The RWTH Aachen Machine Translation System for IWSLT 2011

*Joern Wuebker, Matthias Huck, Saab Mansour, Markus Freitag, Minwei Feng,
Stephan Peitz, Christoph Schmidt 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 statistical machine translation (SMT) systems of RWTH Aachen University developed for the evaluation campaign of the *International Workshop on Spoken Language Translation* (IWSLT) 2011 is presented. We participated in the MT (English-French, Arabic-English, Chinese-English) and SLT (English-French) tracks. Both hierarchical and phrase-based SMT decoders are applied. A number of different techniques are evaluated, including domain adaptation via monolingual and bilingual data selection, phrase training, different lexical smoothing methods, additional reordering models for the hierarchical system, various Arabic and Chinese segmentation methods, punctuation prediction for speech recognition output, and system combination. By application of these methods we can show considerable improvements over the respective baseline systems.

1. Introduction

This work describes the SMT systems developed by RWTH Aachen University for the evaluation campaign of IWSLT 2011 [1]. We participated in the machine translation (MT) track for all three language pairs and the spoken language translation (SLT) track. State-of-the-art phrase-based and hierarchical machine translation systems serve as baseline systems. To improve these baselines, we evaluated several different methods in terms of translation performance as well as efficiency.

We show that monolingual data selection can be used to adapt a translation system to a specific domain while at the same time reducing language model (LM) size. A similar approach is used for bilingual data filtering, which can decrease model size considerably without hurting performance in our experiments. Further, our statistical phrase training technique is also shown to yield a significant reduction in phrase table size on all three language pairs while moderately improving translation quality on two of them. In the hierarchical system, several different lexical smoothing methods as well as additional reordering models are evaluated. For the SLT track, we compare five different kinds of punctuation prediction, including the application of a monotone phrase-based translation decoder as prediction engine. Additionally, different word segmentation methods are tested for both Arabic and Chinese as source language.

This paper is organized as follows. In Section 2 we describe our baseline translation systems. Sections 3 and 4 give an account of the different data selection techniques and the phrase training procedure. Our experiments for each language pair including the applied novel methods are summarized in Section 5. We conclude in Section 6.

2. Baseline SMT systems

For the IWSLT 2011 evaluation RWTH utilized state-of-the-art phrase-based and hierarchical translation systems as well as our in-house system combination framework. GIZA++ [2] was employed to train word alignments, all LMs were created with the SRILM toolkit [3] and are standard 4-gram LMs with interpolated modified Kneser-Ney smoothing, unless stated otherwise. We evaluate in truecase, using the BLEU [%] [4] and TER [%] [5] measures.

2.1. Phrase-based system

The phrase-based SMT system used in this work is an in-house implementation of the state-of-the-art MT decoder described in [6]. 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 [7] or the downhill simplex algorithm [8].

2.2. Hierarchical phrase-based system

For our hierarchical setups, we employed the open source translation toolkit Jane [9], which has been developed at RWTH and is freely available for non-commercial use. In hierarchical phrase-based translation [10], a weighted synchronous context-free grammar is induced from parallel text. In addition to contiguous *lexical* phrases, *hierarchi-*

Table 1: Results for monolingual data selection on the English-French MT task. *Selection* denotes the selected fraction of the shuffled news data for LM training. *Shuffled news* denotes the LM trained on the selected data only, *combined* the LM trained additionally on TED, Europarl and news commentary data sets, which is used in translation. Perplexity (ppl) is computed on **dev**.

selection	**ppl**		**LM size**	**dev**		**test**	
	shuffled news	combined	# *n*-grams	BLEU	TER	BLEU	TER
full	133.3	85.4	106M	25.5	58.6	29.2	52.4
1/2	120.6	84.4	73.6M	25.7	58.5	29.2	52.4
1/4	111.1	83.9	44.5M	25.9	58.4	29.5	52.2
1/8	107.4	84.3	31.2M	25.7	58.5	29.5	51.9
1/16	110.5	86.6	20.3M	25.4	58.9	29.2	52.5
no	-	88.6	14.4M	25.0	59.3	28.5	52.7

cal 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 [11], discriminative word lexicon (DWL) models, triplet lexicon models [12], a discriminative reordering model [13] and several syntactic enhancements like preference grammars and string-to-dependency features [14]. We utilize the cube pruning algorithm [15] for decoding and optimize the model weights with standard MERT [7] on 100-best lists.

2.3. System combination

System combination is used to produce consensus translations from multiple hypotheses generated with different translation engines. The basic concept of RWTH's approach to machine translation system combination is described in [16, 17]. This approach includes an enhanced alignment and reordering framework. A lattice is built from the input hypotheses. The translation with the best score within the lattice according to a couple of statistical models is selected as consensus translation.

3. Domain-specific data selection

3.1. Monolingual data

To reduce the size of our language models (LMs) and adapt them to the domain of the TED talks, we apply the data selection technique introduced in [18]. Starting point are a small in-domain data corpus and a large out-of-domain corpus. Each sentence of the out-of-domain corpus is scored by the difference in cross-entropy between an LM trained from the in-domain data and an LM trained from a similar-sized sample of the out-of-domain data. A threshold value on this score decides whether a sentence will be selected for language model training. In this work we use 2-gram LMs for computation of the cross-entropy and 4-gram LMs for translation.

Table 2: Comparison of the two bilingual data filtering methods on the Arabic-English MT task. *LM-filter* denotes the selection technique based on LM cross-entropy, *combi-filter* denotes the novel combined LM and IBM-1 cross-entropy-based method. For each, the 400K best sentences are selected from the MultiUN data. *All UN* denotes using the full MultiUN corpus. Phrase table (PT) size is given in number of phrases.

system	**PT**	**dev**		**test**	
	# phr.	BLEU	TER	BLEU	TER
TED-only	6M	27.4	54.1	25.2	57.3
+LM-filter	38M	28.7	52.6	25.7	56.7
+combi-filter	32M	28.6	52.8	26.1	56.4
+all UN	387M	28.6	52.9	26.1	56.6

Table 1 shows the effect of the monolingual data selection on the resulting French LMs. Here, the TED data serves as in-domain corpus, the shuffled news data as out-of-domain corpus. We can see that by selecting $\frac{1}{8}$ of the shuffled news data we can reduce perplexity from 133.3 for the complete data set to 107.4. When additionally using the in-domain TED talks as well as the Europarl and news commentary data for LM training (combined LM), the difference is less pronounced, while the LM sizes can be reduced considerably from 106M *n*-grams to 44.5M *n*-grams for the selection of $\frac{1}{4}$ of the shuffled news data.

For the translation experiments, we used a phrase-based system trained on TED data only. When applying the combined LMs, we observe that selecting $\frac{1}{4}$ of the data leads to a moderate improvement of 0.4% BLEU on **dev** and 0.3% BLEU on **test** over the full LM. This indicates that the intended domain adaptation effect is achieved.

Table 3: Phrase training (Forced Alignment, FA) results for the MT tasks English-French (en-fr), Arabic-English (ar-en) and Chinese-English (zh-en), including phrase table (PT) size and translation speed on the **dev** set.

system		**dev**		**test**		**PT size**	**speed**
		BLEU	TER	BLEU	TER	# phrases	words/sec
en-fr	baseline	25.5	58.9	29.3	52.1	151M	6.3
	FA	25.8	58.2	29.4	51.8	16.6M	8.4
ar-en	baseline	27.6	53.5	25.2	57.2	22.1M	6.1
	FA	27.7	53.5	25.3	57.1	3.73M	8.8
zh-en	baseline	8.0	85.3	8.7	81.1	10.9M	7.8
	FA	7.7	85.4	8.5	80.8	1.58M	11.6

3.2. Bilingual data

[19] propose to apply the selection technique described in Section 3.1 to bilingual training data to perform adaptation of the translation model. We adapted and generalized this approach by combining the LM cross-entropy difference with an IBM model 1 (IBM-1) cross-entropy difference [20]. We summarize the bilingual data selection results in Table 2. The table includes a comparison between systems trained on *TED-only*, *TED +all UN* (with all MultiUN data), *TED +LM-filter* (with LM cross-entropy-based data selection) and *TED +combined-filter* (combined LM and IBM-1 cross-entropy-based data selection). The selection methods extract the top 400K sentences from the MultiUN corpus. The LM used for translation in these experiments is the same as in the final evaluation system and applies the monolingual data selection technique described in Section 3.1. We can see that the system trained on the data selected by the combined method performs equal to using the full data, while the phrase table size is reduced by a factor of 12.

4. Phrase training

As an alternative to the heuristic phrase extraction from word-aligned data, [21] propose to train the phrase table with a procedure similar to the EM algorithm. This is done by force-aligning the training data with a modified version of the translation decoder. Leave-one-out is applied to counteract over-fitting effects. We tested this procedure on all three language pairs with our phrase-based translation system. In this work, we apply the *count model* described in [21] with an *n*-best list size of n=100. In addition to the TED talks, the Europarl and news commentary data for English-French and a selection of 200k sentences of the Multi-UN data for Arabic-English were used for training. The results are shown in Table 3. A clear reduction in phrase table size by at least 83% can be observed on all tasks. This results in an increased translation speed of 33% for English-French, 44% for Arabic-English and 49% for Chinese-English. Translation performance improves slightly on the English-French task, shows nearly no change on Arabic-English and a small degradation on Chinese-English. On the former, the original performance is already very low, so we can assume that the phrase alignments produced in training are of inferior quality.

Table 4: Data statistics for the preprocessed parallel training corpora for the English-French (en-fr), Arabic-English (ar-en) and Chinese-English (zh-en) MT tasks. The corpora include TED, Europarl and news commentary for en-fr, TED and 400K sentences selected from MultiUN for ar-en and TED only for zh-en. In the corpora, numerical quantities are replaced by a single category symbol. The ar-en statistics refer to the MADA-TB segmentation scheme, the zh-en statistics to the ldc segmentation.

en-fr	English	French
Sentences	2.0M	
Running words	54.3M	59.9M
Vocabulary	136K	159K
ar-en	Arabic	English
Sentences	512K	
Running words	11.7M	11.6M
Vocabulary	93K	61K
zh-en	Chinese	English
Sentences	105K	
Running words	1.98M	2.04M
Vocabulary	29K	37K

5. Experimental evaluation

5.1. English-French

For the English-French task, the translation models are trained on the TED, Europarl and news commentary data. Statistics on the bilingual data are shown in Table 4. The LMs used on this task are trained on the shuffled news data in addition to the target part of the bilingual training data. We concentrate on the hierarchical decoder after some initial experiments showing that it is slightly superior to the phrase-based paradigm. The hierarchical (HPBT) baseline system is a setup including the standard models as listed in

Table 5: Results for the English-French MT task. The hierarchical phrase-based decoder (HPBT) is incrementally augmented with monolingual data selection (*mooreLM*), alternative lexical smoothing (*IBM-1*, *DWL*), improved LM smoothing (*opt. KN LM*), phrase table and triplet-based adaptation (*TED TM*, *s2t TED triplets*) and additional reordering models (*discrim. RO*).

system	**dev**		**test**	
	BLEU	TER	BLEU	TER
HPBT	25.7	58.6	29.3	52.8
+mooreLM	26.0	58.1	29.6	51.8
+IBM-1	26.3	58.1	30.0	52.0
+DWL	26.3	58.0	30.2	51.8
+opt. KN LM	26.5	57.9	30.3	51.3
+TED TM	27.2	57.2	30.7	51.1
+s2t TED triplets	27.5	57.0	30.8	50.9
+discrim. RO	27.4	57.0	31.1	50.7

Section 2.2. We limit the recursion depth for hierarchical rules with a shallow-1 grammar [22].

In a shallow-1 grammar, the generic non-terminal *X* of the standard hierarchical approach is replaced by two distinct non-terminals *XH* and *XP*. By changing the left-hand sides of the rules, lexical phrases are allowed to be derived from *XP* only, hierarchical phrases from *XH* only. On all right-hand sides of hierarchical rules, the *X* is replaced by *XP*. Gaps within hierarchical phrases can thus solely be filled with purely lexicalized phrases, but not a second time with hierarchical phrases. The initial rule is substituted with

$$\begin{aligned} S &\rightarrow \langle XP^{\sim 0}, XP^{\sim 0} \rangle \\ S &\rightarrow \langle XH^{\sim 0}, XH^{\sim 0} \rangle , \end{aligned} \tag{1}$$

and the glue rule is substituted with

$$\begin{aligned} S &\rightarrow \langle S^{\sim 0} XP^{\sim 1}, S^{\sim 0} XP^{\sim 1} \rangle \\ S &\rightarrow \langle S^{\sim 0} XH^{\sim 1}, S^{\sim 0} XH^{\sim 1} \rangle . \end{aligned} \tag{2}$$

The main benefit of a restriction of the recursion depth is a gain in decoding efficiency, thus allowing us to set up systems more rapidly and to explore more model combinations and more system configurations.

The experimental results are given in Table 5. With several different methods we are able to improve the baseline by +1.8% BLEU and -2.1% TER on the test set. We proceed with individual descriptions of these methods and their effect in BLEU on the test set.

mooreLM (+0.3% BLEU) We select $\frac{1}{4}$ of the shuffled news data for LM training as presented in Section 3.1.

DWL (+0.6% BLEU) Our standard single-word-based model for lexical smoothing of the phrase table is extracted from word-aligned parallel data, in the fashion of [23]. As an alternative, we utilize phrase-level discriminative word lexicons [12] (DWL) in source-to-target and target-to-source direction, which we found to yield the best results among several lexical smoothing methods. For comparison, the result with IBM-1 is given in Table 5 as well.

opt. KN LM (+0.1% BLEU) [24] recently presented a way to optimize the values of the Kneser-Ney discount parameters with the improved RProp algorithm [25]. We apply their novel method to a machine translation task for the first time and train our French LM with optimized smoothing parameters.

TED TM (+0.4% BLEU) One of the main challenges of the 2011 IWSLT evaluation campaign is adaptation to style and topic of the TED talks. We try to tackle the issue by augmenting our system with an additional phrase table trained on in-domain TED data only. The English-French training data as shown in Table 4 contains 107K parallel sentences from TED sources with 2.1M English and 2.2M French running words. Phrases from the TED translation model are marked with a binary feature.

s2t TED triplets (+0.1% BLEU) We also apply a path-aligned triplet lexicon model [26, 27] for style and topic adaptation. The TED triplet model is trained on the same parallel data as the TED TM. This model is integrated in source-to-target direction only. It takes the full source sentence context into account.

discrim. RO (+0.3% BLEU) The modification of the grammar to a shallow-1 version restricts the search space of the decoder and is convenient to prevent overgeneration. In order not to be too restrictive, we reintroduce more flexibility into the search process by extending the grammar with specific reordering rules

$$\begin{aligned} XP &\rightarrow \langle XP^{\sim 0} XP^{\sim 1}, XP^{\sim 1} XP^{\sim 0} \rangle \\ XP &\rightarrow \langle XP^{\sim 0} XP^{\sim 1}, XP^{\sim 0} XP^{\sim 1} \rangle . \end{aligned} \tag{3}$$

The upper rule in Equation (3) is a swap rule that allows adjacent lexical phrases to be transposed, the lower rule is added for symmetry reasons, in particular because sequences assembled with these rules are allowed to fill gaps within hierarchical phrases. Note that we apply a length constraint of 10 to the number of terminals spanned by an *XP*. We introduce two binary indicator features, one for each of the two rules in Equation (3). In addition to adding these rules, a discriminatively trained lexicalized reordering model [13] is applied.

5.2. Arabic-English

Arabic is known for its complex morphology and ambiguous writing system, where one Arabic word often corresponds to more than one word in traditional target languages such as English. To achieve better correspondence between Arabic and English sentences, we perform the well studied solution of Arabic segmentation. Splitting an Arabic word into its corresponding prefixes, stem and suffixes reduces the number of out-of-vocabulary words, resolves some of the ambiguous Arabic words and generates more one-to-one correspondences between the Arabic side and the target language side, which can then more easily be captured by the IBM alignment models.

In this work, we experimented with the following segmenters:

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++[1] 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.

Due to the large amount of training data and the discrepancy between the test set domain (TED) and the out-of-domain corpora, we use adaptation via cross-entropy based filtering for LM and translation model training (cf. Section 3). To build the LM, we use a mixture of all available English corpora, where news-shuffle and giga-fren.en are filtered in the following way. For news-shuffle, we keep the best $\frac{1}{8}$ sentences and for giga-fren.en we keep the best $\frac{1}{32}$ sentences. The fractions are chosen using the best perplexity LM among different portions of the corpus.

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. Statistics on the combined TED and selected MultiUN data, preprocessed with the MADA-TB scheme, are given in Table 4.

We trained phrase-based systems for all different segmentation schemes on this data. Additionally, one system was trained on all available data, preprocessed with MADA-TB. The results are summarized in Table 6. MADA-TB yields the best results among the different segmentations. We already observed in Section 3.2 that the filtered systems can perform comparably to the system using all MultiUN data, while reducing the phrase table size by a factor of 12. The primary submission is a system combination of all listed systems, which yields another improvement of 0.9% BLEU on the test set.

Table 6: Results on Arabic-English for different segmentations. *MADA-TB ALL* is a system using unfiltered bilingual data. The primary submission is a system combination of all listed systems.

system	**dev**		**test**	
	BLEU	TER	BLEU	TER
FST	26.5	54.2	25.1	57.0
SVM	27.8	53.9	25.4	57.4
HMM	28.0	53.3	25.7	56.9
CRF	27.8	53.8	25.7	56.7
MADA-D1	26.5	54.0	24.7	57.1
MADA-D2	27.0	54.2	25.2	57.1
MADA-D3	27.4	54.1	25.4	57.1
MADA-TB	28.6	52.8	26.1	56.4
MADA-TB ALL	28.6	52.9	26.1	56.6
system combination	29.0	51.3	27.0	54.7

5.3. Chinese-English

For the Chinese-English MT task, we experimented with three different Chinese word segmentation tools, namely the ICTCLAS segmenter[2] (*ict*), LDC segmenter[3] (*ldc*) and Stanford segmenter[4] (*stanford*). Corpus statistics for the ldc segmented TED data can be found in Table 4. The LM used is trained on all provided data ($\sim$425M running words). The translation models are trained using the bilingual TED data. In addition, we experimented with leveraging the bilingual MultiUN data set. The primary submission is a system combination of 12 systems, of which some are identical except for the optimized log-linear parameter values (+*retune*).

The results are given in Table 7. The first impression is that the absolute BLEU values are very low. We can not find obvious differences in translation quality generated by the three segmenters. Further, making use of the MultiUN data does not seem to have a visible effect in translation quality on this task. When added to the PBT system with ldc segmentation, we observe a 0.3% gain in BLEU but a 1.8% loss in TER. A general conclusion from the results is that the hierarchical decoder seems to have a small advantage over the phrase-based decoder on this task. For three of the systems, we added identical setups with different log-linear parame-

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

[2] http://ictclas.org/index.html

[3] http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm

[4] http://nlp.stanford.edu/software/segmenter.shtml

Table 7: Results for the Chinese-English MT task. Hierarchical (HPBT) and phrase-based (PBT) decoders are used. Results are shown for three different Chinese segmenters. *+retune* denotes a different optimization run for the log-linear parameters, *+UN* the use of MultiUN data for training. The primary submission is a system combination of all 12 listed systems.

seg.	system	dev		test	
		BLEU	TER	BLEU	TER
ict	**PBT**	9.9	81.9	11.6	77.0
	HPBT	10.4	80.7	11.5	76.7
	+retune	10.6	81.4	12.0	77.3
ldc	**PBT**	10.0	81.0	11.5	75.7
	+UN	10.3	82.4	11.8	77.5
	HPBT +DWL	10.0	80.5	11.9	76.1
	+retune	10.9	81.7	12.2	77.5
	+UN	10.7	80.1	12.1	76.7
stanford	**PBT**	9.9	82.3	11.5	77.5
	+retune	9.9	80.6	11.2	75.8
	HPBT	10.2	79.9	11.6	75.8
	+DWL	10.1	80.4	11.7	76.3
system combination		11.0	78.9	12.6	74.2

ter values for system combination. This is achieved by the partly randomized optimization procedure, and the two parameter sets for the same setup are chosen due to their different balance between BLEU and TER. To add more diversity to the hypotheses, we also apply the DWL model. This way, the system combination improves over the best single system (ldc, HPBT +UN) by 0.3% BLEU and 1.2% TER on dev and 0.5% BLEU and 2.5% TER on test.

5.4. Spoken language translation (SLT)

The input for the translation systems in the SLT track is the automatic transcription provided by the automatic speech recognition (ASR) track. It does not contain punctuation marks, but the output translation is expected to include punctuations. We experimented with automatically predicting the punctuation at three different stages [32].

- **Before** translation. Punctuation is predicted on the source side. A regular text translation system can be used. We denote this as FullPunct.

- **During** translation. Punctuation is predicted implicitly by applying a translation system trained on data without punctuation marks in the source language, but including punctuation in the target language. We denote this as Implicit.

- **After** translation. The translation system is trained on data containing punctuation on neither source nor target side. The punctuation marks are then predicted automatically in the target language. We denote this as NoPunct.

Table 8: Results for the English-French SLT task (en-fr). Punctuation prediction is evaluated at three different stages: Before (FullPunct), during (Implicit) and after (NoPunct) translation. h-ngram denotes punctuation prediction using the SRILM toolkit, ppmt using a monotone translation decoder. The primary submission is a system combination of the 5 listed systems.

system	dev		test	
	BLEU	TER	BLEU	TER
Implicit	18.0	69.5	21.8	62.5
FullPunct (h-ngram)	18.2	69.3	21.1	62.9
FullPunct (ppmt)	18.3	69.2	21.9	62.2
NoPunct (h-ngram)	17.3	67.9	20.4	62.8
NoPunct (ppmt)	17.8	69.0	21.2	62.2
system combination	18.5	68.3	22.3	61.6

In addition to the three stages at which punctuation is predicted, we tested two methods of performing the punctuation prediction.

h-ngram The SRILM toolkit provides functionality to predict missing tokens based on the LM score (hidden-ngram), which was already used in [32].

ppmt We can interpret punctuation prediction as machine translation, where source and target are the same language, but only the target side contains punctuation.

This results in five different setups, one for Implicit and two each for FullPunct and NoPunct. For FullPunct, we applied the *+s2t TED triplets* system from the English-French MT task (cf. Table 5), which was the best available system when we started the final experiments. Its complete training procedure is copied precisely to train the hierarchical translation systems for Implicit and NoPunct,. However, where punctuation is removed from the training data when appropriate. French and English 9-gram LMs, trained on the bilingual data from the MT track, are used for both types of punctuation prediction. To train the monotone phrase-based translation systems for ppmt punctuation prediction, only the source and target side, respectively, of the TED training data from the MT track is used. As development and test set we also used the same data as in the MT track with removed punctuation. The primary submission is a system combination of the five different hypotheses.

Table 8 shows the comparison between the different translation systems and both prediction tools. FullPunct with ppmt performs slightly better than Implicit by 0.1% in

BLEU and 0.3% in TER on test. In both cases, the prediction method PPMT outperforms the systems using the H-NGRAM tool. Using the FULLPUNCT system with PPMT, we get an improvement of 0.8% in BLEU and 0.7% in TER on test compared to FULLPUNCT using the H-NGRAM tool. A similar improvement is obtained using the NOPUNCT systems. Performing punctuation prediction in the source language leads to a better translation quality, compared to performing it on the target side. On the test set, we achieve an improvement of 0.7% in BLEU using FULLPUNCT instead of NOPUNCT. With the system combination of all five systems, we get an additional improvement of 0.4% in BLEU and 0.6% in TER compared to the best single system FULLPUNCT with PPMT. A complete overview and analysis of these experiments is given in [33].

6. Conclusion

RWTH participated in all MT and SLT tracks of the IWSLT 2011 evaluation campaign. Several different techniques were evaluated and yielded considerable improvements over the respective baseline systems. Among these are different Arabic and Chinese word segmentation tools, monolingual and bilingual data filtering techniques, phrase training, additional lexical smoothing and reordering models for the hierarchical system and different punctuation prediction methods for SLT. Also, both the hierarchical and the phrase-based translation paradigm were used. By system combination of a number of different systems we could achieve additional improvements over the best single system. In this way, RWTH was able to achieve the following positions (automatically measured in BLEU) among all participants: 1st in Arabic-English, 2nd in Chinese-English and 3rd in both MT and SLT track for English-French. An overview over the results of the evaluation campaign is given in [1].

7. Acknowledgments

This work was achieved as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

8. References

[1] M. Federico, L. Bentivogli, M. Paul, and S. Stueker, "Overview of the IWSLT 2011 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] E. Matusov, N. Ueffing, and H. Ney, "Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment," in *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006, pp. 33–40.

[17] E. Matusov, G. Leusch, R. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney, "System Combination for Machine Translation of Spoken and Written Language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 7, pp. 1222–1237, 2008.

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

[19] A. Axelrod, X. He, and J. Gao, "Domain adaptation via pseudo in-domain data selection," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., July 2011, pp. 355–362.

[20] S. Mansour, J. Wuebker, and H. Ney, "Combining Translation and Language Model Scoring for Domain-Specific Data Filtering," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011.

[21] 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.

[22] 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.

[23] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," in *Proc. of the Human Language Technology Conf. (HLT-NAACL)*, Edmonton, Canada, May/June 2003, pp. 127–133.

[24] M. Sundermeyer, R. Schlüter, and H. Ney, "On the Estimation of Discount Parameters for Language Model Smoothing," in *Interspeech*, Florence, Italy, Aug. 2011.

[25] C. Igel and M. Hüsken, "Empirical Evaluation of the Improved Rprop Learning Algorithms," *Neurocomputing*, vol. 50, pp. 105–123, 2003.

[26] S. Hasan, J. Ganitkevitch, H. Ney, and J. Andrés-Ferrer, "Triplet Lexicon Models for Statistical Machine Translation," in *Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii, Oct. 2008, pp. 372–381.

[27] M. Huck, M. Ratajczak, P. Lehnen, and H. Ney, "A Comparison of Various Types of Extended Lexicon Models for Statistical Machine Translation," in *Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Denver, CO, Oct./Nov. 2010.

[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] 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.

[33] S. Peitz, M. Freitag, A. Mauser, and H. Ney, "Modeling Punctuation Prediction as Machine Translation," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, Dec. 2011.