Urdu To Punjabi Machine Translation System

Umrinderpal Singh[1] , Vishal Goyal[2] , Gurpreet Singh Lehal[3]
Department of Computer Science GHG Khalsa College Gurusar Sadhar Ludhiana India[1],
Department of Computer Science, Punjabi University Patiala India[2,3]
{umrinderpal[1], vishal.pup[2], gslehal[3]}@gmail.com

Abstract

Machine Translation is a popular area of NLP research field. There are various approaches to develop a machine translation system like Rule-Based, Statistical, Neural and Hybrid. A rule-Based system is based on grammatical rules and uses bilingual lexicons. Statistical and Neural use the large parallel corpus for training the respective models. Where the Hybrid MT system is a mixture of different approaches. In these days the corpus-based machine translation system is quite popular in NLP research area. But these models demands huge parallel corpus. In this research, we have used a hybrid approach to develop Urdu to Punjabi machine translation system. In the developed system, statistical and various sub-system based on the linguistic rule has been used. The system yield 80% accuracy on a different set of the sentence related to domains like Political, Entertainment, Tourism, Sports and Health. The complete system has been developed in a C#.NET programming language.

1. About the system

In this digital era, where different communities across the world are interacting with each other and sharing digital resources. In such kind of digital world natural languages are obstacles in communication. To remove this obstacle from communication, NLP researchers are working to develop Machine Translation systems. These Machine Translation systems can detect various languages and their domains and automatically translate source language text to target-language text. The machine translation system can be developed using various approaches, for example, Rule-Based, Example-Based, Statistical, Neural and various hybrid approaches(Antony P.J 2013). The Rule-based and example-based systems are based on various linguistic rules and a large lexicons dictionaries (Goyal V and G S Lehal 2010). These dictionaries contained parallel word and phrases of source and target language. The statistical system is purely based on some statistical model. These systems required huge parallel corpus to train the model(G S Josan and G S Lehal 2008). The system automatically creates parallel dictionaries and phrase tables from a given parallel corpus (Goyal V and G S Lehal 2010). In this approach researcher's main task is to create or arrange a parallel corpus. Most of the other work simply was given to the machine-like creating phase table and learning the model etc based on the parallel corpus. The neural machine translation is a trending approach these days. In a neural-based approach, deep learning is fast expanding approach for Machine Translation and many other research areas of computing. This approach required a large parallel corpus to train the system and other tasks like creating and learning the translation rules are automatically handled by training algorithms. The Statistical and Neural Machine Translation system can yield excellent results but required a huge parallel corpus for training (Ajit Kumar and Vishal Goyal

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 16–18
Patna, India, December 18 - 21, 2020.

2011). There are many languages in the world which are resource-poor, they don't have any large enough corpus to train the statistical and neural-based system. Urdu and Punjabi languages are one of them. To the best of our knowledge, there is no large enough parallel corpus is available for Urdu and Punjabi language pair to train the statistical and neural-based model. Along with this, Urdu and Punjabi are morphological rich languages. These languages required many other resources like stemmer (Lehal, G. 2009)., lemmatizers, transliteration, spell checker, grammar checker, font detection and conversion tools.

For this demonstration, the system has been developed to translate Urdu to Punjabi Unicode text. As mentioned previously, Urdu and Punjabi are resource-poor languages therefore we have developed various preprocessing, translation and post-processing tools to refine the translated text. In the preprocessing phase, we have developed sentence and word tokenization models for Urdu. The word tokenization system for Urdu is not simple like any other language. In Urdu, segmentation issue (Lehal, Gurpreet Singh. 2010) is the key challenge therefore the sub-system has been developed to handle this issue in preprocessing. In the preprocessing phase, the text classification system has been developed to classify the input Urdu text into various predefined domains like Political, Entertainment, Health and Tourism. The Naive Bayes approached has been used for the text classification system. The reason to use the text classification system to apply the specific knowledge-based on the given input text. By using this clarification module the system can remove various ambiguities in translation. The system used the Hidden Markov Model as a learning module and the Viterbi algorithm has been used as a decoder. Urdu and Punjabi are closely related languages and share grammatical structure and word order. Therefore the system does not require a word reordering module. The system takes manually mapped words and phrases to generate translation probabilities. The language model for translation has been developed using Kneser-Ney smoothing algorithm. Along with the preprocessing, training and translation modules, various sub-system has been developed to refine the output, for example, removing diacritical marks, Izafaat word checking, stemming and creating inflations and transliteration sub-system to handle unknown words. The text classification system's overall accuracy is 96% and complete translation system's accuracy is more than 80% on various domains. The system mainly trained and tested for Political, Sports, Entertainment, Tourism and Health domains. The training data has been collected from BCC Urdu website and TDIL 50000 thousand parallel sentences. On average the system knowledge-based for different domains contains 56023 mapped phrases and words. The system's phrase table incorporates a maximum length of the phrase was four-gram. The total 2088 four-gram phrase used in phrase table. In the translation process, most of the time uni-gram, bi-gram and tri-gram phrases were sufficient to translate any given Urdu input text. The working system is available on http://u2p.learnpunjabi.org/ URL.

References

Ajit Kumar and Vishal Goyal (2011) "Comparative Analysis of Tools Available for Developing Statistical Approach Based Machine Translation System", ICSIL 2011 CCIS 139, pp: 254-260

Antony P.J (2013) "Machine Translation Approachesand Survey for Indian Languages" Computationa Linguistics and Chinese Language Processing Vol.18, No. 1, March 2013, pp: 47-78

G S Josan and G S Lehal (2008), "A Punjabi to Hindi Machine Translation System" in proceeding Coling: Companion volume: Posters and Demonstrations, Manchester, UK, pp: 157-160

Goyal V and G S Lehal (2010), "Web based Hindi to Punjabi machine translation system", J. Emerg. Technol. Web Intell, pp: 148-151

Goyal, V., & Lehal, G. S. (2009). "Evaluation of Hindi to Punjabi Machine Translation System". IJCSI

International Journal of Computer Science, 4(1), 36- 39.

G.S Josan and G.S Lehal "Evaluation of Direct Machine Translation System For Punjabi To Hindi" http://www.learnpunjabi.org/pdf/directtrans.pdf Accessed on: 1/12/2020

Josan G S and G S Lehal (2008) “Punjabi to Hindi machine translation system”, in Proceedings of the 22nd International Conference on Computational Linguistics, MT-Archive, Manchester, UK, pp: 157-160

Lehal, G S. (2010) “A word segmentation system for handling space omission problem in Urdu script”. 23rd International Conference on Computational Linguistics.

Lehal, G. (2009). “A Two Stage Word Segmentation System for Handling Space Insertion Problem in Urdu Script”, World Academy of Science, Engineering and Technology 60.