Challenges in the Development of Domain based Hindi to Punjabi Machine Translation System for Translating Computer Related Text

Manish Kumar
(Department of Computer Science, Baba Farid College, Bathinda)
manish21578@gmail.com

Abstract

A machine translation system can be a general machine translation system or it can be a domain specific machine translation system as well. A domain specific machine translation system is such a system which is developed to translate the text related to some particular domain from one language to another language. While developing a machine translation system which is specific to a particular domain then there are certain challenges related to that domain only. In this paper we have studied the challenges of developing a Hindi to Punjabi Machine Translation System to translate computer related text material.

Introduction

Accuracy level of any general machine translation system decreases while it translates the text of some specific domain. Main reason for this is specific terminology used in that particular domain is some time creates ambiguity while integrated with the general text of the source or target language. Since, long time researchers have contributed by developing various domain based systems to minimise the ambiguity in translation of technical terms of each domain [1-21].

Index Terms – Machine Translation System, Technical ambiguity, Non-Technical ambiguity, Inflection errors, Word out of vocabulary error, disambiguation.

2. Challenges

2.1 Finalisation of Subject domain:

Computer Science itself is a very vast filed where number of subjects are taught at different level of technical education. There is variation in the subjects and level of teaching at different levels. Being a domain specific system, every specific subject might have some different unique issues which required to be resolved for accomplishment of accurate translation of the text. It was very important to first finalise the subjects on which we should consider for our study.

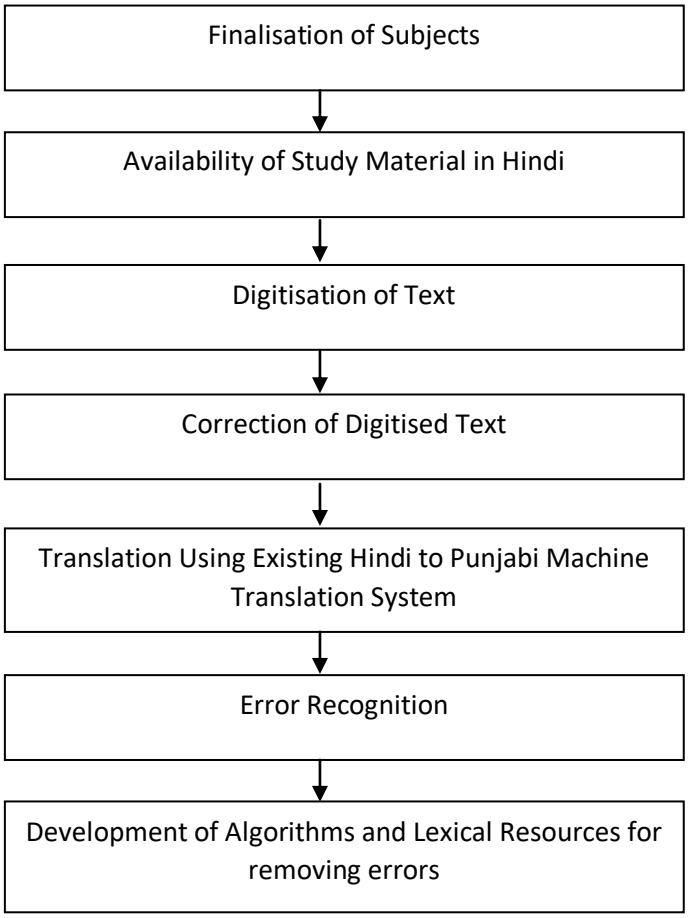

Figure 1

Challenges in the development of system

2.2 Availability of Computer Text Material in Hindi

The first most requirements for developing a domain specific Machine Translation system which can translate the computer related text from Hindi to Punjabi was to acquire study material of computer subjects in Hindi. It is challenge to collect the computer related text material in Hindi which is not available easily in the digitised or printed form.

2.3 Digitisation of Text

Before developing a system to translate computer related text it was required to evaluate the existing Hindi to Punjabi Machine Translation system. To evaluate the existing system, it was necessary to convert all the available study material of computer subjects into digital form. There are mainly two possibilities to convert the text into digitised form. First one is manual by typing the text in Hindi but, this is very long a time consuming process.

There are certain OCR available in the market like Hindi OCR developed by "ind.senz" [22]. Through these OCR printed text can be digitised.

2.4 Correction of Digitised Texts

Due to technical limitations of OCRs, output of Hindi OCR needs to be checked manually to ensure the correct input to Hindi to Punjabi translation system. There could be number of errors which needs to be corrected manually to make the text translatable.

2.5 Translation Using Hindi to Punjabi Machine Translation System

Before developing a domain based translation system it is necessary to evaluate the existing general machine translation system by translating the digitised text from Hindi to Punjabi. During evaluation all the errors can be recognised which needs to be corrected.

2.6 Error Recognition

To revamp the baseline Hindi to Punjabi Machine Translation system into a domain specific system, it is required to first find out the errors generated by baseline system while translating computer related text from Hindi to Punjabi. Again this work has to be done manually and translated text of all the five subjects was manually checked. Different kinds of errors could be identified and all those errors were categorised in four types of errors.

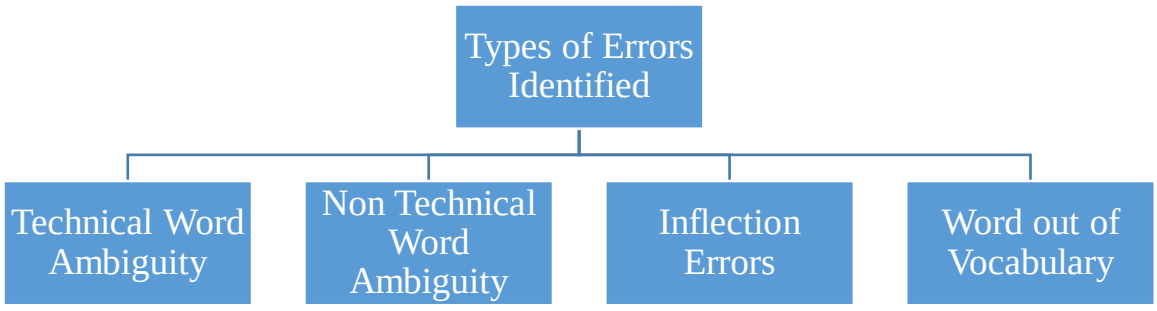

Figure 2

Types of Errors

2.7. Development of Algorithms and Lexical Resources for Removing Errors

All types of errors discussed in section 2.6 were required to be rectified to update the system so that it can translate text related to computer subjects more accurately. There are different kinds of problems which required different method.

2.7.1 Disambiguation

2.7.2 Inflectional Analysis

Hindi and Punjabi are highly inflectional languages. Different researchers have adopted different approaches for resolving the issue of inflectional errors in machine translation. Selecting a approach for the same depends on the degree of similarity between both source and target language. Hindi and Punjabi share almost same inflectional structure of sentences. In this system researcher has selected rule based

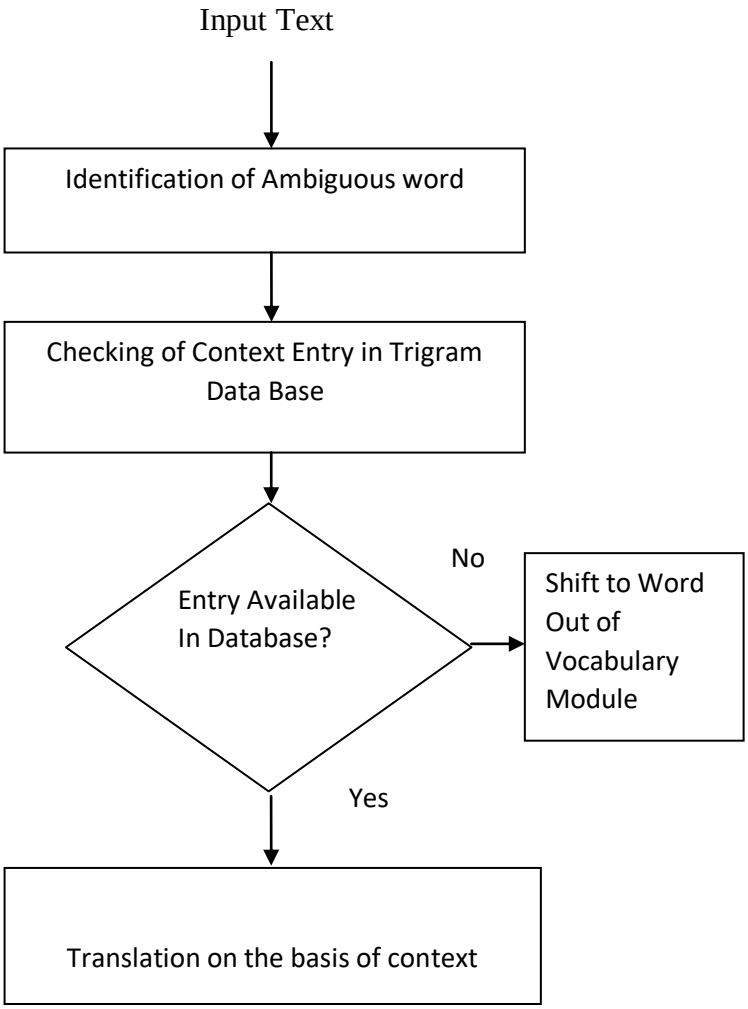

Figure 3

Procedure for Disambiguation

approach for resolving this issue of inflectional error. In this case all the inflectional errors were identified and categorically a rule was implemented to resolve each kind of inflectional error. Those rules were integrated with the rules described earlier in the existing Hindi to Punjabi Machine Translation system. Figure 4 presents the working of inflectional analysis module of the updated system. Every word as a token is send to this sub module of inflectional analysis. It is checked that whether a rule defined matches with the token or not and if a match is found then the same rule is implemented on the token. Translation of the word is generated in Punjabi according to the rule applied. Once the word is generated as output it is checked with the Punjabi corpus to check that whether that generated word is part of Punjabi corpus or not. In case generated word is found in Punjabi corpus it is send as output. But if the Punjabi generated word is not found in the Punjabi corpus of the system it is treated as word out of vocabulary and send to transliteration module.

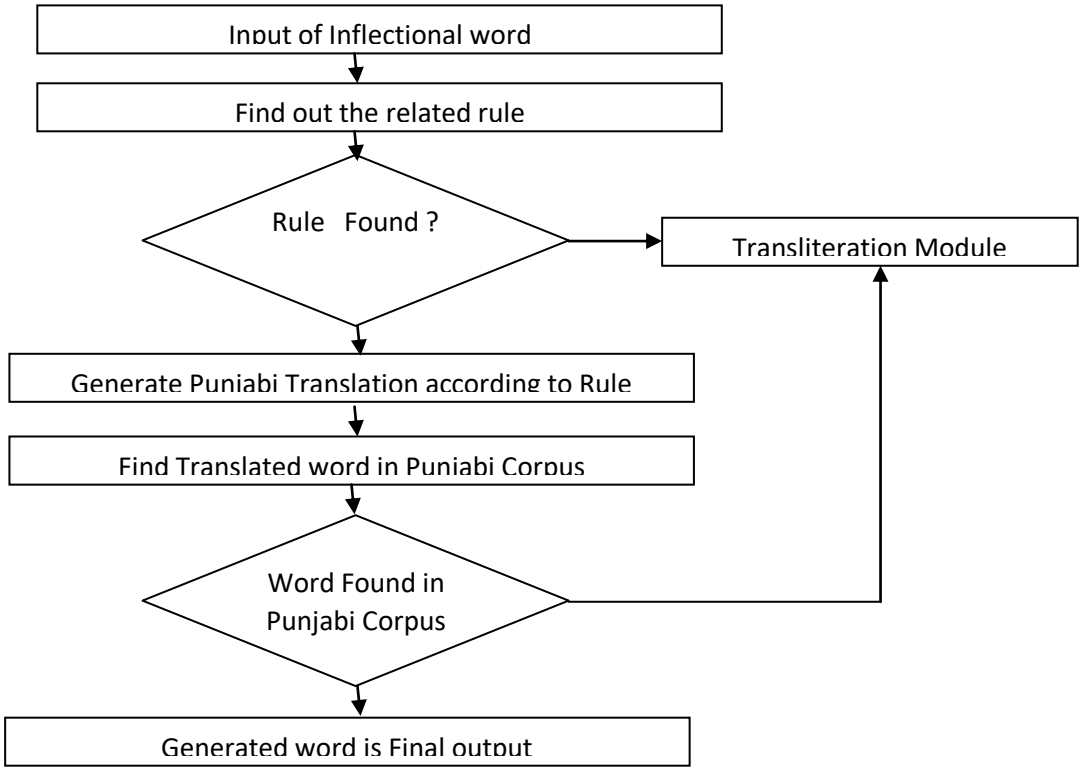

Figure 4
Procedure for Resolving Inflectional Errors

2.7.3 Updating Lexical Resources

To resolve the problem of incomplete lexical resources, terms which are not available needs to be identified and a new bilingual corpus for that need to be developed. That developed corpus than integrated with the existing bilingual corpus of the system to get better results from it.

References

[1] Oubine, I. I., Tikhomirov, B. D., (1982), "Machine Translation System and Computer Dictionaries in the Information Service. Ways of their Development and Operations," Collings 1982, pp. 254-265.

[2] Oren, R., (1986), "Technical Writing and Translating-Language Barriers Part –II –Software /Hardware Adaption Problems," Technical writing and Management, pp 80-86.

[3] Oren, R., (1989), "Technical writing and translating – Language Barriers Part -1 – Mother Tongue & Invented Words," Proceedings of Professional Communication conference , IPCC -89, pp. 80-86.

[4] Lufkin, J. M., (1989), "Current trends in Technical Translation," IEEE 1989, pp. 238 – 244.

[5] Kirk, ST. A., (2000), "Expanding Translation use to improve the Quality of Technical Communication," IEEE Transactions on Professional Communication, 43, pp. 323 – 326.

[6] Fujii, A., Ishikawa, T., (2001), "Japanese/English Cross language Information retrieval : Exploration of query translation and Transliteration," Computer and the Humanities, 35(4), pp. 389-420.

[7] Sakai, T., Kumano, A., Manabe, T., (2002), "Generating Transliteration rules for cross language Information retrieval from Machine Translation Dictionaries," IEEE 2002, pp 555 – 565.

[8] Madhvi, G.,(2005), "OM : One tool for many (Indian) Languages," Journal of Zhejiang University Science, 6(A), pp. 1348-1353.

[9] Hettige, B., Karunananda, A.S., (2007), "Transliteration system for English to Sinhala Machine translation," Second International conference on Industrial and Information Systems, ICIIS, pp 8-11.

[10] Genzel, D, Macherey, K., (2009), "Creating a High Quality Machine Translation System for Low Resource Language : Yiddish," Machine Translation Summit, Ottawa 2009, 2, pp. 346-359.

[11] Chaudhury, S., Rao, A., Sharma, D. M., (2010), "Anusaaraka : An Expert System Based Machine Translation System," IEEE (2010), pp. 189 – 196.

[12] Ren, F., Zhu, J., Wang, H., (2010), "Web based technical terms translation pairs mining for patent documents translation," Natural Language Processing and Knowledge Engineering NLP-KE 2010, pp. 1-8.

[13] Manso, A., Marques, C. G., Dias, P., (2010), "Portogol IDE V3.X – a new environment to teach and learn computer programming," IEEE EDUCON 2010, pp. 1007-1010.

[14] Mitra, B., Raj, A., (2011), " Multilingualism : Mother Tongue as tool for learning in classrooms," Professional communication Conference IPCC 2011, pp 1-5.

[15] Dubey, V., Sharma, H.R., (2011), "New Challenges in Machine Translation of Technical English Digital Text in Hindi," Journal of Computer Science, 7, pp. 899 – 905.

[16] Logacheva, V.K., (2011), "A Method for Generating Rules for Cross-lingual Transliteration," Automatic Documentation and Mathematical Lingustics, 45(5), pp. 239-248.

[17] Laura, R., (2012), "Automatic translations versus human translations in nowadays world," Procedia- Social and Behavioral Science, pp. 1768 – 1777.

[18] Ahmed, F., Nurnberg, A., (2012), "Literature Review of Interactive Cross Language Information Retrieval Tools," The International Arab Journal of Information Technology, Vol 9, pp. 1022-1034.

[19] Donald, J.K., (2013), "Collecting Korean-English Pairs for Translation of Technical Terms," Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 989-999.

[20] Mathur, S., Saxena, V. P., (2014), "Hybrid approach to English-Hindi Name Entity Transliteration," IEEE Students' conference on Electrical, Electronics and Computer Science, pp. 669-678.

[21] Sanjanshree, P., Anand, K. M., (2014), "Joint layer based Deep Learning Framework for Bilingual Machine Transliteration," IEEE (2014) pp. 678 – 688

[22] http://www.indsenz.com/int/index. php Accessed on 04/07/1