Natural Language Processing

Preeti Dubey

Department of Computer Science Govt. College for Women, Parade, Jammu

Abstract: The upcoming field of computational linguists namely natural language processing (NLP) is presented in this paper. Introduction to the basics of NLP, techniques of development and its importance is discussed by the author.

Key words: Computational Linguists, NLP

Introduction

Natural language processing is a branch of computer science that deals with the processing of one human language into another. It involves both computation and language. Natural language processing can be applied to text as well as speech. The development of speech processing systems is more challenging as compared to text processing. Natural language processing is a major technology that can be used to bridge the gap between human communication and digital data. The goal of natural language processing is to design and build software that will analyse, understand, and generate languages that humans use naturally, so that eventually people can address computers as though they were addressing people. NLP has many applications such as Machine Translation, Information Extracton, Summarization, Question Answering etc. In the last decade, most of the effort in this field is inclined towards machine translation. This paper is more focused on the basics of text processing using machine translation.

Machine Translation

MT is a multidisciplinary field of research. It applies the ideas from linguistics, computer science, artificial intelligence, statistics, mathematics, philosophy and many other fields. The text is translated in such a way that the meaning of the source text is preserved during this process. Machine translation systems can be bilingual as well as multilingual. Systems that produce translations between only two particular languages are called bilingual systems and those that produce translations for any given pair of languages are called multilingual systems. Many approaches are available for the development of machine translation systems namely: rule based approach, Corpus based and Hybrid. Present machine systems are also being developed using deep learning methods.

Rule Based Machine Translation Systems are developed to produce output using the linguistic rules. It requires deep understanding of the languages undertaken. It uses machine readable dictionaries.

Corpus Based Machine Translation Systems use large corpora for the development of machine translation systems. The accuracy of these systems depends on both the quantity and quality of corpora used.

Hybrid Machine Translation Systems are developed using more than one technique. A system developed using this method may be using the corpora as well has linguistic rules.

Deep Learning Machine Translation Systems are based on methods based on artificial neural networks. The choice of the approach to be used depends on the languages

*Corresponding author(s):
preetidubey2000@yahoo.com (Preeti Dubey)

undertaken and the availability of computational resources.

Challenges in Machine Translation

The translation process is not mere translation of words but the translated output should be as accurate as translated by some human. It is a challenging task. Some challenges faced in the process of machine translation are:

Non-availability of computational resources required for developing machine translation systems such as dictionaries, corpora, morphological analyzer etc. It requires a lot of finance and time to start from the scratch.

Different Language structures: Language structure can be SVO (Subject- Verb- Object) SOV (Subject- Object- Verb). Some languages have completely different structures. Therefore, they need to be aligned.

Word Sense Ambiguity is another challenge faced by the developers of machine translation systems. Every natural language involves ambiguous words. It is very important to handle ambiguity so as to develop accurate systems.

Discourse Integration is essential for developing efficient systems. The meaning of an individual sentence may depend on the sentences that precede it and may influence the meanings of the sentences that follow it e.g. "He always wanted it". The meaning of 'It' in the sentence depends on the prior discourse context. References such as this are called anaphoric references or anaphora.

Pragmatic Analysis means to correctly interpret the sentence and analyze what was meant. E.g. "My house was broken into last night. They took the jewelry" should be interpret correctly and they should be recognized as referring to thieves.

Key Systems Developed in India

The earliest efforts in Machine Translation in India started from the mid 80s. Some of the prominent contributors in the field of language translators are: **1)** The research and development projects at Indian Institute of Technology, Kanpur, **2)** University of Hyderabad, **3)** National Center for Software Technology, Mumbai, **4)** Center for Development of Advanced Computing (CDAC), **5)** Pune Indian Institute of Technology, Bombay, **6)** International Institute of Information Technology, Hyderabad, **7)** Anna University, **8)** KB Chandrasekhar Research Center, Chennai, **9)** Jadavpur University, Kolkata, **10)** Jawaharlal Nehru University (JNU), **11)** Mahatma Gandhi International Hindi University (MGIHU)

Some machine translation systems developed in India are:

S.No	System	Year	Language Undertaken
1.	AnglaBharti	1991	English to Indian Languages(IL)
2.	Anusaarka	1995	IL to IL
3.	Anubharati	1995	Hindi to English
4.	Mantra	1999	English to Hindi
5.	MAT	2002	English to Kanadda
6.	Vaasannubaada	2002	Bengali - Assameese
7.	AnglaHindi	2003	English to Hindi
8.	AnglaBharti-II	2004	English to Hindi
9.	MaTra	2004	IL to IL
10.	Anubharati-II	2004	IL to IL
11.	Shiva & Shakti	2004	English to Hindi
12.	English -Telugu MTS	2004	English to Telugu
13.	Telugu -Tamil MTS	2004	Telugu to Tamil
14.	Anubaad	2004	English to Bengali
15.	Hinglish	2004	Hindi to English
16.	Punjabi to Hindi MTS	2007	Punjabi to Hindi
17.	English to Malayam MTS	2008	English to Malayam
18.	Sampark	2009	IL-IL
19.	Hindi to Punjabi MTS	2009	Hindi to Punjabi
20.	English to Sanskrit MTS	2010	English to Sanskrit
21.	Hindi to Dogri MTS	2014	Hindi to Dogri

Translation Scenario in J&K

The state of Jammu and Kashmir has a diversity of languages. The languages which are majorly spoken in the state are: Urdu, Dogri and Kashmiri. The translation of Hindi and English text into Urdu is available online. Google has its translation tool where English as well as Hindi can be converted to Urdu. Kashmiri and Dogri are both low resourced languages. Some translation tools for Kashmiri have been developed by researchers in Kashmir University.The Department of Information Technology (DIT), India has got some work done on software localization in Dogri. Some of the software tools that are available in Dogri are: Open Office, Firefox, Thunderbird, Pidgin messenger, Sunbird calendar, scribus page layout application etc. The only tool available for the translation of Hindi text into Dogri is the Hindi-Dogri machine translation software developed by the author. The author of the paper is also working for the development of Dogri-Hindi Machine translation system.

Conclusion

Natural language processing is an important technique to overcome the language barriers. In a multilingual country like India, there is a great need to develop such systems for content localization. The research in this field is now focusing on development of systems to convert from Indian languages to other languages like IL to Japanese, IL to Spanish etc. whereas there are still many Indian languages which do not even have computational resources to develop these tools. A lot of effort in terms of time and finance is required to start from scratch. There is a great need to promote and provide funds for digitization of Indian languages.

References

Bharati, Akshar, Chaitanya, Vineet, Kulkarni, Amba P., Sangal,Rajeev. 1997. Anusaaraka: Machine Translation in stages. *Vivek, A Quarterly in Artificial Intelligence*, 10(3), NCST, Banglore. India, 22-25.

Kommaluri Vijayanand, Sirajul Islam Choudhury, Pranab Ratna. 2002. *VAASAANUBAADA - Automatic Machine Translation of Bilingual Bengali-Assamese News Texts*. Language Engineering Conference. Hyderabad, India. [Internet Source:http://portal.acm.org/citation.cfm?id=788716

Preeti Dubey, *The study and development of Macine translation from Hindi language to Dogri language: An important tool to bridge the digital divide*. Thesis submitted in the Department of Computer Science & IT, University of Jammu, 2014

R. M. K. Sinha, Jain R., Jain A. 2001. *Translation from English to Indian languages: ANGLABHARTI Approach*. In proceedings of Symposium on Translation Support System STRANS 2001.February 15-17, IIT Kanpur, India. pp.167-172.

Renu Jain, R.M.K.Sinha, and Ajai Jain, Anubharti. 2001. *Using Hybrid Example-Based Approach for Machine Translation.* In proceedings of Symposium on Translation Support Systems (SYSTRAN2001), February 15-17, 2001. Kanpur. P.g.123-130