Global Pre-ordering for Improving Sublanguage Translation

12th Dec 2016, WAT 2016

M. Fuji, M. Utiyama, E. Sumita, Y. Matsumoto NICT, NAIST

Summary

- Improve translation of formal documents between distant languages by *global reordering*
- Learn global reordering model from non-annotated bilingual corpora
- Substantial improvement with global reordering + conventional reordering for JE translation

Background

Background

- Large volumes of formal documents are translated in industry
  - Eg. Operation manuals, law docs, patent docs
- Difficult to translate because:
  - Sentences are long
  - Requires “global reordering” between distant languages

Translation between distant languages

- Example: Patent abstract

Sentence pair 1	To provide a communication apparatus capable of performing highly reliable communication. 信頼性の高い通信を行うことができる通信装置を提供すること。
Sentence pair 2	To provide an image formation device which enables high image quality images to be formed. 高画質な画像を形成できる画像形成装置を提供する。

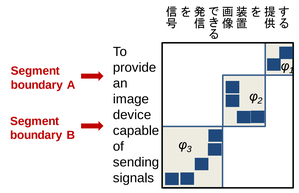
- Features
  - Segment boundaries marked with characteristic strings
  - Global reordering occurring at segment level

Proposed: Global Pre-ordering

1. Detect segments using characteristic strings
2. Globally reorder detected segments
3. Apply conventional reordering to each segment

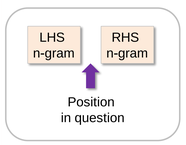
Prepare training data

- Extract sentences containing global reordering
  - Regard sentence to contain global reordering if segments in swap orientation in alignment table



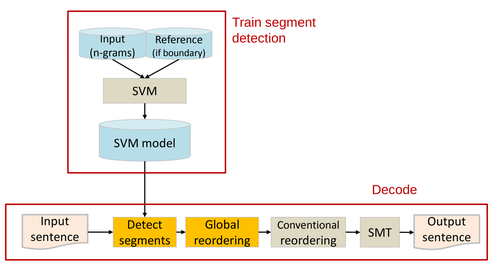
Train segment boundaries

Use n-grams to represent characteristic strings
- Input: n-grams at each position
- Reference: whether position is a segment boundary



Input to SVM		Reference
LHS 2-grams	RHS 2-grams	If segment boundary
<BOS> To	provide an	×
To provide	an image	✓
provide an	image device	×
an image	device capable	×
image device	capable of	✓
device capable	of sending	×

Experiment configuration



Experiment

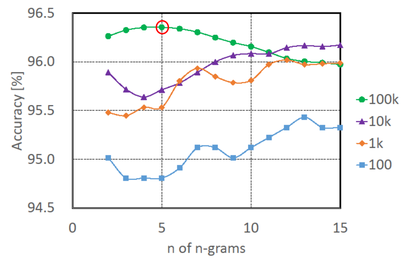
Experiment settings

- Data:
  - Patent Abstract of Japan (PAJ) with its original Japanese patents automatically aligned.
  - 1,000,000 sentence pairs for SMT training. 1,000 for developing and testing.
- Global reordering module:
  - Out of SMT training sentences, 100,000 pairs used.
  - Out of 100,000 pairs, 38,194 pairs required global reordering.
  - 38,194 pairs used to train SVM global reordering model

11

SVM tuning

- 10-fold cross validation: Ja-En



13

Results Ja->En

	Global preord	Syntactic preord	BLEU	RIBES
T1			17.9	44.9
T2	レ		19.3	61.0
T3		レ	25.5	64.9
T4	レ	レ	25.6	72.1

Improvement in RIBES but not in BLEU

14

Analysis

Results Ja->En

- Human eval added

	Global preord	Convent. preord	Preord type	BLEU	RIBES	Structure	Human eval
T1				17.9	44.9	12	15
T2	レ			19.3	61.0	31	33
T3		レ	Syntactic	25.5	64.9	21	36
			TDBTG	22.3	64.6	27	27
T4	レ	レ		25.6	72.1	58	65

RIBES close to human + Proposed method effective

Example translation

Reference: To provide a toner cake layer forming apparatus which forms a toner cake layer having a high solid content and which can be actuated by an electrostatic printing engine.

T1: Solid content of high toner cake layer for generating an electrostatic print engine operates in a toner cake layer forming device.

T2: To provide toner cake layer forming apparatus of the solid content of high toner cake layer for generating an electrostatic print engine can be operated.

T3: For generating toner cake layer having a high solids content and to provide a toner cake layer forming device which can be operated by an electrostatic printing engine.

T4: To provide a toner cake layer forming device for generating toner cake layer having a high solid content, and operable by an electrostatic printing engine.

T4 Segments in correct order + Each segment improved

Results En->Ja

- Human eval added

	Global preord	Convent. preord	Preord type	BLEU	RIBES	Structure	Human eval
T1				32.1	59.1	28	12
T2	レ			29.1	65.3	26	9
T3		レ	Syntactic	36.9	76.1	67	55
			TDBTG	34.9	77.7	59	37
T4	レ	レ		36.5	77.7	68	55

Proposed method NOT so effective for En->Ja