Preordering for Statistical Machine Translation

Automatically Learning Source-Side Reordering Rules

Riashat Islam

University of Cambridge

Outline

Preordering in Machine Translation

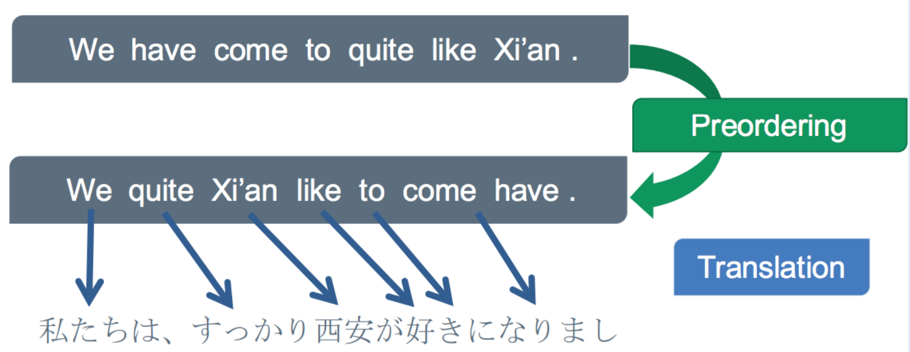
- Reorder the source sentence such that it best resembles the order of the target sentence
- Preprocessing step in phrase-based machine translation
- Transform source-sentence into target-like order

Benefits:

- Better translation models
- Speeds up decoding
- Improved translation quality for language pairs
- Benefifical for real-time commercial systems

Preordering for SMT

Preordering for SMT



Automatically Learning Preordering Rules

Automatically Learning Source-Side Reordering Rules for Large Scale Machine Translation (Genzel, 2010)

Automatically Learning Preordering Rules

- Automatically learn the reordering rules
- Language-independent preordering using source-side language parsers
- Learn rules for 8 different language pairs
- Important word order transformations can be captured using this approach

Goal: To find a method that works for many language pairs regardless of word order transformations needed and without language specific tuning

Related Approaches to Preordering

1. Other preprocessing based reordering approaches
   - Manually written rules for different languages
   - Common language pair, German-English (Collins, 2005)
   - Automatically learning reordering patterns for French-English (Xia and McCord, 2004)
   - Learn reordering rules based on sequences of part-of-speech tags (Rottmann and Vogel, 2007)
2. Automatically aligning source-parsed data
   - Feature-rich logistic regression model (Jehl, 2014)
   - Neural Network to learn node swapping model (Gispert, 2015). Briefly later...

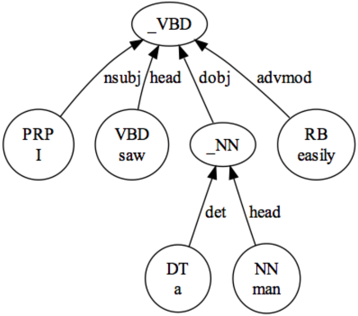
Approach

- Reordering of source side of training and test data
- Produce a parse tree on the source side
- Reordering of nodes in a parse tree of the source sentence
- Each reordering described by series of rules

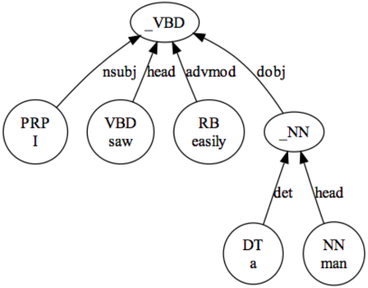
Overview

- Automatically learn a rule series for each language pair
- Tree transformed sequentially
- Reordered sentence read off the leaves of the tree

Approach



(a) A sample parse tree



(b) After reordering (moving RB over _NN)

Figure 1: Parse tree of a sentence and its reordering

Rule Space and Evaluation Metric

Rule Space

- Two pairs of a rule - Conditioning Context and Action
- Node matched against conditioning context
- Each condition is a feature-value pair
- Action on the node for a found match - swap the children of the node

Evaluation Metric

- Evaluate quality of given reordering rule
- Cross alignment links for a given aligned sentence pair
- *Estimated BLEU gain*

Algorithm

- Word-aligned sentence pairs as inputs
- Consider all rules
- Append the best new rule to the sequence according to a metirc
- Consider several variations that produces a different sequence of rules
- Choose the variation that performs best on development set

Several optimization variants for appending rules in the sequence

- Optimizing Crossing Score
- Optimizing Estimated BLEU gain
- Optimizing Estimated BLEU gain in sequence

Algorithm

Algorithm 1 Optimizing alignment links

```
input: A set of aligned sentence pairs
base = <empty sequence>;
for several iterations do
    candidate_rules = GenerateAllCandidateRules(input, base);
    base.append(MinCost(candidate_rules))
end for
```

Evaluation and Results

- Train system from English to 7 other languages
- Evaluate three variants of the algorithm
- Each algorithm outputs a reordering rule sequence which is applied to all the training and test data
- Proceed with phrase-based SMT, complete system trained from scratch - baseline decoder can also apply local reordering of upto 4 words
- BLEU scores for each of the algorithms

Languages Used: SOV languages (Japanese, Korean, Hindi), VSO language (Welsh), Long Distance Verb Movement (German), Noun modifier issues (Russian and Czech)

Evaluation and Results

Language	Base	Manual	Auto-matic	Diff
Hindi	16.85	19.25	19.36	0.11
Japanese	25.91	28.78	29.12	0.34
Korean	23.61	27.99	27.91	-0.08

Figure: Evaluation of Manual vs Automatic Reordering

Evaluation and Results

Language	Google %BLEU	Base %BLEU	Var. 1 gain	Var. 2 gain	Var. 3 gain	Best on dev gain
Czech	16.68	15.35	-0.08	0.13	**0.19**	0.19
German	20.34	18.65	**0.47**	0.30	0.39	0.39
Hindi	19.15	16.85	**2.25**	2.08	0.15	2.08
Japanese	30.74	25.91	**3.05**	2.60	**3.05**	3.05
Korean	27.99	23.61	3.34	3.77	**4.16**	4.16
Russian	16.80	15.33	0.08	**0.10**	**0.10**	0.08
Welsh	27.38	25.48	1.25	0.77	**1.43**	1.43

Figure: Comparison of Results on internal test set for 3 systems

Evaluation and Results

Language	Decoder reordering	No decoder reordering
Czech	0.21	0.08
German	0.72	0.55
Hindi	2.51	2.27
Japanese	3.21	3.21
Korean	4.30	4.15
Russian	0.14	-0.10
Welsh	1.34	0.98

Figure: Difference against baseline system in % BLEU gain

Neural Preordering

Using neural networks for source-side preordering (Gispert, 2015)

- Based on using a logistic regression model (Jehl, 2014)

Neural Networks to learn a node-swapping model

- Estimate node swapping probability in a parse tree using a logistic regression model
- Better modelling capabilites of NNs to achieve better performance at preordering
- Improves translation performance across various language pairs

First steps towards using NNs in preordering for SMT

Comparing Experimental Results

- Baseline (No Preordering)
- Rule-based approach (Genzel, 2010)
- Linear model logistic regression (Jehl, 2014)
- Neural Network based preordering (Gispert, 2015)

Comparison of Results			
System	**eng-jpn**	**eng-kor**	**eng-chi**
Baseline	50.9	28.7	44.8
Genzel(2010)	54.0	30.5	45.4
Jehl(2014)	55.0	33.1	45.8
Gispert(2015)	55.6	33.4	46.5

Summary and Discussion

- (Genzel, 2010) used automatic preordering as a preprocessing step generalized for various language pairs
- Neural Network based preordering obtains best BLEU scores across all language pairs
- Preordering approaches are independent of the MT system used
- Preordering helps obtain faster decoding times
- Improvements in translation quality with better BLEU scores

Thank You. . . Questions?