Ambiguity in Question Paper Translation

Shweta Vikram
Babasaheb Bhimrao Ambedkar University, Lucknow, 226025, India
Email: shwetavikram.2009@rediffmail.com

Sanjay K. Dwivedi
Babasaheb Bhimrao Ambedkar University, Lucknow, 226025, India
Email: skd200@yahoo.com

Abstract—Word sense ambiguity is a prevalent nature of machine translation for various language pairs including English-Hindi language. For example, the word "paper" has several senses which may refer to a question paper, research paper, newspaper, simple paper or a white paper. The specific sense intended is determined by the context in which an instance of the ambiguous word appears. This specific sense which is determined by the context is known as Word Sense Disambiguation (WSD). Translation of question paper is a specific application of MT wherein any type of ambiguity in question may affect the overall meaning of questions. This paper discusses types of ambiguity in the context of question paper translation (English to Hindi) and their impact on translation by analyzing a set of questions taken from National Council of Educational Research and Training (NCERT) and some other resources.

Index Terms—Question paper, Word Sense Disambiguation, Hindi, English, and Translation.

I. INTRODUCTION

India is a multilingual country with a major language spoken is Hindi. Most of the people in India work in Hindi and it is the mother tongue in most of the states [26]. But people who speak Hindi face language problem in machine translation because machines often fail to translate the actual context of a sentence. Sentences sometimes have ambiguous words due to which MT tools usually fail to correctly translate the sentence into the target language. There are many approaches of WSD which use techniques such as tagging, chunking, parsing, name identity recognition, place identity recognition [9-13]. WSD is a very vast area of the research. WSD is not an easy problem and is considered as the NP-complete problem. Past few decades have witnessed researches in word sense disambiguation [7]. Machine Translation and WSD are the complementary or subsidiary of each other. Whenever machine translates from one language to another, it requires the knowledge about certain words which are ambiguous, so that the sentence can be correctly translated. It is done through appropriate WSD algorithm. Manual translation is a very cumbersome

problem as it takes too much time [3]. The fundamental order is lexical, character, syntactic, and semantic features [34]. The complete rule-based expert system has been evaluated for good result. The result of that evaluation is a good and a very good range [35]. For machine translation, query terms are automatically translated from source language to the desired language by using a context [36].

Many approaches have been proposed since 1950 for assigning senses to words in context, although early attempts only served as models for toy systems [15]. Approaches used in WSD can be categorized as supervised, unsupervised, semi-supervised, knowledge-based, bootstrapped, hybrid and dictionary-based approaches [2, 16, 19, 21-24]. The Dictionary based approach is the oldest approach and it was proposed by Karov and Edelman [28]. The supervised approach uses trained data, a major problem with supervised approaches is that it requires a large sense-tagged training set. It is widely used in medical field to get better results [38]. The unsupervised approach does not require trained data as well as a corpus. The main reason for the development of this approach is the complexity of creation of marked corpus and other necessary resources. Hybrid Approach combines two or more than two approaches. Corpus-based machine translation systems have gained much interest in recent few years. It is fully automatic and requires less human labor than other approaches, but they need sentence-aligned parallel text for each language pair. Corpus-based machine translation is classified into statistical machine translation (SMT) and Example-based Machine Translation (EBMT) [14].

II. RELATED WORK

This section reports some significant contribution to the translation of questions and question answering system.

In 2001, Dave and Bhattacharya [10] used interrogative sentences to detect the presence of Wh-word like what, where, why, whom, how etc and also find question mark symbol at the end of the sentences. These interrogative sentences are divided into two categories one is 'wh-questions' and another is 'yes-no' questions. It

is shown that when the Hindi question sentence is written in more than one way by changing the order of words, the meaning of the sentence remains the same.

In 2005, Kumar et al. [8] developed a question answering system for Hindi documents. This paper also gives an idea for question classification, question parsing, question formulation and query expansion. In 2005, Metzler and Croft [30] analyze fact based question classification through statistical method. These fact based questions are different question types.

In 2007, Singh et al [25] introduced the concept of Tense Aspect and Modality (TAM) Marker. It pointed out that many errors occur in MT are due to a wrong translation of TAM markers.

In 2011, Silva et al [31] worked on a question answering system by using question classification from symbolic to sub-symbolic information. Authors also gave the information about last few year work done on supervised machine learning approaches to question classification.

In 2014, Dwivedi and Goyal [26] work on the status of machine translation in India through an experimental analysis of question paper translation. They used BLEU (Bi-Lingual Evaluation Understudy) for evaluating experimental analysis. In 2014, Dwivedi and Singh [29] focus on integrated classification in a higher education domain which is based on rules and pattern matching. In this Wh- questions, considered for question classification.

In 2015, Graesser et al [33] gave some idea for question generation mechanism, question categorization and assumptions behind the questions for question classification scheme.

In 2015, Kamdi and Agrawal [37] give the concept of question answering system for Indian Penal code section and Indian amendment laws by using keywords based closed domain. In this work authors also defined the process of the question in three types as determining the type of question, determining the type of answer and extracting keywords from the question and formulate a query.

In 2017, Dwivedi and Vikram [32] introduced some external resources for machine translation, question classification and discussed some ambiguity related question sentences.

III. TYPES OF AMBIGUITY

Word sense disambiguation techniques can determine the correct sense of ambiguous words with respect to the context. Machine Translation is automated translation and it translates one natural language to another with or without any assistance of a human. Sense ambiguity may be of different types that have been summarized below.

A. Lexical ambiguity

In this, a word or phrase pertains to it, is having more than one meaning [20]. For Example, English WordNet [41] has more than one senses of word "master". Table 1 shows all the senses of this word in Hindi WordNet [4, 42]. Word "Master" has different sense with respect to context, in our example, "मास्टर " has been identified as sense in the question sentences.

1. Explain <u>master</u> method.
   MT (Google): मास्टर विधि बताइए|

2. What is the <u>Master</u>-Slaves flip flop?
   MT (Google): मालिक नौकर फ्लिप फ्लॉप क्या है.\

It is interesting to see that same MT tool translates the words "Master {मास्टर , मालिक}" differently in two examples above.

The WordNet and its Hindi version (Hindi WordNet) provide various senses of the word "Master" as shown in Table1 and Table 2. WordNet is an ongoing lexical resource at Princeton University since the 1980s with a hierarchical structure, where a node is a synset and a link is a relationship between two synsets. Hindi WordNet is a repository of Hindi words connected by lexical and semantic relation along with the browsing interface and associated software. Both there WordNet are considered as machine-readable dictionaries. English WordNet has the collection of all English senses, for a large number of words in English. We collect all the senses of word "master" from English WordNet, which is shown in Table 1, but it does not contain "ekfyd" sense of the word "master".

Both WordNet (Hindi and English) do not have the ekfyd (malik) meaning of the word 'master'. But example1 has the correct sense malik (ekfyd) related to context.

Table 1. Sense of word "Master" in WordNet

	Senses of 'Master' word
The noun master has 10 senses (first 6 from tagged texts)	Maestro
	Overlord, lord
	Victor, superior
	headmaster, schoolmaster
	master copy, original
	captain, sea captain, skipper
	master's degree
	Professional
	passkey, passe-partout, master key
	directs the work of other
The verb master has 4 senses (first 3 from tagged texts)	get the hang
	overcome, get over, subdue, surmount
	Dominate
	Control

Table 2. Sense of word "मास्टर " in WordNet (Hindi)

Gloss	Senses of 'मास्टर ' word
वह व्यक्ति जो विद्यार्थियों को पढ़ाता है	अध्यापक, शिक्षक, आचार्य, उस्ताद, आचार्य, गुरु, मास्टर, मुअल्लिम, स्कंध, स्कन्ध, टीचर, गुरू, वक्ता
वह मानवीकृत वस्तु जो शिक्षा दे या जिससे शिक्षा मिले	शिक्षक, गुरु, मास्टर, अध्यापक, टीचर
वह अधिकारी जिसके पास किसी व्यापारी जहाज को नियंत्रित करने का लाइसेंस होता है	कप्तान, कैप्टन, मास्टर, स्कीपर

B. Syntactical ambiguity

A sentence can be elucidated in more than one way. Often sentences may have more than one meaning because of the structure of the sentence, such as not placing appropriate punctuation [5].

"Panda eats, shoots and leaves" or "Panda eats shoots and leaves." (Comma "," arise ambiguity) [27]. For Example,

There is ambiguity in the above sentence that arises due to 'comma'. The two meanings are inferred from the same sentence but the words are not ambiguous. The human translation (Hindi) also gives two meaning as sown below.

पांडा कोंपले और पत्तियां खाता है.

पांडा खाता है मारता है और चला जाता है.

C. Semantic ambiguity

More than one way of reading a sentence is known as semantic ambiguity [1]. Below example shows semantic ambiguity.

Example: He saw a man on the hill with a telescope. ("with a telescope" arise ambiguity)

Different interpretations are possible from the above sentence. For example

1st Interpretation: There is a man on the hill, and he was watching him with my telescope.

2nd Interpretation: There is a man on the hill, who he was seeing, and he has a telescope.

3rd Interpretation: There is a man, and he is on the hill that also has a telescope on it.

4th Interpretation: He was on the hill, and he saw a man using a telescope.

5th Interpretation: There is a man on the hill, and he was seeing him with a telescope.

All the above Interpretations are possible from the sentence. Such type of ambiguity is very difficult to handle and it requires prior contextual knowledge to get the most appropriate meaning in the context.

D. Lack of information

This problem arises in translation because one language does not have full information in translation. For example, as reported by English newspaper [30]. A Question was asked in some examination and while the Tamil version was asking about "three impacts of solar energy", the English version of the same question has "three environmental impacts of solar energy".

Paper (English language): write the three Impact of solar energy.

Paper (Tamil language): Write the three 'environmental' impact of solar energy.

A Tamil version does not have any ambiguity, but the English sentence missing one word and then it arises ambiguity, English question does not bother for 'environmental' impact of solar energy, it mentions any impact of solar energy.

IV. CHALLENGES FOR QUESTION PAPER TRANSLATION

In the previous section, we discussed the various form of ambiguity. In this section, we try to explain how ambiguity and some other issues can change the translation and meaning of question sentences.

Now, we will discuss the challenges in question paper translation. One of the major problems that occur is ambiguity. Due to this, the sense of the question may change. Along with Ambiguity, we will also discuss some other related issues that may affect the translation of question paper.

A. Ambiguity Issue

Ambiguity is one of the major challenges for Question Paper Translation through the machine. Question sentences may be affected by ambiguity due to individual words or due to the syntax of the question and hence translated meaning may change. For example, let us take a question from the NCERT online book [39] [44]. The question in English has been translated in Hindi by some popular MT tools.

NCERT (English): Explain the causes of the Great Depression.

Reference Translation: महामंदी के कारणों की व्याख्या करें।

MT (Anusaaraka): व्यापक मन्दी के कारण समझाइए.

MT (Babelfish): महान अवसाद के कारणों की व्याख्या |

MT (Babylon): ग्रेट डिप्रेशन के कारणों को स्पष्ट

MT (Bing): महान अवसाद के कारणों की व्याख्या .

MT (Google): ग्रेट डिप्रेशन के कारणों के बारे में बताएं।

Reference sentence represents the question related to economics but MT (Google and Babylon) translate two same and the inferred meaning is somewhat different. However, the other translators MT (Bing and Babelfish) have given word to word (literal) translation of the same question as above. Clearly, these translations also are not appropriate.

Further, another MT Anusaaraka has given the translation of the question similar to reference translation of the same question. Through this translation is not exactly same as reference translation, one can easily understand the meaning as to what is actually being asked in the question. Let us take another example

NCERT (English): What do you understand by "break" in monsoon?
Reference Translation: मानसून में विराम से आप क्या समझते हैं?

MT (Anusaaraka): आप मानसूनी हवा में अवकाश के क्या समझते हैं?

MT (Babelfish): क्या आप द्वारा तोड़ मानसून में समझ आया?

MT (Babylon): क्या आप को समझने के लिए ब्रेक में मानसून? मानसून क्यों है?

MT (Bing): क्या आप द्वारा तोड़ मानसून में समझ में

MT (Google): मानसून को परिभाषित करें। आप तोड़न मानसून में से क्या समझते हैं?

MT (Babylon) translation is very different and vague. The other translators MT (Bing and Babelfish) have translated the word as "break" is "द्वारा तोड़".

However, the other translators MT (Anusaaraka) translate the word "beak" as "अवकाश". As we saw all MT translations are different from reference question sentence. So clearly we can say MT tools failed to understand the question appropriately.

In addition to ambiguity, machine translated sentences face some other problems. Such as ordering problem, gender problem etc.

B. Ordering Issue

When machine translation is used translates one source language to target language, then sometimes it faces ordering problem [17-18] because the parser does not understand the correct parsing order. In brief, tagging means identification of the sentence in term of parts of speech.

As the ordering depends on whether the tagging has been done accurately or not, it also affects the MT process. For example in the following sentence taken from NCERT book.

NCERT English: Compare the advantages and disadvantages of multi-purpose river projects.
Tagged NCERT (English): Compare/VB the/DT advantages/NNS and/CC disadvantages/NNS of/IN multi-purpose/JJ river/NN projects/NNS [45].
Reference Translation: बहुउपेशीय नदी परियोजनाओं से होने वाले लाभ और हानियों की तुलना करें।

Tagged Reference Translation: बहुउपेशीय /JJ नदी /PREP परियोजनाओं /PREP से /PREP होनेs/VNN वालेs/PREP लाभ/NNPC और/NNPC हानियों/PREP की/VFM तुलना/VFM करेंa ।/PUNC [46].

Whereas the same question when translated using MT gave the following translation.

MT (Anusaaraka): बहु प्रयोजक नदी परियोजनाओं के फायदे और नुकसान तुलना कीजिए.

MT (Babelfish): फायदे और नुकसान बहुउद्देश्यीय नदी परियोजनाओं की तुलना करें।

MT (Babylon): लाभ एवं हानि की तुलना में बहु-उद्देशीय नदी है।

MT (Bing): बहु.प्रयोजन नदी परियोजनाओं के फायदे नुकसान की तुलना करें?

MT (Google): बहु.प्रयोजन नदी परियोजनाओं के फायदे और नुकसान की तुलना करें?

If we look at the above translation by MT (Google), it is found that translation faces ordering problem when compared with the NCERT reference translation.

Google has changed the order of word "advantages and disadvantages" in the question sentence, due to which slightly different meaning is inferred.

MT tools (Bing, Babelfish, and Anusaaraka) also have some ordering issue but their meaning is to some extent same as reference translation. MT (Babylon) has given same as reference translation. So we can say that MT is not free from ordering issue. Taking another example to show the same issues,

NCERT (English): What is meant by trade?
Reference Translation: व्यापार से आप क्या समझते हैं?

MT (Anusaaraka): व्यापार से तात्पर्य हो क्या का है?

MT (Babelfish): द्वारा व्यापार का क्या तात्पर्य है?

MT (Babylon): क्यातात्पर्य है? व्यापार

MT (Bing): द्वारा व्यापार का क्या तात्पर्य है?

MT (Google): व्यापार से क्या मतलब है?

In this example, MT Google has given similar to reference translation but all other MT tools face ordering problem that can not easy to understand. We can say that

these MT tools do not clear the meaning of question sentence because they suffer from the ordering.

NCERT (English): Why does India have a monsoon type of climate?
Reference Translation: भारत में मानसूनी प्रकार की जलवायु क्यों है?

MT (Anusaaraka): भारत के एक मानसूनी हवा प्रकार <u>जलवायु</u> का क्यों है?

MT (Babelfish): क्यों भारत की <u>जलवायु</u> मानसून प्रकार किया गया है?

MT (Babylon): भारत में मानसून प्रकार की <u>जलवायु</u> क्यों नहीं है?

MT (Bing): भारत में <u>जलवायु</u> का मानसून प्रकार क्यों है?

MT (Google): क्यों भारत <u>जलवायु</u> की एक मानसून है?

In this example, all MT tools have ordering problem and their translations differ from reference translation. It is evident from the above examples and their translations using different translations that all translator are affected by ordering problem while translating from English to Hindi, through Google translator is slightly better than other it also focused this problem as has been shown in the examples above.

It is, therefore an important issue in MT, many MT translation as we have seen in many cases that varying from the reference translation, changing/modify the meaning in most of the examples.

C. Gender issue

It has been observing that sometimes machine does not identify the correct gender. That may lead to improper translation. Following the example, questions will make the issue clear.

NCERT (English): Name three states having black soil and the crop which is mainly grown in it.
Reference Translation: तीन राज्यों के नाम बताएँ जहाँ काली मृदा पाई जाती है। इस पर मुख्य रूप से कौन सी फसल उगाई जाती है।

MT (Anusaaraka): काली मिट्टी पास होते हुए तीन राज्य और वह पैदावार जो इसमें प्रमुख रूप से <u>बढी जाती है</u> नामकरण कीजिए।

MT (Babelfish): नाम तीन राज्यों में काली मिट्टी और फसल जो में यह मुख्य रूप से <u>उगाया जाता है</u>।

MT (Babylon): काली मिट्टी के तीन राज्यों का नाम और फसल <u>उगाई जाती है</u> जो कि मुख्यत है। :

MT (Bing): नाम तीन राज्यों काली मिट्टी और फसल है मुख्य रूप से इसे में हो रहा है.

MT (Google): के नाम काली मिट्टी और मुख्य रूप से <u>उगाया जाता है</u>। जो फसल होने के तीन राज्यों।

In the above example, the Reference translation and the MT by Google and Babelfish are different in term of gender. The English word "grow" shows two translations "उगाई**Z**" and "उगाया" in Reference Translation and MT translations respectively. MT Babylon has given translation same as reference translation. However, MT Anusaaraka and Bing change the meaning of the word "grown" from the reference translation. MT Anusaaraka translates the word "grow" as "बढी" and MT Bing miss the translation of the word "grown" in translation. That clearly shows gender issue affects the translation [18].

NCERT (English): What type of soil is found in the river deltas of the eastern coast?
Reference Sentence: पूर्वी तट के नदी डेल्टाओं पर किस प्रकार की मृदा पाई जाती है?

MT (Anusaaraka): पूर्वी समुद्र तट के नदी डेल्टा में <u>पाया</u> मिट्टी का क्या प्रकार है?

MT (Babelfish): मिट्टी की किस तरह पूर्वी तट के नदी डेल्टा में <u>पाया</u> जाता है?

MT (Babylon): किस प्रकार की मिट्टी में <u>मिला</u> है।पूर्वी के पुरानेडेल्टा नदी है?

MT (Bing): मिट्टी की किस तरह पूर्वी तट के नदी डेल्टा <u>पाया</u> जाता है?

MT (Google): मिट्टी के प्रकार क्या पूर्वी तट के डेल्टा <u>पाया</u> जाता है?

In the above example, the Reference translation and the MT by all other translators are different in term of gender. The English word "found" shows two translations "पाई **Z**" and "पाया" in NECRT and MT translations respectively. All MT has given translation different from reference translation in gender. Except for MT Babylon, all MT have given the same translation of the word "found" as "पाया" whereas MT Babylon translates the word "found" as "मिला". It clearly shows gender issue affects the translation.

D. Tense-Aspect and Modality (TAM) Issue

Tense, aspect and modality (TAM) is an important part of natural language processing. TAM is necessary for specifying the information about the word which is temporal in nature or tells us something about the status of an action, or about the ability to perform an action [25].

TAM is the prerequisite of every natural language. Table 3 shows the correspondence of some TAM categories (English-Hindi). It often happens that machines do not identify the correct TAM sentence due to which the translation gets vague. For the following sentence,

NCERT (English): How are minerals formed in igneous and metamorphic rocks?

Reference Sentence: आग्नेय तथा कायांतरित चट्टानों में खनिजों का निर्माण कैसे से होता है?

MT (Anusaaraka): खनिज अग्नि सम्बन्धी और चट्टानों में कैसे बनाए गये हैं?

MT (Babelfish): कैसे खनिज आग्नेय तथा रूपांतरित में गठन कर रहे हैं?

MT (Babylon): खनिजों में बनाई गई हैं और अग्नि कायान्तरित अंश तथा डाइक चट्टानों ?

MT (Bing): कैसे खनिज आग्नेय तथा रूपांतरितचट्टानों में गठन कर रहे हैं?

MT (Google): कैसे खनिज आग्नेय और रूपांतरितचट्टानों में बनते हैं?

Reference translation represents the sentence in PRES (Simple Present) such as tA_HE TAM, while the MT (Anusaaraka and Babylon) represent the sentence PRES (Present Indefinite) such as ya, ye. Similarly MT (Babelfish and Bing) related to PRES (Present Continuous) such as tA_HE, while MT Google translation is similar to reference translation. Let us have another example of TAM related issue.

NCERT (English): How do currents affect the temperature?

Reference Sentence: जल धाराएँ तापमान को कैसे करती हैं?

MT (Anusaaraka): प्रवाह तापमान कैसे प्रभावित होते हैं?
MT (Babelfish:) धाराओं के तापमान को प्रभावित कैसे करते हैं?

MT (Babylon): तापमान पर असर लहरों कैसे होगा?

MT (Bing): धाराओं के तापमान को प्रभावित कैसे करते

MT (Google): धाराएं तापमान को कैसे प्रभावित करती हैं?
MT (Google, Anusaaraka, Babelfish, and Bing) have

TAM similar as reference translation and they represent PRES (Simple Present) such as tA_HE, while MT Babylon is different from all other translation as well as reference translation because it represents Future tense.

E. Synonyms Aspect Issue

In addition to the above problems, there can be some other issues which may affect the MT process. Sometimes the machine translates a sentence which might look correct as for the context whereas there could be another translation possible by using different words for the same original word [8].

Due to the morphological structure of natural language, many words in English and Hindi have a number of synonyms. While one synonym of the word in question most suitably fit in the meaning, it is not necessary that after translations the same synonyms is replaced by MT tool. We can understand it by the following example from the same NCERT source.

Table 3. Some TAM categories (English-Hindi): Source [25]

English Tense	Frequent Hindi Senses
Present Tense (PRES)	HE, tA_HE, nA, yA_HE, gA, tA, 0_jATA_HE, yA, ye,yA_jAtA_HE, 0_sakatA_HE, 0)kara, 0_raHA_HE.
Past Tense	yA, thA, tA_thA, HE, yA_thA, 0_gayA, nA, tA_HE, gA, yA_HE, 0_kaHA
To_0	nA, ne_ke_liye, tA_HE, yA 0_sakatA_HE, HE, ye, 0_kara, gA, nA_HE
Is_en	yA_jAtA_HE, tA_HE, HE, yA_gayA_HE, 0_kara.
Ing	nA, tA_HE, 0_kara, HE, tA_HuA, 0_raHA_HE, tA, yA, ne_ke_liye.
Future Tense	gA, HogA, tA_HE, 0_sakatA_HE, HE

NCERT (English): How did the East India Company procure regular supplies of cotton and silk textiles from Indian weavers?

Reference Sentence: ईस्ट इंडिया कंपनी ने भारतीय बुनकरों से सूती और रेशमी कपड़े की नियमित आपूर्ति सुनिश्चित करने के लिए क्या किया।

Anusaaraka: पूर्व दिशा भारत कम्पनी ने इंडियन जुलाहों कपास और रेशम कपडा-उद्योग की नियमित सप्लाई कैसे प्राप्त कीं?

Babelfish: कैसे ईस्ट इंडिया कंपनी से भारतीय बुनकरों एवं रेशमी वस्त्र की नियमित आपूर्ति की खरीद किया?

Babyloan: ईस्ट इंडिया कंपनी की अधिप्राप्ति की आपूर्ति कैसे कपास और रेशम वस्त्रों से भारतीय बुनकर

Bing: भारतीय बुनकरों से कपास और रेशम वस्त्रों के नियमित आपूर्ति की पूर्व -भारत कंपनी की खरीद कैसे

Google: कैसे ईस्ट इंडिया कंपनी से कपास और रेशम उद्योग के लिए नियमित रूप्से आपूर्ति की खरीद की थी भारतीय बुनकरों ।

In the above example, the reference translation and the MT by all other translators are different in term of synonyms. The English word "cotton" shows two translations "सूती" and "कपास" MT tools, whereas reference translation only shows "सूती". All MT has given translation different from reference translation in synonyms. Except for MT Babelfish, all MT have given the same translation of the word "cotton" as "सूती". It shows synonyms issue may affect the translation.

NCERT (English): What is a neutralisation reaction? Give two examples.
Reference Sentence: उदासीनीकरण अभिक्रिया क्या है? दो उदाहरण दीजिए|

Anusaaraka: एक तटस्थीकरण अभिक्रिया क्या है? दो उदाहरण दीजिए

Babelfish: क्या एक neutralisation प्रतिक्रिया है? दो उदाहरण दे।

Babylon: क्या है? प्रतिक्रिया निष्प्रभावी करके दो देते हैं।

Bing: एक neutralisation प्रतिक्रिया क्या है? दो देते हैं.

Google: तटस्थीकरण प्रतिक्रिया क्या है? दो उदाहरण दें

In this example, the reference translation and MT (Anusaaraka and Google) are different in term of synonyms but they translation are same for word "neutralisation" as "तटस्थीकरण ". While MT (Babelfish and Bing) are not translated word "neutralisation" in Hindi. The English word "neutralisation" shows two synonyms "तटस्थीकरण " and "निष्प्रभावी " in Reference and MT translations respectively.

However, there are words for which different synonyms used in translation may not affect the overall meaning of the sentence. Table 4 shows some such words and their synonyms. Our observation using these words and their respective synonyms in translation has not affected the overall meaning of the sentence.

NCERT English: What steps can be taken to control soil erosion in the hilly areas?
Reference Sentence: पहाड़ी क्षेत्रों में मृदा अपरदन की रोकथाम के लिए क्या कदम उठाने चाहिए ।

MT (Anusaaraka): पर्वतीय क्षेत्रों में मिट्टी कटाव वश रखने के लिए ले क्या कदम सकता है?

MT (Babelfish): क्या कदम पहाड़ी क्षेत्रों में मृदा को नियंत्रित करने के लिए लिया जा सकता है?

MT (Babylon): क्या कदम उठाए जा सकते हैं जिससे भूमि कटाव नियंत्रण पर्वतीय क्षेत्रों में है?

MT (Bing): क्या कदम पहाड़ी क्षेत्रों में मृदा अपरदन नियंत्रित करने के लिए लिया जा सकता है?

MT (Google): पहाड़ी क्षेत्रों में मिट्टी का कटाव की के लिए क्या कदम उठाने चाहिए ।

In the above example, the reference translation and all MT are different in terms of synonyms. The English word "soil" shows two synonyms "मृदा" and "मिट्टी " in reference and MT translations respectively. Now, we take another example.

NCERT (English): What is meant by a water divide?
Reference Sentence: जल विभाजक का क्या कार्य है?

MT (Anusaaraka): एक पानी बाँट से तात्पर्य हो क्या का

MT (Babelfish): क्या से एक जल विभाजन का मतलब है?

MT (Babylon): क्या इसका अर्थ है कि जल विभाजन है?

MT (Bing): क्या से एक जल विभाजन का मतलब है?

MT (Google): एक पानी डिवाइड से क्या मतलब है?

In this example, the reference translation and all MT (Anusaaraka and Google) are different in term of synonyms. While MT (Babelfish, Babylon, and Bing) translated question sentences similar as reference translation. The English word "water" shows two synonyms "जल" and "पानी" in Reference and MT translations respectively.

Table 4. Some English words have more than one synonym

Word	Meaning
Plate	छाप, तख्ता , थाली, पट्ट , पठार, पत्तल
cotton	रुई सूती, कपड़ा
outraged	तोड़ना, अत्याचार करना, भंग करना, उल्लंघन करना, उपद्रव करना, नाराज होना, व्यतिक्रम करना
common	जनसाधारण, साधारण, सामान्य , सार्वजनिक , आम, उभयनिष्ठ , मामूली
Predict	भविष्यवाणी करना, पूर्वानुमान करना, प्रागुक्ति
neutralisation	निष्प्रभावीकरण , उदासीनीकरण, तटस्थीकरण
body	मृतिका, शव, वस्तु , शरीर, संस्था, द्रव्यमान , निकाय, पदार्थ , पिंड, पिण्ड

That clearly shows synonyms issue may or may also affect the translation. Table 5 shows some synonyms words those will not affect the MT translation which words taken from Shabdkosh [47] and Hindi WordNet.

Table 5. Some English words have more than one synonym

Word	Meanings
World	संसार, दुनिया
Soil	मृदा, मिट्टी
Erosion	अपरदन, कटाव, भूक्षरण
Water	जल, पानी
Book	किताब, पुस्तक
Advantage	फायदे, लाभ
Disadvantage	हानि, नुक़सान

V. DISCUSSION

In this paper, we discuss the issue of the ambiguity in question paper translation. In the previous section, we discussed types of ambiguity and their machine translation issues. The question sentences are collected from NCERT (English and Hindi) book, whereas NCERT Hindi translation has been taken as reference translation in order to compare with MT tools. Five MT tools namely Anusaaraka, Babelfish, Babylon, Bing, and Google are selected for translation of English question into Hindi.

These tools are considered best translator. WSD has been identified as one of the major challenges in MT translation. It has been one of the main hurdles in translation accuracy for different Natural Languages including Indian languages such as Hindi [6]. Many researchers have identified this issue through their analysis and experiments. Among these are Navigli, R. [2, 9, 11, 19-20], Bhattacharyya, P. [3-6, 10], Sharma, D. M. [18], Yang, C. Y., and Hung, J. C [24].

When it comes to translating question from English to Hindi, correct translation becomes very important as the even slight change in meaning may result in the different interpretation and the answer to the question might change for what should have been the correct answer to that question. We have tried to raise these issues in this paper through various question have shown that how ambiguity may cause the problem in translation.

The translation of the ambiguous word in the original English questions has been shown again in bold. It can be easily observed that in many cases, different translators translate these ambiguous words differently.

Clearly, these translators understand the context of the question differently. That is why the same ambiguous word has been translated differently by these MT tools. After analyzing these translations through MT tools we found that Babelfish has shown the best result of 36% correct translation whereas the other popular translators that are Bing, Anusaaraka, Google, and Babylon have shown 24%, 20%, 12% and 8% respectively. This implies that ambiguity in question is a big hurdle as less than 40% questions are correctly translated even by the best translator in our experiment that is Babelfish. The poorest among these is Babylon. While analyzing these questions through various translators we considered all those translation as correct wherein the overall meaning and its interpretation are similar to the reference translation.

In all, we have taken 25 questions from NCERT and other resources some of the sample questions are shown in table 6 (shows 6 questions). This table has four columns. In the first column, we have identified the ambiguous word in bold for each of the questions, the second column of the reference translation which have been taken from the same resource, in this, the word in bold shows the correct Hindi translation for the corresponding ambiguous word in English. The third column shows the translation for each of these question sentences using five different translators.

VI. CONCLUSION

As we have covered various aspect of ambiguity that is in the count during question papers translation from English to Hindi. It is evident from our study that incorrect interpretation of the translated question may lead to confusion to the candidates which may have the serious impact on the evaluation.

In order to generalize our experiment, we use five most popular online translators and tested over 25 English questions our experimental analysis as clearly shown that none of these are capable of appropriately handling the issue raised in this paper. It has also been observed that in many cases translators while translation the question in Hindi have changed the overall interpretation of the questions. Hence the MT tools need to follow approach so that these issues are reduced. As the change of interpretation may lead to serious consequence and would be more challenging compared to the translation of the normal text.

Table 6. Experimental Questions sentences, reference translation, and MT translation

MT Translation	Anusaaraka: मेरीऐन और जर्मनीऑ कौन थे? **मार्ग** का था कौन सा में वे क्या दर्शाए गये Babelfish: Marianne और Germania कौन थे? क्या **तरीका** है जिसमें वे चित्रित थे का महत्व था? Babylon: Marianne थे और Germania? क्या **जिससे** वे के महत्व है? Bing: Marianne और Germania कौन थे? **जिस तरह** से वे चित्रित किया गया था के महत्व क्या था? Google: मरियन और जर्मेंनिऑ थे\ **रास्ते** में करने का महत्व क्या था जो वे चित्रित कर थे ।	Anusaaraka: किसी दो देशों पर **केंद्र बिंदु** में से, बताइए राष्ट्रों उन्नीसवाँ शताब्दी कैसे विकास किया. Babelfish: किसी भी दो देशों पर **ध्यान केंद्रित** के माध्यम की व्याख्या कैसे उन्नीसवीं सदी से भी अधिक विकसित किया। Babyloan: किसी भी दोनों देशों पर **ध्यान केन्द्रित** के माध्यम विकसित राष्ट्रों बतायेगा कि 19वीं शताब्दी Bing: किसी भी दो देशों पर **ध्यान** करने के माध्यम से, समझा कैसे देशों उन्नीसवीं सदी में विकसित की है. Google: किसी भी दो देशों पर **ध्यान देने** के माध्यम से] जातिय उन्नीसवीं में विकसित की व्याख्या कैसे शताब्दी।	Anusaaraka: तनाव बोल्कन्ज में **निकले**? Babelfish: क्यों बाल्कन में तनाव **उभरने** था? Babyloan: क्यों ? बाल्कन देशों में **उभरने** तनाव Bing: क्यों तनाव बाल्कन में **उभरने**? Google: क्यों तनाव बाल्कन में **उभरने** किया ।	Anusaaraka: साक्ष्य वट करता है हम **अचेतन** विषय से जीवन की उत्पत्ति के लिए पास होते हैं? Babelfish: क्या सबूत हम **निर्जीव** पदार्थ जीवन की उत्पत्ति के लिए है? Babylon: क्या हमने ऐसा साक्ष्य के मूल **मात्रक** जीवन की? Bing: क्या सबूत हम **चेतन** बात से जीवन के मूल के लिए है? Google: **निर्जीव** से जीवन की उत्पत्ति लिए हमारे पास क्या प्रमाण है?	Anusaaraka: एक अम्ल का जलीय हल विद्युत क्यों च**लाता** है? Babelfish: क्यों एक एसिड की एक जलीय घोल बिजली **संचालन** करता है? Babylon: एक मूलभूत निर्माण इकाइयों का **समाधान** क्यों नहीं? बिजली आचरण Bing: क्यों एक आचरण बिजली का एक जलीय **समाधान** करता है? Google: एसिड **चालन** बिजली का जलीय समाधान क्यों होता है?	Anusaaraka: वियतनाम में विद्यार्थियों की तिहाई school-leaving परीक्षाएँ **उत्तीर्ण** करेगी . Babelfish: वियतनाम में छात्रों में से केवल एक तिहाई स्कूल लीविंग परीक्षा **पास** होता। Babyloan: केवल एक तिहाई विद्यालय छोडने वाले छात्रों वियतनाम में के **पारित** होने का भरोसा Bing: केवल एक वियतनाम में छात्रों तीसरे स्कूल परीक्षा छोड़ने से **गुजरें** होगा- Google: वियतनाम में छात्रों के केवल एक तिहाई **पारित होगा** स्कूल छोडने परिक्षाओं।
NCERT Hindi Translation	मारीआन और जर्मेनिया कौन थे\ <u>जिस तरह</u> उन्हें चित्रित किया गया उसका क्या महत्व था ।	किन्हीं दो देशों पर ध्यान केंद्रित करते हुए बताए! कि उन्नीसवीं सदी में किस प्रकार विकसित हुए।	बाल्कन देशों में राष्ट्रवादी तनाव क्यों पनपा ।	किन प्रमाणों के पर हम कह सकते हैं कि जीवन की **अजैविक** पदार्थों से है?	अम्ल का जलीय विलयन क्यों विधुत **चालन** करता है?	वियतनाम केवल एक तिहाई विद्यार्थी ही स्कूली पढ़ाई <u>सपफलतापूर्वक</u> पूरी कर पाते थे।
NCERT English Translation	Who were Marianne and Germania? What was the importance of the **way** in which they were portrayed?	Through a **focus** on any two countries, explain how nations developed over the nineteenth century.	Why did nationalist tensions **emerge** in the Balkans?	What evidence do we have for the origin of life from inanimate matter?	Why does an aqueous solution of an acid **conduct** electricity?	Only one-third of the students in Vietnam would pass the school-leaving examinations.
S. No.	1	2	3	4	5	6

I.J. Modern Education and Computer Science, 2018, 1, 13-23

REFERENCES

[1] T. Hao, D. Hu, L. Wenyin, and Q. Zeng, "Semantic patterns for user interactive question answering", *Concurrency and Computation: Practice and Experience, 20(7)*, pp.783-799, 2008

[2] R. Navigli, "Meaningful clustering of senses helps boost word sense disambiguation performance", In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, pp. 105-112, 2006

[3] P. Dungarwal, R. Chatterjee, A. Mishra, A. Kunchukuttan, R. M. Shah, and P. Bhattacharyya, "The IIT Bombay Hindi-English Translation System at WMT 2014". In WMT@ ACL, pp. 90-96, 2014

[4] D. S. Chaplot, S. Bhingardive, and P. Bhattacharyya, "IndoWordnet visualizer: A graphical user interface for browsing and exploring wordnets of Indian languages", In Global WordNet Conference (GWC 2014), 2014

[5] D. Chakrabarti and P. Bhattacharya, "Syntactic Alternations of Hindi Verbs with Reference to the Morphological Paradigm", Language Engineering Conference (LEC 2002), Hyderabad, India 2002.

[6] M. Sinha, M. Kumar, P. Pande, L. Kashyap, and P. Bhattacharyya, Hindi "word sense disambiguation. In International Symposium on Machine Translation", Natural Language Processing and Translation Support Systems, Delhi, India, 2004

[7] R. V. Bhala, and S. Abirami, S. "Trends in word sense disambiguation. Artificial Intelligence Review", 42(2), pp 159-171, 2014

[8] P. Kumar, S. Kashyap, A. Mittal, and S. Gupta, "A Hindi question answering system for E-learning documents", In Intelligent Sensing and Information Processing, 2005. ICISIP 2005. Third International Conference on (pp. 80-85). IEEE, 2005

[9] R. Navigli, S. Faralli, A. Soroa, O. De Lacalle, and E. Agirre, "Two birds with one stone: learning semantic models for text categorization and word sense disambiguation. In Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 2317-2320, ACM. 2011

[10] S. Dave, and P. Bhattacharyya, "Knowledge extraction from Hindi text". IETE Technical Review, 18(4), pp 323-331, 2001

[11] R. Navigli, and P. Velardi, "Structural semantic interconnections: a knowledge-based approach to word sense disambiguation". IEEE Transactions on pattern analysis and machine intelligence, 27(7), pp 1075-1086, 2005

[12] L. Li, B. Roth, and C. Sporleder, "Topic models for word sense disambiguation and token-based idiom detection", In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 1138-1147, 2010

[13] F. Mandreoli, and R. Martoglia, "Knowledge-based sense disambiguation (almost) for all structures". Information Systems, 36(2), pp 406-430, 2011

[14] L. R. Nair, and S. David Peter, "Machine translation systems for indian languages", International Journal of Computer Applications pp 0975–8887, 2012

[15] A. Montoyo, A. Suárez, G. Rigau, and M. Palomar, "Combining Knowledge-and Corpus-based Word-Sense-Disambiguation Methods". J. Artif. Intell. Res.(JAIR), pp 299-330, 2005

[16] H. C. Seo, H. Chung, H. C. Rim, S. H. Myaeng, and S. H. Kim, "Unsupervised word sense disambiguation using WordNet relatives", Computer Speech & Language, pp 253-273, 2004

[17] R. Tromble, and J. Eisner, "Learning linear ordering problems for better translation". In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2, Association for Computational Linguistics, pp. 1007-1016, 2009

[18] S. Chaudhury, A. Rao, and D. M. Sharma, "Anusaaraka: An expert system based machine translation system" In Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on IEEE. pp. 1-6, 2010

[19] R. Navigli, "A quick tour of word sense disambiguation, induction and related approaches", SOFSEM 2012: Theory and practice of computer science, pp 115-129, 2012

[20] R. Navigli, "Word sense disambiguation: A survey". ACM Computing Surveys (CSUR), 2009

[21] F. Mandreoli, and R. Martoglia, "Knowledge-based sense disambiguation (almost) for all structures". Information Systems, 36(2), pp 406-430, 2011

[22] H. Li, and C. Li, "Word translation disambiguation using bilingual bootstrapping". Computational Linguistics, pp pp 1-22, 2004

[23] B. Broda, and M. Piasecki, "Semi-supervised word sense disambiguation based on weakly controlled sense induction", In Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on IEEE. pp. 17-24, 2009

[24] C. Y. Yang, and J. C. Hung, "Word sense determination using wordnet and sense co-occurrence", In Advanced Information Networking and Applications2006. AINA 2006. 20th International Conference on Vol. 1, IEEE. pp. 779-784, 2006

[25] A. K. Singh, S. Husain, H. Surana, J. Gorla, D. M. Sharma, and C. Guggilla, "Disambiguating tense, aspect and modality markers for correcting machine translation errors". In Proceedings of RANLP, 2007

[26] S. Dwivedi, and A. Goyal, "Machine Translation status in India", Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies - ICTCS '14, 2014

[27] M. Bryant, "Eats, Shoots and Leaves: The Zero Tolerance Approach to Punctuation". Law Now, 29, pp. 94. 2004.

[28] Karov, Yael, and Shimon Edelman. "Similarity-based word sense disambiguation." Computational linguistics 24, no. 1, pp 41-59. 1998

[29] S. K. Dwivedi and V. Singh, "Integrated question classification based on rules and pattern matching". In Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies pp. 39. ACM. 2014

[30] D. Metzler and B. W. Croft, "Analysis of statistical question classification for fact-based questions". Information Retrieval, 8(3), pp. 481-504.2005

[31] J. Silva et al., "From symbolic to sub-symbolic information in question classification". Artificial Intelligence Review, pp. 137-154. 2011

[32] S. K. Dwivedi and S. Vikram, "Word Sense Ambiguity in Question Sentence Translation: A Review", In International Conference on Information and Communication Technology for Intelligent Systems pp. 64-71. Springer, Cham, 2017

[33] A. Graesser et al., "Question classification schemes". In Proc. of the Workshop on Question Generation. 2008

[34] I. S. Abuhaiba and M. F. Eltibi, " Author Attribution of Arabic Texts Using Extended Probabilistic Context Free Grammar Language Model", International Journal of Intelligent Systems and Applications, 8(6), 27. 2016

[35] M. A. Kadhim et al. "A Multi-intelligent Agent System for Automatic Construction of Rule-based Expert System. International Journal of Intelligent Systems and Applications, 8(9), 62. 2016

[36] G. Chandra and S. K. Dwivedi, "Assessing Query Translation Quality Using Back Translation in Hindi-English CLIR", International Journal of Intelligent Systems and Applications, 9(3), 51. 2017

[37] P. R. Kamdi et al., "Keywords based closed domain question answering system for indian penal code sections and indian amendment laws", International Journal of Intelligent Systems and Applications, 7(12), 54. 2014

[38] A. S. Medjahed et al., "Urinary System Diseases Diagnosis Using Machine Learning Techniques", International Journal of Intelligent Systems and Applications, 7(5), 1. 2015

[39] http://epathshala.nic.in/e-pathshala-4/flipbook/

[40] http://timesofindia.indiatimes.com/city/chennai/Class-12-girl-cites-ambiguity-in-biology-paper-seeks-full-marks/articleshow /33146287.cms

[41] English WordNet http://wordnetweb.princeton.edu/perl/webwn

[42] Hindi WordNet: http://www.cfilt.iitb.ac.in/wordnet/webhwn/wn.php

[43] https://drive.google.com/file/d/0B3HSpNixd2_YVUM5U3d2ZjlldHc/view

[44] NCERT: http://ncert.nic.in/NCERTS/textbook/textbook.htm English Parser

[45] http://nlp.stanford.edu:8080/parser/

[46] Hindi Tagging: http://text-processing.com/demo/tag/

[47] English-HindiDictionary: http://www.shabdkosh.com/

Authors' Profiles

Shweta Vikram received her M. Tech. degree from Banasthali Vidyapith, Rajasthan India and is pursuing her Ph.D. degree in Department of Computer Science at Babasaheb Bhimrao Ambedkar Central University, Lucknow, India. Her major research interests include WSD and Machine Translation. She is approachable at shwetavikram.2009@rediffmail.com

Prof. Sanjay K. Dwivedi is working as Professor, Department of Computer Science at Babasaheb Bhimrao Ambedkar Central University, Lucknow, India. His research interest includes Artificial Intelligence, Information Retrieval, Web Mining, NLP and WSD. He has published a number of research papers in reputed journals and conferences. He is approachable at skd200@yahoo.com.