Improving English-Vietnamese Statistical Machine Translation Using Preprocessing Dependency Syntactic

Viet Hong Tran[1,2] , Vinh Van Nguyen[2] and Minh Le Nguyen[3]

[1]University of Economic and Technical Industries

[2]University of Engineering and Technology

[3]Japan Advanced Institute of Science and Technology

INTRODUCTION

Phrase-based Statistical Machine Translation

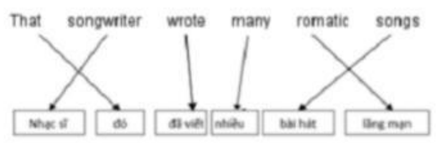
- The state of the art model
- Distance base word reordering
- Short distance and local context

A new method to solve word-reordering problem

- Using dependency parsing for preprocessing with training and testing
- Applying transformation rules to reorder the source sentence
- An English-Vietnamese Machine Translation System.

For example:

English sentence: that songwriter wrote many romatic song.
In Vietnamese order: songwriter that wrote many song romantic.
Vietnamese sentence: nhạc_sĩ đó đã_viết nhiều bài_hát lãng_mạn.



RELATED WORKS

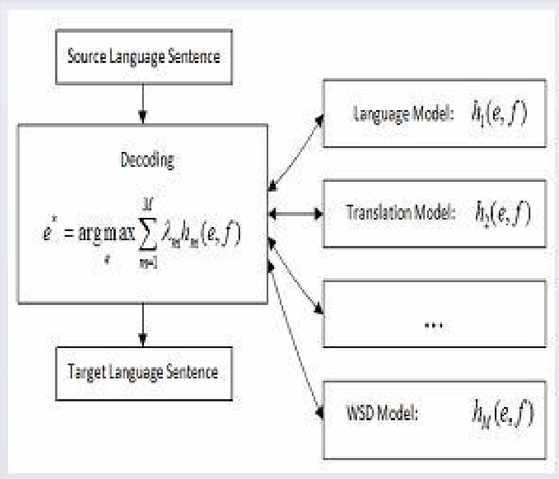
Reorder word during preprocessing step

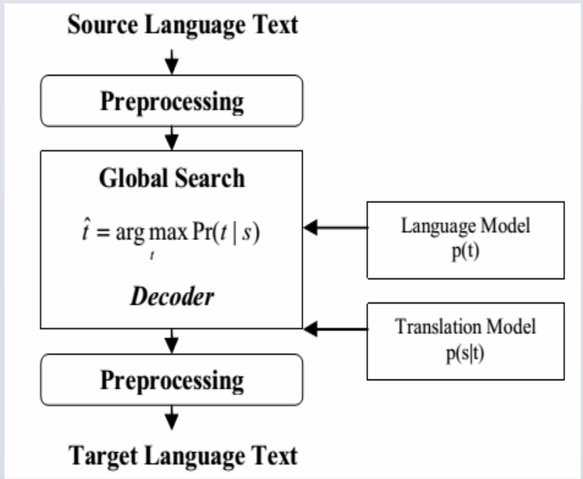
- Automatic extracted rules
- Manually written rules

Using dependent parse tree and flexible rule

- Rule transformation: lexical CFG rules
- Precedence ranking: Assign a precedence score to each clause and sort
- Tree transformation
- Classifier: predict the target word order by treating each permutation as a label in a multi class classifier.

SYSTEM OVERVIEW

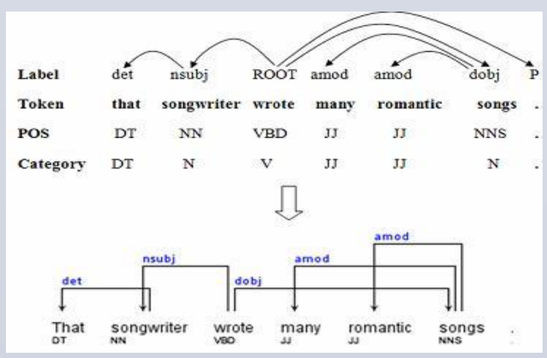


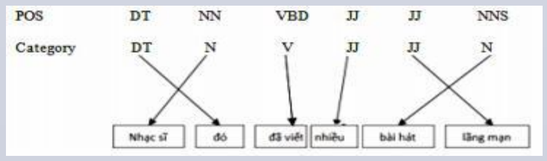


TRANSFORMATION RULES

Using the dependency grammars and the differences of word order between English and Vietnamese to create the set of the reordering rules

Source sentence:
that songwriter wrote many romantic songs .
Tagging:
that/DT songwriter/NN wrote/VBD many/JJ romantic/JJ songs/NNS ./.
Parse:
(ROOT
(S
(NP (DT that) (NN songwriter))
(VP (VBD wrote)
(NP (JJ many) (JJ romantic) (NNS songs)))
(. .)))
det(songwriter-2, that-1)
nsubj(wrote-3, songwriter-2)
root(ROOT-0, wrote-3)
amod(songs-6, many-4)
amod(songs-6, romantic-5)
dobj(wrote-3, songs-6)





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

Handwritten rules for reordering English to Vietnamese using dependency syntactic preprocessing

EXPERIMENTS

Corpus	Sentence Pairs	Training Set	Development Set	Test Set
General	55341	54642	200	499
			English	Vietnamese
Training	Sentences		54620	
	Average Length		11.2	10.6
	Word		615478	580754
	Vocabulary		23804	24097
Development	Sentences		200	
	Average Length		11.1	10.7
	Word		2221	2141
	Vocabulary		825	831
Test	Sentences		499	
	Average Length		11.2	10.5
	Word		5620	6240
	Vocabulary		1844	1851

SIZE OF PHRASE TABLE

Name	Size of phrase-table
Baseline	1152216
Applying rules with category JJ or JJS or JJR	1218676
Applying rules with category NN or NNS	1228187
Applying rules with category IN or TO	1231365

BLEU SCORE

Name	Bleu score (%)
Baseline	36.97
Applying rules with category JJ or JJS or JJR	36.75
Applying rules with category NN or NNS	37.51
Applying rules with category IN or TO	37.71

CONCLUSION

- An English-Vietnamese Automatic Translation System.
- Using preprocessing approach based on a dependency parser and applying for systems translate English to Vietnamese.
- Improvement over 0.74 BLEU point is valuable because baseline system is the strong phrase-based SMT
- This approach has the potential and high performance: transformation rules are flexible and cover many linguistic reordering phenomena.