CASIA Statistical Machine Translation System for IWSLT 2008

Chengqing Zong

NLPR, Institute of Automation
Chinese Academy of Sciences
cqzong@nlpr.ia.ac.cn

No.95, Zhongguancun East Road
Beijing 100190, China

http://www.nlpr.ia.ac.cn
Tel. No.: +86-10-6255 4263

Outline

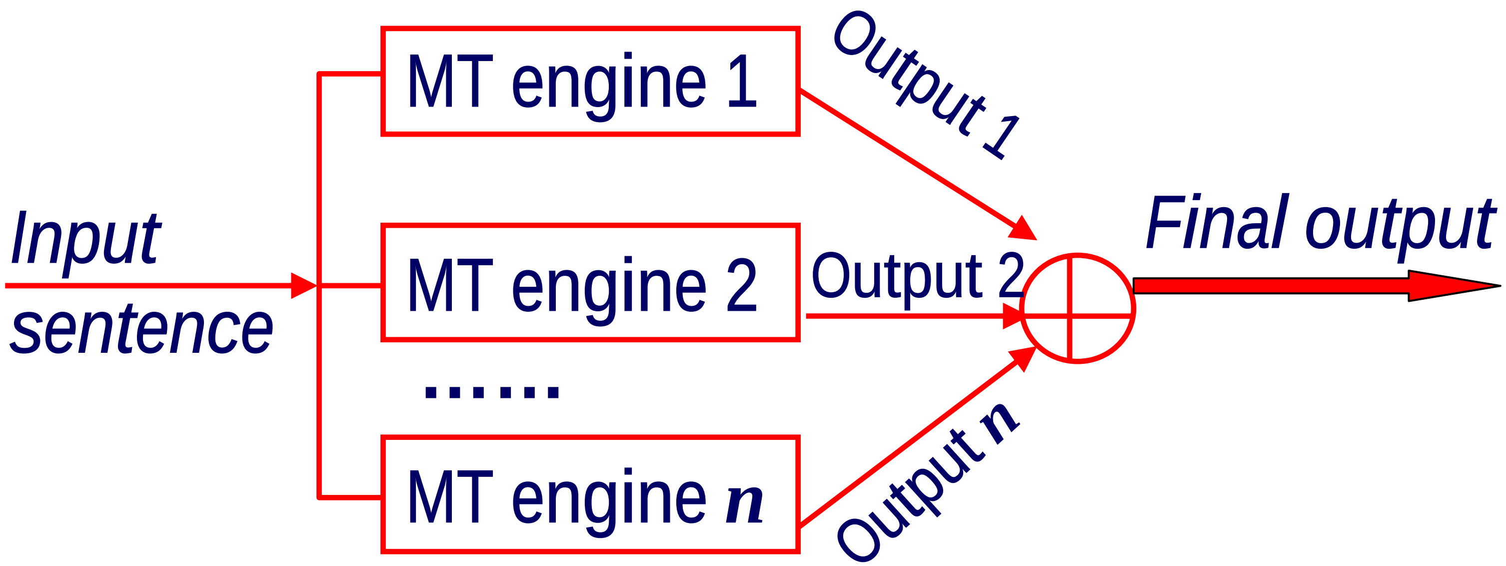
- Our tasks
- System overview
- Technical modules
  - Preprocessing
  - Multiple translation engines
  - System combination
  - Rescoring
  - Post-processing
- Experiments
- Conclusions

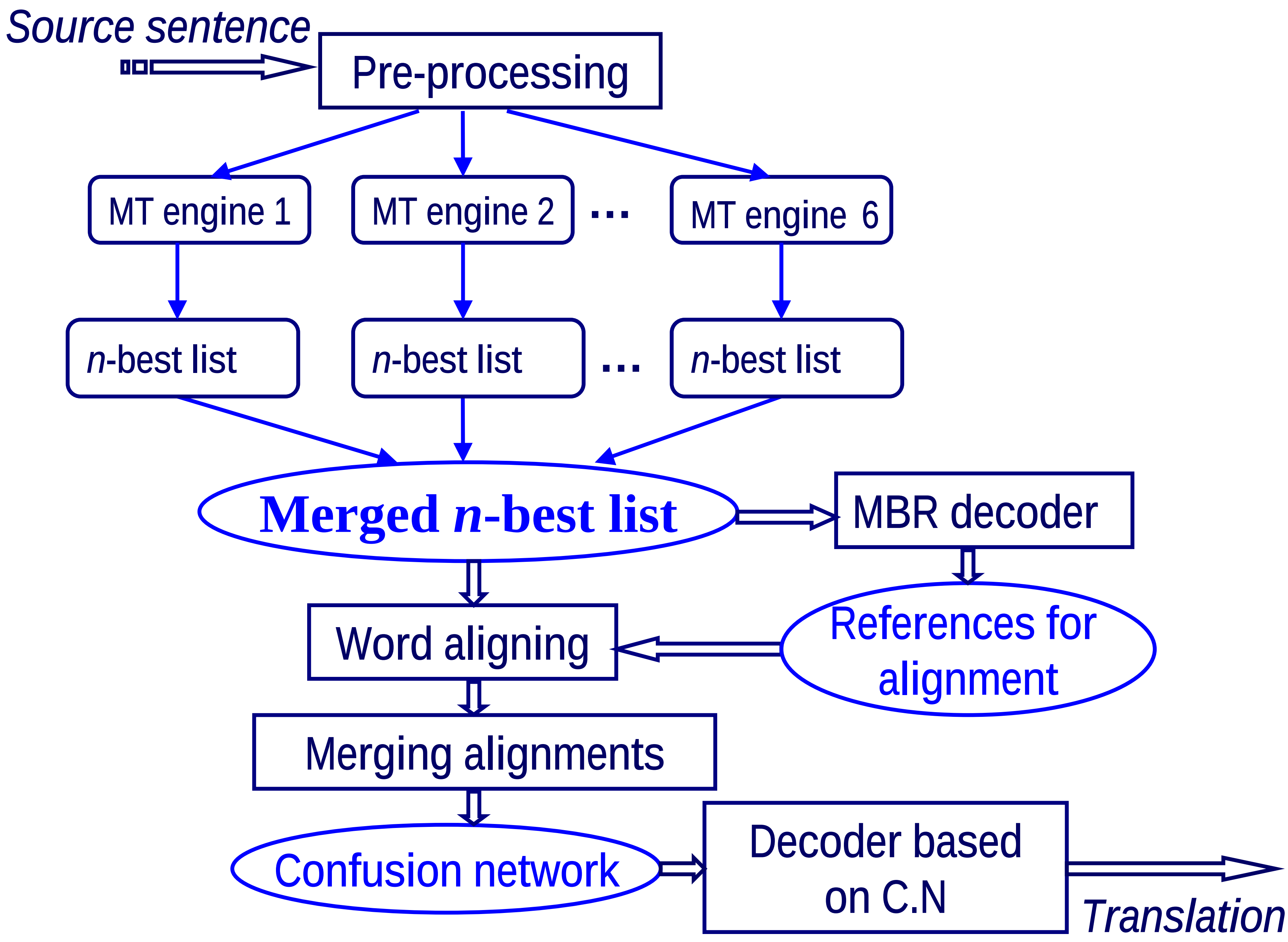
Our tasks

- We participated in:
  1. Challenge task for Chinese-English
  2. Challenge task for English-Chinese
  3. BTEC task for Chinese-English.

System overview

- **Using multiple translation engines**
- **Rescore the combination results to get the final translation outputs**





Source sentence
Pre-processing
MT engine 1
MT engine 2
...
MT engine 6
n-best list
n-best list
...
n-best list
Merged n-best list
MBR decoder
References for alignment
Word aligning
Merging alignments
Confusion network
Decoder based on C.N
Translation

Technical modules

- **Preprocessing**
  - **Chinese**
    - **Chinese word segmentation**
    - **Transforming the Sexagesimal to Binary Converter (SBC) to Decimal to Binary Converter (DBC)**
  - **English**
    - **Tokenization of the English words - separates the punctuations with the English words;**
    - **Transforming the uppercase into lowercase.**

Technical modules

- **Phrase-based translation engines modeled in log-linear model**

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

✓Phrase translation probability ;
✓Lexical phrase translation probability ;
✓Inversed phrase translation probability ;
✓Inversed lexical phrase translation probability ;
✓English language model based on 3-gram ;
✓English sentence length penalty ;
✓Chinese phrase count penalty.

Technical modules

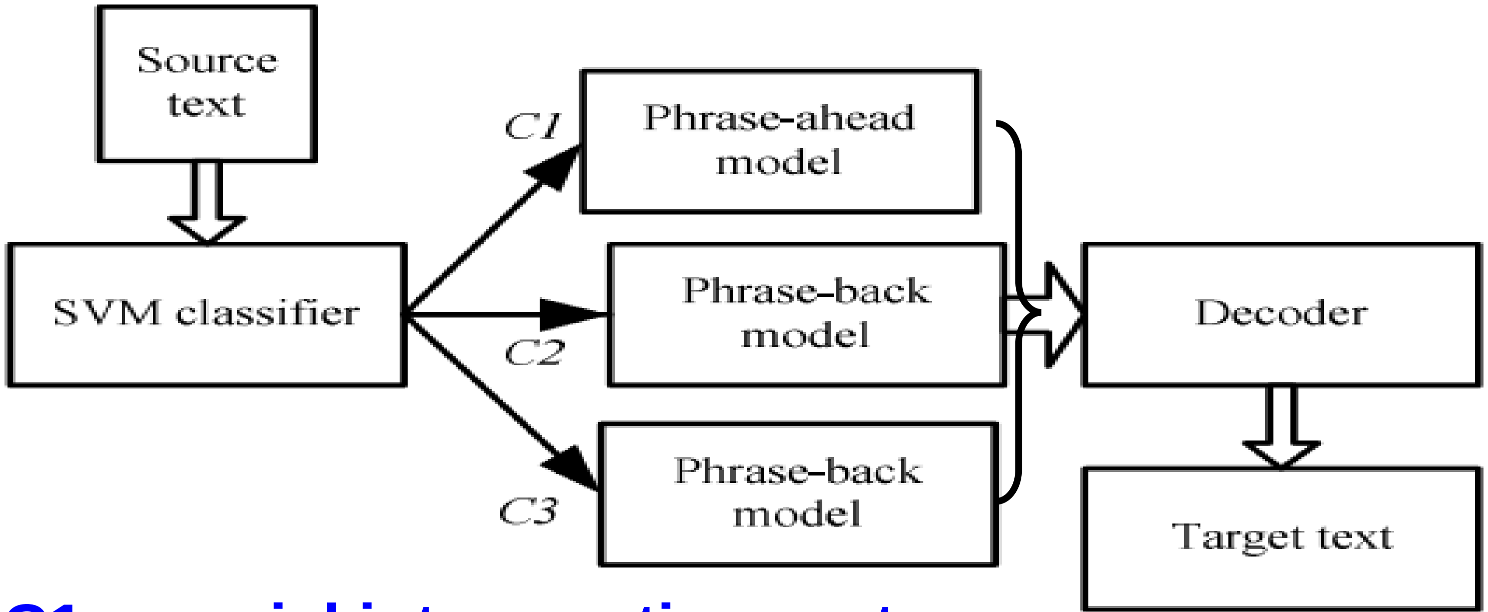
- We use three phrase-based SMT:
  - In-home developed phrase-based decoder (baseline)
  - Moses decoder
  - Bandore: A sentence type based reordering decoder

Technical modules- Bandore

- Preprocessing for PB SMT engine
  - SVM is employed to divide the source (Chinese) sentences into three types, different types of sentences are reordered using different models
  - Three types:
    - Special interrogative sentences
    - Other interrogative sentences
    - Non-question sentences

Technical modules- Bandore

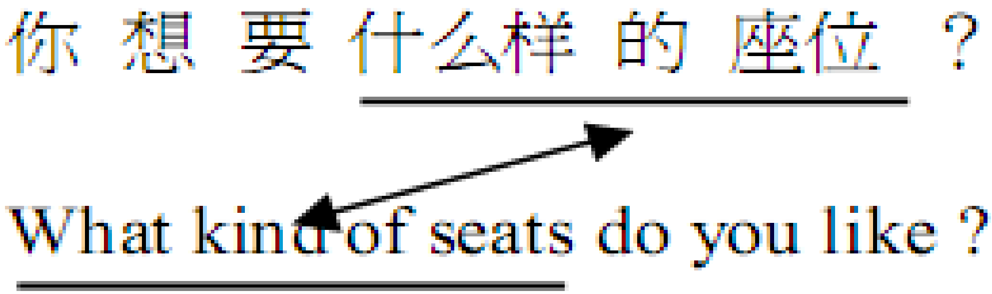
- **Architecture**



- **C1: special interrogative sentences**
- **C2: other interrogative sentences**
- **C3: non-question sentences**

Technical modules- Bandore

- **Special interrogative sentences**
  - There is a fixed question phrase at the end of Chinese sentence, which is moved to the first position in the English translation. (We call the question phrase as Special Question Phrase)



Technical modules- Bandore

- **Phrase-ahead reordering model**

moves the SQP to the frontal position in Chinese sentence

- **Two problems:**
  - Identification of SQP
  - What position should SQP be moved to

Technical modules- Bandore

- **Special words in SQP**
  - **Some Chinese words indicate the sentence is a special interrogative sentence**
  - **Close set: 什么(what)、哪(where)、多 (多长、多久) (how long)、怎(how)、谁(who, whom, whose)、几(how many)、为什么(why)、何 (when)**

Technical modules- Bandore

- **Definition of SQP:**
  - **The syntactic component containing a special word in the close set**
- **Identification:**
  - **Use a shallow parsing toolkit (FlexCrf) (http://flexCRF.sourceforge.net)**

Technical modules- Bandore

- Where should SQP be moved to?

  - Three possible positions:
    - The beginning of the sentence
    - After the rightmost punctuation before the SQP
    - After a regular phrase such as “请问(May I ask)” and “你 知道(Do you know)”

Technical modules- Bandore

这 道 菜 怎么 样 ? **How about this dish?**

你好，去 海滩 怎么 走 ? **Hello, how can I get to the beach?**

你 知道 到 那里 需要 多长 时间 ? **Do you know how long it takes us to there?**

Technical modules- Bandore

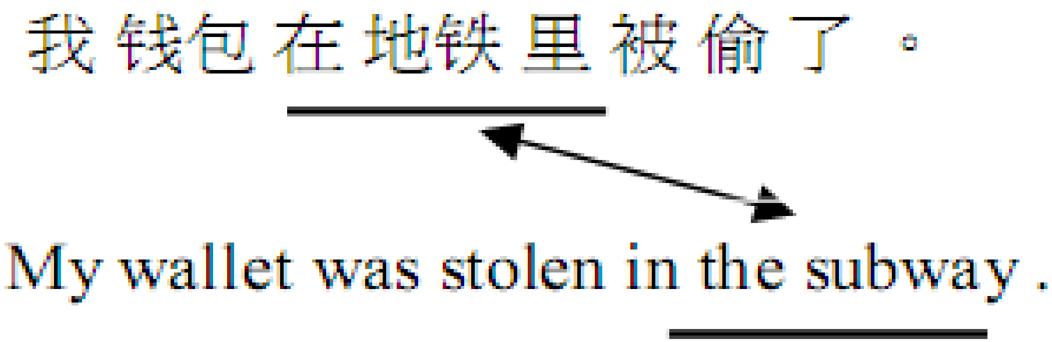
If we have known the *SQP*, *S* becomes S^0 *SQP* S^1, where S^0 is the left part of the sentence before *SQP*, and S^1 is the right part of the sentence after *SQP*. Therefore, we have learned the reordering templates from bilingual corpus to find the right position in S^0 where *SQP* will be moved to.

Technical modules- Bandore

- Other interrogative sentences
  - Some specific Chinese words like “会、能、可以” are simply translated into “Can …”, “Do …” or “May …” at the beginning of the English sentence.
  - This case is easy to process. So, we treat it as the no-question sentences

Technical modules- Bandore

- **Non-question sentences**
  - **Some phrases are usually moved back during translation**
  - **Three types of Chinese phrases are usually moved after the verb phrase in English sentence: (1)Prepositional phrase (PP), (2) Temporal phrase, and (3)Spatial phrase (SP)**



Technical modules- Bandore

➢ **For the other interrogative sentences and non-question sentences, the phrase-back reordering model has been designed to move some phrases to the back positions**

- **Two problems:**
  - **Identification of PP, TP, SP and VP**
  - **Reordering rules**

Technical modules- Bandore

- **Identification**

  Use a shallow parsing toolkit (http://flexCRF.sourceforge.net)

- **Reordering rules**
  - **Maximum entropy model is employed to decide whether a PP, TP or SP is moved back after VP**

Technical modules- Bandore

We develop a probabilistic reordering model to alleviate the impact of the errors caused by the parser when recognizing *PP*s, *TP*s, *SP*s and *VP*s. The form of phrase-back reordering rules:

$$A: \quad A_1 X A_2 \Rightarrow \begin{cases} A_1 X A_2 & straight \\ X A_2 A_1 & inverted \end{cases}$$

$A_1 \in \{PP, TP, SP\}$, $A_2 \in \{VP, FVP\}$

X is any phrases between A_1 and A_2.

Technical modules- Bandore

A Maximum Entropy Model is trained from bilingual spoken language corpus to determine whether *A_1* should be moved after *A_2*:

$$P(O \mid A) = \frac{\exp(\sum_i \lambda_i h_i(O, A))}{\sum_O \exp(\sum_i \lambda_i h_i(O, A))}$$

$O \in \{straignt, inverted\}$, $h_i(O, A)$ is a feature, and λ_i is the weight.

The features include the leftmost, rightmost, and the POSs of A_1 and A_2.

Technical modules

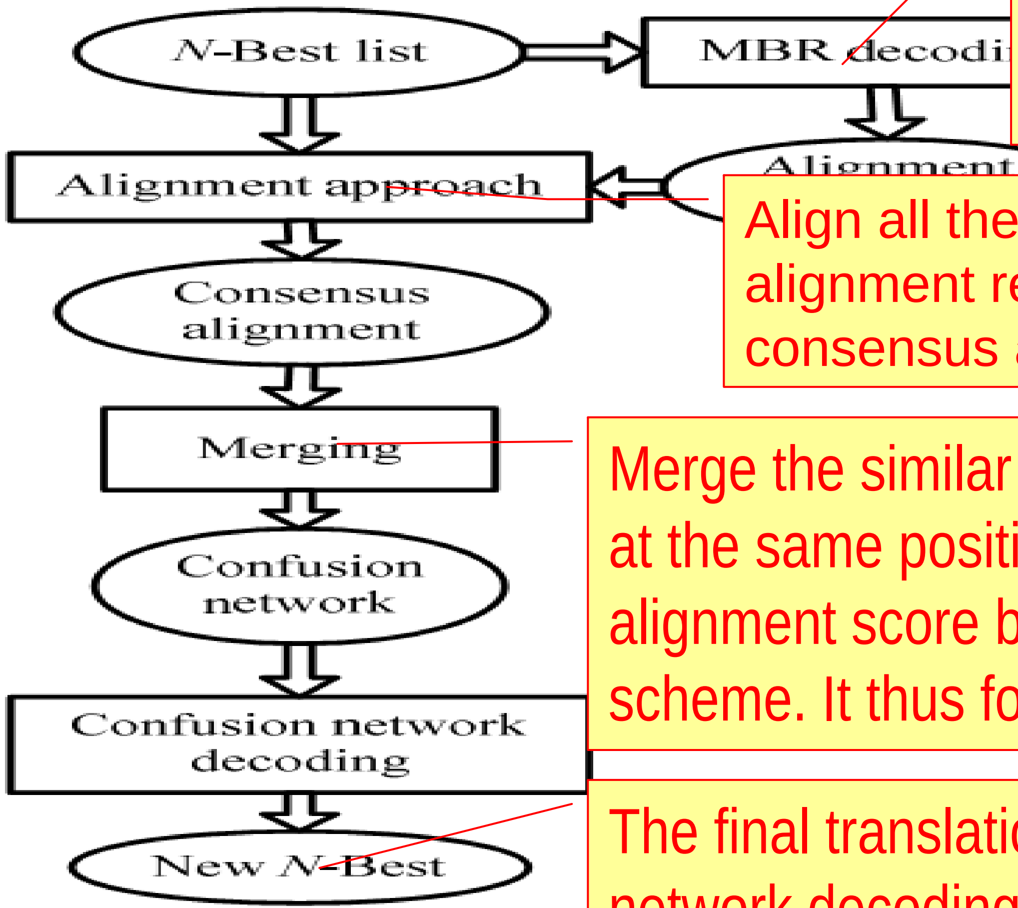
Other translation engines:

- Two formal syntax-based SMT engines:
  - HPB: A hierarchical phrase-based model
  - MEBTG : A maximum entropy-based reordering model
- A linguistically syntax-based SMT:
  - SAMT: A syntax-augmented machine translation decoder

System combination

- **We implement system combination on *N*-Best list from multiple translation engines.**

System combination



Find a hypothesis as the alignment reference with the minimum Bayesian risk

Align all the hypotheses against the alignment reference and forms a consensus alignment

Merge the similar words being aligned together at the same position and assign each word an alignment score based on a simple voting scheme. It thus forms a confusion network.

The final translation is found by the confusion network decoding with the language model feature and word penalty introduced.

Rescoring

- Use global feature functions to score the new *n*-best list
  - Direct and inverse IBM model 1 and model 3
  - 2, 4, 5-gram target language model
  - 3, 4, 5-gram target pos language model
  - Bi-word language model
  - Length ratio between source and target sentence
  - Question feature
  - Frequency of its *n*-gram (*n*=1, 2, 3, 4) within *n*-best translations
  - *n*-gram posterior probabilities within *n*-best translations.
  - Sentence length posterior probabilities.

Post-processing

- **The post-processing for the output results mainly includes:**
  - Case restoration in English words
  - Recombination the separated punctuations with its left closest English words
  - Segmenting the Chinese output into characters

Experiments

- **Corpus**
  - Besides the training data provided by IWSLT 2008, we collected all the data from the website of IWSLT2008.
  - We extract the bilingual data which are highly correlative with the training data of each track.
  - We also filter some development sentences and their reference sentences from all the released development data of the track as our development data according to the similarity calculation.

Experiments

- **The detailed statistics of our corpus for development set**

Track	Data		Sen.	Running words	Voc.
CT CE CRR	Train set	Chi	324,626	2.4M	11,214
		Eng	324,626	2.57M	9,488
	Dev set	Chi	534	3,163	649
		Eng	3,204	22,861	1,132
CT EC CRR	Train set	Chi	311,438	2.28M	11,113
		Eng	311,438	2.42M	9,370
	Dev set	Chi	2275	15,266	797
		Eng	325	2,061	404
BTEC CE CRR	Train set	Chi	321,770	2.38M	11,202
		Eng	321,770	2.51M	9,493
	Dev set	Chi	764	4,899	910
		Eng	4,584	34,310	1,536

Experiments

- **ASR translation**
  - We first translate the ASR *n*-best list.
    - For our experiments the value n=5 is used
  - We pass the translation results into our combination module and rescore all the translation hypotheses
    - With the feature functions of translation hypotheses plus the features of ASR

Experiments

- **Results of development set for CT_CE track**

	CRR		ASR	
	BLEU	NIST	BLEU	NIST
PB	0.4505	7.4649	0.4732	7.4777
MOSES	0.5048	7.9175	0.4980	7.7488
Bandore	0.5033	8.0267	0.4651	7.4983
MEBTG	0.4571	7.6887	0.4969	7.8267
HPB	0.4412	6.8600	0.4536	7.4474
COM	0.5109	8.1780	0.5093	8.0045
Rescore	0.5741	8.3162	0.5787	8.7570

Experiments

- **Results of development set for BTEC_CE track**

	CRR		ASR	
	BLEU	NIST	BLEU	NIST
PB	0.4659	7.9333	0.4831	7.8623
MOSES	0.5100	8.0298	0.4870	7.4720
Bandore	0.5127	8.3513	0.4856	7.7699
MEBTG	0.4717	7.8045	0.4915	7.7357
HPB	0.4764	6.5603	0.4445	5.9105
COM	0.5308	8.5689	0.5087	8.0778
Rescore	0.6100	8.7823	0.5235	8.2364

Experiments

- **Results of development set for CT_EC track**

	CRR		ASR	
	BLEU	NIST	BLEU	NIST
PB	0.4385	7.0469	0.4350	7.3629
MEBTG	0.4399	7.5303	0.4569	7.5691
MOSES	0.4522	7.3626	0.4676	7.5165
HPB	0.4298	7.0914	0.4544	7.5165
COM	0.4555	7.6200	0.4578	7.5600
Rescore	0.5242	7.7361	0.5011	7.9627

Experiments

- **Engines for combination on development set**

	CT_CE		CT_EC		BTEC_CE	
	CRR	ASR	CRR	ASR	CRR	ASR
PB			√	√		√
MOSES	√	√	√	√	√	√
Bandore	√	√			√	
MEBTG	√	√	√	√	√	√
HPB					√	√

Experiments

- Results of test set for each track
  - Con1 : our system combination
  - Con2 : the rescoring module
  - Primary: we RE-rescore “Con1” and “Con2” by using the feature of the prior probability of the length-ratio of source sentence to target sentence.

Experiments

Track	System	CRR		ASR	
		BLEU	NIST	BLEU	NIST
CT CE	Primary	0.4844	7.5859	0.4066	6.6384
	Con1	0.4803	7.4277	0.3750	6.3134
	Con2	0.4767	7.4237	0.4067	6.5887
CT EC	Primary	0.5122	7.3513	0.4312	6.6867
	Con1	0.4968	7.1525	0.4172	6.4864
	Con2	0.4817	6.7254	0.4162	6.4713
BTEC CE	Primary	0.5077	8.5389	0.4339	7.7247
	Con1	0.4842	8.4094	0.4303	7.6550
	Con2	0.5162	8.2884	0.4318	7.6203

Experiments

- The best performance relatively compared with PB decoder among the scores on development set.

System	Compared with PB
Bandore	11.72%
MEBTG	5.03%
HPB	4.45%

Conclusions

- **Our system combines the output results of multiple machine translation engines and by using some global features we rescore the combination results to get the final translation outputs.**

Conclusions

- **In all the translation engines, Moses has a performance with considerable robust**
- **Bandore has an outstanding performance among the three engines**
  - It uses Moses as its decoder
  - The reordering model of Bandore aims at the spoken language. It has an effective ability to translation in the domain of IWSLT.

Thanks

谢谢！