The RWTH Phrase-based Statistical Machine Translation System

Richard Zens, Oliver Bender, Saša Hasan, Shahram Khadivi, Evgeny Matusov, Jia Xu, Yuqi Zhang, Hermann Ney

International Workshop on Spoken Language Translation
Pittsburgh, PA – October 24-25, 2005

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik VI
Computer Science Department
RWTH Aachen University, Germany

Overview

- Introduction: Statistical Machine Translation
- Phrase-based Approach
- Search: Summary of Models
- Rescoring Models
- Integrating ASR and MT
- Experimental Results
- Conclusions

Related Work

- R. Zens, F. J. Och and H. Ney: Phrase-based statistical machine translation. KI2002, Aachen, Germany.
- R. Zens and H. Ney: Improvements in phrase-based statistical machine translation. HLT-NAACL2004, Boston, MA.
- E. Matusov and H. Ney: Phrase-based translation of speech recognizer word lattices using log-linear model combination. ASRU2005, Cancun, Mexico.
- F. J. Och and H. Ney: Discriminative training and maximum entropy models for statistical machine translation. ACL2002, Philadelphia, PA.
- A. Stolcke: SRILM - an extensible language modeling toolkit. ICSLP2002, Denver, CO.

Introduction: Statistical Machine Translation

- source string $f_1^J = f_1 ... f_j ... f_J$ to be translated into a target string $e_1^I = e_1 ... e_i ... e_I$.
- classical source-channel approach:

$$\begin{aligned} \hat{e}_1^I &= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I | f_1^J) \right\} \\ &= \underset{e_1^I}{\operatorname{argmax}} \left\{ Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \right\} \end{aligned}$$

- $Pr(f_1^J | e_1^I)$: translation model (usually can be further decomposed into alignment and lexicon model)
- $Pr(e_1^I)$: language model

Log-linear Models

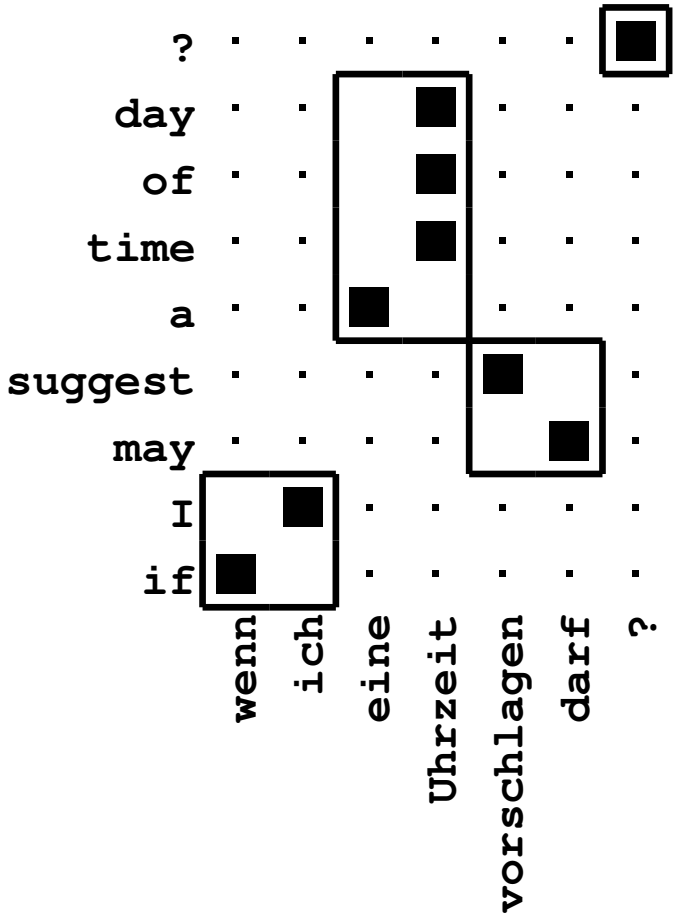
- **alternative: direct modeling of the posterior probability $Pr(e_1^I|f_1^J)$**
- **use a log-linear model (Och and Ney 2002):**

$$Pr(e_1^I|f_1^J) = p_{\lambda_1^M}(e_1^I|f_1^J) = \frac{\exp\left[\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J)\right]}{\sum_{e'^I_1} \exp\left[\sum_{m=1}^{M} \lambda_m h_m(e'^I_1, f_1^J)\right]}$$

  - **models $h_m(e_1^I, f_1^J)$**
  - **model scaling factors λ_m**
- **search criterion:**

$$\hat{e}_1^I = \operatorname*{argmax}_{e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\}$$

Phrase-based Translation: Idea



idea:

- **segment source sentence into phrases**
- **translate each phrase**
- **concatenate these phrase translations**

Phrase Extraction

given: a sentence-aligned training corpus

goal: find phrase pairs $(\tilde{f}, \tilde{e})$ which are translations of each other

idea:

- **make use of word alignment**
- **symmetrize source-to-target and target-to-source alignment**
- **$(\tilde{f}, \tilde{e})$ is a bilingual phrase, if:**
  - **$\tilde{f}$ and $\tilde{e}$ are contiguous**
  - **all words in $\tilde{f}$ are aligned only to words in $\tilde{e}$**
  - **all words in $\tilde{e}$ are aligned only to words in $\tilde{f}$**
- **remember ALL seen phrase pairs**
  - **no length constraint**
  - **restrict to test corpus**

Phrase-based System: Search Algorithms

the RWTH PBT system provides two alternative search strategies:

- **translating a source language word graph in a monotone way**
  - **e.g. word graphs representing different reordering, ASR lattices, etc.**
- **source cardinality synchronous search**
  - **search proceeds synchronously with the cardinality of already translated source positions**
  - **pruning is carried out jointly for all coverage sets with the same cardinality**

- **both search algorithms generate word graphs containing the most likely translation hypotheses**
- **extraction of N-best lists out of these word graphs and rescoring**

Phrase-based System: Models Used in Search

translation models:

- **phrase translation model:** $p(\tilde{f}|\tilde{e})$ **and** $p(\tilde{e}|\tilde{f})$
  estimated via relative frequencies
- **word translation model:** $p(f|e)$ **and** $p(e|f)$
  used for IBM1 within the phrase pairs

additional models:

- n**-gram language model**
- **word and phrase penalty: constant cost per produced words/phrases**
- **deletion model: count the number of deletions, i.e. source words for which no target words exist with a probability higher than a given threshold**
- **reordering model: simple costs based on the jump width**

Optimization of Model Scaling Factors

situation:

- M-dimensional optimization problem
- objective function: e.g. BLEU, NIST, etc.
- NOT smooth ($\rightarrow$ no derivative)
- many local optima

our solution: use Downhill Simplex algorithm (Numerical Recipes)

- no derivative needed
- (almost) no assumption on functional form
- many function evaluations (in our case about 200 iterations)

Rescoring Models

- **clustered language models:**
  clustering of hypotheses based on regular expressions and application of cluster-specific (i.e. sentence-type-specific) language models
- **IBM model 1:**
  captures lexical co-occurrences, helpful for translation adequacy
- **IBM1 deletion model:**
  count the number of source words, for which the IBM-1 translation probability given any of the target words in the hypothesis is below α
- **hidden Markov alignment model:**
  use the so-called homogeneous HMM to compute the log-likelihood of a sentence pair
- **word penalties:**
  simple heuristics that affect the generated hypotheses length, longer sentences are to be favored

Integrating ASR and MT

the RWTH PBT system for ASR output:

- use the ASR lattices as input and translate them as source language word graph
- problem: search is only monotone
- remedy: extension of the search algorithm
  - while traversing the input lattice, a phrase can be skipped and processed later
- in addition, the acoustic model and the source language model scores are integrated into the search simultaneously to the base models and the scaling factors are also optimized

out-of-vocabulary (OOV) problem:

- a significant obstacle for integrating ASR and MT is the mismatch between the vocabularies of the two systems
- in IWSLT, the OOV rates were rather high and thereby affected the obtained results adversely

Experimental Results

task:

- **experiments were carried out on the *Basic Travel Expression Corpus* (BTEC)**
- **as additional training resources for the C-Star track, we used the full BTEC for Japanese-English and the *Spoken Language DataBase* (SLDB)**

experiments:

- **C-Star'03 corpus was used as development corpus**
- **IWSLT'04 test set was used as blind test corpus**
- **text translation experiments were performed on Arabic-English, Chinese-English, English-Chinese and Japanese-English supplied data tracks plus Japanese-English C-Star track**
- **speech translation experiments were performed on Chinese-English and Japanese-English supplied data tracks**

Corpus Statistics

		Supplied Data Track				C-Star Track	
		Arabic	Chinese	Japanese	English	Japanese	English
Train	Sentences	20 000				240 672	
	Running Words	180 075	176 199	198 453	189 927	1 951 311	1 775 213
	Vocabulary	15 371	8 687	9 277	6 870	26 036	14 120
	Singletons	8 319	4 006	4 431	2 888	8 975	3 538
C-Star'03	Sentences	506					
	Running Words	3 552	3 630	4 130	3 823	4 130	3 823
	OOVs	133	114	61	65	34	–
IWSLT'04	Sentences	500					
	Running Words	3 597	3 681	4 131	3 837	4 131	3 837
	OOVs	142	83	71	58	36	–
IWSLT'05	Sentences	506					
	Running Words	3 562	3 918	4 226	3 909	4 226	3 909
	OOVs	146	90	293	69	10	–

Official Evaluation Results

Data Track	Input	Translation Direction	Accuracy Measures				Error Rates	
			BLEU [%]	NIST	Meteor [%]	GTM [%]	WER [%]	PER [%]
Supplied	Manual	Arabic-English	54.7	9.78	70.8	65.6	37.1	31.9
		Chinese-English	51.1	9.57	66.5	60.1	42.8	35.8
		English-Chinese	20.0	5.09	12.6	55.2	61.2	52.7
		Japanese-English	40.8	7.86	58.6	48.6	53.6	44.4
	ASR	Chinese-English	38.3	7.39	54.0	48.8	56.5	47.2
		Japanese-English	42.7	8.53	62.0	49.6	51.2	41.2
C-Star	Manual	Japanese-English	77.6	12.91	85.4	78.7	24.3	18.6

- for the English-Chinese task only one reference was used, therefore the scores are in a different range
- for the Japanese-English supplied data track our system suffers from the high number of OOVs (cf. results for the C-Star track)

Translation of Chinese ASR Lattices

statistics for the Chinese ASR lattices:

Test Set	WER [%]	GER [%]	Density
C-Star'03	41.4	16.9	13
IWSLT'04	44.5	20.2	13
IWSLT'05	42.0	18.2	14

translation results (* indicates late submissions):

System		Input	BLEU [%]	NIST	WER [%]	PER [%]
Graph	Mon*	1-Best	31.1	6.18	62.1	52.7
		Lattice	34.1	7.20	58.3	48.1
	Skip	1-Best	33.1	6.51	61.3	51.7
		Lattice	35.1	7.53	57.7	47.2
SCSS (primary)		1-Best	38.3	7.39	56.5	47.2
+Rescoring*			40.2	7.33	55.1	46.5

Translation Examples for ASR Input

Input	Translation
1-Best	*Is there a pair of room with a bath*
Lattice	*I would like a twin room with a bath*
Reference	*A double room including a bath*
1-Best	*Please take a picture of our*
Lattice	*May I take a picture here*
Reference	*Am I permitted to take photos here*
1-Best	*I'm in a does the interesting*
Lattice	*I'm in an interesting movie*
Reference	*A good movie is on*

- comparison of 1-best and lattice translations in the Chinese-English supplied data track
- recognition errors that occur in the single-best ASR hypotheses are often corrected when lattices are used

Translation Examples for the Japanese-English Tracks

Data Track	Translation
Supplied	*What would you like*
C-Star	*What would you like for the main course*
Reference	*What would you like for the main course*
Supplied	*Is that flight two seats available*
C-Star	*Are there two seats available on that flight*
Reference	*Are there two seats available on that flight*
Supplied	*Have a good I anything new*
C-Star	*I prefer something different*
Reference	*I prefer something different*

- the C-Star track system is able to produce one of the reference translation
- the supplied data track examples show the effect of a single unknown, poor word order and entire incomprehensibility

Rescoring Improvements

System	BLEU [%]	NIST	WER [%]	PER [%]
Baseline	45.1	8.56	48.9	40.1
+CLM	45.9	8.24	48.6	40.7
+IBM1	45.9	8.48	47.8	39.7
+WP	45.4	8.91	47.8	39.4
+Del	46.0	8.71	47.8	39.6
+HMM	46.3	8.73	47.4	39.7

- effect of successively adding models for the Chinese-English IWSLT'04 test set
- on the development set (C-Star'03) all models gradually enhanced the performance
- the results on the IWSLT'04 blind test set are not as smooth but still improvements on all evaluation criteria are achieved

Effect of Rescoring: Translation Examples

System	Translation
Baseline	*Your coffee or tea*
+Rescoring	*Would you like coffee or tea*
Reference	*Would you like coffee or tea*
Baseline	*A room with a bath*
+Rescoring	*I would like a twin room with a bath*
Reference	*A twin room with bath*
Baseline	*How much is that will be that room*
+Rescoring	*How much is that room including tax*
Reference	*How much is the room including tax*
Baseline	*Onions*
+Rescoring	*I would like onion*
Reference	*I would like onions please*

Conclusions

- phrase-based translation system based on log-linear model combination
- two pass approach:
  - generate an N-best list using a dynamic programming beam search algorithm
  - rescoring and re-ranking of this N-best list
- direct optimization of base models using minimum error training of model scaling factors
- translations of good quality were produced (significant improvements compared to the RWTH IWSLT04 system)
- translation of ASR lattices yields significant improvements over the translation of ASR single-best hypotheses

Thank you for your attention

Oliver Bender

bender@cs.rwth-aachen.de

http://www-i6.informatik.rwth-aachen.de