بسم الله الرحمن الرحیم

Syntactic Reordering of Source Sentences for Statistical Machine Translation

Mohammad Sadegh Rasooli

Columbia University

rasooli@cs.columbia.edu

April 9, 2013

Overview

1. First Paper: Collins, et al. (2005)
   - The Role of Syntax in SMT
   - Syntactic Preprocessing Approaches
   - Clause Restructing
   - Experiments
   - Discussion

2. Second Paper: P. Xu, et al., (2009).
   - Approaches to Syntactic Reordering
   - Translation Between SVO and SOV Languages
   - Precedence Reordering Based on a Dependency Parser
   - Experiments
   - Discussion

First Paper

M. Collins, et al.: Clause Restructuring for Statistical Machine Translation. ACL 2005.

The Role of Syntax in SMT

- In the original phrase-base SMT, syntax is not taken into acount.
- Phrase-based systems have limited potential to model word-order differences between languages.
  - The word order differences between languages are considered as distortion.
  - Each reordering rule adds distortion penalties to the overall score of the translation model.

Example: German vs. English Word Order

English

I will pass on to you the corresponding comments, so that you can adopt them perhaps in the vote.

German

I will to you the corresponding comments pass on, so that you them perhaps in the vote adopt can.

Research on Syntax in MT

- Changing the word order of one of the languages or both, to make their word order more similar to each other.
- Syntax-Based MT Approaches
  - Make use of bitext grammars to parse both parts.
  - Change the syntax of target language alone.
    - Transform the translation problem into a parsing problem.
  - Reranking methods
    - Select between N-best results of the phrase-based system, using syntactic information.
- Preprocessing Approaches
  - The source language sentences are modified before translation.
  - This approach is used in this paper.

Syntactic Preprocessing Approaches

English

I will pass on to you the corresponding comments, so that you can adopt them perhaps in the vote.

German

I will to you the corresponding comments pass on, so that you them perhaps in the vote adopt can.

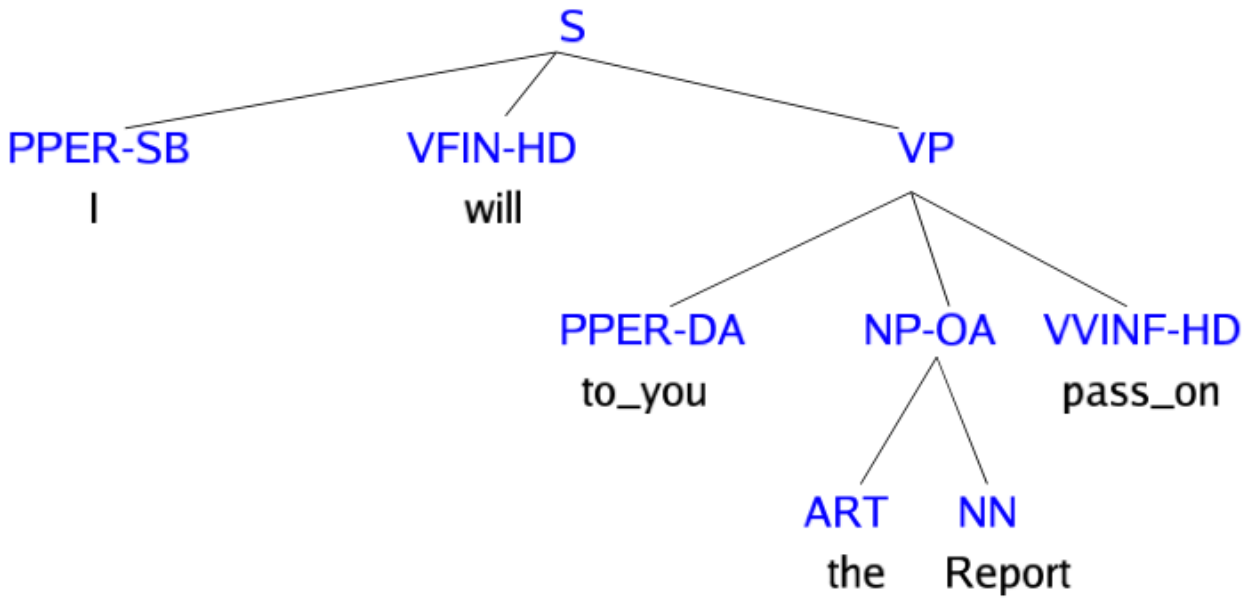
German (Preprocessed)

I will pass on to you the corresponding comments, so that you can adopt them perhaps in the vote.

Clause Restructing

- Steps (both in training and decoding)
  1. Parse the source sentence.
  2. Apply reordering rules on the source sentence.
  3. Use phrase-based models.

Example Parse Tree



Six Reordering Rules in German

Transformation	Example
Verb Initial	Before: Ich werde Ihnen die entsprechenden Anmerkungen **aushaendigen**, . . . After: Ich werde **aushaendigen** Ihnen die entsprechenden Anmerkungen, . . . I shall be passing on to you some comments, . . .
Verb 2nd	Before: . . . damit Sie uebernehmen das eventuell bei der Abstimmung **koennen**. After: . . . damit **koennen** Sie uebernehmen das eventuell bei der Abstimmung . . . . so that could you adopt this perhaps in the voting.
Move Subject	Before: . . . damit koennen **Sie** uebernehmen das eventuell bei der Abstimmung. After: . . . damit **Sie** koennen uebernehmen das eventuell bei der Abstimmung . . . . so that you could adopt this perhaps in the voting.
Particles	Before: Wir fordern das Praesidium **auf**, . . . After: Wir **auf** fordern das Praesidium, . . . We ask the Bureau, . . .
Infinitives	Before: Ich werde der Sache **nachgehen** dann, . . . After: Ich werde **nachgehen** der Sache dann, . . . I will look into the matter then, . . .
Negation	Before: Wir konnten einreichen es **nicht** mehr rechtzeitig, . . . After: Wir konnten **nicht** einreichen es mehr rechtzeitig, . . . We could not hand it in in time, . . .

Table 1: Examples for each of the reordering steps. In each case the item that is moved is underlined.

Experiments

- Experimental setup
  - Data: Europarl Corpus.
  - 751,088 parallel sentence.
  - Evaluation on 2000 sentences.
  - Average sentence length: 28 words
  - Baseline: no reordering phrase-based system.
- Results (BLEU score)
  - Basline: 25.2%
  - Reordering: 26.8%

Human Translation Judgments

- Two annotators judged 100 sentences (10 to 20 words in length; chosen at random).
- Three versions: Human, baseline, reordered.
- Judgments: Worse/better or equal.

	Better	Equal	Worse
Annotator 1	40%	40%	20%
Annotator 2	44%	37%	19%

Example Output

Human

i think it is wrong in principle to have such measures in the european union.

Reordered

i believe that it is wrong in principle **to take** such measures in the european union.

Baseline

i believe that it is wrong in principle such measures in the european union **to take**.

BLEU Statistical Significance

- Authors use sign test for statistical significance.
  - f(x) is $+$ if better than baseline, f(x) is - if worse; and f(x) is $=$ if equal
  - p_+: probability of (f(x) is $+$) and p_-: probability of f(x) is minus
- BLEU does not have per-sentence evaluation.
- Authors create an artificial comparison:
  - s baseline BLEU score
  - s_i baseline BLEU score except the sentence i translated by the reordered model.
  - f(x) is $+$ is $s_i > s$; f(x) is - is $s_i < s$.
- 52.85% improved, 36.4% worse than baseline and 10.75% equal.
- With 95% confidence, this method improves the baseline.

Discussion

- The method clearly improves the baseline.
- The rules are language-specific (even cannot be used for English to German translation).
- The authors did not try to learn reordering rules automatically.

Second Paper

P. Xu, et al., Using a dependency parser to improve SMT for subject-object-verb languages. NAACL 2009.

Approaches to Syntactic Reordering

- Explicitly model phrase reordering distances; e.g. distance based distortion models.
- Syntactic analysis of the target language into both modeling and decoding.
- Reordering source sentences based on syntactic analysis
  - This paper uses this approach

Translation Between SVO and SOV Languages

- Subject-Verb-Object (SVO) and Subject-Object-Verb (SOV) are two common word order in the world languages.
- English is SVO and Korean is SOV.
  - "John hit the ball." vs. "John the ball hit."
- When the sentences get longer, the cost of moving structures during decoding (in phrase-based models) can be quite high.
  - "English is used as the first or second language in many countries around the world."
    $\rightarrow$ "is used" should skip 13 words to go to the end of the sentence.

Precedence Reordering Based on a Dependency Parser

- The children of each word have some relative ordering.
- A **Precedence reordering rule** is a mapping from T to a set of tuples $\{(L, W, O)\}$
  - T: POS tag
  - L: Dependency label
  - W: Weight indicating the order (highest to lowest)
    - Children with the same weights are ordered according to the order defined in the rule.
    - Why not explicitly pre-define unequal weights?
  - O: order type
    - NORMAL: preserve the original order
    - RESERVE: flip the order
- If a node is not listed in the rules, $W = 0$ and $O = NORMAL$
- Use **self** to refer to the head node itself.
- Punctuations and conjugations disallow movements across them.

Precedence Reordering Based on a Dependency Parser

T	(L, W, O)
VB*	(advcl, 1, NORMAL) (nsubj, 0, NORMAL) (prep, 0, NORMAL) (dobj, -1, NORMAL) (prt, -2, REVERSE) (aux, -2, REVERSE) (auxpass, -2, REVERSE) (neg, -2, REVERSE) (self, -2, REVERSE)
JJ or JJS or JJR	(advcl, 1, NORMAL) (self, -1, NORMAL) (aux, -2, REVERSE) (auxpass, -2, REVERSE) (neg, -2, REVERSE) (cop, -2, REVERSE)
NN or NNS	(prep, 2, NORMAL) (rcmod, 1, NORMAL) (self, 0, NORMAL)
IN or TO	(pobj, 1, NORMAL) (self, -1, NORMAL)

Table 1: Precedence Rules to Reorder English to SOV Language Order (These rules were extracted manually by a bilingual speaker after looking at some text book examples in English and Korean, and the dependency parse trees of the English examples.)

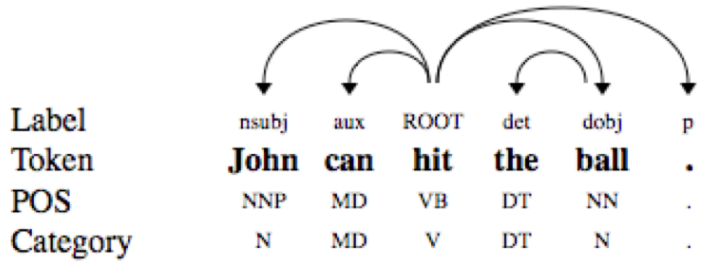


Figure 2: Dependency Parse Tree of an Example English Sentence

After apply precedence rule, this will be: John the ball hit can.

Novelties in This Work

1. This model is more efficient than its counterpart.
2. Outperforms the state-of-the-art (stronger baseline).
3. It is not restricted to one language pair.
4. It is possible to automatically learn precedence rules.
5. They use dependency parse trees rather than constituency trees.

Experiments

- English to 5 SOV languages.
- Baseline: Maximum entropy based lexicalized phrase reordering model.
  - Maximum allowed reordering: 10.
- Parser: Deterministic transition-based dependency parser.
  - Parses in linear time.
- Another baseline: Hierarchical phrase-based system.
  - Can capture long distance reordering by using a PCFG model.
  - Uses chart parsing during decoding: slower than deterministic dependency parser.
- 9.5K English sentences (from web) as evaluation data.
  - 3,500 sentences for *dev* (to perform MERT).
  - 1,000 sentences for *test*.
  - 5,000 sentences for *blind test*.

Experiments

System	Source	Target
English→Korean	303M	267M
English→Japanese	316M	350M
English→Hindi	16M	17M
English→Urdu	17M	19M
English→Turkish	83M	76M

Table 2: Training Corpus Statistics (#words) of Systems for 5 SOV Languages

Results

Language	System	dev	test	blind
Korean	BL	25.8	27.0	26.2
	-LR	24.7	25.6	25.1
	-LR+PR	27.3	28.3	27.5**
	+PR	27.8	28.7	27.9**
Japanese	BL	29.5	29.3	29.3
	-LR	29.2	29.0	29.0
	-LR+PR	30.3	31.0	30.6**
	+PR	30.7	31.2	31.1**
Hindi	BL	19.1	18.9	18.3
	-LR	17.4	17.1	16.4
	-LR+PR	19.6	18.8	18.7**
	+PR	19.9	18.9	18.8**
Urdu	BL	9.7	9.5	8.9
	-LR	9.1	8.6	8.2
	-LR+PR	10.0	9.6	9.6**
	+PR	10.0	9.8	9.6**
Turkish	BL	10.0	10.5	9.8
	-LR	9.1	10.0	9.0
	-LR+PR	10.5	11.0	10.3**
	+PR	10.5	10.9	10.4**

Table 3: BLEU Scores on Dev, Test and Blindtest for English to 5 SOV Languages with Various Reordering Options (BL means baseline, LR means maximum entropy based lexialized phrase reordering model, PR means precedence rules based preprocessing reordering.)

Results

Language	System	dev	test	blind
Korean	PR	27.8	28.7	27.9
	Hier	27.4	27.7	27.9
	PR+Hier	28.5	29.1	28.8**
Japanese	PR	30.7	31.2	31.1**
	Hier	30.5	30.6	30.5
	PR+Hier	31.0	31.3	31.1**
Hindi	PR	19.9	18.9	18.8
	Hier	20.3	20.3	19.3
	PR+Hier	20.0	19.7	19.3
Urdu	PR	10.0	9.8	9.6
	Hier	10.4	10.3	10.0
	PR+Hier	11.2	10.7	10.7**
Turkish	PR	10.5	10.9	10.4
	Hier	11.0	11.8	10.5
	PR+Hier	11.1	11.6	10.9**

Table 4: BLEU Scores on Dev, Test and Blindtest for English to 5 SOV Languages in Hierarchical Phrase-based Systems (PR is precedence rules based preprocessing reordering, same as in Table 3, while Hier is the hierarchical system.)

Discussion

- Reordering of languages with different word orders is essential.
- The method seems to work fine for 5 languages.
- Although authors claim that the rule can be extracted automatically, they did not try.
- The improvement of the basic over hierarchical phrase-based is not significant.

Thanks!