Improvements in Corpus Quality for Statistical Machine Translation

Shikha Maheshwar[1] Himanshu Sharma [2]
[1] JECRC, Jaipur [2]JECRC-UDML College of Engineering, Jaipur

.

Abstract— In this paper, we tried to explore what data quality means for parallel corpuses. This work is motivated by our attempts to understand the factors which can affect the quality of corpus for statistical machine translation.

Key words: Statistical Machine Translation, IBM models, parallel corpuses .

I. INTRODUCTION

Machine translation is one of the earliest areas of research in natural language processing. With 26 constitutionally recognized languages, India is, no doubt, a highly Multilanguage country. Still, English is understood by, less than 3% of Indian population and therefore, machine translation is required for breaking language barrier within the sociological structure of the country. Here comes the parallel corpus into the scene playing an important role in machine translation. A parallel corpus can be defined as a collection of text, paired with translations into another language.

As we are very well aware that English is a highly positional language with rudimentary morphology with the default sentence structure as "subject-verb-object". In contrast, the Indian languages are highly instructional, with such morphology, selectively free word order, and default sentence structure as "subject-object-verb". Apart from this; there are many stylistic differences too.

As the parallel content of Source and Target language is made available with increased capacity of memory & high processing speed, the trend is moving towards Statistical Machine Translation. SMT models rely heavily on the available bilingual corpus. As it is already known to us that the corpus quality plays a significant role in improving statistical machine translation quality, for this parallel corpora is developed generally as a collection of English corpus of various domain from various resources, or by generating multiple references for each sentence by getting it translated by different expert translators. However, the large amount of data causes will cast more computational resources too. Therefore, a need for compact, clean, normalized corpus is expected, which in turn also improves BLEU scores as comported to raw data.

II. OVERVIEW OF SMT

Brown et al [2] practically initiated the statistical approach to machine translation which is presented to the world in the form of IBM models 1 to 5, giving a completed mathematical formulation [5].In SMT, basically, it is given a source language sentence set S which is to be translated in target language sentences set T. SMT is based on a noisy channel model & requires a parallel corpus, in which each sentence given in S is aligned to its translation in T. Here, it is considered T as the target of communication channel & S as the source of the channel. System is able to generate multiple translation problem identifies the best translation sentence T for the source sentence S. Therefore, the machine translation tasks become the recovery of the source, from the target, and that's why the need to maximize P(T/S) arises. According to the Bayes Rule:-

$$t^* = \arg\max P(T|S) \qquad (2.1)$$

$$= \arg\max P(S|T) * P(T) / P(S) \qquad (2.2)$$

As P(S) is constant,

$$t^* = \arg\max P(S|T) * P(T) \qquad (2.3)$$

Here in (2.3), P(S|T) represents Translation Model & P(T) represents language model.

Translation model plays the role of ensuring translation faithfulness& language model to ensure the fluency of translated output. Here, a very large collection of sentences aligned to their corresponding translation is required by an algorithm to learn translation parameters. However, many experiments have been carried with source resource language pairs with modest & complete collection.

III. LITERATURE REVIEW

Till date, many researchers have focused on data collection for training data and development data. Resnik& Smith (2003) has extracted parallel sentences from web resources as much focus was given in large collection of parallel data for training. Eck et.al (2005) used unseen n-gram contained in the sentences for measuring the importance of the sentence. However, using unseen n-gram coverage they only considered its quantity. Weight was not taken in to account for this research work. Lü et.al (2007) has applied the Information Retrieval methods for data collection with the assumption that the target test data must be previously known before building ant translation model. But the limitation of this method was that the test text must be known earlier. Snover et.al (2008) has used comparable corpora for improving the performance of translation. Yasuda et.al (2008) selected parallel translation pair from out-of-domain corpus using perplexity as the measure. They have also done a certain amount of work for integrating the translation model using linear interpolation. Matsoukas et.al (2009) assigned a weight for each sentence in the given training data using discriminative training method and hence limited the negative effects of low quantity training data. Liu et.al (2010) considered the estimation of weight of phrases from test data for data selection for development set. However this method is totally dependent on test data which is a limitation.

As mentioned above, most of the work focused on the training data and little attention is paid to development set. In this research work, for improving the quality of corpus, it is proposed to work with both training as well as development data. The high quality sentences will be chosen for constructing the translation model and for tuning the translation parameters.

IV. RESOURCE REQUIREMENTS

For exposing the meaning of quality of data for bilingual parallel corpus, we have used English-Hindi-Parallel data from the EMILLE corpus for our experiments. EMILLE corpus is electronic collection of 63 million words of south Asian languages, especially spoken as minority languages in UK. It contains around 1,20,000 words of parallel data in each of English, Hindi, Gujarati, Sinhala & Tamil (Baker etc at 2004). Generally, the possibilities & parameters can be made more accurate & better by using more data for training & tuning SMT.

For this Moses [5] toolkit along with GIZA++ (a software for word/phases alignment) & a utility for making bilingual word classes, mkcls are used for training. For tuning MERT [10] script was used while BLEU [8] was used for testing.

V. IMPROVING PARALLEL CORPUS QUALITY

In SMT, the quality of a corpus is improved usually by removing noise present in data. The noise is classified in both source and target language as format noise or semantic noise. The format noise includes the HTML/XML tags, wrongly encoded words/characters, multi-bytes symbols in English language such as Greek symbols, currency symbols, full-width and half-width letters, numbers and punctuations etc. For vocative case, punctuation sign may be used in source language but not necessarily such symbols is detected on the target language every time, similarly, there may be mismatch of colon, bullets numbering & paragraphing.

However, the semantic noise is consist of the misaligned pairs of sentences in source-target language, length wise mismatched pairs, wrongly swapped pairs in both languages etc. In this research work, much focus in thrown in handling the noise related problem of second category.

Source of some corpora is from web and the sentence pairs are aligned automatically by using alignment tools. Therefore, it was expected to contain some misaligned sentences in it. As a measure for this problem, a parallel lexical dictionary can be created using relatively clean data sets just to find out whether the meaning in source and target language matches or not. To improve the accuracy, it is tried to keep only real words in the lexical dictionary. This also reduced the negative impact, if any, induced by the prepositions.

The problem of length wise mismatched pairs indicates that the length ratio between source and target language sentences is not reasonable, i.e. one side is too much longer than the other side in terms of numbers of words or characters which will decrease the alignment accuracy. Also a limitation is applied that the length of each sentence in both the languages must be longer than 5 words because it is assumed that some short sentences are only composed of abbreviations. The problem of wrongly swapped pairs indicates that some source language sentences are wrongly appeared in the target language and vice -verse. It is easily solved by detecting the encoding characters.

The sentence length is also limited for GIZA++ training. The default setting of maximum sentence length for GIZA++ is 100 words which would relatively slow down the alignment speed and increase the alignment complexity. In order to speed up the alignment process and have a better word alignment result, the Perl script wrapped in the Moses toolkit is used to limit the sentence length no more than 60 words.

If these problems of semantic noise are overcome, the data is then classified as “clean & normalized data” containing consistent data values, Apart from this if the same word or phases has been consistently used when same concept is referred throughout the corpus, it is then said attainment of value consistency the terms of statistical machine translation. SMT also supports the feature of data currency if the sentence translated from source to target years ago, it will still give the same translation results in an up-to-date data. If the training data contains all the information for a successful translation of source to target, the data is said to be complete for SMT.

VI. CONCLUSION

There is always need of more parallel text for appropriate learning of translation parameters. In this paper, we have tried to classify the noise which may present in different ways in the English-Hindi corpus.

REFERENCES

[1] PF Brown, J Cocke, S A D Pietra, V J D Pietra, F Jelinek, J D Lafferty, R L Mercer, P S Roossin: *A Statistical Approach to Machine Translation*. Computational linguistics Volume 16, Number 2, June 1990.

[2] Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della, and Robert L. Mercer: *The Mathematics of statistical machine translations: parameter estimation.* Computational Linguistics, 19(2): 263-311, 1993.

[3] P F Brown, S A D Pietra, V J D Pietra, R L Mercer: *The Mathematics of Statistical Machine Translation: Parameter Estimation.* ACL 1993. Computational Linguistics Volume 19, Issue 2, June 1993, Pages: 263 -311.

[4] J Hutchins: *Research methods and system designs in machine translation – a ten - year review*, 1984-1994. International conference ‘Machine Translation: ten year on’, Cranfield University, England, 12-14 November 1994.

[5] D Jurafsky, J H Martin: *Speech and Language Processing*. 2nd Edition. May 2008.ISBN-10: 0131873210

[6] P Koehn, H Huang, A Birch, C Callison -Burch, M Federico, N Bertoldi, B Cowan, W Shen, C Moran, R Zens, C Dyer, O Bojar, A Constantin, E Herbst. *Moses: Open Source Toolkit for Statistical Machine Translation.* ACL Demos.

[7] A Lopez. Statistical Machine Translation. ACM Computing Surveys (C Sur), Volume 40, Issue 3, Article No. 8, August 2008.

[8] K. Papineni, S. Roukos, T. Ward and W-J. Zhu, “*BLEU: a method for automatics evaluation of*

Machine Translation", in proc. Of 40th ACL, Philadelphia, Pennsylvania, USA, 2002, pp. 311-318

[9] F J Och, H Ney: *A Syntactic Comparison of Various Statistical Alignment Models*. Computational Linguistics Volume 29, number 1, pp 19-51, March 2003.

[10] F J Och: *Minimum error rate training in statistical machine translation.* Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, p.160-167, July 07-12, 2003, Sapporo, Japan.

[11] Raghavendra U, T A Faruquie: *An English –Hindi Statistical Machine Translation System.* IJCNLP 2004, LNAI 3248, pp. 254-262.

[12] Thomas C. Redman, *Data Quality for the Information Age* . Boston: Artech House, 1996.

[13] Aasim Ali, Shahid Siddiq, Muhammad Kamran Malik : "Development of Parallel Corpus and english to Urdu Statistical Machine Translation" IJET-IJENS Vol 10 No. 05

[14] Jinhua Du, Sha Wang: "*XAUT Statistical Machine Translation Systems for CWMT2011"*2011

[15] Sauleh Eetemadi, Hayder Radha: " Effects of Parallel Corpus Selection on Statistical Machine Translation Quality"