IJCSC 0973-7391 Vol- 6 • Issue-1 Sep - Mar 2015 pp.145-147 **Impact Factor-2.5** available at www.csjournalss.com

Tuning of Moses Decoder for Dogri SMT

Shubhnandan S Jamwal, Sunil Dutt
Department of Computer Science & IT, University of Jammu, Jammu, India
Jamwalsnj@gmail.com

Abstract: Moses is Statistical Machine Translation (SMT) System widely used as open source SMT in Research. Tuning is the process by which Speed and Quality of translation is improved. There are various parameters that affect the performance of Moses SMT. Translation Table size, Stack size, Language model, reordering model and word penalty are the key parameters which affect the performance the Moses Decoder. In this research paper we have tested these parameters for better speed and quality of the Moses SMT.
Keywords: Moses, Language Model, Stack, Translation table, Decoder, Speed and Quality.

Introduction

Machine Translation (MT) is the process of translating text from Source Language to a Target Language. There are various approaches to MT. These are Rule-Based Machine Translation, Transfer-based Machine Translation, Interlingual Machine Translation, Statistical Machine Translation and Example- based Machine Translation. SMT is the dominant approach in the field of machine translation. In SMT translation machine is trained on large quantities of parallel data. Collection of sentences in two different languages are known as parallel data, which is sentence-aligned, i.e. sentence in one source language is corresponding to translated sentence in the target language which is also known as a bitext. There are two approaches to SMT, i.e. Phrase-based MT and Syntactic MT or Factored MT. In this paper we are experimenting with Phrase-based MT. In Phrase-based machine translation system, sequences of words known as phrases are used for translation between two languages of interest and the job of the Moses decoder is to find the highest scoring sentence (phrase) to a given source sentence (phrase) in the target language.

Tuning is the process by which Speed and Quality of translation is improved. We can improve quality and speed of translation by tuning the decoder. The probability i. e. assigned to a translation is a product of probability costs of phrase translation table, language model, reordering model and word penalty.

- The phrase translation table ensures that source language phrases are good translation of target language phrases.
- The language model ensures that the output is fluent in target language.
- The reordering (distortion) model allows for reordering of source sentence.
- The word penalty ensures that don’t get too short or too long.

Review of Literature

Adam Lopez [1] in 2008 published a comprehensive survey of Statistical Machine Translation at University of Edinburgh. According to this report, statistical systems can be word based and phrase based.

Peter E. Brown and et al. [2] described the five statistical models of the translation process and proposed the algorithms for estimating the parameters of the models. On the bases of this research the IBM developed model 4 and it is described as target-to-source model. It produces the source sentence f from the target sentence e. The model can be explained in three steps

- Each target word e, select a fertility n and copies itself n times.
- Each copy of each target word is translated to a single source word.
- The source words are reordered into their final positions.

Model 4 alignment is asymmetric. Each source word can align to exactly one target word or the null word. However, a target word can link to arbitrary source words, as defined by fertility. The major problem with this model is of word reordering.

Koehn et al. [3] described an open source toolkit for statistical machine translation system known as Moses and it includes wide variety of tools for training, tuning and applying the system many translation tasks.

Vishal Goyal et al. [4] discussed large number of Machine Translation System developed for non-Indian and Indian Languages. A few Indian systems are found to use Statistical Approach partially. Anglabharti-II(2004) uses statistical language model for automatic post editing.

Unnikrishhanan et al. [5] explained the development of machine translation system using statistical approach for translating English to South Dravidian language like Malayalam and Kannada. The various tools used for the development of the said system are: SRILM for creating language model, GIZA++ for training translation model and MOSES decoder for translating English to Malayalam (or Kannada). Other tools used at various level of translation process are: The Stanford statistical parser, Roman to Unicode and Unicode to Roman converter, Morphological analyzer and generators, English morphological analyzer, Malayalam and Kannada morphological analyzers, Malayalam and Kannada morphological generators and Transfer rule file. The architecture of the system is:
The main ideas were implemented and proven very effective for English to south Dravidian languages. The SMT system proposed is elaborated in the following steps

- Reordering of the English source sentence according to Dravidian syntax,
- Using the root suffix separation on both English and Dravidian words and
- Use of morphological information for improving the quality of translation.

Experimentation

Quality

Quality of translation can be improved by tuning of the model parameters which are as follows:-

- phrase translation table,
- language model,
- reordering model, and
- word penalty.

Each parameter can be given a weight that sets its importance. Mathematically, the cost of translation is:

$$p(e|f) = PT(f|e)^\wedge weight\_t * LM(e)^\wedge weight\_l * DM(e,f)^\wedge weight\_d * WP(e)^\wedge weight\_w$$

Where, the probability $p(e|f)$ of the English translation e given the foreign input f, $PT(f|e)$ is phrase translation, $LM(e)$ is language model, $DM(e,f)$ is distortion model, and word penalty $WP(e)$.
weight-t, weight-l, weight-d, and weight-w are the weights provided to the the four parameters. 1, 1, 1, and 0 are the default values in the configuration file moses.ini. Setting these weights to the right values can improve translation quality. For example see the table below setting the distortion weight to 0, get the right translation.

Input(German)	Weight_t	Weight_l	Weight_d	Weight_w	Output(English)	Result
ein haus ist das	1	1	1	0	a house is the	Incorrect
ein haus ist das	1	1	0	0	this is the house	Correct

Speed

Speed can be improved by limiting the search space. There are two ways to reduce the search space i.e. limiting the phrase translation table size and reducing the stack size to some extent.

Translation Table Size

One strategy to limit the search space is by reducing the number of phrase translation table entries that are retrieved. See the table below which shows the translation output with no translation table limit and a limit of 1 for input *'das ist ein kleines haus'* .

Ttable_Lim	Tt_Options	T_hypo_gen	N_recomb	N_pruned	N_discard	B_trans	Log_Prob
No limit	12	453	69	0	272	this is a small house	-28.923
1	6	127	8	0	61	it is a small house	-30.327

Reducing the number of translation options to only one per phrase, had a number of effects:
- Only 6 translation options instead of previously collected12 translation options.
- Total hypothesis generated 127 instead 442, and no hypotheses were pruned out.
- Recombined hypothesis are only 8 instead 69
- The translation changed, and the output now has lower log-probability: -30.327 vs. -28.923.

Hypothesis Stack Size

Other way to reduce the search space is by reducing hypothesis stacks size and by this change search will be faster, because of less number of hypotheses generated. See the table below which shows output for stack size 1000, 100 (the default), 10, and 1. This shows that the number of hypothesis generated getting smaller with the stack size: 453, 453, 208, and 29. In this example, translation changed with a stack size of 1 is a worse translation, with worse scoring of (-30.991 vs. -28.923). But achieving, best scoring translation is the job of good decoder. If decoder gives us good quality with worse scoring, then there is a problem of the model, which can be resolved by better modeling.

Stack_Lim	T_hypo_gen	N_recomb	N_pruned	N_discard	B_trans	Log_Prob
1000	453	69	0	272	this is a small house	-28.923
100	453	69	0	272	this is a small house	-28.923
10	208	23	42	103	this is a small house	-28.923
1	29	0	4	19	this is a little house	-30.991

Conclusion/Future Work

It has been observed that if the size of the stack limited upto 10 the output remains same with log probability -28.923 which is the best scoring by our model with best translation. This shows that by increasing the stack won't improve the quality of translations which clear from the output of stack size 100 and 1000. So we can conclude that speed of translation can be improved by some extent without compromising the quality of translation. In future work we would find the effects other parameters on the speed and translation quality of this model.

References

[1] Lopez, A., "Statistical machine translation". *ACM Computing Surveys.*, 40, 3, Article 8, 2008, 49 pages. http://doi.acm.org/10.1145/1380584.1380586

[2] Peter F. Brown, Stephan A. Della Pietra, Vincert J. Della Pietra, and Robert L. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation", IBM T.J, Watson Research Center, Yorktown Heights, NY 10598, *Association for Computational Linguistics*, 19, 2, 1993, pp 263-311.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Richard Zens, Christine Moran, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, "MOSES: Open Source Tool Kit for Statistical Machine Translation", *Proceedings of Association for Computational Linguistics,* Prague, Czech Republic, 2007, pp 177-180.

[4] Vishal Goyal, Gurpreet Singh Lehal, "Advances in Machine Translation Systems", *Languages in India*. Vol. 9, ISSN 1930-2940, 11 November 2009, pp.138-150

[5] Unnikrishnan P, Antony PJ and Soman KP, 2010. "A Novel Approach for English to South Dravidian Language Statistical Machine Translation System", *International Journal on Computer Science and Engineering(IJCSE)* 2, 8, pp 2749-2759.

[6]. www.statmt.org