2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC) | 978-1-6654-9710-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAAIC53929.2022.9792969

Natural Language Processing based Refining Hindi to English Machine Translation

Paridhi Gupta
Military College of Telecommunication Engineering,
Mhow, India
E-mail: paridhirkagrawal@gmail.com

Brijendra Kumar Joshi
Military College of Telecommunication Engineering,
Mhow, India
E-mail: brijendrajoshi@yahoo.com

Abstract—India is a country full of diversity in culture and languages. Hindi is one of the most spoken languages in India and translation of Hindi text into English text is the need of the hour. This paper proposes a seven-module system architecture that translates some Hindi sentences into English sentences correctly that Google translate fails to do so. There are about ten sentences to cross-check the results and it is found that the proposed architecture outperforms the Google Translator.

Keywords Machine translation, Translation-Rules, and Word-Net in English and Hindi

I. INTRODUCTION

Multilingualism and multiculturalism are hallmarks of India. In various parts of the world, people speak distinct dialects of the same language. In India, it has been found that the language varies significantly every 50 kilometers or so. Hindi is the most widely spoken language among the country's 23 official languages, including English. Different languages are used by state governments to conduct business and perform other official duties. In addition to English and Hindi, the federal government releases directives and other papers. A machine translation system (MTS) that can translate from English to Hindi and vice versa, as well as Hindi to other regional and worldwide languages, is seen to be very beneficial [1][2].

Though MT research has been ongoing for decades, achieving high-quality MT remains a challenge. In India, MTS has the most customers [3]. Figure 1 is a block schematic of a basic MTS.

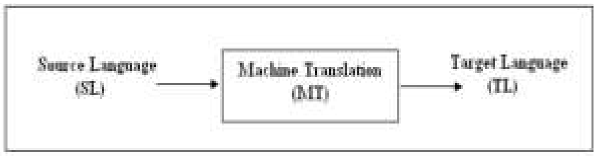

Figure 1. A simple MTS

Numerous Natural Language Processing (NLP) issues arise while dealing with machine translation (MT), including issues related to phonetics and phonological, morphological, syntactic, semantic, and discourse issues. Ambiguity is the most important of them (Semantics).Aside from that, the different languages may have an issue with linguistic variety (also known as translation divergence). Under the umbrella of NLP, MTS deal with ambiguity and language variety issues [4].

In India, we believe that Hindi-English and Hindi-Regional languages are the most essential MTs.

Due to its peculiarities, Hindi-to-English MT is more complicated. Depending on the context, anything written in Hindi might have many meanings. People may speak any statement in the Indian language in a different order [5] [6]. A block diagram for a Hindi-to-English MTS is shown in Figure 2.

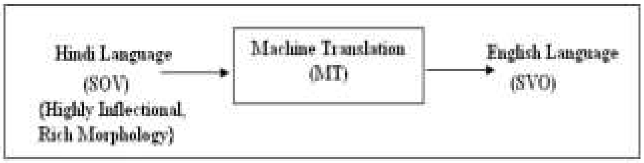

Figure2. Hindi-to-EnglishMT

A large amount of verb movement is the result of MT from English to Hindi. Hindi, on the other hand, adheres to the gender agreement, unlike English. Adding linguistic characteristics to the source side may help with the morphological issues. Resources in English There are five and six. Figure 3 is a block diagram for an English-to-Hindi MT system.

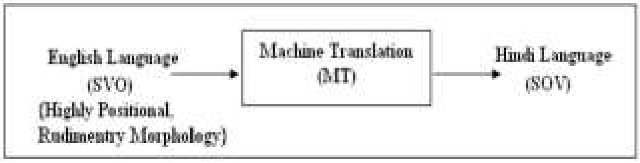

Figure 3. English-to-Hindi MT

By using a method known as Word Sense Disambiguation (WSD), the Hindi-to-English may be enhanced. Knowledge/dictionary-based, supervised, semi-supervised, and unsupervised techniques are all types of WSD algorithms. Using singles or combos, on the other hand, has no bounds. Previously, the combinations have had positive outcomes [7] [8].

II. Literature Review

MT has been the subject of a great deal of study and development in India during the last three decades. Even though they have generated some excellent MTSs, they all have their own set of benefits, drawbacks, and limits. As a result, there is still room for greater research in this field. Many studies and research efforts are also underway to overcome these drawbacks and restrictions. These objectives motivate students and researchers to learn about MT from an Indian perspective. [9].

Many surveys are conducted in the field of MT from an Indian perspective. The first part of the survey is about MT resources, services, and tools in India. This survey provides a thorough examination of the Indian viewpoint. [10]. Second, the survey incorporates a WSD technique that can be utilized to improve MT [11]. Knowledge-based approaches, for example, supervised approaches, unsupervised approaches, hybrid approaches, and so on are covered in this section. The Word-Net approach is also covered here, as well as new techniques under these approaches. Lastly, the review examines a variety of methods for building systems [12]-[15]. There are several Indian-created varieties included in the study. Name, year of creation, people and/or organizations involved, funding source(s), location of the development, system domains/applications, approaches/techniques and resources/tools used, and so on are all included in these surveys. Links to these MT technologies are included in all surveys and may be accessed online.

The MTS has to deal with ambiguity and divergence issues at all levels of NLP [4][8]. In multilingual systems, resource restrictions like WordNet, which is expensive and time-consuming, have been observed, If you're looking to translate between English and Indian languages, you may use Anglabharti. A methodology based on pseudo-interlingua rules is used. The system has a positive effect on the outcome. However, this isn't always the case when translating from English to another language. The rule-based translation is used in the Personnel Administration part of Mantra. Other facilities may benefit from the additional topics of study generated by this strategy. The direct (word-to-word) method was used to build the Anusaaraka system, which transforms texts from one Indian language to another [21]. This method is also effective, but it has serious drawbacks if it is used broadly. A system that utilizes the Universal Networking Language (UNK) (UNL). Although the English-to-Indian language translation process is a good system, UNL outputs are impacted by linguistic divergence problems between the source and destination languages [22].

Translation from English to Hindi is handled by AnglaHindi, a part of the Anglabharti project [23]. Rules and examples are used in a hybrid method to develop it. MaTra is a general-purpose English-to-Hindi machine translation system that is fully automated [24]. It's made with rule-based (transfer-based) approaches. Google Translator, Bing Translator, Worldlingo, and IBM Server are statistical-based MT from Google, Microsoft, Worldlingo, and IBM, respectively.

A lot of internet apps for Hindi-to-English MT are available and accessible. Table 1 provides a detailed breakdown of the effectiveness of such applications. For example, utilizing the web apps listed in the table, the Hindi language statement "मेरे लाल का रंग काला है" has been transformed into the English language. By examining the output, it is clear that the majority of the apps failed to provide the anticipated results. "My Red color is Black" is the only positive result from "Google Translate." However, it is unable to detect the ‘Noun” as a result of which it produces "Red" instead of "Son’s" which is the synonym of the Hindi word “लाल”. The remaining apps give incorrect results. Consequently, an upgraded and acceptable version of the Hindi to English machine translator is needed to provide better and more suitable results. When searching for words, you may use Word-Net, an online lexical database for the English language, to find a synonym set for each of the four basic parts of speech (PoS): noun, verb, adverb, and adjectival adjective [26]. Indo Word-Net is an interconnected network of the most important Word-Nets on the Indian subcontinent.[27]

Supervised, semi-supervised, unsupervised, and hybrid methods for WSD techniques and applications based on knowledge/dictionaries [7] Each has its own set of strengths, weaknesses, and limitations. [28] Selecting MTS-improving approaches is made easier with the help of the critical assessment [28]. Experimentation with graph connection led to the invention of Unsupervised WSD [29].

Creating a concept map to aid in MT improvement is critical because it enables the combination of concepts and data that are linked in some way. This creates a logical connection between two concepts or strands of information. A concept map can be used to link concepts from the same domain [30, 31].

Study pathways are provided for various sorts of similar translations, such as the Hindi English Sign Language Translation System [32], in the proposed system for Chinese-Japanese Sign Language Translation. The study of MT between Hindi and English (Hinglish) is critical in the quest to distinguish between a mixed collection of languages and pure component languages [33].

The key measure is BLEU (Bilingual Assessment Understudy), however, other metrics are also useful in the autonomous evaluation of MTS. BLEU includes several approaches that are useful in evaluating MTS [6], [34].

There is a richness of classical literature in Hindi. In the 15th century, a script is known as the "Devanagari lipi (script)" was devised specifically for these texts. There are a large number of books published in the Hindi script. Today's world needs a great deal of English translation. These studies have been more intense in the last several decades.

Poetry translation is one of the most difficult types of MT. In Hindi, there are several poetries. This relocation has taken a lot of time and effort. For poetry translation into English, the current approach requires a better MTS [36].

III. Possible Approach

Any MTS has a tough time resolving ambiguities and translation divergences (TD). At this point, we're trying to come up with an effective strategy for dealing with these issues and producing a high-quality translation. Our Hindi-to-English MT technique will include the following modules and the characteristics shown in Figure 4. Here is a breakdown of the modules:

a. Statistical (MT)Methods/Models are

i. Finite-state Transducer Models
- Word-based Models
- Phrase-based Models

ii. Synchronous context-Free-grammar Models
- Bracketing Grammars
- Hierarchical phrase-based Translation
- Syntactic phrase-based Models
- Synchronous dependency Grammar

iii. Tree adjoining Grammar[25]

b. Machine Translation using Semantic Rules

i. Rule-based MT
- Direct approaches
- Transfer based approaches
- Interlingua based approaches

ii. Corpus-based MT
- Statistical approaches
- Example based approaches

iii. Hybrid MT

IV. System Architecture

This paper outlines an efficient approach for translating from English to Hindi. New rules have been introduced to the proposed system in order to speed up the translation process. The performance of the suggested technique has been significantly enhanced by the addition of the elements shown in figure 4. The hybrid mechanism we may investigate might be applicable to the English-to-Hindi language in general since Indian languages have very rich morphology and follow the SOV order. New morphological reordering rules may be developed for more Indian languages to improve the suggested technique.

a. Input Language (IL): English/Hindi
b. Output Language (OL): Hindi/English

An introduction to Natural Language Processing (NLP).

- Simplification & splitting
- Part Of Speech Tagger
- Morphological Analyzer
- Semantics Parser
- Reordering of structure
- Discourse [4] [7]

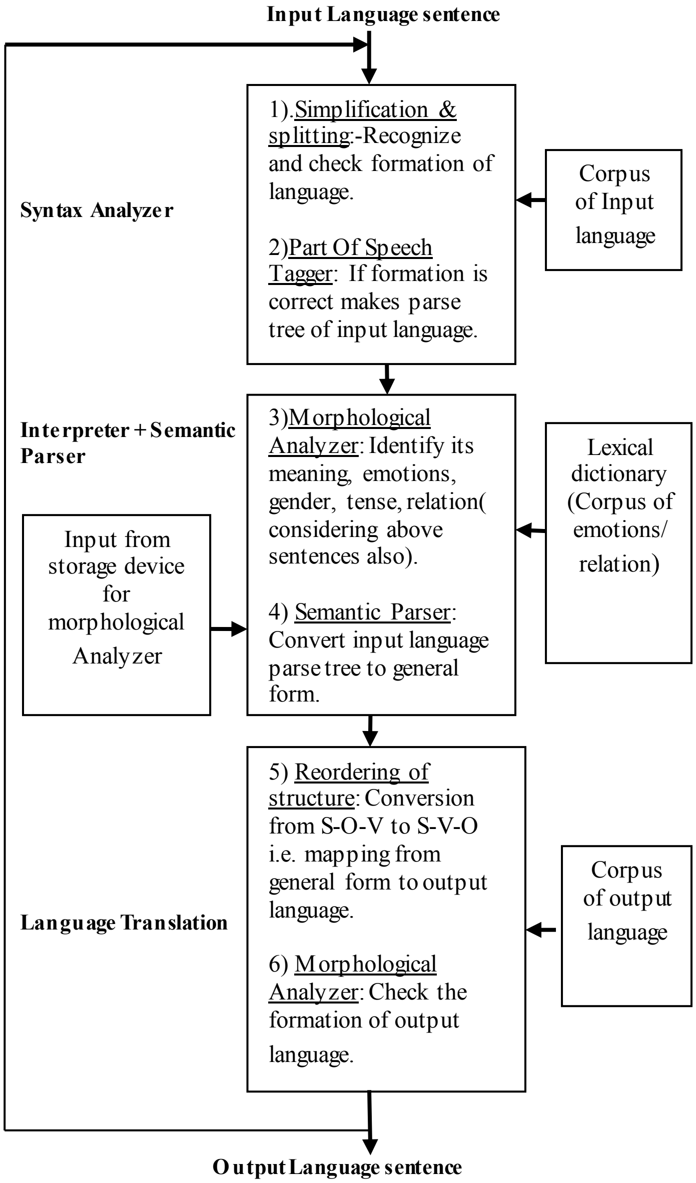
Input Language sentence
Syntax Analyzer
1).Simplification & splitting:-Recognize and check formation of language.
2)Part Of Speech Tagger: If formation is correct makes parse tree of input language.
Corpus of Input language
Interpreter + Semantic Parser
Input from storage device for morphological Analyzer
3)Morphological Analyzer: Identify its meaning, emotions, gender, tense, relation(considering above sentences also).
4) Semantic Parser: Convert input language parse tree to general form.
Lexical dictionary (Corpus of emotions/ relation)
Language Translation
5) Reordering of structure: Conversion from S-O-V to S-V-O i.e. mapping from general form to output language.
6) Morphological Analyzer: Check the formation of output language.
Corpus of output language
Output Language sentence

Figure4. Structure of Proposed System Hindi-to-English MT

The proposed system has effectively applied over 10 English-to-Hindi sentences. The exact accuracy of this MT is only defined when we apply the procedure over a large dataset and can able to find out its performance over another translation system.

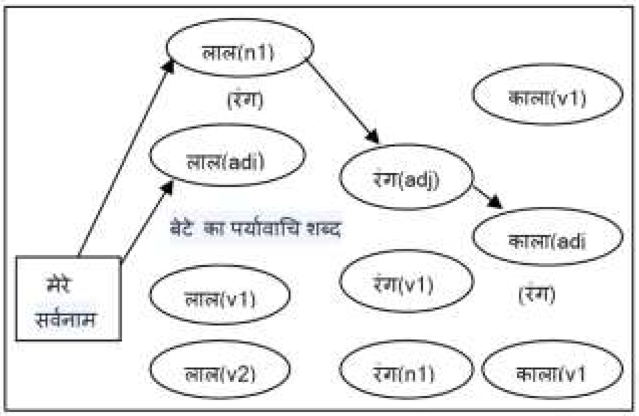

Figure 5. Constitutes of Hindi Sentence "मेरे लाल का रंग काला है".

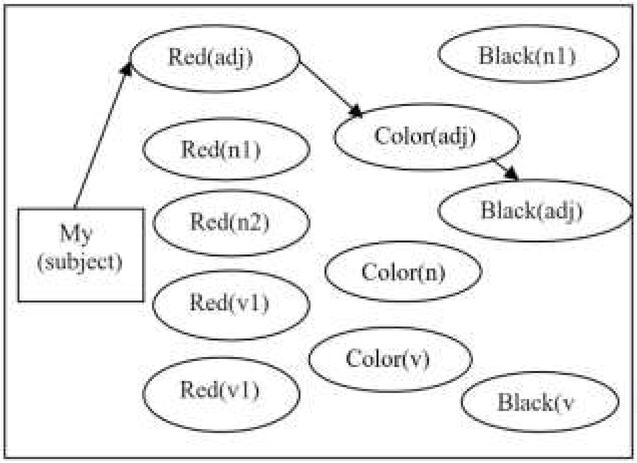

Figure 6. Constitutes of English sentence “My Red color is Black”

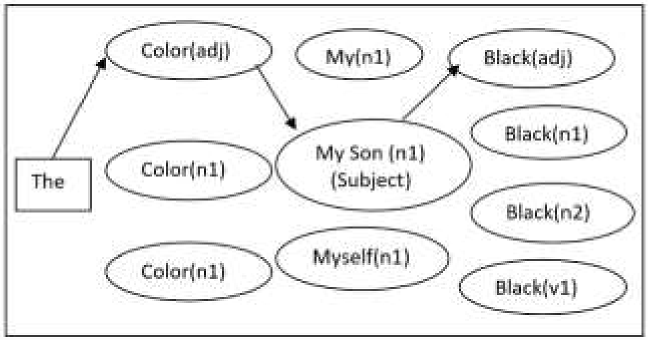

Figure 7. Constitutes of English Sentence “The color of my Son is Black”.

Example WSD algorithms and Concept-Maps for a suitable Hindi-to-English MT are shown in Figures 5, 6, and 7. We used Word-Net and Hindi WordNet to build graphs for both English and Hindi. We may deduce from these graphs that WSD algorithms and Concept-graph Construction would be extremely beneficial in Hindi-to-English Machine Translations. Figures 5, and 6 are results from online MT and Figure 7 shows the result for the proposed system. The system results when applied over large datasets it outperforms better than other MTS.

V. CONCLUSION

Making things accessible to everyone is the goal of building a translation system like this one. An Indian-to-English voice translation system might be created using this technology in conjunction with Automatic Speech Recognition (ASR) and Text To Speech (TTS) technologies.

There were 37 papers evaluated for different projects in India and throughout the globe for Hindi-to-English and other MT initiatives. MTS development has been well researched in terms of techniques, procedures, and resources. First, a number of different translation systems were tested to see how well they performed. Different MT models have been thoroughly examined and architecture has been described that may help enhance current Machine Translation systems. With this approach, it is feasible to get the following results:

Higher and increased accuracy in the translation of Hindi-to-English sentences from a computer.

Large-scale, shallow-depth poetry and/or literature may now be translated automatically.

High-quality Hindi to English MT techniques.

VI. FUTURE SCOPE

This study thoroughly examines the need for MT and proposes a method for a successful Hindi-to-English MT that is both cost-effective and easy to implement. There have been several attempts to find a machine translation system that can translate from Hindi to English both inside and outside of the country. These Hindi-to-English MTS may be implemented and upgraded using a variety of resources, methodologies, and tools. Improved Hindi-to-English MTS is possible using the suggested architecture and approach. Poem translation will be used to illustrate the usefulness of the Hindi-to-English MTS in Primary Education. There has been a lot of research done on the structure of poetry and literature in general. There may be a high-quality Hindi to English translation of poetry that is regarded as the best in the world.

Table 1. Input statement analysis of existing Hindi-to-English MT system“मेरे लाल का रंग काला है”

S. N .	Name of the Program or Tool	Online sources	Output Sequence	Observation
1	Google Translate	https://translate.google.com/?hl=en	My Red color is Black	Good but couldn’t understand Noun.
2	Google Translate	https://translate.google.com/?hl=en	The color my Red is Black	Good but couldn’t understand the adjective.
3	Bing Translator	https://www.bing.com/translator/	My Red is Black	Improper Output
4	Word lingo	http://www.worldlingo.com/en/products_services/worldlingo_translator.html	The color Red is Black	Improper Output
5	TDIL	http://www.tdil-dc.in/components/com_mtsystem/CommonUI/homeMT.php	Doesn’t Support	---
6	Anusaaraka	http://anusaaraka.iiit.ac.in/drupal/node/32	Doesn’t Support	---
7	Mantra	https://mantra-rajbhasha.rb-aai.in/RegisterFirst.do?function=init	Doesn’t Support	---
8	Machine Translation	http://www.cfilt.iitb.ac.in/machine-translation/eng-hindi-mt/	Doesn’t Support	---
9	Anuvad	http://kbcs.in:8080/anuvad/	Doesn’t Support	---
10	Language Translator	http://www.ibm.com/smarterplanet/us/en/ibmwatson/developercloud/machine-translation.html	Doesn’t Support	---

REFERENCES

[1] NaskarS,BandyopadhyayS.UseofMachineTranslationin India:CurrentStatus.Thail and Phuket ProceedingsofMT SUMMIT X. 2005 September; p.465-70.

[2] GarjeGV, Karate GK. Survey of Machine Translation Systems in India. International Journal on Natural Language Computing (IJNLC).2013October;2(4):47-67.Crossref

[3] Latha RN, David PS. Machine Translation Systems for Indian Languages. International Journal of Computer Applications (IJCA). 2012 February;39(1):25-31.

[4] JurafskyD, Martin JH. Speech and language processing. 2nd and. Pearson Education India.2002.

[5] Rao D. Machine Translation in India. Bangalore: SCALA 2001Conference: A Brief Survey.2001;p.1-6.

[6]

[7] DungarwalP, Chatterjee R, Mishra A, KunchukuttanA, Shah R, Bhattacharyya P. The IIT BombayHindi⇔English Translation System at WMT 2014. Association of Computational Linguistics (ACL). 2014; p. 1-7.

[8] Agirre E, Edmonds PG. Word sense disambiguation. Algorithms and applications.Springer Science and Business Media. 2007;33(1):255-58.

[9] Das A, Sarkar S. ICON.201: 3Word SenseDisambiguation in Bengali applied to Bengali-Hindi Machine Translation.2013; p.1-10.

[10] Bandyopadhyay S. Teaching MT - An Indian Perspective. UK: Manchester: Proceedings of the 6th EAMT Workshop on Teaching Machine Translation. 2002;p.13-22.

[11] Badodekar S. Translation resources, services, and tools for Indian languages. Computer Science and Engineering Department. Indian Institute of Technology, Mumbai, 2003. Date accessed: 21/03/2016: Available from: http://www.cfilt.iitb.ac.in/Translationsurvey/survey.pdf.

[12] Shallum, Gupta. a survey of Word-sense Disambiguation Effective Techniques and Methods for Indian Languages. Journal of Emerging

Technologies in Web Intelligence. 2013 November; 5(4):354-60.Crossref

[13] Tripathi S, Sarkhel JK. Approaches to machine translation. Annals of library and information studies.2010;57(1):388- 93.

[14] Sanyal S, Borgohain R.Machine TranslationSystems in India. 2013; p. 5. Available from: arXiv preprint arXiv.1304.7728.

[15] Antony PJ. Machine translation approaches and surveys for Indian languages. International Journal of Computational Linguistics and Chinese Language Processing.2013March; 18(1):47-78.

[16] Godse A, Govilkar S. Machine Translation Development for Indian Languages and its Approaches. Date accessed: 16/06/2016: Available from: http://airccse.org/journal/ijnlc/papers/4215ijnlc05.pdf.

[17] Goyal V, Lehal GS. Advances in Machine Translation Systems.Language in India.2009;9(11):1-13.

[18] Dwivedi SK, SukhadevePP. Machine Translation System in Indian perspectives. Journal of computer science. 2010; 6(10):1111-16.Crossref

[19] Bhattacharyya P. Natural language processing: A perspective from computation in presence of ambiguity, resource constraint and multilingualism. CSI Journal of Computing. 2012;1(2):1-13.

[20] Sinha RMK, Sivaraman K, Agrawal A, Jain R, Srivastava R,JainA.ANGLABHARTI, a multilingual machine-aided translation project on translation from English to Indian languages.IEEEInternationalConferenceonSystemMAN and Cybernetics. 1995; 2(1):1609-14.Crossref

[21] Darbari H. Computer-assisted translation system–an Indian perspective. Machine Translation Summit VII.1999 September; p.80-85.

[22] Bharati A, Chaitanya V, Kulkarni AP, Sangal R, Rao GU. Anusaaraka, overcoming the language barrier in India. arXiv preprint cs/0308018. 2003; p.1-19.

[23] Dave S, Parikh J, Bhattacharyya P. Interlingua-based English-Hindi machine translation and language divergence.MachineTranslation.2001;16(4):251-304.Crossref

[24] Sinha RMK, Jain A. AnglaHindi, an English to Hindi machine-aided translation system.USA: New Orleans: MT Summit IX. 2003; p.494-97.

[25] AnanthakrishnanR, Kavitha M, Jayprasad JH, ShekharRS, Bade SM. MaTra. A practical approach to fully-automatic indicative English-Hindi machine translation. Symposium on Modeling and Shallow Parsing of Indian Languages(MSPIL'06). 2006; p.1-8.

[26] Bhattacharyya P. Machine Translation. USA: CRC Press: Taylor and Francis Group.2015.

[27] Miller G. A. WordNet, a lexical database for English. CommunicationsoftheACM.1995November;38(11):39-41. Crossref

[28] Bhattacharyya P. IndoWordNet, Proceedings of LREC-10. 2010; p.1-10.

[29] Dwivedi SK, Rastogi P.CriticalanalysisofWSDalgorithms. Proceedings of the International Conference on Advances in Computing. ACM: Communication and Control. 2009 January; 3:62-7.Crossref

[30] Navigli R, Lapata M. An experimental study of graph connectivity forum supervised word sense disambiguation.IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010April;32(4):678-92.CrossrefPMid:20224123

[31] Novak JD, CanasAJ. The theory underlying concept maps and how to construct them. Florida Institute for Human and Machine Cognition. 2006;1:1-6.

[32] Ca-asAJ,ValerioA,Lalinde-PulidoJ,CarvalhoM,ArguedasM. Using WordNet for Word Sense Disambiguation to support Concept Map construction. Springer Berlin Heidelberg: String Processing and Information Retrieval. January 2003; p. 350-59.

[33] LiuZ,ZhangX,KatoJ.ResearchonChinese-JapaneseSign Language Translation System. IEEE Fifth International Conference on Frontier of Computer Science and Technology(FCST).2010August;p.640-45.Crossref

[34] Sinha R.M.K, Thakur A. Machine Translation of bi-lingual Hindi-English (Hinglish) text. Thailand: Phuket: Tenth Machine Translation summit. 2006; p.149-56.

[35] Ananthakrishnan R, Bhattacharyya P, Sasikumar M, Shah RM.SomeissuesinautomaticevaluationofEnglish-Hindi MT,morebluesforBLUE.ICON-2017.2007;p.1-8.

[36] Sawant DG. Translation literature in India, 2012. Date accessed: 18/07/2016: Available from: https://www.Research.gate.net/publication/230814146.

[37] HariyantoS.ProblemsinTranslatingPoetry.Dateaccessed: 19/09/2016: Available from: http://www.translation directory.com/article640.htm.

[38] Ali OM, GadAlla M, Abdelwahab MS. Word Sense DisambiguationinMachineTranslationusingMonolingual Corpus.ProceedingsoftheEighthConferenceonLanguage Engineering. Egypt: Cairo: Ain Shams University. 2008; p.141-51.