Chapter 3

Problems in Machine Translation

While most theories of translation are skeptical about the possibility of translation, it is a fact that works of translations are carried out. The process of translation involves decoding of a text in one language followed by encoding it into another. The theories of translation talk about the divergence between languages at several levels. When we turn from human to machine processing of translation, various problems hitherto unexpected surface. Several problems at decoding level that are not noticed by human translators are encountered by machine, challenging the task of Machine Translation. We give below the reasons as to why decoding is not easy for a machine. Further, we discuss a number of problems at the level of cross lingual transfer of information that demand a series of resolutions at the level of encoding.

3.1 Why is decoding not easy?

In any language there is always a tension between brevity and precision. Inherent linguistic process constantly strives to avoid ambiguity and bring in precision. If one states everything explicitly then the text not only becomes very lengthy, it also leads to loss in the focus due to cluttering of information, whereas brevity helps in focusing the attention. It is a natural tendency to go for brevity leaving the resolution of
ambiguity to the context.

Look at how different languages view the activity of *cigarette smoking*. The smoking of a cigarette is a complex activity which involves many sub-activities ranging from inhaling the smoke to release of the smoke through the nostrils. Hindi expresses it either as ‘sigāreta pīnā’ or ‘dhūṃrapāna karana’. The latter comes close to the inhaling, whereas the first one is more an idiomatic one. English speaker would use *to smoke* rather than *to inhale*. Thus we see that languages use an expression that describes one of the sub-activities of the complex activity to describe the whole complex activity. This introduces an ambiguity. Look at another example: *shelve the books* to mean *put the books on the shelf*. *Butter the bread* to mean *spread the butter on a slice of a bread*. *To butter* and *to shelf* are the verbs derived from the corresponding nouns *butter* and *shelf*, but the meanings conveyed by them are totally different. In the former case, it means *to spread* and in the later case it means *to put*. However, the semantics is so clear for a native that s/he does not find any difficulty in grasping the meaning of these verbs.

Brevity leads to overloading at the word level as well as at the structural level. In the process of interpretation, a reader uses various extra-linguistic resources such as common sense, world knowledge, language conventions, cultural background, domain specific knowledge, etc. These resources, as yet, are either not available to the computer, or even if such a knowledge is available in electronic form, it is still not clear how to use such information computationally.

The main reasons of difficulty in decoding are:

- Languages may code information only partially, and
- Languages may code information at arbitrarily long distance.
3.1.1 Languages code information only partially

In this section, we give some examples illustrating how brevity leads to overloading at various levels of coding introducing an ambiguity.

Syntactic Ambiguity

- A word may belong to more than one POS category.

Consider the sentence

Time flies like an arrow.

Any normal English parser would produce several parses of this sentence, whereas human being does not even ‘see’ any ambiguity in this sentence unless it is pointed out. The most natural parse corresponds to the meaning - time passes away speedily. What are the other parses which machine makes explicit? English overloads a form to achieve brevity. For example, it is a very natural phenomenon, in English, to use nouns as verbs\(^1\). In the above example also, the words time, flies and like may have different part of speech in different contexts as shown in table 3.1. Thus there are \(3 \times 2 \times 2 (=12)\) possible parses. Though many of these parses are rare, one can imagine situations where each of these parses is meaningful. The implication is that we can not translate a sentences in isolation. The situation or the context “controls” the translation.

<table>
<thead>
<tr>
<th>Time</th>
<th>flies</th>
<th>like</th>
<th>an</th>
<th>arrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/V/A</td>
<td>N/V</td>
<td>Prep/V</td>
<td>Det</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 3.1: ambiguity

- The relations are not expressed with fine granularity.

Consider a pair of sentences, from Sanskrit, with structural ambiguity.

\(^1\)E.g. to butter, to shelf, etc. as seen above.
Chapter 3: Problems in Machine Translation

San: pitā putreṇa saha grāmam gacchati. (1)
gloss: father{nom.} son with village{acc.} goes. (1a)
Eng : Father goes to a village with son. (1b)

San: rāmaḥ dugdhena saha annam khādati. (2)
gloss: Rama{nom.} milk with rice{acc.} eats. (2a)
Eng: Rama eats rice with milk. (2b)

Here the word ‘saha’ assigns a role of associative kāraka to the head words viz. putra and dugdha. Whether this kāraka role is that of kartā or karma is not marked. Thus putren. a saha and dugdhena saha are saha kartā and saha karma respectively can be decided only by appealing to the extra-linguistic knowledge.

- Overloading the markers with more than one relation.

English uses the same preposition to mark both a noun-noun relation as well as a noun-verb relation, and as such, English sentences are much more ambiguous than the corresponding sentences in Sanskrit or in any Indian languages which use different markers to mark these two types of relations. For example, consider an often-quoted example from English

He saw a man on the hill with a telescope. (3)

The problem here is of the prepositional phrase attachment. Depending on where the prepositional phrase with a telescope is attached - to saw, to a man or to the hill, the meaning changes. In one case it is a kāraka relation, and in other two cases it is modifying a noun.
Semantic Ambiguity

The concepts are language independent and are infinite. We express these concepts by the words in a specific language which are denumerably finite. Overloading a word with related meanings is unavoidable. This leads to polysemy. Further, borrowing from other languages, shift in the meaning over a period of time, etc. may sometimes introduce homonymy. Homonymy and polysemy introduce the semantic ambiguities.

Consider the sentences

San: saindhavam ānaya.  \[4\]

Gloss: horse/salt bring.  \[4a\]

Eng: He played well.  \[5\]

(a violin / cricket / in a drama?)

Eng: He went to the bank.  \[6\]

(river bank / money bank? )

To resolve the ambiguities originating from polysemy and/or homonymy, machine needs the context, world knowledge, common sense, cultural background, etc. Moreover one can never say ‘this much’ extra-linguistic information is sufficient for disambiguation, since potentially there will always be exceptions. Further it is still not clear to the computational linguists\(^2\) how to organise the world knowledge so that machines can use it effectively.\(^2\)

\(^2\)There had been efforts under the CYC [96] project to organize the encyclopaedic knowledge so that machine can do reasoning. However, still the effort is far from realistic use in any Machine Translation system.
3.1.2 Information is coded at arbitrarily long distance

In the previous section we saw examples where the information for ambiguity resolution is available within a sentence. The anaphora resolution may require processing of more than one sentences. But it is not clear a priori how much text needs to be processed. Sometimes a text as much as a complete novel may have to be processed. For example, Hindi has a famous story by Guleri [36], entitled ‘usane kahā thā’(s/he said), and only after reading the complete story, one gets an answer to the question ‘who’ said it.

Thus we see that, when a machine is translating a text, even at the level of decoding several problems crop in, which otherwise are less problematic from a human translator’s point of view.

In the next section we look at the problems involved in transferring the concepts from one language to the other in the context of encoding.

3.2 Cross-lingual information transformation: Problems

Once the information in the source language is decoded, the next step is to map it appropriately into the target language. The divergences between the source language and the target language lead to the problems in this transfer. There have been studies in the divergences by many – both in the context of Machine Translation as well as in the context of Human Translation [38], [84], [30]. These divergence studies look at the differences from various aspects of encoding such as distribution and use of grammatical features viz. gender, number, person, etc., divergences at the level of syntax, semantics, pragmatics, etc.
We study the divergences from the point of view of coding of information. Words and sentences are the basic building blocks of a language. Various grammatical features get realised through them. When a language encodes information partially, the world knowledge, domain knowledge, etc. help a reader understand the text. Assuming that the same world knowledge and domain knowledge is available to the target language reader, the ‘gloss’ of the source language words in the target language should help a TL reader to understand the text. But this does not happen. It is because the two languages need not code the information the same way. They may differ in the manner they code it, what they code, and where they code. To illustrate it further, look at the following English sentence and its Hindi translation.

Eng: He is there. (7)
gloss: \{3 per sg masc pron\} be\{pres, sg, 3 per\} there. (7a)

Hin: vaha vahāin hai. (8)
gloss: \{3 per sg pron\} there be\{pres, sg, 3 per\} (8a)

Though these two sentences are taken as translational equivalents of each other, the information content in both these sentences is not the same. This causes a problem, when one has to translate from Hindi into English. Because, English requires some information which is not explicitly coded in Hindi either at the word or sentence level. Or in other words, a sentence level analyser is not enough to represent the information of pronouns. What is required is a pronoun reference handler. The discrepancy at the higher level of coding demands higher level of processing which in turn requires more knowledge resources, and hence is costly.

We look at the discrepancy in coding at various levels. At the word level, they may
differ in labeling and packaging of concepts, at the sentence level, they may differ in the syntax, language conventions, etc. The difference in culture may get reflected at any of the levels from the word level to pragmatics. In what follows we discuss the divergences at the word and sentence level.

### 3.2.1 Divergence at Word Level

Sapir (1929) asserted that “the ‘real world’ is to a large extent unconsciously built up on the language habits of the group. [...] The worlds in which different societies live are distinct worlds, not merely the same world with different labels attached.” [88](91). The different perception and mental organisation of reality leads to various kinds of mappings between the words in one language to the words in another language. We give below some examples illustrating these points.

Given two languages, one can imagine the following possibilities:

- **Lexical Gap:**
  A lexical gap is an absence of a word to express a specific concept in another language. The concept may correspond to a content word or to a functional word. For example, the technical words such as ‘transducer’, ‘electricity’, ‘genes’, etc., the function words such as determiners in English, or the adverbial marker in Telugu. All these do not have their counterparts in Hindi. In other words, Hindi has a Lexical Gap corresponding to the English technical words, determiners in English and the adverbial marker in Telugu.

- **One to One mapping:**
  This is the most comfortable situation from translation point of view. Typically the proper nouns, words belonging to body parts, number words etc. fall under this category.

- **One to many and many to one mappings:**
  If the direction of translation is changed, then one to many mapping changes to
many to one and hence we have clubbed them together. This situation arises when two languages differ in the granularity.

The word ‘uncle’ in English may refer to father’s (elder or younger) brother, mother’s brother, or aunt’s (mother’s sister’s or father’s sister’s) husband. But Hindi distinguishes between each of these relations, and has a distinct label for them. Another often quoted example may be cited from the Eskimo which distinguishes between several states of ice (formation), whereas Hindi does not even distinguish between snow and ice.

One may think that many to one mapping is not at all a problem from the translation point of view, however, it is not so. English has 3 distinct third person singular pronouns corresponding to 3 genders, Hindi has only one. Now consider the following sentence

Eng: He gave her a book. \hspace{1cm} (9)

Its Hindi translation would be

Hin: usane usako pustaka di. \hspace{1cm} (10)
gloss: (S)he/it him/her/it book gave. \hspace{1cm} (10a)

Since Hindi pronoun does not mark the gender, an unambiguous sentence in English has become ambiguous in Hindi, which is not a comfortable situation.

- Overlapped regions:

  This is a very common phenomenon, especially when the two languages belong to two different families, and/or belong to two different cultures. The English verb \textit{play} means \textit{bajānā} when its object is a musical instrument as in \textit{play a violin}. It is \textit{khelanā} when its object is a game, and it means \textit{abhinaya karanā} when it refers to playing a role in a drama. The Hindi verbs \textit{bajānā}, \textit{khelanā}
and *abhinaya karunā* have their own domains, and are used in other contexts as well, where the English verb *play* can not be used. For example, *to ring the bell*, Hindi uses *bajānā*.

Thus we see that these differences are either because of the absence of a concept or due to the differences at the level of labeling and packaging of concepts. The cases where there is an absence of concept is comparatively easier to handle than the cases where the domains overlap. Unless the correct shade of meaning is captured, there are more chances of leading such situations to catastrophe or even to loss in communication.

### 3.2.2 Divergence at Sentence Level

To study the divergence at the sentence level, we need to look into the dynamics of information coding across the languages. We seek answers to the questions like, where does a language code information? and what is the manner in which it codes the information? If the two languages differ in any of these aspects, it leads to difficulty in translation.

1. Where is the information coded?

Two languages may code the same information in different ways. Consider the English sentence,

Eng: Rats kill cats.  \(\text{(11)}\)

Hin: cūhe billiyom ko mārate haim.  \(\text{(12)}\)

gloss: Rats cats{acc.} kill.  \(\text{(12a)}\)

English codes the information about the relation of a verb to a noun in position. That is the information that ‘rats’ are the killers and ‘cats’ are being killed is coded in position. Hindi, on the other hand, requires an explicit *vibhakti* marker to code this information.
Hindi also has instances where it does not use explicit vibhakti markers as in:
Hin: rāma phala khātā hai. (13)
gloss: Rama fruit eats. (13a)
Eng: Rama eats fruits. (14)

Here the information that the fruit is eaten by Rama is not explicitly coded by any morpheme. Still a native Hindi speaker does not find it ambiguous as he uses the world knowledge to get the preferred reading. In chapter 6 we show how the information about where does a language code information helps us in understanding the divergences between English and Hindi.

2. How is the information coded?

Different languages may follow different conventions of sharing the information within or across sentences. In Sanskrit the verbal suffix ‘ktvā’ marks the sharing of ‘kartā’. Consider the following sentence:

San: rāmaḥ dugdham pītvā sālām gacchati. (15)
gloss: Rama milk having drunk school to goes. (15a)

The fact that the kartā of the activity of drinking is same as the kartā of the activity of going is coded in a language convention, and not through any morpheme. In other words this information is implicit.

The ‘i’ suffix in Telugu comes very close to the ‘ktvā’ of Sanskrit. The Sanskrit language convention of the marking of sharing of kartā has a counterpart in Telugu but it is overloaded as it also indicates the cause-effect as in the following Telugu sentence.

Telugu: pāmu karici bāludu caccipoyādu. (16)
gloss: snake having bitten child died. (16a)
Chapter 3: Problems in Machine Translation

Eng: The child died because of a snake bite. (17)

Similarly consider the following English sentence

Eng: Mohan dropped the melon and burst. (18)

What is expressed through English language convention is Mohan burst. This phenomenon, usually known as gapping, is not present in Hindi. When the two languages differ in their conventions of information coding, i.e., one codes the information implicitly whereas the other codes it explicitly, one has to make the implicit information explicit. This is not an easy task, unless one appeals to the extra-linguistic information. Hence the cases where two languages differ in their ‘language conventions’ of coding information, are the cases of translation failure.

3.3 Problems at the encoding level

Translation involves understanding the original text and presenting it in another language. The presentation part involves creativity. Thus, for a given text there is no ‘one specific way’ of translation. What exists is a spectrum of translations and it is the translator who translates, what (s)he feels is the most appropriate one. Needs and likes of different people may demand different translations.

All these issues make the task of translation difficult and machine translation more difficult.

As we notice, the brevity necessarily implies partial encoding of the information leading to ambiguity. This calls for the use of extra-linguistic information at the level of
decoding. To discover the sources of information for extra-linguistic analysis, and the additional resources that are needed may take some time. At this stage, it may not be feasible to provide such a ‘Knowledge Base’. The differences between two languages in information coding make the task more complex. Assuming that the extra-linguistic knowledge (except the cultural knowledge) needed by different languages is the same, it suffices to handle the differences between the languages at various levels.

In the next chapter, we look at the current MT architecture and provide a fresh look to the problem, resulting into a new architecture. Chapter 6 discusses the syntactic divergences between English and Hindi and provides reasons behind these divergences. In chapter 7 we see how anusaaraka provides a solution to various divergences discussed by Dorr [32]. In chapter 8 we discuss the concept of śabdasiṣṭra to handle the divergences at the lexical level.