

An Innovative Open Framework for Multilingual Translation with Ambiguity Resolution

Shreya N. Patankar

Department of Computer Engineering
Datta Meghe college of Engineering
Airoli, NaviMumbai
snp.cm.dmce@gmail.com

Madhura M. Phadke

Department of Computer Engineering
Datta Meghe college of Engineering
Airoli, NaviMumbai
mmp.cm.dmce@gmail.com

Satish R. Devane

Department of Information Technology
Datta Meghe college of Engineering
Airoli, NaviMumbai
satish @dmce.ac.in

Abstract—Language is the core of Information exchange, where world has about 5000 languages in the spoken and written form, also available is infinite documented information written and stored in these languages. It's a present day need in this Information technology era to make available this information to the entire world in their understandable form i.e. in their natural language. Presently machine translation is attempted by many researchers for one to one language translation or for more than one language pair and with this to translate from any to any language requires huge parallel programs and corpus and major issue of huge training and testing period with various complexities like languages in concern exhibit a different structure, not all the words in one language have equivalent words in other languages, different ways in which sentences are put together, words with multiple meanings which results into ambiguity etc. This paper proposes an open frame work for multilingual machine translation with ambiguity resolution.

Keywords—Neutral language, ambiguity, multilingual, machine translation.

I. INTRODUCTION

Multilingual machine translation is a need of today's technological society where huge information is available usually in the form of documents. Most of the information is available in a particular language. It is an important concern in today's world to process these documents and obtain meaningful information where different linguist background officials can work together. Many businesses and organizations seeking insight into global market require multilingual translation systems for translating huge amount of information available in multiple languages. Human translators with high level of confidentiality and professional experts perform translation in one particular language. Manual translation being expensive and coping up with the growing requirement of cost competence in translation of various languages has paved the way for multilingual machine translation.

The basic requirement for machine translation (MT) is availability of lexical resources and thorough understanding of the intended language pair, which will be available from the grammar rules and its underlying structure. However all the rules are not formally or clearly defined [4]. Complexity of the system again increases whenever it needs to be translated into multiple languages. Multilingual Machine translation from one language to many languages and vice versa is a complex task and comes with its share of challenges categorized as

A. Language structure

- Scarcity of lexical resources

- Dissimilarity in representation methods
- Structural differences
- Lexical gap

B. Context meaning

- Word dependency
- Polysemy words
- Parts of speech inflexion
- Different ambiguities

Multilingual machine translation exhibits various major issues of huge training and testing period with different languages in concern exhibiting a different structure, not all the words in one language have equivalent words in other languages, different ways in which sentences are put together, etc. One of the major concerns of any translated text is to preserve the correct meaning of the text, which gives scope for intelligent disambiguation process. These ambiguities occur in the context where there are words with multiple senses and decrypting the intrinsic meaning of the words in the context becomes difficult [6].

Taking into consideration the vogue of the languages used and dissemination of the information currently present in variety of languages its needed that issues related to word ambiguity should be resolved so that the translated information is in the correct form of meaning, grammar and sense.

Relationship among words in the sentence or context describes the appropriate meaning of the intended word. Recognizing the relationship between the words is achieved by observing the dependencies between them. Language structure (subject-object verb) defines the dependencies between various words. Due to language diversity, finding these dependencies is a complex task which makes word sense disambiguation a challenging problem among researchers. Our approach targets to design a multilingual translation framework which alleviate ambiguity by finding these dependencies. The task our framework address is the disambiguation of polysemy words and is based on the use of contextual information, ontological information and dependency observation.

This paper proposes an innovative open framework by introducing a neutral language, designing repository and ambiguity resolution using the dependency check. Repository provides a convenient way of integrating information of different language words and their forms. Each word in the repository in any of the form is assigned a

unique code which is universally same for any language across the globe. The innovative approach for generating uniqueness of the code in the repository derived from the meaning and sense of the source language will ease the multilingual translations and will act as milestone in the development of language translation.

The proposed multilingual open framework initially progresses using three languages, Hindi as national language, marathi as regional language and English being global auxiliary language for research purpose but the framework is adaptable for future development using any language. Rest of the paper is presented as follows. Section II focusses on related work in Word sense disambiguation and Machine translation process. Section III presents the innovative framework of neutral language repository design and dependency checking. Section IV sums up the conclusion and future work.

II. LITERATURE REVIEW

Natural language processing's ability to change our technological society has gained momentum. One promising application of natural language processing is Machine translation and more recently Neural Machine Translation(NMT) is considered highly active approach [5]. NMT requires huge amount of parallel corpus which results in extensive training time. Language structure which is an important aspect of translation is overlooked in NMT. Navigli R. in [1] has beautifully presented a survey of Word sense disambiguation (WSD) highlighting the motivation for solving the ambiguity of words and providing description of the task. Multilingual translation and word sense disambiguation survey was presented in [2] which highlights various types of ambiguities and discusses what way word sense disambiguation is helpful in multilingual machine translation.

For a multilingual country like India which is largest democratic country in whole world, there is a big requirement of automatic machine translation system. With the advent of Information Technology many documents and web pages are coming up in a local language so there is a large need of good MT systems to address all these issues in order to establish a proper communication between states and union governments to exchange information amongst the people of different states. Phadke and Devane in [3] focusses on analytical study of multilingual machine translation. It mainly discusses about major available framework like CICC, Google Translator and Anglabharati.

The work in MT supports three approaches namely Direct, Transfer and Interlingua. The Direct approach is a word-for-word substitution between language pairs using a large bilingual dictionary. The Transfer approach operates over three stages: analysis, transfer and generation. The Interlingua approach, involve an understanding of the content of texts. Recent development in interlingua MT is use of Universal Networking Language(UNL) which can be used to translate one language into various languages without writing code for set of all these languages. A lot of research is going on UNL all over the world. Currently it supports English, Japanese and Chinese [8] [9].

It is clearly evident from the literature review that mass of languages makes it necessary to have innovatory multilingual machine translation systems. One language pair where two languages are concerned in itself adds to the complexity of the system. As more languages are added, complexities tend to multiply in the multilingual scenario. Traditional multilingual machine translation tools like google translate lacks accurate translation as context is least considered for ambiguity resolution and poor quality in target language generation. Hence there is a need to develop a system which is ambiguity free and provides better precision. It is also observed that huge amount of parallel corpus is being used which consumes the training time. Since the machine lacks the extensive experience of concept perception as that of human, interpreting the correct meaning of the polysemy word becomes a difficult problem. Survey done on Word sense disambiguation and machine translation motivates us to design a new framework which focusses on retrieving the appropriate meaning of the ambiguous word from the context using neutral language repository and perform translation for multilingual systems.

III. PROPOSED FRAMEWORK

Multilingual MT is a complex task and to simplify it, the framework is divided into two parts namely word sense disambiguation and Machine translation as shown in Fig. 1.

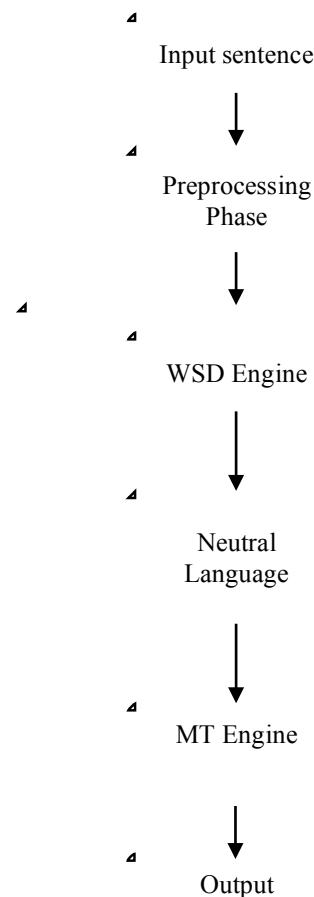


Fig. 1. Proposed multilingual MT framework

Existing Multilingual MT framework suffers from maintaining multiple parallel corpus along with the two way

efforts to support language pair. Let us consider the efforts required to translate from one language to other is E1 which includes various processing steps of machine translation. It is calculated as follows.

E1 = part of speech tagging + morphological analysis + syntactic analysis + semantic analysis + ambiguity resolution + others processing required.

Also maintaining the parallel corpus for the same. Visa versa will be E2. As shown in Fig. 2, as we add a new language, bidirectional efforts are needed which adds complexity to the system which increases with multilingual scenario.

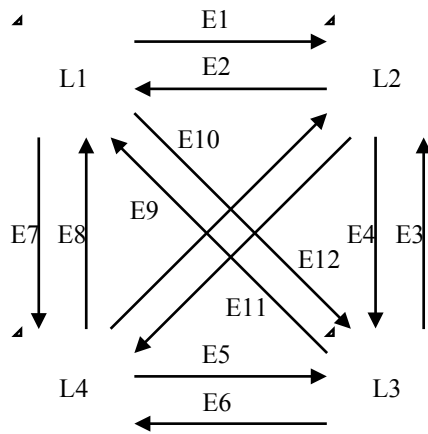


Fig. 2. Existing multilingual MT framework

With intent to support ambiguity free multilingual translation, this paper proposes an innovative open framework that makes efficient contribution by designing a Neutral language along with multilingual repository which includes not only nouns but all parts of speech and its different forms like gender, singular, plural, tenses etc. It provides a convenient way of integrating information of different language words and their forms which makes it different from multilingual dictionary. Each word in the repository in any of the form is assigned a unique code which is universally same for any language. Uniquely identified, the neutral language code is used to produce efficient and meaningful target language translation. The idea of framework is represented in Fig. 3 below.

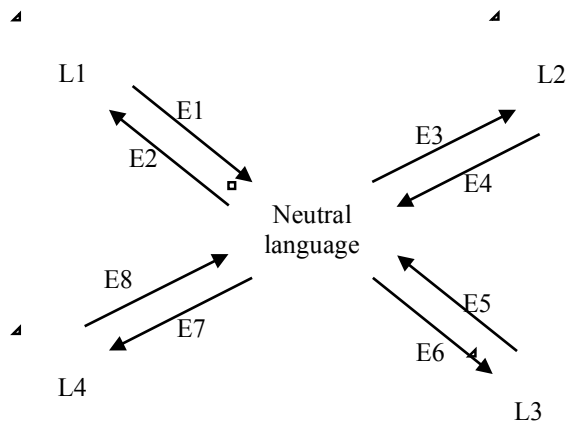


Fig. 3. Neutral language

It is clearly evident from the proposed framework that there is gradual reduction in the efforts required to handle a new language in multilingual framework which is also represented in table I below. It compares the existing multilingual system with proposed multilingual MT framework.

TABLE I. EFFORTS IN MULTILINGUAL MT FRAMEWORK

No. of Languages	Efforts in existing system	Efforts in proposed system
3	6	6
4	12	8
5	20	10
6	30	12
7	42	14
8	56	16
9	72	18
10	90	20

Number of efforts required in existing framework is calculated as

$$nP_r = \frac{n!}{(n-r)!} \quad (1)$$

where n is number of languages used.

where as the efforts required in proposed framework is twice the number of languages in concern.

$$\text{efforts} = 2 * n \quad (2)$$

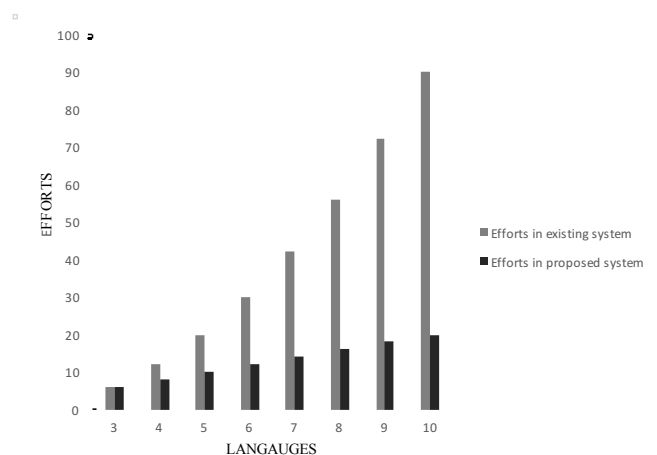


Fig. 4. Efforts

As seen from the graph in Fig. 4, as the number of languages (n) increases, the inter-translation of sentences amongst languages becomes exponentially complex and time consuming. Efforts does not add up while inter-translating between 3 or more languages resulting in separate efforts each time. The idea is to reduce this complexity and efforts required thereon by introducing the concept of Neutral Language Hence when inter-translation of languages is required, neutral language acts as central indexed

resource thus reducing efforts and complexity for inter-translating languages

A. Word sense disambiguation

Design of neutral language includes forming a repository having expandable, multilingual, multidimensional features in which ambiguities can be efficiently handled leading to successful multilingual translation. Design of neutral language includes each word represented uniquely. Context in any language will be reduced to words for which codes will be assigned which is universally same for any language. This code is independent of any language and the idea behind designing such a neutral language code stem from the byte code concept of java virtual machine which is independent of any platform. For example, sense of *bat* in English, *वटवाघुळ* in Marathi and *चमगादड़* in Hindi corresponds to a single unique code. It also handles various forms of a given word. To eliminate ambiguities from the text, we perform dependency check by providing world knowledge to the word sense disambiguation(WSD) engine. World knowledge will be influenced by our knowledge repository and other sources.

World knowledge helps to determine the correct sense of the word in the given context. It provides various properties of a given word helpful for finding the dependency between the various words in the context. Neutral language code and world knowledge is fed to the WSD engine. Depending upon the context and world knowledge provided, the engine will select the appropriate sense of the ambiguous word. For example

Crane loves eating fish

Let us consider the ambiguous word *crane* having more than one meaning(machine, bird, etc). World knowledge says bird as a living species. Living species with properties like flying, eating, standing, etc marks crane as a bird sense. This scenario is just a glimpse of the uphill battle. The actual work need lots of digging and learning to retrieve the world knowledge, facts that are available from the available repositories. Grouping of relevant properties of the given entity helps in resolving this issue. Word sense disambiguation engine returns the appropriate non ambiguous neutral language code for the words in the input sentence. This codes are further used for Machine Translation (MT) process.

B. Multilingual machine translation

Multilingual machine translation is the dream of researchers to eliminate the translation load of the human translators. The idea is to have a system to provide translated text in required language for the given input in a specific language. By considering the language diversity handling multilingual translation is a effortful task. Indian languages are rich in varied structure. The complexity is increased with the use of languages with different morphological features. The framework deals with two most popular Indian languages viz Hindi and Marathi along with the global language English. Marathi is grammatically rich language and more efforts are required to handle the peculiar forms of the words. Hindi on the other side is similar to Marathi in some aspects but needs extra attempts [7].

Upon receiving the neutral language code from the WSD engine, it is expected from MT framework to produce the meaningful translation as shown in Fig. 5. To achieve this,

language grammar rules are used. For a single word in English , Marathi and Hindi may go with several different forms depending upon the context.

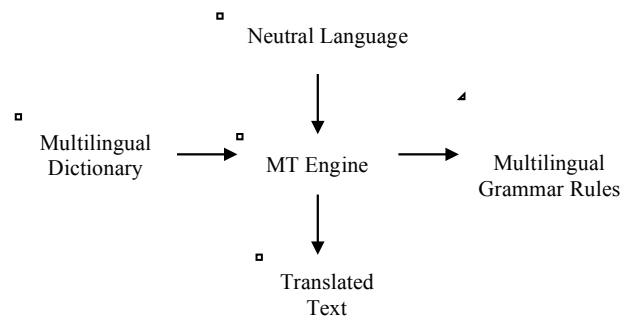


Fig. 5. Multilingual MT framework

Use of Multilingual dictionary makes for more accurate translation. Grammar structure of all the three languages being different, makes the translation process more complex. Hence designing rules by considering all possible forms from the multilingual repository contributes towards efficient machine translation.

IV. CONCLUSION

Multilingual translation is the need of society to make sense of the huge information available in the form of documents in a particular language and translated into a language required by the user. Proposed intelligent framework is an important concern which will process this data and obtain useful information for translation between any to any language as human translation is a time consuming task. We introduce the concept of neutral language independent of any language for the evaluation of word sense disambiguation and multilingual machine translation task. This framework is adaptable and open for natural language processing users to use it for translation task for any required language.

REFERENCES

- [1] Navigli Roberto, " Word sense disambiguation: A survey", ACM Computing surveys}, Vol. 34, No. 2, Article 10, 2009
- [2] Patankar S., Devane S., "Issues in handling word sense disambiguation for multilingual translation ", International conference on intelligent computing and control systems (ICICCS), IEEE, 2017
- [3] Phadke M., Devane S., "Multilingual Machine translation: An analytical study", International conference on intelligent computing and control systems (ICICCS), IEEE, 2017
- [4] Garje G., Kharate G.; "Survey of machine translation systems in India.", International Journal on Natural Language computing. Vol. 02, No.4, October 2013.
- [5] Firal O, Cho K., Bengio Y; " Multiway multilingual neural machine translation with a shared attention mechanism", Proceedings of the NAACL-HLT 2016, pp. 866-875.
- [6] John Hutchins, "Current commercial machine translation systems and computer-based translation tools: system types and their uses", *International Journal of Translation*, 2005, vol.17, no.1-2, pp.5-38.
- [7] Latha R. Nair & David Peter S., "Machine Translation Systems for Indian Languages", *International Journal of Computer Applications* (0975 – 8887), 2012, Volume 39– No.1
- [8] S. Dave, J. Parikh, P. Bhattacharyya, "Interlingua-based English–Hindi Machine Translation and Language Divergence" , *Machine Translation*, December 2001, Volume 16, Issue 4, pp 251–304
- [9] www.undl.org

