Study of hybrid machine translation system for different language pairs

Jyoti Sheoran
Department of Computer Science
University of Victoria
Victoria, Canada
jsheoran@uvic.ca

I. INTRODUCTION

Machine Translation (MT) is a branch of Natural Language Processing. We need efficient MT systems that will eliminate the need of a human translator. India being a diverse country has different regional languages. Central government of India works in both Hindi and English while state governments are free to use regional languages. A good MT for Indian regional languages can help automate the process of converting the government documents from one language to other. MT systems can be used for translation of text, websites and e-mail messages. A portable MT system (installed on a portable device) can be used by travellers rather than using a bilingual dictionary.

There has been a lot of research in MT systems based on rule-based machine translation systems (RBMT) and Statistical Machine Translation (SMT) systems. However, the accuracy of these systems is only up to 60%, which is not good enough to be considered as a fully automated system. These systems require post-processing by human translators. RBMT and SMT systems are complementary to each other. There is increasing interest in the hybrid systems that utilize a combination of RBMT and SMT systems. Many hybrid systems are designed which perform better with specific language pair or when evaluated with a specific method. A multi-layer Linguistically augmented Statistical Terminology Extraction (LiSTEX) hybrid MT system architecture shows a promising improvement in the translation quality [1]. It employs statistical and RBMT techniques at transfer phase. The system was tested on English-German and English-French language pairs and the results show significant improvement in translation quality and reduction in error rate. It would be interesting to see the performance gain by this system on Indian language pairs (both related languages and un-related languages). Such an analysis would help us understanding the feasibility of using this model for Indian languages. It could also help us generalize the system for other language pairs. Also, it will help us understand the key elements that need further attention to improve the system for related language and un-related language pairs.

II. METHDOLOGY

The methodology for this research will be incremental oracle experiments with different language pairs and evaluation of the results of the hybrid MT system. The model will be tested with 3 related-language pairs and 3 unrelated-language pair. A tool will be developed to implement the LiSTEX MT system on a Linux machine. Training data set will be taken from some publicly available data. We will test the data set incrementally with same language pairs (i.e test with different size of data-sets). The scoring system's implementation will consist of 3 evaluation metrics - BLEU, Translation Error Rate and Word Error Rate to evaluate the results of LiSTEX MT system. Testing will be done in both in-domain and out-of-domain data sets. We will also test the same data with some freely available commercial MT system (Google translate) and evaluate the results. A comparative analysis will be done on the results of LiSTEX MT system and Google translator.

III. EXPECTED RESULTS

The expected results will be an analysis of the results obtained from various evaluation matrixes. The results will help us identify the feasibility of using the LiSTEX MT systems for Indian languages. The accuracy of this system is expected to be significantly better than current systems that are used for Indian languages.

IV. CONCLUSION

This research will have significant contribution to the machine translation community. A comparative study of LiSTEX with different language pairs and with other freely available good SMT systems will show the strengths and weakness of this model. The research results can be used in` the development of hybrid MT systems for Indian languages. This research can be further extended to propose improvements in the LiSTEX MT system for better performance. Comparison of related and un-related languages will help in the improvement of the LiSTEX MT system to be used for any pair of languages.

References

[1] Wolf, P.,Bernardi, U., Federmann, C., & Hunsicker, S. (2011). From Statistical Term Extractionto Hybrid Machine Translation. In Proceedings of the 15th Annual Conference of the European Association for Machine Translation (pp. 379–419)