System Combination of RBMT plus SPE and Preordering plus SMT

Terumasa EHARA

Ehara NLP Research Laboratory
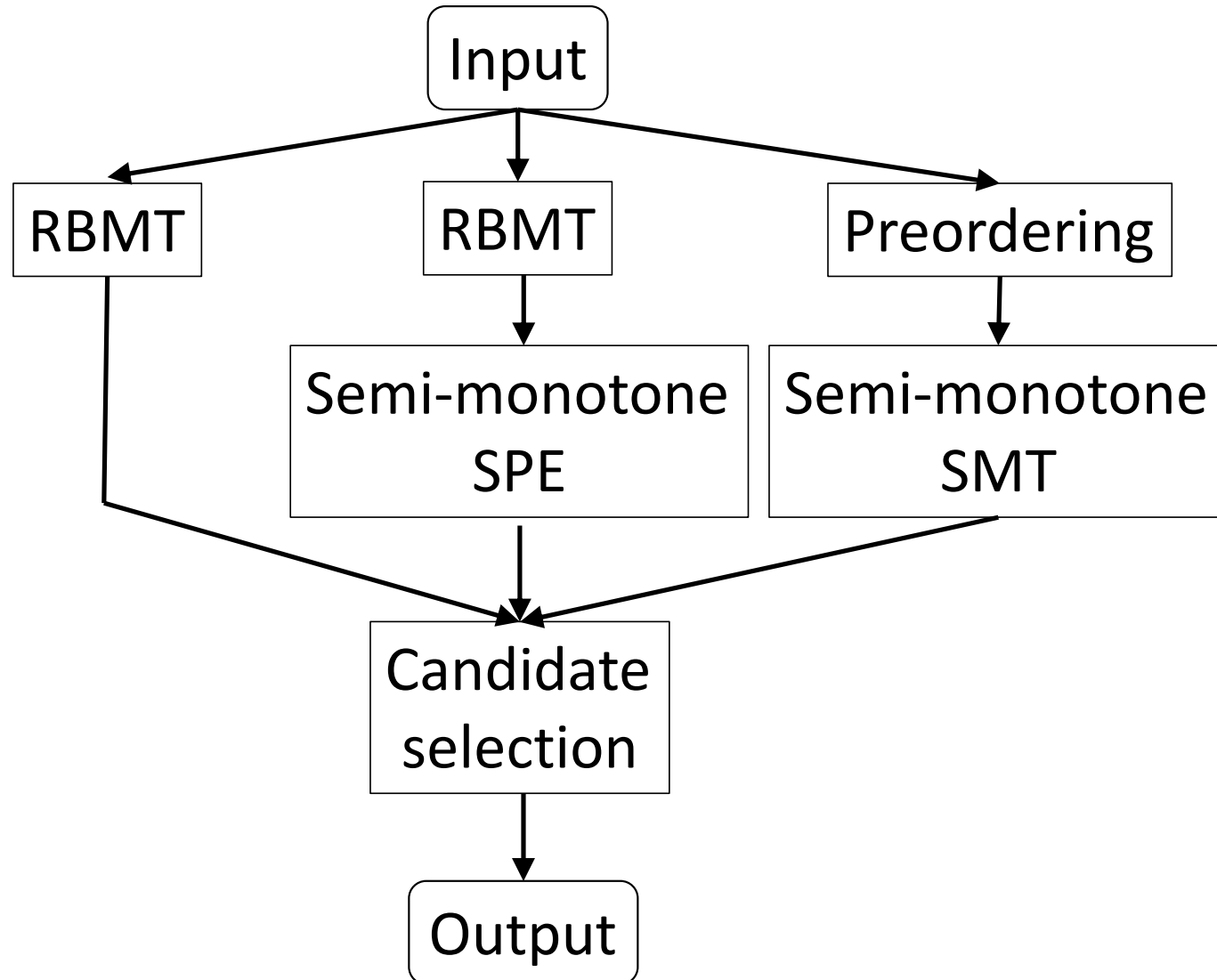
http://www.ne.jp/asahi/eharate/eharate/

2015. 10. 16

WAT2015

Introduction

- Hybrid system combining rule-based method (RBMT, preordering) and statistical method (SMT, SPE) may make MT more accurate.

- Several improvements are implemented.

System architecture

RBMT part

Task	Engine	User dictionary	Sentence pattern dic.
En-ja	Commercial	--	--
Zh-ja	Commercial	JPO dic 1,463,265	--
JPO zh-ja	Commercial	JPO dic 1,463,265	13
JPO ko-ja	Commercial	434,334	--

SPE part

- Phrase based Moses
- LM for en-ja and zh-ja:
  3.6M sent., KenLM, order 6
- LM for JPOzh-ja and JPOko-ja:
  5M sent. (with NTCIR-10 data),
  KenLM, order 6
- TM size: see later
- Distortion limit: 6 (semi-monotone)

Preordering part (1)

- En-ja, zh-ja, JPOzh-ja task only
- Stanford segmenter
- Berkeley parser
- Rule-based preordering

VP ⇒ ADVP VBN PP 2 0 1 (widely utilized in many fields)

VP ⇒ VV NP IP 1 2 0 (使 各个 电动机 13 旋转 驱动)

Preordering part (2)

- Improvements of the parser

  ✓ Chinese grammar improvement

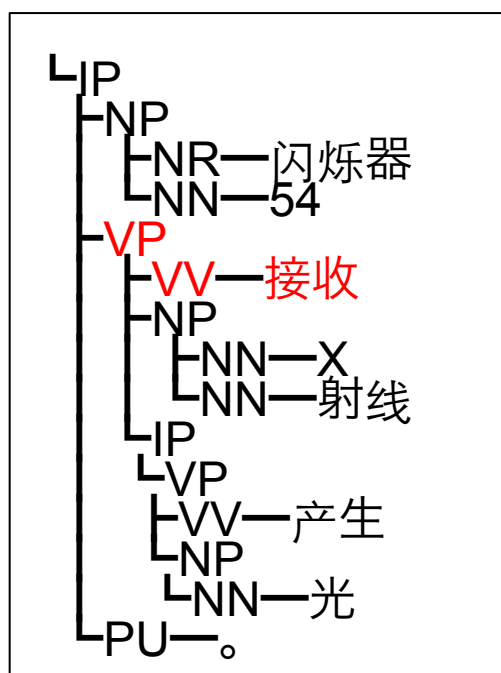 ✓ Reranking of k-best parse trees

Chinese grammar improvement（1）

- Determine gold standard of Chinese sentence head

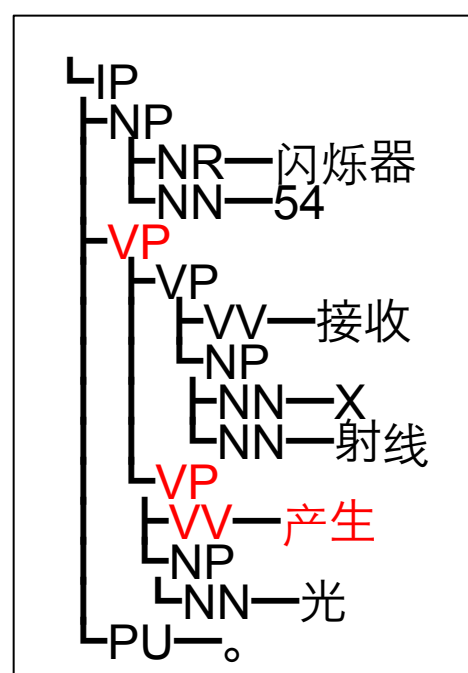シンチレータ 54 は 、 X線 を 受けて 光 を 発生 する 。

閃烁器 54 接收 X 射线 产生 光 。

- Select best parse tree from 100 parse trees

Chinese grammar improvement (2)

- Select the best parse tree ⇒ (b)

```
└IP
 ├NP
 │ ├NR─闪烁器
 │ └NN─54
 ├VP
 │ ├VV─接收
 │ ├NP
 │ │ ├NN─X
 │ │ └NN─射线
 │ └IP
 │   └VP
 │     ├VV─产生
 │     └NP
 │       └NN─光
 └PU─。
```
(a)

```
└IP
 ├NP
 │ ├NR─闪烁器
 │ └NN─54
 ├VP
 │ ├VP
 │ │ ├VV─接收
 │ │ └NP
 │ │   ├NN─X
 │ │   └NN─射线
 │ └VP
 │   ├VV─产生
 │   └NP
 │     └NN─光
 └PU─。
```
(b)

- Retraining the grammar

9

Chinese grammar improvement (3)

- Sentence head agreement rate with the gold standard (devtest data)

Grammar	Agreement rate (%)
Original (chn_sm5.gr)	50.5
Improved (chn_po.gr)	63.0

Reranking of k-best parse trees (1)

- Reranking with the reordering accuracy (WER, for training data)

Gold:闪烁器　54　X　射线　接收　光　产生　。

P(a):闪烁器　54　X　射线　光　产生　接收　。

P(b):闪烁器　54　X　射线　接收　光　产生　。

Reranking of k-best parse trees (2)

- Reranking with the LM score for dev, devtest and test data

  LM: trained by the reordered training data (1M)

  LM score = output of query[1] / # of words

  1. ~/mosesdecoder/bin/query

SMT part

- Phrase based Moses
- LM: same as SPE's LM
- TM:

Task	Corpus size
JPC ko-ja	994,998
JPC zh-ja	995,385
zh-ja	668,468
en-ja	2,351,575

- Distortion limit: 6 (semi-monotone SMT)

Candidate selection part

- Selection with the LM score

Number of selected candidates

Task	RBMT	RBMT+SPE	Preordering +SMT	Total
JPC ko-ja	25	1177	798	2000
JPC zh-ja	2	875	1123	2000
zh-ja	9	1270	828	2107
en-ja	5	658	1149	1812

JPOzh-ja results for devtest data

System	Additional feature	BLEU	RIBES
RBMT		16.55	0.7192
RBMT	User dictionary	23.54	0.7510
RBMT+SPE	User dictionary	41.35	0.8203
SMT	(DL=10)	40.86	0.8071
Preordering+SMT	original grammar	40.15	0.8164
Preordering+SMT	improved grammar	41.84	0.8218
Preordering+SMT	improved grammar + reranking of parse trees	42.75	0.8237
System combination		43.01	0.8265

Conclusion and future works

- Combination of rule-based method and statistical method.

- Human evaluation results:

Task	Crowd	JPO adequacy
En-ja	11/15	--
Zh-ja	2/11	3/3
JPO zh-ja	3/12	2/3
JPO ko-ja	4/13	--

- To improve parsing accuracy.