, "Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking," in *Proceedings of ACL-08: HLT, Short Papers*, Columbus, Ohio, June 2008, pp. 117–120.

[32] T. Watanabe and E. Sumita, "Bidirectional decoding for statistical machine translation," in *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, ser. COLING '02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1–7.

[33] A. Finch and E. Sumita, "Bidirectional phrase-based statistical machine translation," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, ser. EMNLP '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 1124–1132.

[34] P. Koehn and K. Knight, "Empirical Methods for Compound Splitting," in *Proceedings of European Chapter of the ACL (EACL 2009)*, 2003, pp. 187–194.

[35] R. Moore and W. Lewis, "Intelligent Selection of Language Model Training Data," in *ACL (Short Papers)*, Uppsala, Sweden, July 2010, pp. 220–224.

[36] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *International Workshop on Spoken Language Translation*, Kyoto, Japan, Nov. 2006, pp. 158–165.

[37] S. Peitz, S. Wiesler, M. Nussbaum-Thom, and H. Ney, "Spoken language translation using automatically transcribed text in training," in *International Workshop on Spoken Language Translation*, Hongkong, Dec. 2012, to appear.