Challenges in building Dogri-Hindi Statistical MT System

[1]Manu Raj Moudgil, [2]Preeti Dubey, [3]Ajit Kumar
[1]Research Scholar (Ph.D), [2,3]Assistant Professor
[1]JJT University, Jhunjhunu, [2]J&K Higher Education Dept, Jammu, [3]Multani Mal Modi College,Patiala
[1]manu.moudgil@gmail.com, [2]preetidubey2000@yahoo.com, [3]ajit8671@gmail.com

ABSTRACT

In present time to overcome the problem of language barrier in communication, lots of researches have been done in the field of language translation. The system which is developed for translation of languages by the researchers is known as machine translation system. There are different types of Machine Translation systems are developed with the change of time and for the need of more accuracy in the translation. No doubt the development of MT systems proved to be great achievement in the field of science and technology. But such achievement was not that easy. Different challenges and issues were faced by various researchers during the development of their MT systems. This paper focuses on the challenges and problems faced during the development of Dogri-Hindi SMT system. This paper also discusses the steps that are taken under consideration to overcome that challenges.

Keywords: *Statistical machine translation, parallel corpus, Translation,*

INTRODUCTION TO DOGRI-HINDI SMT SYSTEM

Dogri is composed utilizing Devanāgarī content and has thirty eight segmental and five supra segmental phonemes. Segmental phonemes have been isolated into two general gatherings i.e. vowels and consonants. It has 10 vowel phonemes and 28 consonant phonemes. It is believed by researchers that it is easy to develop machine translation systems for closely related languages. Dogri and Hindi are closely related languages. These languages are written from left to right. Dogri and Hindi are both written in Devnagri script. These languages have their origin in Sanskrit and have same sentence structure i.e. Karta, Karam, Kria (SOV). In both languages, sentence is comprised of subject and predicate. In both languages, the basic elements are Kaaraka. Both have eight numbers of Kaaraka (Karta, Karam, Karan, Sampardaan, Apadaan, Sambandh, Adhikaran, Sambodhan) which by combining with each other create a sentence. The general sequence for transitive sentence is Karta, Karam, Kria and for intransitive sentence is Karta, Kriya. In both languages the relation between Kaarka's are shown by postpositions. Total eight parts-of-speech are recognized in both Dogri and Hindi. Beside this, both have same types of nouns, genders, number, persons, tenses and cases. The close relationship between Hindi and Dogri is established by a study by Preeti Dubey, Devanand and. et. al. [1] The authors have also concluded that Hindi and Dogri are closely related languages. The close relationship between Hindi and Punjabi is established by a study by Josan and Lehal [2] and Goyal V. et. al. [3]

NEED OF DOGRI-HINDI SMT SYSTEM

Dogri is a prominent regional language of Jammu and Kashmir. A large population of the state specially the Jammu division speaks the Dogri language. It is also a constitutional language of India. It is a new language in the map of natural language processing. There is no any such system developed, that can translate Dogri to Hindi. Presently, there is only one MT system which is based on direct approach that can only translate Hindi into Dogri. But there is no system at present that can translate Dogri into Hindi, so there is a great need to develop the machine translation system that can translate Dogri text into Hindi text. This task has been undertaken in this research and the statistical approach of machine translation has been chosen for this purpose. By this development the conversion of Dogri to Hindi language is easily possible and any person either officially or personally can use this system.

LITERATURE SURVEY

The primary requirement to build the Dogri to Hindi statistical machine translation system was the Dogri-Hindi parallel corpus developed during this research work. The foremost activities involved in this task are corpus cleaning and tokenization and some pruning of our training data. The primary goal of statistical machine translation (SMT) is to generate a target sentence (e) from a source sentence (f) i.e. to produce a Hindi sentence from the given Dogri sentence in the work undertaken. The Dogri-Hindi MTS uses phrases as translation units [4][5] and a log linear framework, which implies it appears the form of function whose log is a linear combination of model which makes it is to apply linear regression. The log linear framework in order to introduce several models explaining the translation process:

$e^{*}=$ argmaxe $p(e\backslash f\,)$
$e =$ argmaxe $\{\exp\,(\sum \lambda i\ hi(e,f\,))\}$ ....(1)
i

The feature functions
hi are the system models [6]
λi weights are typically optimized to maximize a scoring function on a development set

These components capacities incorporate lexical translation probabilities in two directions, seven features for the lexicalized distortion model, a word, a phrase penalty and a target language model (TLM).

As it has been observed in available literature, there are various techniques for creating monolingual corpus which include picking text from web, manual text typing, OCR for extraction of text from scanned documents, etc. These approaches suffer from their own limitations, for example, manual typing of text is time consuming but despite involving a lot of time, labour, and cost, its accuracy is much more than other techniques. Using OCR for extraction of text from scanned documents has the limitations of OCR. Moreover, for many Indian languages, OCRs do not exist. Even if OCRs are available for some languages, the accuracy of text generated is not satisfactory and manual checking of the same is required. Further, the problem gets aggravated because of non-availability of the spell checker for rectifying the corpus. At the first glance, the option of picking text from the web appears to be practical but there are certain problems which are faced doing so. These are- variety in coding techniques, comparatively small reservoir of text than English and other languages of Europe, use of PDF documents, various fonts and font sizes and copy right of the text etc.

Baker et. al. (2004) explained the origination of EMILLE Corpus that is comprised of monolingual corpora in 14 languages of South Asia. This corpus also contains an annotated component viz. partof-speech tagged Urdu data, along with 20 Hindi corpus files (written) which are annotated for showing the nature of demonstrative usage in Hindi. [7]

The main approach used for developing corpus is manual typing of those documents which are at hand in printed, PDF GIF or JPEG format. Use of OCR for extraction of text from these documents and conversion of TTF documents into Unicode also come under this approach. For creating the parallel corpus, English documents are translated into other languages. In the case of non-availability of machine translation, manual translators are hired and documents are typed after translation and saved as Unicode formats.

Arora et. al. (2010) introduced an approach for automatic creation of parallel corpus (Hindi-Punjabi) from comparable/corresponding Hindi-Punjabi corpus. In their opinion, the comparable/corresponding documents are processed for finding sentence boundaries and tokenization of sentences is done at the word level. Afterwards, the sentences are aligned making use of weight assignment techniques and POS tagger. [8]

Kumar et. al. (2010) made use of on-line Hindi to Punjabi machine translation tool which is available at http://h2p.learnpunjabi.org for developing Hindi-Punjabi parallel corpus. The authors believe that monolingual Hindi corpus is already there and it can be translated to comparable Punjabi corpus for development of Hindi-Punjabi parallel corpus. They employed this approach for developing the corpus and faced several problems. First of all, only a limited number of about 100 sentences can be translated at a time. Secondly, several Hindi words are transliterated instead of translated. Moreover, minor errors related to spelling and incorrectly translated words are shown in output as a result of which the output is required to be manually edited. [9]

Antonova et. al. (2011) presented the techniques which are used for developing parallel corpus for Russian-English Language pair. The authors opine that the web is a source of parallel documents. While using this approach, collection of comparable documents is done from the web and their processing is undertaken for the purpose of sentence wise alignment. [10]

Ali et. al. (2010) highlighted the problems which they faced while developing the English-Urdu parallel corpus. The major difficulties faced by them were- lack of parallel text in Urdu and English, punctuation marks, sentence alignments, and translations issues. They accomplished the development of a parallel corpus consisting 6000 lines by translating manually the English sentences into Urdu and implemented Moses for developing Statistical Machine Translation System yielding BLEU score of 9.035. [11]

Ghayoomi et. al. (2010) revealed some of the common difficulties which they faced experimentally while developing a Persian language corpus from written text and explained some crude solutions for fixing them. In their viewpoint, the source of problems might be the mixing of Persian script with Arabic script; the typists' typing style; Persian orthography; having different linguistic styles and creativity in the language; code pages of the control characters in the operating systems and word processors. They observed that before processing of the Persian corpus, pre-processing of the raw data manually, automatically, and in combination of both by spending time and energy is required. [12]

ISSUES IN MACHINE TRANSLATION

Machine translation is the component of large sphere of pure research in computer based NLP which covers linguistic as well as artificial intelligence. It examines the basic structure of languages as well as mind by modelling and simulating in computer programs. It also provides platform for more research on large scale for techniques and theories created by tests in computational linguistic and computerized reasoning. In computers, the real problem in translation is linguistic not computational. Some linguistic issues to be handled are lexical ambiguity, syntactic problem, difference between the languages, construction of words, finding out meaning for text and signs, producing sentences in another language with an equivalent meaning. So MT has to update with the advances in linguistic researches so that above discussed problems can be minimized.

CHALLENGES IN DOGRI-HINDI SMT SYSTEM

Though the statistical machine translation approach is not a very new approach now-a-days, but some challenges in the development of statistical Dogri to Hindi machine Translation System are:

Dogri is the regional language of jammu, but still there is not much availability of online text, journals and literature of Dogri language. Also printed textbooks are in less in number.. To get the Dogri text in soft form is a big challenge.

Basically, till today there is very less work done on Dogri, and not having much research papers. so to gather the literature and knowledge about the language is also is big task.

The freely available tools like Moses, GIZA++, IRSTLM, SRILM etc. have been successfully used for English and other European languages, which has entirely different sentence structure from Dogri and Hindi. Therefore, a lot of resources need to be developed for making these tools usable for Dogri and Hindi. As mentioned earlier, in case, these tools cannot be used for our system, they need to be developed from scratch.

The Lack of resources for example parallel corpus and the digital bilingual dictionary for Dogri and Hindi languages. The development of large Dogri-Hindi parallel corpus in itself is a big challenge,

For the creation of Dogri language from Hindi language to make a parallel corpus, A Hindi-Dogri MT system is used. But it is a very big challenge as the corpus which is created by this system contains the mistakes and needs a manual editing and to make the corrections, we need a Dogri linguist, who has the sufficient knowledge of Dogri. This step is very complex and time consuming.

Another challenge is Sentence Alignment. In parallel corpora single sentences in one language can be converted into several sentences in the other and vice versa. Sentence aligning can be performed through the Gale-Church alignment algorithm.

Ambiguity is a challenge faced in developing machine translation systems for most languages. Even in Dogri, there are several words that have numerous meanings related with the context in which the word is available in the sentence. The software has to handle ambiguity to provide accurate translations.

Data Dilution: - This is a typical abnormality caused when endeavouring to build another statistical model (engine) to represent a distinct terminology (for a particular corporate brand or area).

Training sets utilized from elective sources to the particular brand to adjust for a restricted amount of brand specific corpora may dilute' mark phrasing, selection of words, text format and style.

Another challenge is Idioms. Depending on the corpora used, idioms may not translate "idiomatically".

Identifying Named Entities introduce in the content like the word S. Parkash Singh Badal, State Bank of India, Parkash , Dr.Vishal Goyal and so on as it must remain same after translation.

A Collection of expressions that can't be translated word-by-word and also these have distinctive meaning in collection than in the individual.

PROCESSES INVOLVED TO OVERCOME THESE CHALLENGES

All the challenges in building the Dogri-Hindi SMT is primarily focused on the Dogri-Hindi parallel corpus. To overcome this problem, the following are the processes that are used for the development of the parallel corpus of Dogri-Hindi Languages:

Cleaning of Hindi Corpus: Cleaning of corpus is beginning step but plays an important role. The Hindi raw data which is taken from the Gyan Nidhi and other web portals are cleaned properly to remove all the symbols and other language words.

Convert Sentence wise: This is also important step as the Hindi corpus which is cleaned must have a proper sentence format, so it can be used for translation purpose to give optimum results. The cleaned documents are then tokenized at the sentence level to make document sentences-wise.

Spell Check: After making the Hindi corpus sentence wise, now work starts for the spelling checking and grammatical corrections in the corpus. In this step checking of Hindi corpus is performed manually and also the spell check facility of Open-Office is used on Ubuntu platform.

Hindi-Dogri MTS: This machine translation system is used to make Dogri corpus with the input of Hindi language corpus. All the above steps i.e cleaning and spelling check for Dogri language is performed manually.

Manual Correction by linguist

Further the output generated by Hindi-Dogri MTS is given to Dogri language expert to make the proper corrections which are required in the parallel corpus. This task is time consuming but very important as the accuracy of the system is totally depends upon the accuracy of the corpus.

After the successful completion of all the above steps, the parallel corpus for both Hindi and Dogri languages are developed and ready to be used for the development of Dogri-Hindi SMT system.

CONCLUSION

The development of Dogri-Hindi machine translation system is challenging task particularly for resource deficient language like Dogri, the development of parallel corpus in itself is a big challenge. We have developed the required resource from scratches and developed a fully functional

Dogri-Hindi machine translation system by overcoming all the challenges described in the paper. The resources developed through this research work can be used in further research in Dogri language.

ACKNOWLEDGEMENT

My sincere thanks to Dr. Vishal Goyal, Associate Professor, Department of Computer Science, Punjabi University, Patiala, for his continuous guidance.

REFERENCES

[1] Preeti Dubey, Shashi Pathania, Devanand, Comparative Study of Hindi and Dogri Languages with regard to Machine Translation, Language In India, Volume 11:10 October 2011,ISSN 1930-2940

[2] G S Josan and G S Lehal,(2008), A Punjabi to Hindi Machine Translation System. COLING: Companion volume: Posters and Demonstrations, Manchester, UK, pp. 157-160

[3] Vishal Goyal, Gurpreet Singh Lehal, (2008), Comparative Study of Hindi and Punjabi Language Scripts, Napalese Linguistics, Journal of the Linguistics Society of Nepal, Volume 23, pp 67-82

[4] Philipp Koehn, Franz Josef Och, and Daniel Marcu. (2003), Statistical phrased-based machine translation, In HLT/NACL, pages 127–133

[5] Franz Josef Och and Hermann Ney. (2003), A systematic comparison of various statistical alignment models , Computational Linguistics, 29(1):19–51

[6] Franz Josef Och and Hermann Ney. (2002), Discriminative training and maximum entropy models for statistical machine translation , In ACL, pages 295–302

[7] Paul Baker, Andrew Hardie, Tony McEnery, Richard Xiao, Kalina Bontcheva, Hamish Cunningham, Robert Gaizauskas, Oana Hamza, Diana Maynard, Valentin Tablan (2004), Corpus Linguistics and South Asian Languages: Corpus Creation and Tool Development, Lit Linguist Computing 19 (4): 509-524, DOI:https://doi.org/10.1093/llc/19.4.509, Published:01 November 2004

[8] Sunita Arora, Rajni Tyagi, Somi Ram Singla, Creation of Parallel Corpus from comparable Corpus, Proceedings of ASCNT – 2010, CDAC, Noida, India, pp. 77 – 83.

[9] Pardeep Kumar, Vishal Goyal, (2010), Development of Hindi-Punjabi Parallel Corpus Using Existing Hindi-Punjabi Machine Translation System and Using Sentence Alignments, International Journal of Computer Applications (0975 – 8887), Volume 5– No.9, pp.15-19

[10] Alexandra Antonova, Alexey Misyurev, Building a Web-based parallel corpus and filtering out machinetranslated text, Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 136–144, 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 24 June 2011, Association for Computational Linguistics

[11] Ali, Aasim, Shahid Siddiq and M. Kamran Malik, (2010), Development of parallel corpus and English to urdu statistical machine translation, International Journal of Engineering and Technology/IJENS, 1: 30-33, ISSN 2077-1185.

[12] Masood Ghayoomi, Stefan Muller, PerGram: A TRALE Implementation of an HPSG Fragment of Persian, Proceedings of the International Multiconference on Computer Science and Information Technology, pp. 461–467, ISBN 978-83-60810-22-4, ISSN 1896-7094