The RWTH Phrase-based Statistical Machine Translation System

Author：Richard Zens, Oliver Bender, Sasa Hasan, Sharam Khadivi, Evgeny Matusov, Jia Xu, Xuqi Zhang, Hermann Ney

Professor：陳嘉平

Reporter：陳逸昌

Outline

- Introduction
- Review the statistical approach
- Models used during search
- Rescoring models
- Integrating ASR and MT
- Tasks and corpora
- Experimental results

Introduction

- An overview of the RWTH phrase-based statistical machine translation system

- Two pass approach
  - Generate
  - Rescoring and reranking

Review the statistical approach

- Source-channel approach to SMT
  - Source language $f_1^J = f_1 \dots f_j \dots f_J$
  - Target language $e_1^I = e_1 \dots e_i \dots e_I$

$$\hat{e}_1^{\hat{I}} = \arg\max_{I, e_1^I} \{\Pr(e_1^I \mid f_1^J)\}$$

$$= \arg\max \{\Pr(e_1^I) \cdot \Pr(f_1^J \mid e_1^I)\}$$

Review the statistical approach

- Log-linear model

$$\Pr(e_1^I \mid f_1^J) = \frac{\exp(\sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J))}{\sum_{e_1'^{I'}} \exp(\sum_{m=1}^{M} \lambda_m h_m(e_1'^{I'}, f_1^J))}$$

- $h(\bullet)$ feature
- λ_1^M scaling factor

Review the statistical approach

- Phrase-based approach
  - Segment the give source sentence into phrases
    - EX: sentence pair $\left(f_1^J, e_1^I \right)$ into K blocks

$$k \to s_k := \left(i_k ; b_k , j_k \right), \text{ for } k = 1 \ldots K$$

    - i_k target phrase positions
    - b_k source phrase start positions
    - j_k source phrase end positions

Review the statistical approach

$$\tilde{e}_k := e_{i_{k-1}+1} \ldots e_{i_k}$$
$$\tilde{f}_k := f_{b_k} \ldots f_{j_k}$$

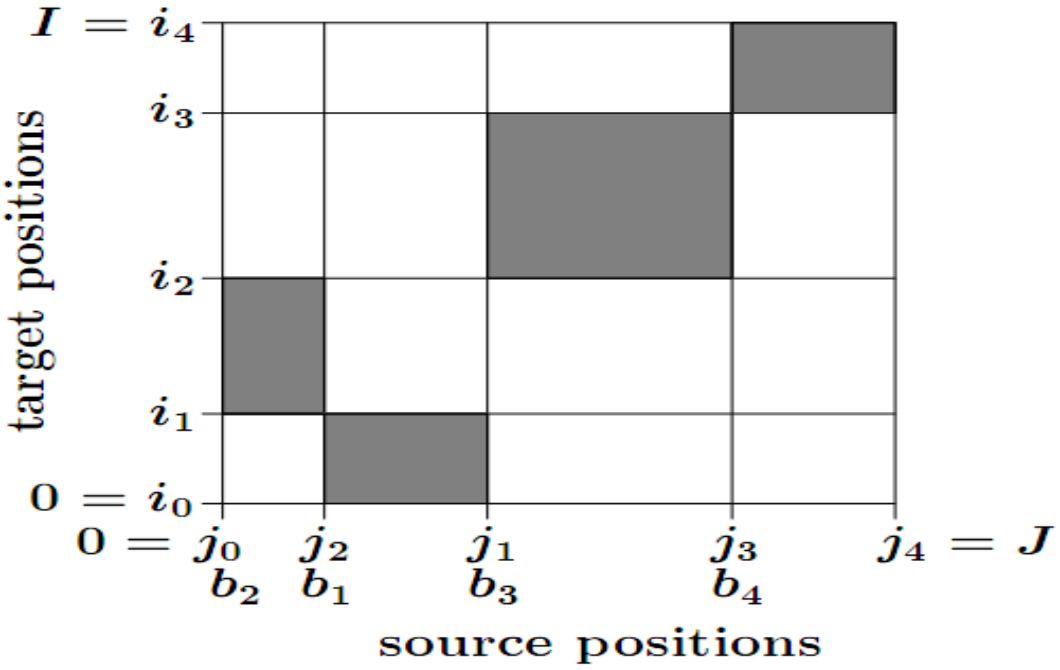

Figure 1: Illustration of the phrase segmentation.

Models used during search

- Phrase-based model

$$p(\tilde{f} \mid \tilde{e}) = \frac{N(\tilde{f}, \tilde{e})}{N(\tilde{e})}$$

  - $N\left(\tilde{f}, \tilde{e}\right)$ the number of co-occurrences

$$h_{Phr}\left(f_1^J, e_1^I, s_1^K\right) = \log \prod_{k=1}^{K} p\left(\tilde{f}_k \mid \tilde{e}_k\right)$$

Models used during search

- Word-based lexicon model

$$h_{Lex}\left(f_1^J, e_1^I, s_1^K\right) = \log \prod_{k=1}^{K} \prod_{j=b_k}^{j_k} \sum_{i=i_{k-1}+1}^{i_k} p\left(f_i \mid e_i\right)$$

Models used during search

- Deletion model

$$h_{Del}\left(f_1^J, e_1^I, s_1^K\right) = \sum_{k=1}^{K} \sum_{j=b_k}^{j_k} \prod_{i=i_{k-1}+1}^{i_k} \left[p\left(f_j \mid e_i\right) < \tau \right]$$

  - τ threshold

Models used during search

- Word and phrase penalty model

$$h_{WP}\left(f_1^J, e_1^I, s_1^K\right) = I$$

$$h_{pp}\left(f_1^J, e_1^I, s_1^K\right) = K$$

Models used during search

- Target language model

$$h_{LM}\left(f_1^J, e_1^I, s_1^K\right) = \log \prod_{i=1}^{I} p\left(e_i \mid e_{i-n+1}^{i-1}\right)$$

Models used during search

- Reordering model

$$h_{RM}\left(f_1^J, e_1^I, s_1^K\right) = \sum_{k=1}^{K} \left| b_k - j_{k-1} - 1 \right| + J - j_k$$

Rescoring models

- Clustered language model

$$h_{CLM}\left(f_1^J, e_1^I\right) = \log \sum_c \left[\Re_c\left(e_1^I\right)\right]\left(\alpha_c p_c\left(e_1^I\right) + \left(1-\alpha_c\right) p_g\left(e_1^I\right)\right)$$

  - $p_g\left(e_1^I\right)$ global language model
  - $p_c\left(e_1^I\right)$ cluster-specific language model
  - $\left[\Re_c\left(e_1^I\right)\right]$ if the regular expression matches the target sentence then 1 or 0 for otherwise

Rescoring models

- IBM model 1

$$h_{IBM1}\left(f_1^J, e_1^I\right) = \log\left(\frac{1}{(I+1)^J}\prod_{j=1}^{J}\sum_{i=0}^{I} p\left(f_j \mid e_i\right)\right)$$

Rescoring models

- IBM1 deletion model

$$h_{Del}\left(f_1^J, e_1^I\right) = \sum_{j=1}^{J} \prod_{i=0}^{I} \left[p\left(f_j \mid e_i\right) < \tau \right]$$

Rescoring models

- Hidden Markov alignment model

$$h_{HMM}\left(f_1^J, e_1^I\right) = \log \sum_{a_1^J} \prod_{j=1}^{J} \left(p\left(a_j \mid a_{j-1}, I\right) \cdot p\left(f_j \mid e_{aj}\right)\right)$$

Rescoring models

- Word penalties

$$h_{WP}\left(f_1^J, e_1^I\right) = \begin{cases} I \\ I / J \\ 2\left|I - J\right| / (I + J) \end{cases}$$

Integrating ASR and MT

- Add acoustic model and the source language model in the log-linear model
- In the IWSLT, the vocabulary of the recognition system is not the subset of the translation system source vocabulary.

Tasks and corpora

Table 1: Corpus statistics after preprocessing.

		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 (Running Words)	133	114	61	65	34	–
IWSLT'04	Sentences	500					
	Running Words	3 597	3 681	4 131	3 837	4 131	3 837
	OOVs (Running Words)	142	83	71	58	36	–
IWSLT'05	Sentences	506					
	Running Words	3 562	3 918	4 226	3 909	4 226	3 909
	OOVs (Running Words)	146	90	293	69	10	–

Experimental results

Table 3: Official results for the RWTH primary submissions on the IWSLT'05 test set.

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

Experimental results

Table 6: Rescoring: effect of successively adding models for the Chinese-English IWSLT'04 test set.

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

Experimental results

Table 8: Translation results for ASR input in the Chinese-English supplied data track on the IWSLT'05 test set (*: 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