Status of Dogri with context to Natural Language Processing

Preeti Dubey, Assistant Professor
Department of Computer Applications, GCWParade, Jammu
Preetidubey2000@yahoo.com

Abstract

This paper is a brief introduction to Natural Language Processing (NLP) with a focus on its need in today's time. In the era of Information and Communication Technology, there occurs a digital divide or a digital exclusion of a section of society due to language barrier, which is where the importance of natural language processing lies and aids to bridge the digital divide. NLP has also played a key role in preserving low resource languages. This paper discusses the challenges of machine translation of Dogri and its present status on the map of natural language processing.

Keywords: Machine Translation, Dogri, Natural Language Processing, Digital Divide

I. Introduction

Natural Language Processing is a branch of artificial intelligence that aims to process natural languages i.e. languages spoken by humans. An important application of natural language processing is Machine Translation (MT). Machine translation is defined as automatic conversion of text, speech or content in any other from one natural language to another. It is an interesting and a complex NLP problem that requires knowledge of artificial intelligence and linguistics. With the advent of ICT, there is a rise in the Digital Divide. Studies have shown Language to be one of the factors for the digital divide. NLP can play a vital role in *bridging this digital divide.* Development of computational resources for languages will aid in *preserving low resourced language*. Machine translation of text and voice from one language into another will enable *understanding of unknown languages* and will aid *sharing of literature* among various languages and support cross lingual research. With the development of computerised tools, there shall not only be reduction of manual work or human effort but will also ease the use of technology to all. Machine translation enables the availability of content in one's regional language or any other language known to a person. Some

of the application areas for machine translation are *Governance, Commerce, Education, Travel, Customer Support, Manuals, Menu* etc

II. Machine Translation Approaches:

There are many approaches for Machine translation such as rule based, sstatistical, example-based, hybrid approach, but the system based on Machine learning is the state-of -art machine translation systems. Machine learning based systems are trained using large amount of parallel corpora. Parallel Corpora is the basic and most required resource of machine learning. If the corpora contains translations of text of two languages then it's called bilingual corpora, and if it contains translations in multi languages then its multilingual corpus. The results of the MT rely on factors like size, type of sentences in the corpora. Larger and Varied the Corpus, better are the results. Following is an example of Bilingual Parallel Corpora:

Source	**Target**
When is my order arriving?	मेरा आडर कुसलै आवा दा ऐ?

III. Challenges in Machine Translation

The major expectation from a machine translation is the quality or we can say accuracy of the output. There are many factors that influence the output of a machine translation system. The output of a machine translation should not only be a mere translation of words, but it should be grammatically correct, maintain the sense of the source text. Some of the challenges faced in machine translation are discussed below:

- **Language Diversity:** The beauty of any language is its diversity; but this diversity is a challenge for machine translation. The machine translation software should be able to handle the variety of the language e.g. synonyms, ambiguity (kaise, kaneya and kinyah), standardization of spellings (e.g. बंदर, बन्दर), translation of phrases, Idioms. handling irony and sarcasm, structural dissimilarity, different Word order languages e.g SVO (Subject- Verb- Object) or SOV pattern (Subject- Object- Verb)
- Ambiguity: Presence of more than one meaning for a particular word is called ambiguity

in a language. Disambiguation of the word with respect to the sense of the word is very important to get quality output. Two types of ambiguity is seen in most languages namely Lexical and structural.

- **Lexical Ambiguity** occurs when a word has many forms and its usage varies depending upon the context of the sentence. If the correct meaning of the words is not preserved during the translation the entire sense of the sentence may change. For example is the word 'से' in Hindi can have approximately 6 to 7 variations depending on the context of the sentence. Another example is shown below:

  Hindi : आप कैसे है?

  Dogri: तुस किं'या हो?

- **Structural Ambiguity** has a potential of multiple interpretations of a sentence e.g.

  In this class I have learned how to distinguish which part of the brain has been damaged by asking some simple questions.[5]

- **Variation of words in formal and casual usage** is another challenge. In many Indian languages language there are various forms of a word as per their usage e.g the usage of the word 'sit' in Dogri is different when we are addressing a friend to a person elder or younger to you.If this variation is not taken care of then it may result in depletion of many friendly words from the automated text leading to the loss of emotion in the sentence. Consider the following Dogri variations:

  Sit : बैठ, बैठिये,

  You : आप, तू

- **Lack of computational resources** : Dogri is a computationally low resourced language, therefore development of tools from the scratch is expensive in terms of time as well as cost.

- **Identifications of Proper Nouns or Noun Entity Recognition (NER):** NER adds to the quality of the output. Most names have some meaning; therefore take must be taken to recognize the word

as a proper noun, so that such words are not translated. Given below are some examples of such words:

बाग सिंह should not be translated as *बगीचा* ***सिंह***

Reserve Bank of India should be translated as *भारतीय रिजर्व बैंक*

- **Collocations:** Many combinations of words have a meaning which is different from their individual meanings, such words are called collocations. Handling accurate translation of collocations is very crucial for efficient output. For example उतर प्रदेश which is the name of a Indian state if translated word by word may become जबाब राज in Dogri output , similarly नाग पंचमी which is the name of an Indian festival may be translated into सप्प पंजमी by translating individual words. Hence, it's very important to recognize collocation and translate it accordingly.

IV. Status of Dogri on the map of NLP

The presence of Dogri on the map of NLP is minimal. The development of a more sophisticated translation system for the Dogri language has been difficult for the researchers. This is mostly due to the language being relatively unexplored and the lack of digital linguistics tools. Dogri is under developed in the field of natural language processing. A *rule based Hindi to Dogri machine translation system* is developed by the author that can translate Hindi text into Dogri text with above 90% accuracy. With the inclusion of under the 'National Translation Mission: Bhashini', funded by Meity, there shall be tremendous growth of Dogri in this field.

V. The Hindi Dogri Machine Translation System

This system is first text translation system developed to convert Hindi text into Dogri. It is a rule based system, therefore it does not require parallel corpora, and it requires knowledge of the Morphology, lexicon of the language. Rule based systems provide accurate results but are slower as compared to the

neural MTS'. The above system has a small sized dictionary of 18k words. It handles collocations, named entities and ambiguity of some words only. It has 66 morphological rules and also handles the peculiarities of Dogri like handling of Kar, Raha, laga etc by framing rules. Some use cases are shown below:

चढ़कर - चढ़ियै, **चढ़- चढ़कर** –चढ़ी- चढ़ियै, कर सकते हैं- करी सकदे न

The above variations of 'kar' show the importance of handling words as they appear before and after **कर** to get accurate results. Below are some sample translations of the first Hindi-Dogri Machine Translation System, the errors in Dogri translation are marked bold, underlined and italicized:

Hindi	**Dogri**
जेल में बंद शातिर अपराधी योगेश मेरठ के भदौड़ा गांव का रहने वाला है।	जेह्ल च बंद शातर अपराधी योगेश मेरठ दे भदौड़ा ग्रां दा रौह्ने आह्ला ऐ।
पेट साफ करने वाला कोई भी **चूर्ण आपके** पाचन **स्वास्थ्य** में सुधार कर सकते हैं।	ढिड्ड साफ करने आह्ला कोई बी ***चूना*** तुं'दे पाचन ***स्वास्थ्य*** च सधार करी सकदे न।
लहरों से डरकर नौका पार नहीं होती, कोशिश करने वालों की कभी हार नहीं होती। नन्हीं चींटी जब दाना लेकर चलती है, चढ़ती दीवारों पर, सौ बार फिसलती है। मन का विश्वास रगों में साहस भरता है, चढ़कर गिरना, गिरकर चढ़ना न अखरता है। आखिर उसकी मेहनत बेकार नहीं होती, कोशिश करने वालों की कभी हार नहीं होती।	लहरें कोला डरियै किश्ती पार नेईं होंदी, कोशश करने आह्लें दी कदें हार नेईं होंदी। नन्हीं चींटी जिसलै दाना लेइयै चलदी ऐ, चढ़दी दीवारें पर, सौ बारी तलैह्टदी ऐ। मन दा विश्वास रगें च हिम्मत ***भरोचदा*** ऐ, चढ़ियै डिग्गना, डिग्गियै चढ़ना न ***अखरता*** ऐ। आखर ओह्दी मैह्नत बेकार नेईं होंदी, कोशश करने आह्लें दी कदें हार नेईं होंदी।

VI. Limitations of the Hindi Dogri Machine Translation System

Although the system provides output with accuracy above 90%, following are some limitations of the system:

- Limited Lexicon: The system uses a a small sized dictionary containing 18k words.
- Limited rules for Inflectional analysis: The system uses only 66 rules for Inflectional analysis.
- Word Sense Ambiguity Module needed to be enhanced to incorporate more words.

- Gender Disagreement: Errors have been observed due to gender disagreement e.g.

  Hindi: मेरी बहन की शादी हो गई।

  Dogri: मेरी भैन <u>दा ब्याह होई</u> गेआ ।

- **Feminine nature of the language**: Addition of 'ी' at the end of main verbs as seen in below examples:

  Hindi : मैं यह कर सकता हूं।

  Dogri: में एह् <u>करी</u> सकनां ।

  Hindi: वह चल पड़ा।

  Dogri: ओह् <u>चली</u> पेआ।

Conclusion: The author brings to light the prevalent digital divide in the society which has been created with the advent of ICT due to language barrier. The paper lays emphasis on the need to develop machine translation tools as it not only bridge the digital divide but also preserve the languages. The status of Dogri on the map of NLP is also discussed with a focus on the first Hindi to Dogri Machine Translation System. It paper concludes with a need to develop more machine translation systems based on newer and state of art research methods like deep learning, because the limitations of the rule based system can be very well addressed with deep learning techniques ,therefore there is need to take up research and development of linguistic tools for Dogri using the newer techniques.

References:

1. Preeti Dubey, et.al, "A Study to Examine the Digital Divide Factors: Jammu and Kashmir Perspective", 2011
2. https://www.cse.iitb.ac.in/~anoopk/publications/presentations/icon_2013_smt_tutorial_slides.pdf
3. Harold Somers "Machine Translation". Chapter 13 of R Dale, H Moisl & H Somers (eds) Handbook of Natural Language Processing, New York (2000): Marcel Dekker
4. John Hutchins "Machine translation: general overview". Chapter 27 of R Mitkov (ed.) The Oxford Handbook of Computational Linguistics, Oxford (2004): OUP
5. https://people.ku.edu/~sjpa/Classes/CBS516/BasicSyntax/structural-ambiguity.html